MSc Dissertation

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

An Application of Survival Analysis in Credit Risk Management

By

Patrick Mthisi

0413210D

Supervisors

Dr. J Majakwara

Dr. H Chipoyera

A dissertation submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science

October 19, 2022

Dedication

I dedicate this dissertation to God Almighty for giving strength, determination, and zest to pursue this project.

Acknowledgement

First and foremost, I am grateful to my supervisors, Dr. H Chipoyera and Dr. J Majakwara, for their invaluable advice, unwavering support, and patience throughout my research. Their vast knowledge and experience have encouraged me in all the time of my academic research. I would also like to express my gratitude to the School of Statistics and Actuarial Science for granting me the opportunity to pursue this degree. Finally, I would like to thank my sister, Mrs. P Al Eman, and her husband, Dr. O Al Eman, for their help. Without their tremendous understanding, encouragement, and financial support during the year, it would have been difficult for me to finish my studies.

Declaration

I, Patrick Mthisi, declare that this dissertation is my own, unaided work. It is being submitted for a Masters of Science in Mathematical Statistics degree at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.

October 19, 2022

Abstract

Banks, credit unions, and finance companies use credit risk models to estimate the degree of credit risk associated with lending money to borrowers. Credit scoring models form the preponderance of these models, and their function is to calculate the probability of default (PD) for borrowers' credit facilities. The global financial crisis that lasted from mid-2007 to early 2009 gave birth to a new capital provisioning approach that requires the use of forward-looking PD estimates. This research focuses on applying survival analysis to provide PD estimates throughout the lifetime of a loan facility. The study covers a variety of survival analysis models that encompass semi-parametric, parametric, and mixture cure models. A large sample of personal loans from Prosper, a peer-to-peer lending marketplace was analysed. Loans are monitored monthly from entry to exit. The study reveals that survival models are suitable for providing point-in-time (PIT) PD estimates. The findings from the top-performing models, namely, Log-normal, Logistic-Log-normal mixture cure, and stratified Cox regression, were satisfactory based on their probabilistic statistical measures, ability to discriminate prognosis in the future, and calibration performance.

Keywords: Credit risk management, Survival analysis, Probability of default, Point-in-time, Capital provisioning.

List of Acronyms

AFT	Accelerated Failure Time
AIC	Akaike's Information Criterion
BCBS	Basel Committee of Banking Supervisors
BIC	Bayesian Information Criterion
BLR	Bayesian Linear Regression
BS	Brier Score
CAIC	Consistent Akaike's information criterion
CIF	Cumulative Incidence Function
EAD	Exposure At Default
ECAI	External Credit Assessment Institution
ECL	Expected Credit Loss
EL	Expected Loss
EM	Expectation Maximisation
FASB	Financial Accounting Standard Board
GAM	Generalised Additive Model
GLM	Generalised Linear Model
GOF	Goodness-of-Fit
IAS	International Accounting Standards
IASB	International Accounting Standards Board
IBS	Integrated Brier Score
IFRS	International Financial Reporting Standard
IM	Informative Missing
IPA	Index of Prediction Accuracy
IRB	Internal Ratings-Based
IV	Information Value
KM	Kaplan-Meier
KS	Kolmogorov-Smirnov
LGD	Loss Given Default
M	Maturity
MAR	Missing At Random
MCAR	Missing Completely At Random
MICE	Multivariate Imputation by Chained Equations
MLE	Maximum Likelihood Estimate
NCA	National Credit Act
PD	Probability of Default
PH	Propotional Hazard
PIT	Point-In-Time
PMM	Predictive Mean Matching
P2P	Peer-to-Peer
RF	Random Forest
ROC	Receiver-Operating Characteristic
TTC	Through-The-Cycle
UL	Unexpected Losses
VIF	Variance Inflation Factor
WOE	Weight Of Evidence

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Banks and finance companies loan money to their customers through credit facilities such as credit cards, personal loans, or fund purchase assets such as real estate, motor vehicles, or office equipment. The risk of loss to the bank if a borrower fails to make principal payments on the loan is known as credit risk [Leo and Sharma, 2019]. An institution's credit risk management function involves determining the adequacy of its capital and loan loss reserve at any particular time in order to mitigate financial losses. Regulators expect banks to allocate capital that reflects the level of risk their business units face. Under the aegis of the Basel Committee on Banking Supervision (BCBS), the Basel II accord sets out a framework for credit risk quantification and capital provisioning. The BCBS gives banks the option of calculating capital requirements using one of two methods. The first method is the Standardised Approach, which evaluates credit risk via external credit assessments offered by an external credit assessment institution (ECAI), whose duties fall under the ambit of the national supervisors. ECAIs are required to meet a set of eligibility criteria, including objectivity, independence, transparency, credibility, and disclosure. The Internal Ratings-Based (IRB) approach, which is the second method, allows banks to evaluate credit risk using their own internal rating systems. A bank may choose either the foundation or advanced approach, which are both part of the IRB framework, depending on the number of risk parameters that a bank must evaluate.

Effective credit risk management requires a comprehensive understanding of a lender's overall credit risk by analysing risk at the customer and portfolio levels. Credit risk models are employed to evaluate the degree of credit risk. A majority of these models are credit scoring models that calculate probability of default (PD) associated with the extension of credit to customers. PD is the probability that a borrower will not be able to meet debt obligations before the end of a predetermined period. Altman [1968] introduced one of the earliest credit scoring models called the Altman's Z-score model which is based on discriminant analysis. Since then, several quantitative credit scoring models have been proposed over the years. Binary classification using logistic regression, a popular Generalised Linear Model (GLM), is the most extensively used approach by banks to inform the granting of credit to new borrowers or monitoring the behaviour of current borrowers [Keramati and Yousefi, 2011; Leo and Sharma, 2019]. As a binary classification model, logistic regression classifies borrowers as "bad" if they exhibit poor performance during a fixed observation period, while the remaining borrowers receive a "good" status. Logistic regression is popular because it performs classification and estimates class probabilities which are calibrated to the PDs used by banks. This approach is suitable for banks concerned with modelling PD over a fixed horizon, and it fits well within Basel's IRB methodology [Jiang et al., 2019; Noh et al., 2005].

For the last 20 years, banks have been reliant on models that provide one-year PD estimates for Basel II regulatory requirements [Bellini, 2019]. The Basel II capital provisioning framework relies on the through-the-cycle (TTC) approach to estimate PDs. TTC PDs are one-year PDs that are largely unaffected by the credit cycle. GLMs, as evidenced by their widespread use, form the cornerstone of models used to calculate these PDs. The 2007 – 2009 global financial crisis and the subsequent credit crunch sparked a series of concerns around the propriety of TTC PDs in capital provisioning. As a result, the International Accounting Standard Board (IASB) and the Financial Accounting Standard Board (FASB) collaborated to revise accounting standards in favour of the Expected Credit Loss (ECL) paradigm. ECL requires the use of dynamic forward-looking PDs over the credit life cycle of borrowers [Bellini, 2019]. Thus, credit scoring models had to be adapted to embed a forward-looking perspective. This research provides one of the adaptations in the form of survival modelling.

1.1 Background

The International Financial Reporting Standard (IFRS 9), instituted by the IASB, gained momentum in 2014. Impairment provision under IFRS 9, known as ECL, is calculated based on the expected economic loss of a loan or asset under consideration. IFRS 9 adopts a forward-looking approach, requiring provisions to be made based on the assumption of credit losses at the time assets are first recognised [Frykström and Li, 2018]. In contrast, the International Accounting Standard (IAS) 39, introduced in January 2001, adopts a backward-looking approach based on losses incurred during the period. In addition, provisions are made when asset impairment is evident at the time of reporting [Hague, 2004]. IAS 39 has played a crucial role in establishing standards for the recognition of financial assets and impairments; this guideline forms the basis of Basel's IRB approach.

As of January 2018, in South Africa, IFRS 9 supplanted IAS 39, meaning that banks and finance companies have adopted different iterations of the GLM framework such as the Lifetime GLM, transition matrix modelling, or machine learning models, to account-level panel data. Survival analysis is used in this study to fit lifetime PD since it is a straightforward method. Unlike transition matrix modelling and Lifetime GLM (and panel data-based machine learning modelling), which involve multi-state modelling and multi-step processes, respectively, survival analysis estimates lifetime PDs all at once by utilising a survivor function.

1.2 Statement of the Problem

The recent reporting mandate under IFRS 9 requires banks to conduct a more in-depth assessment of borrowers in order to predict their behaviour throughout the credit life and provide ECL estimates continually, rather than just annually, as the IRB framework requires [Frykström and Li, 2018]. Logistic regression, on its own as a framework for estimating PD, is inadequate because it gives a static PD snapshot for a fixed outcome period (such as twelve months, in the IRB framework) [Bellini, 2019]. Thus, there is a need for an alternative credit scoring approach that models both the credit lifetime of borrowers and the time to default. Survival analysis lends itself more readily to such

applications since it is easily adaptable to model time to default on a forward-looking basis, allowing prediction of credit customers' future dynamic behaviour.

1.3 Aim and Objectives

The aim of the study is to estimate PD and predict default times within a personal loan portfolio as part of the bank's reporting under IFRS 9. The specific objectives are as follows:

1. Address the problem of missing customer information in the bank's data using data imputation.
2. Employ survival analysis techniques to compute annual and lifetime PDs given borrowers' application and behavioural characteristics.
3. Systematically identify top-performing survival models using evaluation metrics.
4. Compute an estimate of the cure rate based on the best-performing survival model.

1.4 Limitations and Assumptions

The following assumptions are made throughout the current investigation:

1. The occurrence of default is the only event of interest.
2. Competing risks are outside the scope of our study, which only focuses on PD estimation for the IFRS 9 purposes. The IFRS 9 requires point-in-time (PIT) PD estimation to determine stage allocation; hence the entire loan portfolio, including accounts that fall off due to competing risk, is considered to avoid obtaining biased estimates.
3. The borrower does not return to non-default status after failing to redress missed payments, negotiate debt re-arrangement, or the credit provider has terminated

debt review after sixty business days according to Section 86(10) of NCA [Renke et al., 2007; Roestoff et al., 2009]. In other words, default is non-recurring.

1.5 Dissertation Outline

This research has the following outline:

- Chapter 1: gives the aims, objectives, and limitations of the study.
- Chapter 2: provides a synopsis of prior literature, as well as a detailed theoretical literature review of each of the models utilised in the research.
- Chapter 3: provides a detailed discussion of the research methodology in terms of the data, variable selection methods, and analytical metrics used in the study.
- Chapter 4: presents and discusses results of the research.
- Chapter 5: provides a summary of the research findings, gives recommendations for future research, and concludes the study.

Chapter 2

Literature Review

2.1 Introduction

Event times to events such as recidivism, death following a significant operation, or time in remission for Leukemia patients are examples of problems explored using survival analysis. Historically, survival analysis has been limited to mortality. Recent years have seen it applied to a variety of disciplines, including credit risk. The contribution of Kaplan and Meier [1958] to the calculation of non-parametric survival probabilities and hazard rates resulted in significant advances in survival analysis. An alternate non-parametric estimator of the hazard rate and the resulting survivor curve is the Altschuler-Nelson-Aalen estimator proposed by Altshuler [1970], Nelson [1972] and Aalen [1976]. The main criticism levelled at these earlier methods is that they do not differentiate between individual survival experiences within a sample, i.e., they result in homogeneous survivor curves [Harrell et al., 2001; Kleinbaum and Klein, 2010]. The proportional hazard (PH) and accelerated failure time (AFT) models, on the other hand, make it relatively simple to link explanatory variables to failure times. Both of these models produce heterogeneous failure time distributions that are far better at summarising individual survival experiences. The implicit assumption of PH models is that explanatory factors have a multiplier effect on the baseline hazard function. On the other hand, AFT models assume that explanatory

variables can either accelerate or decelerate failure time. The Exponential and Weibull distributions, which are compatible with both the AFT and PH assumptions, are two parametric survival models belonging to a family of Generalised Gamma distributions that obscure the distinction between the PH and AFT model types [Kleinbaum and Klein, 2010]. Another significant contribution to survival analysis is the semi-parametric Cox PH by Cox [1972]. This model is categorised as semi-parametric because it comprises a time-dependent non-parametric baseline, which explains how risk evolves, and a parametric time-independent exponential multiplier function of a linear combination of explanatory variables.

2.2 Survival Analysis in Credit Risk Scoring

The main appeal for applying survival analysis to retail credit risk scoring is that it can handle censored data. Narain [1992], one of the pioneers to apply survival analysis to credit risk scoring, investigated the relationship between the probability of classifying a borrower as "bad" and the time the borrower has been in the books using an AFT Exponential model. The data used comprised 1,242 applicants approved for 24-month loans covering the mid-1986 to mid-1988 data range. An account that has missed three consecutive payments is assigned a "bad" status (uncensored), while the rest are right-censored. Explanatory variables such as time with an employer, residential status, etc., are incorporated. The investigation provides an additional dimension outside what classical approaches such as logistic regression offer since it models time to default. The author concludes that credit scores that are backed by survival times result in better credit-granting decisions. Subsequent studies to Narain [1992] made significant advances and expanded the application of survival analysis in credit risk prognosis. Noh et al. [2005] use AFT models to predict default times and PDs using credit card data. It cannot be determined from the study whether or not survival analysis outperforms logistic regression and neural networks. However, it is observed that survival analysis is superior at predicting "bad" customers, thereby reducing the risk of erroneous credit-granting decisions. In other words, survival analysis is better at identifying "bad" customers than classical binary approaches.

Thomas et al. [2001] propose the Markov chain model and PH model for profit and behavioural scoring of new and existing consumer loans, respectively. Markov chains model consumer behaviour by utilising scoring bands to identify consumers that are more likely to default. On the other hand, survival analysis models time to the occurrence of default. Stepanova and Thomas [2002] employ the Cox PH model to estimate default time, as well as early settlement time as a competing risk using personal loan data from the United Kingdom. The study proposes a novel approach involving classifying predictive variables using the Cox PH model. The idea behind this approach is to group variables based on the type of risk, which can either be default or early settlement risk, in order to ensure that the credit scoring systems are robust. Default and early settlement risks are estimated using either the Cox PH model or logistic regression. The study concludes that the Cox PH is competitive with the industry standard, the logistic regression.

Survival analysis has also gained prominence in wholesale banking, where it has replaced the classic Merton firm value approach. The Merton model, developed by Merton [1974], is a mathematical formula that evaluates a company's structural credit risk by modeling its equity as a call option on its assets. The model is prevalent among stock analysts and commercial loan officers, who use it to calculate PDs for companies. Beran and Djaïdja [2007] use the Logistic-Exponential mixture model, while Carling et al. [2007] employs the Cox PH model as a substitute for the Merton model. Both studies identify various advantages to modeling default time with survival analysis: survival models accommodate censored data and incorporate time-varying explanatory variables, resulting in unbiased estimates and more predictive power. Baesens et al. [2005] and Sarlija et al. [2009] explore and provide a comparison between neural network-based credit risk models and the Cox PH model. Although the neural network-based approach is better at handling complex terms such as interaction terms between inputs, quadratic terms, and so forth, it did not significantly outperform the Cox PH model.

According to the studies discussed thus far, survival analysis is at least as good at classification as traditional credit scoring approaches. However, survival analysis adds an extra dimension that allows the estimation of default times. Forecasts of default levels as a function of time are crucial for debt provisioning and profit scoring.

In addition, in the presence of censored data, traditional approaches such as Merton's firm value method for corporate defaults or logistic regression for retail consumer defaults are not applicable. Several studies have proposed techniques to enhance the performance of survival analysis-based credit-scoring models by focusing on the quality of explanatory variables.

Bellotti and Crook [2009] explore the hypothesis that the PDs of credit card customers are affected by general economic conditions. In their study, input to the Cox PH regression comprises time-varying macroeconomic variables and their interactions. On the contrary, for classical credit scoring approaches, which yield static PD snapshots for a specific period, incorporating time-varying covariates is impractical. According to the findings, the Cox PH model with time-varying macroeconomic variables significantly improves predictive performance. Luo et al. [2016] and Dirick et al. [2017], in their studies to investigate default and attrition rates for a portfolio of credit card customers, use regression splines to incorporate non-linear effects of continuous predictors in the Cox PH model. The Cox regression model augmented with regression splines outperformed the standard Cox PH model in both studies. Djeundje and Crook [2019] investigate whether standard survival models can be improved through Generalised Additive Models (GAMs). In their study, the authors use GAMs to formulate predictors as combinations of "flexible univariate functions of risk factors" for credit card accounts. In the study, GAMs are parameterised using penalised splines. According to the authors, GAMs improve the performance of survival models significantly, resulting in more accurate predictions.

Retail credit data is frequently associated with a low number of defaults [Florez-Lopez and Ramon-Jeronimo, 2014]. Due to the limited amount of data, binary classification approaches may have difficulty determining the default distribution. Furthermore, regardless of the time frame considered, a substantial fraction of borrowers will not default. This subgroup of long-term survivors is called the immune or "cured" fraction [Peng, 2003]. Mixture cure models are used to investigate the heterogeneity between borrowers that are "cured" of default and those that are not [Kuk and Chen, 1992]. The survivor function of these models is a mixture of two models, namely, the incidence model and the latency model. Whether a

borrower is non-susceptible (long-term survivor) or susceptible is indicated and defined by the susceptibility binary variable [Farewell, 1986]. This binary variable adds an extra dimension to the default classification. The incidence model estimates the proportion of susceptible borrowers. A binary regression model having either the Logit, Probit, or log-log link function is used as an incidence model. The latency model predicts when default occurs. In other words, it calculates the conditional probability that a borrower will survive past a certain point, given that default eventually occurs. Latency can be modelled parametrically or semi-parametrically. Farewell [1982] reanalysed a toxicological study by postulating a fraction of long-term survivors. The author employs the Logistic-Weibull mixture cure model in the study, which models latency parametrically using the Weibull distribution and incidence using logistic regression. Other studies that model latency parametrically include Yamaguchi [1992] and Peng et al. [1998] which consider the Generalised Gamma models and the Generalised F distribution, respectively.

Another class of mixture cure models that have received much attention in the literature models latency semi-parametrically. The most widely used model is the Logistic-Cox PH model, which keeps the logistic form for incidence while using the Cox PH model for latency. This model is implemented in several studies for retail credit scoring with immunes. More notably, Tong et al. [2012] and Dirick et al. [2017] use this model for scoring personal loan and credit card customers, respectively. These studies show that the Logistic-Cox PH model outperforms the Cox PH, AFT, and logistic regression models.

Survival analysis has gone through a number of changes since its initial application to credit scoring by Narain [1992]. The majority of the current literature does not tie survival analysis to credit management on a holistic level. Plenty of existing research focuses on credit scoring as a stand-alone method of managing credit risk. In this study, modern regulatory framework and cutting-edge loan loss provisioning, in conjunction with survival analysis, are used to establish a template for PIT PD prediction that credit providers can adopt on a holistic level in credit risk management. The entire loan portfolio is scored, and the borrowers' survival track record is assessed from the moment their loans are funded to when they mature.

The data for this study was obtained from Prosper, one of the top peer-to-peer (P2P) loan platforms which account for a sizable portion of the money lending sector. For instance, since the end of 2019, Prosper and another P2P platform, Lending Club, have lent a total of $76 billion [Fitzpatrick and Mues, 2021]. As with traditional retail scoring, P2P lending platforms assess potential borrowers against their own approval standards. Borrowers are scored and assigned a risk rating based on their attributes, loan amount requested, and credit history following approval, and the loan is then listed on the platform. P2P lending institutions are differentiated from traditional lending banks in that credit is extended by P2P investors, who assume all the risks and rewards.

Due to the expanding P2P lending market, which has drawn a large amount of interest in that sector, a corpus of research for P2P loan profit scoring is emerging. Berkovich [2011] has examined data from Prosper in a study to investigate investors' herding effects in P2P lending. A herding behaviour occurs when an individual decides to follow others and emulate group behaviour rather than making independent and atomistic decisions based on their private information. By analysing the variance of the difference between the price and the ex-post return on Prosper loans, the findings reveal indications of herding behaviour. Bachmann et al. [2011] utilise Prosper data for research on the determinants of P2P lending and the likelihood of successful funding for borrowers. Based on an analysis of Prosper transaction data, Chen et al. [2016] evaluated the relationship between individuals' group social capital and their loan outcomes in the online P2P financial credit market. According to the findings of their study, the borrower's general group social capital and relational capital have conflicting effects, and the borrower's structural social capital has a negative impact on funding and payment performance. Mariotto [2016] explored Prosper and Lending Club's competitive strategies and contrasted how they differ from one another and retail banks. The investigation is facilitated by publicly available data from the Prosper and Lending Club websites. The author concludes that these platforms substitute one another and are direct competitors.

This study employs survival analysis to calculate PIT PDs under IFRS 9 for a Prosper loan portfolio. As far as we know, no analogous published work has used the same models on this data.

2.3 Overview of Credit Loss Reporting

The expected loss (EL) is the amount of money a bank or a lender can expect to lose on average over a set period when it extends credit to its customers. The volatility of credit losses around EL is called unexpected loss (UL). To maintain its long-term fiscal viability, it is imperative that a lender's capital reserves are commensurate with the level of credit risk it is taking. Banking supervision standards calculate regulatory capital as the sum of EL and UL. Banks that adopt the IRB approach use their own internal assessments of EL and UL. The computation of both of these losses requires the estimation of risk parameters, in addition to PD. These include loss-given default (LGD), exposure at default (EAD), and effective maturity (M).

EAD measures the expected gross exposure of the credit facility upon default, while M is the contractual maturity of a credit facility. LGD reflects the percentage loss of the total exposure in the event of default. When default happens, the bank has recourse to the market value of any collateral pledged by the borrower. According to the BCBS, LGD calculations must be based on historical recovery rates to account for banks' possible failure to take ownership of the collateral and liquidate the collateral quickly. Additional adjustments, typically tied to currency mismatch or economic downturn, are incorporated to obtain final LGD estimates. The foundation approach requires banks to provide only internal PD estimates and consult an ECAI for the other risk parameters. By contrast, the advanced approach requires banks to determine all risk parameters internally.

The IFRS 9 framework has two methods for calculating ECL: the General approach and the Simplified approach. The General approach, which is by far the most intricate and extensively adopted, uses a three-stage ECL framework for the recognition of impairments on a forward-looking basis [Frykström and Li, 2018]. The Stage 1 bucket comprises loans that are currently performing. These loans have not experienced a significant increase in credit risk since their initial recognition or have low credit risk at the reporting date. Banks are required to reserve one-year ECL for these loans. A one-year ECL is a loss resulting from default events occurring twelve months after the reporting date. Those loans that have shown a significant increase in credit risk since initial recognition but no objective evidence of impairment belong to

the Stage 2 bucket. Lifetime ECL is reserved for these loans. Lifetime ECL considers all potential default events throughout the expected life of the loan facility. The Stage 3 bucket comprises loans that have shown objective evidence of impairment as of the reporting date. A lifetime ECL is set aside for these loans [Bellini, 2019].

Table 2.1 provides a comprehensive summary of the stage allocation and the respective ECL requirements under the General approach. The Simplified approach disregards

Table 2.1: IFRS 9 General approach.

Description	**Stage 1**	**Stage 2**	**Stage 3**
Loss allowance updated at each reporting date	**12-month expected credit losses**	**Lifetime expected credit losses**	
Life expected credit losses criterion (and adjust for change in credit risk since initial recognition)		**Credit risk has increased significantly since initial recognition**	
	individual or collective basis		**credit impaired**
Interest revenue calculated based on	Effective interest rate on gross carrying amount	Effective interest rate on gross carrying amount	Effective interest rate on gross carrying amount less allowance

stage allocations and offers lifetime ECL on initial recognition of these assets, making it more suited for short-term receivables that do not require tracking of movement in credit risk. Note that both Basel's IRB approach and IFRS 9's General approach (in particular Stage 1) require estimation of PD based on a twelve-month horizon [Bellini, 2019; Hashim et al., 2016]. The final PD is calibrated somewhere between the point-in-time (PIT) and TTC estimates under the IRB approach, while only PIT adjustments are relevant under IFRS 9 [Edwards, 2014]. For loans that fall into Stage 2 and Stage 3 buckets, IFRS 9 requires the calculation of ECL over the lifetime of these financial assets. The lifetime ECL is given by:

$$\text{ECL}(\mathbf{x}^*) = \sum_{i=1}^{n_{pt}} \left[\sum_{t=1}^{T_i} \left\{ v^t \cdot P_i(T=t) \cdot \text{LGD}_{i,t} \cdot \text{EAD}_{i,t} | \mathbf{x}_t^* \right\} \right], \tag{2.1}$$

where n_{pt} is the number of loans in a given portfolio, $\mathbf{x}^*$ is a vector of macroeconomic scenarios, v is the discount factor, $t = 1, \ldots, T_i$ are the remaining years of a loan until

maturity T_i, and $P_i(T = t)$ is the probability that a borrower's default time T is in year t, implying that he or she has not defaulted in the years $j = 1, \ldots, t-1$ [Bellini, 2019]. The calculation of ECL for IFRS 9 shown in Equation 2.1 requires credit scoring methods that take into account the loan's lifetime and provide PD estimates, $P_i(T = t)$, throughout the remaining term of the loan.

South African banks abide by the guidelines that are stipulated by the National Credit Act (NCA) when issuing loans to customers. The NCA fosters "a fair, transparent, competitive, responsible, and accessible credit market and industry" and entrenches several fundamental rights aimed at consumer protection [Renke et al., 2007]. However, the NCA does not provide conditions that constitute default. As a result, South African banks adhere to the BCBS's definition of default which stipulates that a borrower has defaulted when either one of the following two situations arise:

1. The bank has assessed the borrower cannot repay all loan commitments in full.
2. The borrower has missed three consecutive monthly payments.

Under Section B5.5.37 of the IASB (in relation to IFRS 9), default is not defined explicitly, but instead, banks are required to accord with internal credit risk management. The FASB asserts that "Estimating expected credit losses is highly judgemental and generally will require an entity to make specific judgments. Those judgements may include any of the following: The definition of default for default-based statistics ..." [Bellini, 2019]. This thesis adheres to the default guidelines provided by BCBS, and PD estimates are determined internally as part of a bank's internal credit risk management.

2.4 Theoretical Review of Survival Analysis

Censoring transpires when information about the borrower's behaviour is only partially known. It can occur when a study participant does not experience the event of interest during the study period, is "lost to follow-up" or exits the study (due to competing risks). These three scenarios constitute right censoring, meaning that the individual's "survival time becomes incomplete at the right side of the follow-up period" [Kleinbaum and Klein, 2010]. There are two other types of censorship: left

censoring and interval censoring. Left censoring refers to a situation where an individual's actual survival time is at most equal to the follow-up time. Individuals with survival times which fall within a specified interval are subject to interval censoring. Right censoring is ubiquitous in the context of consumer credit [Dirick et al., 2017]. In this context, let $(0, \tau]$ be the study period. For a retail loan portfolio, Figure 2.1 depicts what can occur for a borrower under right censoring:

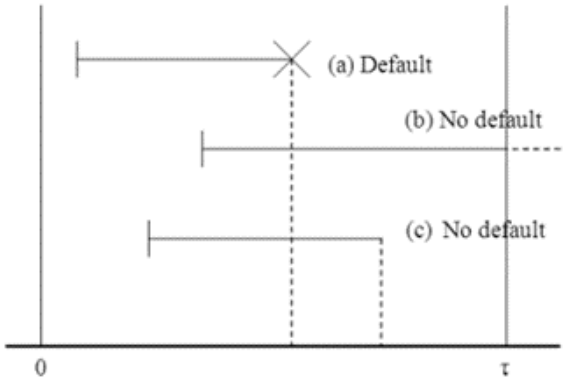


Figure 2.1: Right censoring

Scenario (a) depicts a borrower defaulting at an arbitrary time $T \in (0, \tau]$, called time to default, and is referred to as uncensored. Scenario (b) depicts a borrower that does not default during the study period, while scenario (c) illustrates a case where the borrower ceases to be a customer because of competing risks such as early repayment, changing the lender, or death. Scenarios (b) and (c) are censored and collectively denoted by a random variable C.

Let T be the response variable that denotes time to default for a loan facility. The survivor function, $S(t)$, given by:

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

is the probability that a borrower has not defaulted up to time t and $f(t)$ is the probability density function of T evaluated at t. The survivor function, $S(t)$, is a monotonically decreasing function of t and is always equal to 1 at the time of credit

facility initiation, as the borrower has not had enough time to default. Let Δt be a small interval in the neighbourhood of t. The hazard function, $h(t)$, given by:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P\left(t < T \leq t + \Delta t | \, T > t\right)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(T \leq t + \Delta t) - P(T < t)}{\Delta t \cdot P(T > t)} \\ &= \frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \\ &= \frac{1}{S(t)} \frac{dF(t)}{dt} \\ &= \frac{f(t)}{S(t)}, \end{aligned} \tag{2.2}$$

is the instantaneous PD, assuming the borrower has not defaulted until time t. While $S(t)$ is right continuous and non-increasing, $h(t)$ can take on various shapes in response to risk fluctuations [Moore, 2016]. The cumulative hazard until time t is calculated as follows:

$$H(t) = \int_0^t h(u)du = \int_0^t \frac{f(u)}{S(u)}du = -\int_0^t \left(\frac{\frac{dS(u)}{du}}{S(u)}\right) du = -\ln S(t),$$

giving:

$$S(t) = \exp\left[-H(t)\right] = \exp\left[-\int_0^t h(u)du\right]. \tag{2.3}$$

The hazard function can be used as a vehicle for implementing mathematical modelling of survival data. Given either one of the functions $S(t)$, $h(t)$, or $H(t)$, the other two can be derived. The survival probability is estimated either non-parametrically, semi-parametrically, or parametrically [Kleinbaum and Klein, 2010].

By re-arranging Equation 2.2, note that $f(t) = h(t)S(t)$, giving:

$$\mathbb{E}[H(t)] = \int_0^\infty H(t)h(t)\exp\left[-H(t)\right]dt = \int_0^\infty u\exp\left(-u\right)du = 1.$$

The time by which q percent of borrowers default is the q^{th} quantile of the population, t_q, calculated as:

$$t_q = S^{-1}(1 - q).$$

Using the notation for the random default time and censoring time, T and C, respectively, let the values t_i and c_i be their corresponding realisations in the data sample. The default status for the i^{th} borrower is denoted by a binary variable:

$$\delta_i = \begin{cases} 1, & \text{if borrower defaults } (t_i \leq c_i), \\ 0, & \text{if borrower does not default } (t_i > c_i). \end{cases} \tag{2.4}$$

The earliest event time, which is either default time or censoring time, is denoted by the observed response variable $y_i = \min(t_i, c_i)$. If a' is any cut-off time, the expected value of the cumulative hazard function is:

$$\begin{aligned} \mathbb{E}[H(\min(T, a'))] &= \mathbb{E}[1_{\{T \leq a'\}} H(t) + 1_{\{T > a'\}} H(a')] \\ &= \mathbb{E}[1_{\{T \leq a'\}} H(t)] + H(a')S(a'), \end{aligned} \tag{2.5}$$

where:

$$\begin{aligned} \mathbb{E}[1_{\{T \leq a'\}} H(t)] &= \int_0^\infty H(t)[t \leq a']h(t) \exp\left[-h(t)\right]dt \\ &= \int_0^{a'} H(t)h(t) \exp\left[-H(t)\right]dt \\ &= -\left[u \exp\left(-u\right) + \exp\left(-u\right)\right] |_0^{H(a')} \\ &= 1 - S(a')\left[H(a') + 1\right]. \end{aligned}$$

For the response information $\{(y_i, \delta_i); i = 1, \ldots, n\}$, an estimate of expected number of defaults until cut-off time a' is given by $\sum_{i=1}^n H(\min(t_i, a'))$ [Harrell et al., 2001].

2.4.1 The Likelihood Function for Censored Data

Let $t_1, t_2, \ldots, t_n$ denote default times for a sample of n borrowers. Suppose there is no censoring, i.e., $c_i = 0$, then $y_i = t_i$ for all $i = 1, 2, \ldots, n$. The likelihood function of the observed data is given by:

$$\mathcal{L}(t_1, t_2, \ldots, t_n) = f(t_1) \cdot f(t_2) \ldots f(t_n) = \prod_{i=1}^n f(t_i).$$

Modifications to the likelihood function are applied if some observations are right-censored. In that case, the survivor function forms part of the likelihood function to indicate that survival time is only known to exceed a specific value. As a result, the likelihood function becomes:

$$\mathcal{L}(t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i). \tag{2.6}$$

2.4.2 Kaplan-Meier Estimator

The exact functional form of the population survivor function is rarely known. It can, however, be approximated without relying on any distributional assumptions about default times. The non-parametric Kaplan-Meier (KM) estimator, also known as the product-limit estimator, is the simplest model for the survivor function. For a total of D defaults occurring at distinct ordered default times $t_1, t_2, \ldots, t_D$, the KM estimator for the survivor function is given by:

$$S_{\text{KM}}(t) = \prod_{i:t_i \leq t}^{D} \left(1 - \frac{d_i}{n_i}\right), \tag{2.7}$$

where d_i is the number of defaults at t_i and n_i is the number of borrowers at risk at time t_i [Harrell et al., 2001]. The estimator in Equation 2.7 is a non-increasing step function with discontinuities at default times. Thus, the KM survivor curve is a useful graphical representation of the survival data because it illustrates when default and censoring happen. Using the relation established by Equation 2.3, the KM estimator of $H(t)$ is $H_{\text{KM}}(t) = -\ln S_{\text{KM}}(t)$.

2.4.2.1 Variance Estimation

Greenwood et al. [1926] and Klein and Moeschberger [2003] use the delta method to construct a confidence interval of the product-limit estimator, firstly by computing the

variance of $\ln\left[S_{\text{KM}}(t)\right]$ as:

$$\widehat{\sigma}^2\left(\ln\left[S_{\text{KM}}(t)\right]\right) \approx \sum_{i:t_i \leq t}^{D} \frac{d_i}{n_i(n_i - d_i)}.$$

Using the delta approach once more, the Greenwood variance estimator for a KM survivor curve is approximated by:

$$\widehat{\sigma}^2\left(S_{\text{KM}}(t)\right) \approx \left[S_{\text{KM}}(t)\right]^2 \sum_{i:t_i \leq t}^{D} \frac{d_i}{n_i(n_i - d_i)}.$$

The confidence interval of the survivor curve is obtained by assuming asymptotic normality for the KM survivor function [Moore, 2016]. The main disadvantage of this method is that the confidence intervals may stretch beyond the $(0, 1)$ interval. Kalbfleisch and Prentice [2011] recommend using the complementary log-log transform of $S_{\text{KM}}(t)$ to compute the variance estimator as:

$$\widehat{\sigma}^2\left(\ln\left[-\ln S_{\text{KM}}(t)\right]\right) = \frac{\sum_{i:t_i \leq t}^{D} \frac{d_i}{n_i(n_i - d_i)}}{\{\sum_{i:t_i \leq t}^{D} \ln\left[\frac{n_i - d_i}{n_i}\right]\}^2} \approx \frac{1}{\left[\ln S_{\text{KM}}(t)\right]^2} \sum_{i:t_i \leq t}^{D} \frac{d_i}{n_i(n_i - d_i)},$$

which ensures that confidence limits are within the bounds.

2.4.2.2 Altschuler-Aalen-Nelson Estimator

An alternative nonparametric estimator of the cumulative hazard function, $H(t)$, is the Altschuler-Aalen-Nelson estimator, proposed by Altshuler [1970], Nelson [1972], and Aalen [1976], which is calculated as:

$$\widehat{H}_{\text{A}}(t) = \sum_{i:t_i \leq t}^{D} \frac{d_i}{n_i} = \sum_{i:t_i \leq t}^{D} h_{\text{A},i}(t),$$

giving the survivor function estimator $S_{\text{A}}(t) = e^{-\widehat{H}_{\text{A}}(t)}$. It is worth noting that the Altschuler-Aalen-Nelson estimate is equivalent to the expected number of defaults, i.e., $\sum_{i=1}^{n} H_{\text{A}}(y_i) = \sum_{i=1}^{n} \delta_i$. Moreover, the variances of $S_{\text{KM}}(t)$ and $S_{\text{A}}(t)$ are

assymptotically equal [Harrell et al., 2001].

2.4.2.3 Using Kernels to obtain Smooth Hazard Estimate

The true shape of the hazard function can be revealed by applying a kernel function $K(u;\vartheta)$ centred at each default time [Moore, 2016]. Given D ordered, distinct default times $t_1, t_2, \ldots, t_D$ and the smoothing hyperparameter ϑ, the kernel-based hazard function estimator is given by:

$$\widehat{H}(t) = \frac{1}{\vartheta} \sum_{i=1}^{D} K\left(\frac{t - t_i}{\vartheta}\right) \frac{d_i}{n_i}. \tag{2.8}$$

Note that when $K(u;\vartheta) = 1$, the hazard function in Equation 2.8 becomes the Altschuler-Aalen-Nelson cumulative hazard estimator. By smoothing the hazard function, one can obtain a smooth estimate of the survivor curve. Several specialised applications require a smoothed survivor curve, typically using the Epanechnikov kernel, $K(u) = \frac{3}{4}(1 - u^2)$ for $|u| \leq 1$ [Epanechnikov, 1969]. However, the usual step function S_{KM} (and S_{A}) is preferred since it is a maximum likelihood estimator of the survivor function. The main criticism of the KM estimator is that it assumes borrowers in the data sample have the same survivor function, thus failing to incorporate heterogeneity amongst borrowers; hence it is typically used only as a diagnostic tool.

2.4.2.4 Comparing Survival Curves

When there are two survival groups, it may be worthwhile to assess how distinct they are from one another. The most prevalent tests are nonparametric log-rank tests such as the Mantel-Haenszel test and Wilcoxon-Gehan test. For each of these, the test statistic is calculated based on the ranks of survival times [Emmert-Streib and Dehmer, 2019]. Let $n_{0,i}$ and $n_{1,i}$ be the borrowers at risk at time t_i for group 0 and group 1, respectively, and $d_{0,i}$ and $d_{1,i}$ be the corresponding number of defaults such that $n_i = n_{0,i} + n_{1,i}$ and

$d_i = d_{0,i} + d_{1,i}$. The distribution of $d_{0,i}$ follows a hypergeometric distribution:

$$p(d_{0,i}|n_{0,i}, n_{1,i}, d_{1,i}) = \frac{({}^{n_{0,i}}C_{d_{0,i}})({}^{n_{1,i}}C_{d_{1,i}})}{{}^{n_i}C_{d_i}},$$

where ${}^nC_d = \frac{n!}{d!(n-d)!}$ is the number of d-combinations from a set of n subjects. The mean and variance of $d_{0,i}$ are given by:

$$e_{0,i} = \mathbb{E}[d_{0,i}] = \frac{n_{0,i}d_i}{n_i} = \left[\frac{n_{0,i}}{n_{0,i} + n_{1,i}}\right](d_{0,i} + d_{1,i}) = \text{Prob}\left\{\text{group } 0\right\} \times (d_{0,i} + d_{1,i}),$$

and

$$v_{0,i} = \sigma^2(d_{0,i}) = \frac{n_{0,i}n_{1,i}d_i(n_i - d_i)}{n_i^2(n_i - 1)},$$

respectively. Fleming and Harrington [1984] define the log-rank test, $G(\rho)$, using a series of n weights, $w_i = n\{S_{KM}(t_i)\}^\rho$, as:

$$G(\rho) = \frac{U_0^2}{V_0} \sim \chi_1^2,$$

where $U_0(w) = \sum_{i=1}^N w_i\,(d_{0,1} - e_{0,i})$ and $V_0(w) = \sum_{i=1}^N w_i v_{0,i}$. Fleming and Harrington [1984] states that the test statistic used for conducting the log-rank test, $G(\rho)$, follows a Chi-square distribution with 1 degree of freedom. There are several $G(\rho)$ variants, depending on the value of ρ. For instance, the log-rank test becomes the Mantel-Haenszel test when $\rho = 0$, while the Prentice modification (also known as Peto-Peto) of the Wilcoxon-Gehan test uses $\rho = 1$.

2.4.3 Cox Proportional Hazards Model

The Cox PH model, developed by Cox [1972], is the most extensively utilised semi-parametric survival analysis model. It comprises a non-parametric baseline hazard function, as well as a parametric component that takes into account the effect of covariates. The covariates in credit scoring are either application, behavioural, or, in some cases, external macroeconomic variables. Under Cox PH, the hazard function

is given by:

$$h(t|\mathbf{x}) = h_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{x}), \tag{2.9}$$

where $h_0(t)$ is the time-dependent baseline or underlying hazard function, while $\mathbf{x} = (x_1, x_2, \ldots, x_p) \in \mathbb{R}^p$ is a p-dimensional vector of predictors. The baseline hazard function, $h_0(t)$, is not subject to any parametric assumption. The parametric component, on the other hand, is a linear function of covariate information in a p-dimensional vector $\mathbf{x}$. In particular, covariates have the effect of multiplying the baseline hazard by a factor $\exp(\boldsymbol{\beta}^\top \mathbf{x})$. The regression coefficient for $x_j \in \mathbf{x}$, given by:

$$\begin{aligned}\beta_j &= \ln\left[\frac{h(t|x_1, x_2, \ldots, x_j + 1, x_{j+1} \ldots, x_p)}{h(t|x_1, x_2, \ldots, x_j, \ldots x_p)}\right] \\ &= \ln h(t|x_1, x_2, \ldots, x_j + 1, x_{j+1} \ldots, x_p) - \ln h(t|x_1, x_2, \ldots, x_j, \ldots x_p),\end{aligned}$$

is the increase in log hazard at any fixed point in time for a unit increase in x_j while keeping all other predictors constant.

Similarly, denoting the baseline survivor function by $S_0(t)$, the borrower-specific survivor function is modelled taking into account covariates as:

$$S(t|\mathbf{x}) = \exp\left[-H_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{x})\right] = S_0(t)^{\exp(\boldsymbol{\beta}^\top \mathbf{x})}.$$

It is worth noting that the parameter $\psi = \exp(\boldsymbol{\beta}^\top \mathbf{x})$ depicts the differences between two survivor distributions through $S(t|\mathbf{x}) = [S_0(t)]^\psi$. In the scenario where the only covariate is a binary predictor, $S_0(t)$ represents the survivor function of the control subgroup, whereas $S_1(t)$ represents the survivor function of the treatment subgroup. As part of a bank's customer retention policy, one might be interested in knowing the effectiveness of a marketing campaign. In such a case, the predictor may equal 1 for borrowers targeted in the marketing campaign or 0 for those not targeted.

2.4.3.1 Proportional Hazards Assumption

In PH models, the hazard ratio, η, for a borrower whose attributes are $\mathbf{x}^*$ compared to a borrower with attributes $\mathbf{x}$ is given by:

$$\eta = \frac{h(t|\mathbf{x}^*)}{h(t|\mathbf{x})} = \frac{h_0(t)\exp\left(\boldsymbol{\beta}^\top \mathbf{x}^*\right)}{h_0(t)\exp\left(\boldsymbol{\beta}^\top \mathbf{x}\right)} = \exp\left[\boldsymbol{\beta}^\top(\mathbf{x}^* - \mathbf{x})\right].$$

Equivalently, the relation between these borrowers can be expressed as $h(t|\mathbf{x}^*)\exp\left(\boldsymbol{\beta}^\top \mathbf{x}^*\right) = \eta h(t|\mathbf{x})\exp\left(\boldsymbol{\beta}^\top \mathbf{x}\right)$. The PH assumption states that predictors have the same effect across all default times, i.e., η is time-invariant [Kleinbaum and Klein, 2010]. The PH assumption is crucial in the formulation of the likelihood function since it allows the baseline hazard function to cancel out from the likelihood function. If this assumption is violated, the Cox PH regression model is rendered invalid, necessitating the use of more complex approaches such as stratified Cox regression or extended Cox regression. These models can be adjusted to include variables that violated the PH assumption or model non-proportional hazards.

2.4.3.2 Partial Likelihood Estimation

Let t_j denote the j^{th} default time and $R(t_j)$ be a set of borrowers at risk of default just before time t_j. The hazard function of borrower i at default time t_j under the PH assumption is $h_i(t_j) = h_0(t_j)\exp\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right)$. Given a sample $\{(t_i, \delta_i); i = 1, \ldots, n\}$, the probability that borrower i defaults at default time t_j is given by the conditional probability:

$$\begin{aligned}
p_j &= \text{Prob}\left\{\text{borrower } i \text{ defaults at time } t_j | R(t_j) \text{ and one default at } t_j\right\} \\
&= \frac{\text{Prob}\{\text{borrower } i \text{ defaults at } t_j | R(t_j)\}}{\text{Prob}\{\text{one default at } t_j | R(t_j)\}} \\
&= \frac{h_i(t_j)}{\sum_{k\in R(t_j)} h_k(t_j)} \\
&= \frac{h_0(t_j)\exp\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right)}{\sum_{k\in R(t_j)} h_0(t_j)\exp\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)} \\
&= \frac{\exp\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right)}{\sum_{k\in R(t_j)} \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)},
\end{aligned}$$

which is not dependent on the baseline hazard function, $h_0(t_j)$. This approach of computing p_j, which does not prejudge the form of the baseline hazard function, is widely used in credit scoring [Thomas et al., 2001]. After defaulting at time t_j, borrower i, as well as censored borrowers in the interval $(t_j, t_{j+1}]$, are dropped from the risk set $R(t_j)$. Note that, if predictors have no impact, the conditional probability simplifies to $p_j = \frac{1}{n_j}$, where n_j is $|R(t_j)|$, the cardinality of $R(t_j)$. Moreover, the p_js are conditionally independent for different default times [Harrell et al., 2001]. The likelihood is given by:

$$\mathcal{L}(\boldsymbol{\beta}; t_1, t_2, \ldots, t_D) = \prod_{y_i \text{ uncensored}} p_j = \prod_{j=1}^{D} p_j = \prod_{j=1}^{D} \frac{\exp\left(\boldsymbol{\beta}^\top \mathbf{x}_j\right)}{\sum_{k \in R(t_j)} \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)}. \tag{2.10}$$

This likelihood is known as a partial likelihood since it only includes information about uncensored borrowers. For an accurate maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$, Cox [1975] recommends treating the partial likelihood as if it were an ordinary likelihood. Kalbfleisch and Prentice [2011] show that in the absence of predictors, Equation 2.10 is effectively the marginal likelihood of the ranks of the observations, where the order in which defaults occur is considered instead of the actual times. The partial log-likelihood function is given by:

$$\ell(\boldsymbol{\beta}; t_1, t_2, \ldots, t_D) = \sum_{j=1}^{D} \left\{ \boldsymbol{\beta}^\top \mathbf{x}_j - \ln \left[\sum_{k \in R(t_j)} \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right) \right] \right\}.$$

Fisher's score vector $\mathbf{U}(\boldsymbol{\beta})$, which is the first derivative of $\ell(\boldsymbol{\beta})$, has p components, one for each of the p explanatory variables. The maximum partial log-likelihood estimate of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$, is estimated via $\arg\max \ell(\boldsymbol{\beta}; t_1, t_2, \ldots, t_D)$ or by solving for the partial log-likelihood estimate of $\boldsymbol{\beta}$ such that:

$$\ell'(\boldsymbol{\beta}; t_1, t_2, \ldots, t_D) = \mathbf{U}(\boldsymbol{\beta}) = \sum_{j=1}^{D} \left\{ \mathbf{x}_j - \frac{\sum_{k \in R(t_j)} \mathbf{x}_j \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_j\right)}{\sum_{k \in R(t_j)} \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)} \right\} = 0,$$

using an iterative algorithm such as Newton-Raphson. If the log partial likelihood is concave, then $\mathbf{U}(\widehat{\boldsymbol{\beta}}) = (0, 0, \ldots, 0)$. Given all pairwise combinations of the covariates,

Fisher's information matrix or Hessian matrix is given by:

$$I(\boldsymbol{\beta}; \mathbf{x}) = \mathbb{E}_{\boldsymbol{\beta}} \left[\mathbf{U}(\boldsymbol{\beta})^\top \mathbf{U}(\boldsymbol{\beta}) \right] = \left\{ -\mathbb{E}_{\boldsymbol{\beta}} \left[\frac{\partial^2}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j} \ell(\boldsymbol{\beta}) \right] \right\}_{i,j=1,2,\ldots,p}.$$

The maximum partial likelihood estimate, $\widehat{\boldsymbol{\beta}}$, is consistent and asymptotically follows a Gaussian distribution, with the covariance matrix given by the inverse of the observed information $I^{-1}(\widehat{\boldsymbol{\beta}}; \mathbf{x})$.

2.4.3.3 Goodness of Fit Assessment

Hypotheses regarding log hazard ratios are tested using a number of goodness-of-fit (GOF) metrics [Harrell et al., 2001; Moore, 2016]. The Wald test is based on a large sample distribution of partial likelihood estimates and is arguably the most extensively used test. The Wald test statistic for testing the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ is given by:

$$W = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right)^\top I(\widehat{\boldsymbol{\beta}}; \mathbf{x}) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right) \sim \chi^2_{p-1},$$

where p is number of covariates in the model.

Another prevalent test is the Score test, and its test statistic is given by:

$$S(\boldsymbol{\beta}_0) = \mathbf{U}^\top(\boldsymbol{\beta}_0) I^{-1}(\boldsymbol{\beta}_0; \mathbf{x}) \mathbf{U}(\boldsymbol{\beta}_0) \sim \chi^2_{p-1},$$

which does not depended on $\widehat{\boldsymbol{\beta}}$. The Score test for evaluating the significance of a binary covariate is comparable to the Mantel-Haenszel test for investigating similitudes between two survivor functions if the model comprises a single binary predictor that characterises two groups.

The likelihood ratio test is another widely used metric, and its test statistic is given by:

$$\lambda_{\text{LR}} = 2 \left\{ \ell(\widehat{\boldsymbol{\beta}}) - \ell(\boldsymbol{\beta}_0) \right\} \sim \chi^2_{p-1}. \tag{2.11}$$

In Equation 2.11, $\ell(\boldsymbol{\beta}_0)$ is the reduced model where predictors have no effect. This metric is invariant to the various transformation of $\boldsymbol{\beta}$, meaning that hypothesis testing computed in terms of $\boldsymbol{\beta}$ or $\exp \boldsymbol{\beta}$ does not influence p-values.

2.4.3.4 Handling Ties

There are two possible ways in which tied event times may occur. If the underlying data is continuous, rounding may cause ties. Another manner is via event times that are genuinely discrete [Moore, 2016]. If the sample comprises tied default times, the actual partial likelihood function contains permutations of tied events, which might be computationally expensive. Let $D(t_i)$ denote a set of borrowers defaulting at time t_i and $\alpha_i = \sum_{j \in D(t_i)} \mathbf{x}_j$. If the sample contains tied default times, then $k < n$, yielding $t_1, t_2, \ldots, t_k$ distinct ordered default times. Breslow [1972] proposes an approximate log partial likelihood function given by:

$$\ell(\boldsymbol{\beta}; t_1, t_2, \ldots, t_k) = \sum_{i=1}^{k} \left\{ \boldsymbol{\beta}^\top \alpha_i - d_i \ln \left[\sum_{y_j \geq t_i} \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_j\right) \right] \right\},$$

where $d_i = |D(t_i)|$. Essentially, this approximation treats tied default times as coming from the same risk set. Breslow [1972] also provides a generalisation of the Altschuler-Nelson-Aalen estimator that incorporates covariates, giving the baseline cumulative hazard function:

$$\widehat{H}_0(t) = \sum_{i:t_i<t} \frac{d_i}{\sum_{y_j \geq t_i} \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_j\right)}.$$

Thus, for any p-dimensional covariate vector $\mathbf{x}$, the semi-parametric estimator of the cumulative hazard function, $\widehat{H}(t, \mathbf{x})$, and survival probability, $\widehat{S}(t, \mathbf{x})$, are given by:

$$\widehat{H}(t|\mathbf{x}) = \widehat{H}_0(t) \exp\left(\widehat{\boldsymbol{\beta}}^\top \mathbf{x}\right),$$

and

$$\widehat{S}(t|\mathbf{x}) = \exp\left[-\widehat{H}_0(t) \exp\left(\widehat{\boldsymbol{\beta}}^\top \mathbf{x}\right)\right],$$

respectively. An even more accurate approximation to the actual log partial likelihood is proposed by Efron [1977]:

$$\ell(\boldsymbol{\beta}; t_1, t_2, \ldots, t_k) = \sum_{i=1}^{k} \left\{ \boldsymbol{\beta}^\top \sum_{j \in D(t_i)} \mathbf{x}_j - \sum_{j=1}^{d_i} \ln \left[\sum_{y_j \geq t_i} \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_j\right) - \frac{j-1}{d_i} \sum_{l \in D(t_i)} \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_l\right) \right] \right\}.$$

This approximation considers all possible ways to break ties and adjusts the risk level accordingly. The results provided by this approach are remarkably similar to those obtained using the more time-consuming permutation method.

2.4.4 Stratified Cox Model

The "stratified Cox model" is a variant of the Cox PH model in which covariates that do not satisfy the PH assumption are used for stratification [Kleinbaum and Klein, 2010]. The model includes covariates that satisfy the PH assumption but exclude stratification variables. Assuming j covariates meet the PH assumption, whereas k covariates do not, the covariate vector $\mathbf{x}$ can be factored into $(\mathbf{v}, \mathbf{z})$, where $\mathbf{v} = (v_i, v_2, \ldots, v_j)$ is a vector of covariates that satisfy the PH assumption and $\mathbf{z} = (z_1, z_2, \ldots, z_k)$ is a vector of covariates that do not. If z' is the new stratification variable derived from combinations of z variables, then variable z' has k' categories, where k' is the total number of strata. Within the framework of stratified Cox regression, there are "interaction" and "no-interaction" models [Ata and Sözer, 2007].

2.4.4.1 No-Interaction Stratified Cox Model

The no-interaction model is defined by a set of k' hazard functions:

$$\begin{aligned} h_1(t|\mathbf{v}) &= h_{0,1}(t)\exp\left(\boldsymbol{\beta}^\top \mathbf{v}\right) \\ h_2(t|\mathbf{v}) &= h_{0,2}(t)\exp\left(\boldsymbol{\beta}^\top \mathbf{v}\right) \\ &\vdots \\ h_{k'}(t|\mathbf{v}) &= h_{0,k'}(t)\exp\left(\boldsymbol{\beta}^\top \mathbf{v}\right). \end{aligned}$$

The strata comprise different categorisations of the variable z' that do not form part of the regression model. Despite the baseline hazard function, $h_{0,g}(t)$ for $g = 1, \ldots, k'$, being different for each stratum, log hazard ratios are identical across the strata. Note that log hazard ratio estimates are derived by maximising the likelihood function, which, in this case, is a product of the likelihood functions of each stratum.

The interaction model differs from the no-interaction model by containing product terms [Ata and Sözer, 2007]. The regression coefficients in the no-interaction stratified Cox regression model do not differ between strata. This property is called the no-interaction assumption. For each stratum, different coefficients are derived when interactions are considered. The likelihood ratio test is used to investigate the no-interaction assumption.

2.4.4.2 Interaction Stratified Cox Model

There are two ways in which the interaction model can be developed. Suppose only one variable $z = (0, 1, 2, \ldots, k-1)$ violates the PH assumption. The dataset is stratified into k strata according to that variable, giving:

$$h_g(t|\mathbf{v}) = h_{0,g}(t)\exp\left[\sum_{i=1}^{j}\beta_{i,g}v_i\right],\ g = 1, 2, \ldots, k. \tag{2.12}$$

Alternatively, covariates that meet the PH assumption and their products with the covariate that does not meet the PH assumption can be included in the model. In this

case, the alternative interaction model is given by:

$$h_g(t|\mathbf{v}) = h_{0,g}(t) \exp\left[\sum_{i=1}^{j} \beta_i' v_i + \sum_{q=1}^{2j} \beta_{j+q}'(v_q z)\right] \ , \ g = 1, 2, \ldots, k.$$

There is a relationship between the coefficients of the interaction model and its alternate form, and it is given by:

$$\beta_{1,k} = \begin{cases} \beta_j' \ , \text{ if } k = 1, \\ (k-1)\left(\beta_j' + \beta_{2j}'\right) \ , \text{ if } k > 1. \end{cases}$$

When at least two variables do not meet the PH assumption, a new stratification variable is defined by $z' = (0, 1, 2, \ldots, k' - 1)$. In this case, k' represents the combination of the levels of the variables that do not meet the PH assumption. Alternatively, the dataset can be partitioned into k' strata based on z'. This alternative version is similar to the model in Equation 2.12 but with $g = 1, 2, \ldots, k'$.

2.4.5 Extended Cox Regression Model

The inherent assumption in the partial likelihood theory outlined in Section 2.4.3.2 is that covariate values are determined at inception when the borrower enters the study and stay constant throughout the study [Fisher and Lin, 1999; Moore, 2016]. Failing to account for data that evolve over time causes some variables to violate the PH assumption. An extended Cox regression model provides a way to model non-proportional hazards. Using time-dependent variables, this model specifies a formulation for the time by predictor interaction [Harrell et al., 2001]. A time-dependent variable is one that changes over time for a particular subject. Kalbfleisch and Prentice [2011] provide a distinction between external and internal time-dependent covariates. This classification aids in the interpretation of time-dependent covariate regression models and results. An external covariate is not directly related to the failure mechanism, whereas an internal covariate relates to a variable generated over time by the individual under the investigation.

Time-dependent covariates and their product with a function of time form part of the

extended Cox regression model. Given j and k predictors that meet and do not meet the PH assumption, respectively. Let $\mathbf{v} = (v_1, v_2, \ldots, v_j)$ be a vector of time-independent covariates, $\mathbf{z}(t) = (z_1(t), z_2(t), \ldots, z_k(t))$ be a vector of time-dependent covariates and $\mathbf{x}(t) = (\mathbf{v}, \mathbf{z}) = (v_1, v_2, \ldots, v_j, z_1(t), z_2(t), \ldots, z_k(t))$. The hazard function of the extended Cox regression model is given by:

$$\begin{aligned} h(t|\mathbf{x}(t)) &= \text{Prob}\left\{T \in [t, t+dt) | T \geq t, \mathbf{x}(t)\right\} \\ &= h_0(t) \exp\left[\boldsymbol{\beta}^\top \mathbf{x}(t)\right] \\ &= h_0(t) \exp\left[\boldsymbol{\gamma}^\top \mathbf{v} + \boldsymbol{\theta}^\top \mathbf{z}(t)\right], \end{aligned}$$

where γ and θ are coefficients vectors of covariates $\mathbf{v}$ and $\mathbf{z}(t)$, respectively. According to Klein and Moeschberger [2003], the modified partial likelihood is given by:

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\theta}) = \prod_{i=1}^{D} \frac{\exp\left[\boldsymbol{\gamma}^\top \mathbf{v}_i + \boldsymbol{\theta}^\top \mathbf{z}_i(t)\right]}{\sum_{l \in R(t_i)} \exp\left[\boldsymbol{\gamma}^\top \mathbf{v}_l + \boldsymbol{\theta}^\top \mathbf{z}_l(t)\right]}.$$

Since the borrower's characteristics may change between successive default times, the entire denominator of this partial likelihood function is reviewed rather than merely modifying the risk set at each default time.

Let $\mathbf{x}(t) = (\mathbf{v}, \mathbf{z}(t))$ and $\mathbf{x}^*(t) = (\mathbf{v}^*, \mathbf{z}^*(t))$ be two different realisations for the combined set of covariates at time t. For the extended Cox regression model, the hazard ratio is:

$$\eta = \exp\left[\boldsymbol{\gamma}^\top (\mathbf{v}^* - \mathbf{v}) + \boldsymbol{\theta}^\top (\mathbf{z}^*(t) - \mathbf{z}(t))\right].$$

The Cox regression model can be adjusted to handle time-dependent covariates by converting data into a start-stop format and allowing time by covariate interaction in the model [Therneau et al., 2017]. The counting process theory of partial likelihoods underpins the validity of this technique. Alternatively, non-proportional hazards can be modelled by allowing the time-dependent regression coefficients. In this situation, the model used to investigate the PH assumption is given by:

$$h(t|\mathbf{x}(t)) = h_0(t) \exp\left[\boldsymbol{\beta}^\top \mathbf{v} + \boldsymbol{\theta}(t)^\top \mathbf{z}\right],$$

where $\theta(t)$ now varies with time. A new time-dependent covariate, $g(t)$, is defined to accommodate the changing hazard ratios. According to Klein and Moeschberger [2003], Ata and Sözer [2007] and several other authors, possible forms of $g(t)$ include $g(t) = z \cdot t$, $g(t) = z \cdot \log t$, and the step function.

2.4.6 Cox PH Model with Splines

Regression splines can be implemented in survival analysis to accommodate non-linear relationships [Luo et al., 2016]. A spline of order k is a $k-1$ degree piecewise polynomial function of a continuous covariate stitched together at points called knots [Dirick et al., 2017]. By using splines, the log-linear model in Equation 2.9 can be modified to incorporate non-linear effects of a continuous covariate. A smooth covariate function is derived from a basis of functions whose gradients are continuous. A spline function is essentially a linear combination of spline basis functions. The piecewise polynomial basis of an s-spline, for an arbitrary $s \geq 1$, with K knots given by $\kappa_1, \kappa_2, \ldots, \kappa_K$ in the interval $\mathcal{I} = [x_{\min}, x_{\max}]$, is given by:

$$\phi_i(x) = \begin{cases} x^i, & \text{if } 0 \leq i \leq s, \\ \max\left\{0, (x - \kappa_{i-s})^s\right\}, & \text{if } s+1 \leq i \leq K. \end{cases} \tag{2.13}$$

Let q be the number of continuous covariates modelled by an s-spline approximation. The linear predictor function in Equation 2.9 can be rewritten as:

$$\boldsymbol{\beta}^\top \mathbf{x} = \sum_{j=1}^{q} \sum_{i=0}^{K} \beta_i \phi_i(x_j) + \sum_{j=q+1}^{p} \beta_j x_j = \sum_{j=1}^{q} f_j(x_j) + \sum_{j=q+1}^{p} \beta_j x_j,$$

where $f_j(x_j)$ denotes an s-spline approximation for a continuous covariate x_j. The number of knots impacts the estimated shape of a spline function: a large number of knots implies high flexibility, but it also adds more variation than might be justified by the data. A small number of knots, on the other hand, may result in an under-fit bias [Perperoglou et al., 2019]. Several approaches are used to determine an optimal number of knots that result in an intermediate amount of smoothing that offsets under- or over-fitting bias. This study uses penalised splines, also known as P-splines,

to facilitate choice between knot positions and the number of knots. Instead of a polynomial piecewise basis in Equation 2.13, the B-spline basis is applied to the P-spline model since it is more stable [Hall and Opsomer, 2005]. Let $\kappa_0, \kappa_1, \kappa_2, \ldots$ be a knot sequence; B-spline of order one is given by:

$$B_{i,1}(x) = \begin{cases} 1, & \text{if } \kappa_i \leq x \leq \kappa_{i+1}, \\ 0, & \text{otherwise.} \end{cases}$$

De Boor and De Boor [1978] propose the Cox-de Boor recursion formula to derive higher-order B-splines as follows:

$$B_{i,k+1}(x) = \left[\frac{x - \kappa_i}{\kappa_{i+k} - \kappa_i} B_{i,k}(x)\right] + \left[\frac{\kappa_{i+1+k} - x}{\kappa_{i+1+k} - \kappa_{i+1}}\right] B_{i+1,k}(x).$$

According to O'Sullivan [1986], by applying a slight modification to the the partial likelihood function in Equation 2.10, the maximum likelihood estimator of β is computed as:

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left[\mathcal{L}(\boldsymbol{\beta}; t_1, t_2, \ldots, t_k) - \frac{\lambda}{2} \int_{\mathfrak{I}} \left\{\sum_{j=1}^{q} f_j''(x)\right\}^2 dx\right],$$

where $\lambda \geq 0$ is a regularisation parameter that determines the amount of smoothing as well as model fit.

2.4.7 Mixture Cure Models

A common characteristic of bank retail data is the existence of a subgroup of borrowers that are immune to default [Jiang et al., 2019; Tong et al., 2012]. By distinguishing between two subpopulations, mixture cure models can estimate parameters that determine susceptibility to default as well as those that influence borrowers' time-to-default. Let ψ be the susceptibility to default indicator defined by a binary random variable that takes the value of 1 when the borrower is susceptible to the event of default and 0 otherwise. Let the bank data consist of a set of n loans. The

complete survival data with a fraction of long-term survivors is of the form:

$$\{(t_i, \mathbf{z}_i, \mathbf{x}_i, \delta_i, \psi_i)\,; i = 1, \ldots, n\} = (t, \mathbf{z}, \mathbf{x}, \delta, \psi)\,,$$

where $\mathbf{z} = (z_1, z_2, \ldots, z_q)$ is a q-dimensional vector of covariates for the incidence model and $\psi = (\psi_1, \psi_2, \ldots, \psi_n)$ denotes indicators of susceptibility to default. The combination of the default indicator and the susceptibility indicator yields the following states:

Table 2.2: Mixture cure state description.

δ_i	ψ_i	**State Description**
0	0	Borrower i is censored and not susceptible to the default event.
0	1	Borrower i is censored and susceptible to the default event.
1	1	Borrower i is uncensored and has experienced default.

Since the borrower cannot default if he/she is not susceptible to default, $\delta_i = 1$ cannot occur when $\psi_i = 0$. It is important to note that the susceptibility to default is only observed to be $\psi_i = 1$ for borrowers that have defaulted, $\delta_i = 1$, otherwise it is latent, taking a value equal to either 0 or 1 for censored borrowers. As a result, for a portfolio consisting of n retail loans, the vector $\boldsymbol{\psi} = (\psi_1, \psi_2, \ldots, \psi_n)$ is partially missing information. In contrast, the default indicator, given by the vector $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)$, is fully observed. The survivor function for the mixture cure model is given by:

$$S(t|\mathbf{z}, \mathbf{x}) = \pi(\mathbf{z})S(t|\psi = 1, \mathbf{x}) + 1 - \pi(\mathbf{z}),$$

where the latency model, $S(t|\psi = 1, \mathbf{x}) = P(T > t|\psi = 1, \mathbf{x})$, is the survivor function of the susceptible borrowers given the covariate vector $\mathbf{x}$, whereas $\pi(\mathbf{z}) = P(\psi = 1|\mathbf{z})$, the incidence model, also referred to as the link function, computes the probability that a loan defaults given its covariate vector $\mathbf{z}$ [Peng, 2003]. In widely used mixture cure models, the link function depends on $\mathbf{z}$ by the log-odds ratio form given by:

$$\ln\left[\frac{\pi(\mathbf{z})}{1 - \pi(\mathbf{z})}\right] = \boldsymbol{\gamma}^\top(1, \mathbf{z}),$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_q)$ is a vector of unknown regression coefficients associated with the q-covariate vector $\mathbf{z}$. The density and survivor functions for the non-susceptible borrowers are set equal to zero and one, respectively, for all finite values of t because that subpopulation of borrowers will never experience default. Put differently, for sufficiently long follow-up period, $S(t|\mathbf{z}, \mathbf{x}) \to 1 - \pi(\mathbf{z})$. The KM estimator for survival data that contains long-term survivors shows a long and stable plateau as the follow-up period increases, and the level at which the survivor function plateaus is equivalent to the proportion of cured borrowers in the data [Amico and Van Keilegom, 2018]. The conditional survival probability of borrowers who are at risk of default, i.e., the latency model, is modelled using either semi-parametric or parametric models. The most common semi-parametric mixture cure model is the Logistic-Cox PH model, proposed by Kuk and Chen [1992], which models latency as:

$$S(t|\psi = 1, \mathbf{x}) = S_0(t|\psi = 1)^{\exp(\boldsymbol{\beta}^\top \mathbf{x})} = \exp\left[-H_0(t|\psi = 1)\exp(\boldsymbol{\beta}^\top \mathbf{x})\right].$$

Using a parametric distribution to model latency results in a variety of parametric mixture cure models, such as the Logistic-Exponential mixture cure model, in which latency is modelled by an Exponential distribution or a constant hazard model, i.e., $f(t;\lambda) = \lambda e^{-\lambda t}$. The parameter λ is called the shape parameter. Notably, the Exponential distribution is a special case of the Weibull distribution, given by:

$$f(t;\lambda,\sigma) = \frac{1}{\sigma}\lambda^{\frac{1}{\sigma}} t^{\frac{1}{\sigma}-1} \exp\left[-(\lambda t)^{\frac{1}{\sigma}}\right] \; ; \; t > 0,$$

obtained by setting the scale parameter, σ, to 1. Among the other types of parametric models used to model latency are the Log-Logistic and Log-Normal distributions, which are based on:

$$f(t;\lambda,\sigma) = \frac{\frac{\lambda}{\sigma}\left(\frac{t}{\sigma}\right)^{\lambda-1}}{\left(1 + (\frac{t}{\sigma})^{\lambda}\right)^2} \; ; \; t > 0,$$

and:

$$f(t;\lambda,\sigma) = \frac{1}{\sqrt{2\pi}\sigma t} \exp\left[-\frac{(\ln t - \lambda)^2}{2\sigma^2}\right] \; ; \; t > 0.$$

respectively. The corresponding survivor and hazard functions are discussed in more detail in Section 2.4.8.

2.4.7.1 Likelihood Function Parametric Mixture Cure Models

According to Chen and Wang [2000] and Qing Chen [2001], incorporating covariates result in the AFT assumption which can expressed via the survivor function:

$$S(t|\psi = 1, \mathbf{x}) = S_0(t \exp(\boldsymbol{\beta}^\top \mathbf{x}))^{\exp(-\boldsymbol{\beta}^\top \mathbf{x})},$$

or equivalently, using the hazard function:

$$h(t|\psi = 1, \mathbf{x}) = h_0(t \exp(\boldsymbol{\beta}^\top \mathbf{x})),$$

The unknown parameters of the AFT mixture cure model are $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and h_0. Given the censoring vector $(\delta_1, \delta_2, \ldots, \delta_n)$ and indicators of susceptibility to default $\psi = (\psi_1, \psi_2, \ldots, \psi_n)$, the complete log-likelihood function of AFT mixture cure models that considers the contribution of loan i is given by:

$$\ell_C(\boldsymbol{\beta}, \boldsymbol{\gamma}, H_0; \psi) = \ell_1(\gamma; \psi) + \ell_2(\boldsymbol{\beta}, h_0; \psi), \tag{2.14}$$

with:

$$\ell_1(\boldsymbol{\gamma}; \psi) = \sum_{i=1}^{n} \psi_i \ln[\pi(z_i)] + (1 - \psi_i) \ln[1 - \pi(z_i)], \tag{2.15}$$

and

$$\ell_2(\boldsymbol{\beta}, h_0; \psi) = \sum_{i=1}^{n} \delta_i \psi_i \ln[h_0(t_i \exp(\boldsymbol{\beta}^\top \mathbf{x}_i))] + \psi_i \ln[S_0(t_i \exp(\boldsymbol{\beta}^\top \mathbf{x}_i))^{\exp(-\boldsymbol{\beta}^\top \mathbf{x}_i)}]. \tag{2.16}$$

The log-likelihood function in Equation 2.14 factors into a sum of two different types of contributions: the log-likelihood function of the logistic regression, which models if borrower i defaults, and the AFT component to estimate when default occurs as given

by Equation 2.15 and 2.16, respectively. Note that if the susceptibility status vector ψ is completely observable, the maximisation of ℓ_C is straightforward, i.e., given the parameter space, $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, h_0)$, the unknown parameters are computed as:

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \widehat{h}_0) = \widehat{\theta} = \arg\max_{\theta \in \Theta} \ell_C(\boldsymbol{\beta}, \boldsymbol{\gamma}, h_0; \boldsymbol{\psi}).$$

In more realistic applications, the vector ψ is partially missing information, and as a result, the maximisation is not as straight as forward.

2.4.7.2 Expectation Maximisation

It is worth noting that there is no clear differentiation between susceptible and non-susceptible borrowers in mixture cure models. The susceptibility to default indicator comprises information for loans that defaulted during the observation period; this information is missing for other loans, making estimation of regression parameters challenging. In this dissertation, the Expectation Maximisation (EM) algorithm [Dempster et al., 1977] proposed by Zhang and Peng [2009] is applied. The primary motivation for this methodology is that the susceptibility/cure status is modelled through the latent variable ψ. In addition, the EM algorithm maximises latency and incidence separately [Amico and Van Keilegom, 2018]. The EM algorithm is an iterative method that performs two primary steps during each iteration: expectation (E-step) and maximisation (M-step). The algorithm first estimates the expected value of the complete log-likelihood given by Equation 2.14. The estimated values are obtained by replacing each susceptibility status with its expectation given observed data and parameters. Starting with initial parameters, $\theta^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)}, h_0^{(0)})$, the EM algorithm iterates the two steps until the likelihood function converges.

Expectation (E-step):
Let $\theta^{(r)} = (\boldsymbol{\beta}^{(r)}, \boldsymbol{\gamma}^{(r)}, h_0^{(r)})$ be a vector of current parameter estimates at the $(r+1)^{th}$ iteration and $\mathcal{O} = \{\text{observed } \psi_i\text{s}, (t_i, \delta_i, \mathbf{v}_i, \mathbf{x}_i); i = 1, \ldots, n\}$ be the observed data. The E-step is proceeds as follows:

1. For censored cases, the ψ_is are linear in terms of log-likelihood of the complete data, thus, the only calculation performed is:

$$\begin{aligned} e_i^{(r)} &= \mathbb{E}\left[\psi_i|\theta^{(r)}, \mathcal{O}\right] \\ &= P\left(\psi_i = 1|T > t_i, \delta_i = 0, \mathbf{x}_i; \theta^{(r)}, \mathcal{O}\right) \\ &= \frac{\pi(\mathbf{z}_i)S_0(t_i \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right))^{e^{-\boldsymbol{\beta}^\top \mathbf{x}_i}}}{\left[1 - \pi(\mathbf{z}_i) + \pi(\mathbf{z}_i)S_0(t_i \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right))^{\exp\left(-\boldsymbol{\beta}^\top \mathbf{x}_i\right)}\right]}, \end{aligned} \tag{2.17}$$

2. For the uncensored cases, $e_i^{(r)} = \mathbb{E}\left[\psi_i|\eta^{(r)}, \mathcal{O}\right] = \psi_i = 1$.

Essentially, the E-step replaces ψ_is with $w_i^{(r)}$s, where:

$$w_i^{(r)} = \begin{cases} e_i^{(r)}, & \text{if borrower } i \text{ is censored,} \\ 1, & \text{if borrower } i \text{ is uncensored.} \end{cases}$$

Using Equations 2.15 and 2.16, the E-step computes the conditional expectation of the complete log-likelihood function given current estimates, $\theta^{(r)}$, and the observed data, $\mathcal{O}$ [Cai et al., 2012; Peng, 2003], giving:

$$\mathbb{E}[\ell_1(\boldsymbol{\gamma}; w^{(r)})] = \tilde{\ell}_1(\boldsymbol{\gamma}; w^{(r)}) = \sum_{i=1}^{n} w_i^{(r)} \ln\left[\pi(z_i)\right] + (1 - w_i^{(r)}) \ln\left[1 - \pi(z_i)\right],$$

and

$$\begin{aligned} \tilde{\ell}_2(\boldsymbol{\beta}, h_0; w^{(r)}) = \sum_{i=1}^{n} \delta_i \ln\left[w_i^{(r)} h_0(t_i \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right))\right] \\ + w_i^{(r)} \ln\left[S_0(t_i \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right))^{\exp\left(-\boldsymbol{\beta}^\top \mathbf{x}_i\right)}\right], \end{aligned} \tag{2.18}$$

where $\tilde{\ell}_2(\boldsymbol{\beta}, h_0; w^{(r)}) = \mathbb{E}[\ell_2(\boldsymbol{\beta}, h_0; w^{(r)})]$. For a censored borrower i, the weight $w_i^{(r)}$ represents a fractional allocation to susceptible borrowers. Note that in the E-step, for censored borrowers, missing values of ψ are imputed based on $w^{(r)}$, where $w^{(r)} = \left\{w_i^{(r)} : i = 1, \ldots, n\right\}$, thus allowing the closed-form complete likelihood function.

Maximisation (M-step):
The M-step in the $(r+1)^{th}$ iteration maximises $\tilde{\ell}_1(\boldsymbol{\gamma}; w^{(r)})$ and $\tilde{\ell}_2(\boldsymbol{\beta}, H_0; w^{(r)})$ separately with regards to the model parameters. Equation 2.15 is the likelihood function of the logistic regression model. The parameters of the incidence model are estimated using the Newton-Raphson algorithm. Zhang and Peng [2007] propose a rank-type estimation equation of $\boldsymbol{\beta}$ given by:

$$\Psi(\boldsymbol{\beta}; k(.)) = \sum_{i=1}^{n} \delta_i k(t_i \exp(\boldsymbol{\beta}^\top \mathbf{x}_i)) \times \left(\mathbf{x}_i - \frac{\sum_{j=1}^{n} \mathbf{x}_j w_j^{(r)} \exp(-\boldsymbol{\beta}^\top \mathbf{x}_j) 1_{\{t_j \exp(\boldsymbol{\beta}^\top \mathbf{x}_j) \geq t_i \exp(\boldsymbol{\beta}^\top \mathbf{x}_i)\}}}{\sum_{j=1}^{n} w_j^{(r)} \exp(-\boldsymbol{\beta}^\top \mathbf{x}_j) 1_{\{t_j \exp(\boldsymbol{\beta}^\top \mathbf{x}_j) \geq t_i \exp(\boldsymbol{\beta}^\top \mathbf{x}_i)\}}} \right),$$

where $k(.)$ is a weight function. In order to proceed with the M-step, for the latency, Zhang and Peng [2009] and Kalbfleisch and Prentice [2011] propose an approach that relies heavily on the work of Breslow [1972]. Essentially, an estimate of $H_0(t|\psi = 1)$ is based on the residual $t_i \exp({\boldsymbol{\beta}^{(r)}}^\top \mathbf{x}_i)$, where $\boldsymbol{\beta}^{(r)}$ is an updated estimate of $\boldsymbol{\beta}$. Let d_{t_j} be the number of borrowers experiencing default at time t_j and $R(t_j)$ be the risk set just before default time t_j. The Breslow-type estimator used to estimate the conditional cumulative hazard function is given by:

$$\widehat{H}_0^{(r)}(t|\psi = 1) = \exp\left[- \sum_{j:t_j \leq t} \left(\frac{d_{t_j}}{\sum_{k \in R(t_j)} w_k^{(r)} \exp(-{\boldsymbol{\beta}^{(r)}}^\top \mathbf{x}_k)} \right) \right].$$

Thus;

$$\widehat{S}_0^{(r)}\left(t_i {\boldsymbol{\beta}^{(r)}}^\top \mathbf{x}_i \right) = \exp\left[\widehat{H}_0^{(r)}(t_i|\psi = 1) \left(t_i \exp({\boldsymbol{\beta}^{(t)}}^\top \mathbf{x}_i) \right) \right].$$

Given $\boldsymbol{\beta}^{(r)}$ and $\widehat{S}_0^{(r)}\left(t_i {\boldsymbol{\beta}^{(r)}}^\top \mathbf{x}_i \right)$, $e_i^{(r)}$ in Equation 2.17 can be updated and the EM algorithm will proceed until convergence. Note that the EM algorithm, in this case, is exclusively used to impute missing values of ψ to allow $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in the incidence and latency, respectively, to be estimated.

Sy and Taylor [2000] propose another technique based on the product-limit-type

estimator that estimates the baseline conditional survivor function non-parametrically. Peng and Dear [2000] put forward a method similar to Kuk and Chen [1992] that uses the marginal log-likelihood obtained by integrating Equation 2.18. For the further exposition of these methods and many more, the reader can refer to Amico and Van Keilegom [2018].

2.4.8 Parametric Survival Models

Parametric survival models assume that the time to default follows a distribution, in other words, the survivor function of parametric survival models has a specific functional form. Specifying the functional form of the hazard or survivor function makes it relatively more straightforward to compute selected quantiles or expected failure time. Parametric survival models are much easier to implement than non-parametric or semi-parametric models because they have a minimal and fixed set of unknown parameters. However, restricting the time to default to a specific distribution limits the flexibility to accommodate a wide range of hazards.

The Generalised Gamma distribution facilitates a comprehensive parametric family that includes commonly used parametric survival models [Cox et al., 2007; Prentice, 1974]. A distribution belonging to this parametric family is defined by the location, scale, and shape parameters, denoted by λ, σ, and ϑ, respectively. The survivor function of this distribution is given by:

$$S(t) = \begin{cases} 1 - \Gamma\left[\vartheta^{-2}(\lambda t)^{\frac{\vartheta}{\sigma}}; \vartheta^{-2}\right], & \text{if } \vartheta > 0, \\ \Gamma\left[\vartheta^{-2}(\lambda t)^{\frac{\vartheta}{\sigma}}; \vartheta^{-2}\right], & \text{if } \vartheta < 0, \end{cases}$$

where:

$$\Gamma[t; \upsilon] = \frac{\int_0^t x^{\upsilon-1} \exp(-x) dx}{\Gamma(\upsilon)},$$

is the cumulative distribution for a particular case of the Gamma distribution with mean and variance equal to $\upsilon > 0$.

This family of distributions includes all four of the most common types of hazard functions: monotonically increasing and decreasing, as well as bathtub and arc-shaped

hazards [Glaser, 1980]. Figure 2.2 shows a taxonomy of these hazard functions:

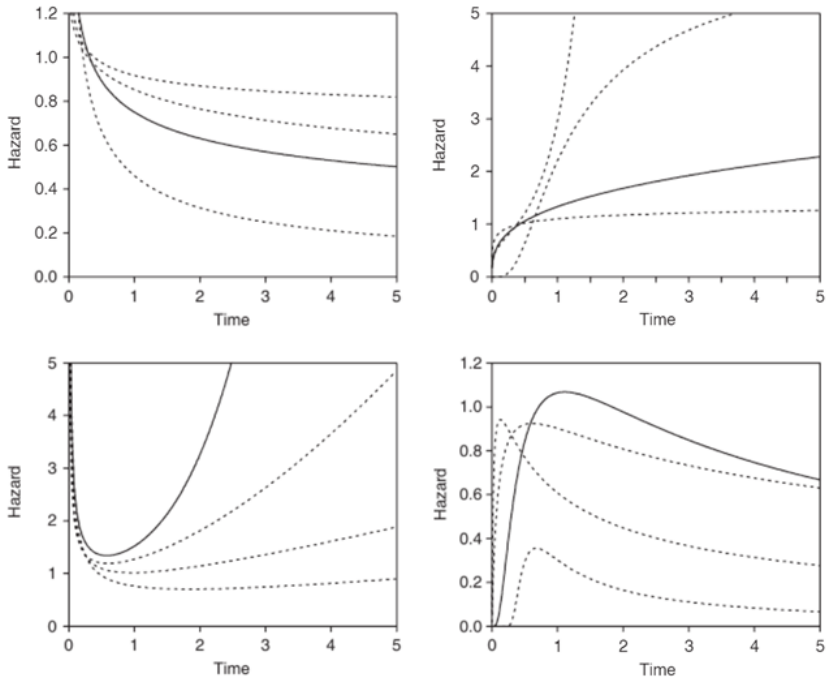


Figure 2.2: Taxonomy of hazard functions [Cox et al., 2007]

Using the relation $f(t) = -\frac{dS(t)}{dt}$, the probability density function of the Generalised Gamma distribution is given by:

$$f(t) = \frac{|\vartheta|}{\sigma t \Gamma(\vartheta^{-2})} \left(\vartheta^{-2}(\lambda t)^{\frac{\vartheta}{\sigma}}\right)^{\vartheta^{-2}} \exp\left[-\vartheta^{-2}(\lambda t)\right]^{\frac{\vartheta}{\sigma}}.$$

The Exponential and Weibull distributions are the only two sub-models of the Generalised Gamma distribution that satisfy both the AFT and PH assumptions [Kleinbaum and Klein, 2010]. In the statistical community, both the Exponential and Weibull distributions and their respective properties/moments are well known. Other members belonging to the parametric family accommodate the AFT assumption but not the PH assumption, for instance, the Log-Logistic distribution [Cox et al., 2007]. Explanatory variables are incorporated into the survivor function by re-parameterising λ in terms of the covariates.

2.4.8.1 Exponential Distribution

The Exponential distribution is derived from the Generalised Gamma distribution by setting the scale parameter, σ, and the shape parameter, ϑ, to 1, giving:

$$f(t) = \lambda \exp(-\lambda t) \; ; \; t > 0.$$

The associated survivor function and hazard function are $S(t) = \exp(-\lambda t)$ and $h(t) = \lambda$, respectively. The log-likelihood function of the Exponential distribution is given by:

$$\ell(\lambda; t_1, t_2, \ldots, t_n) = d \ln \lambda + \lambda \sum_{i=1}^{n} t_i,$$

where $d = \sum_{i=1}^{n} \delta_i$ is the total number of defaults. The maximum likelihood estimate $\widehat{\lambda}$ is obtained by setting the score function equal to zero and solving for λ. The estimated baseline survivor function is given by $\widehat{S}_0(t) = \exp(-\widehat{\lambda} t)$. Covariates are incorporated by letting λ depend on a vector of covariates using a log-linear model $\ln \lambda = \boldsymbol{\beta}^\top \mathbf{x}$. Because the Exponential distribution has a constant hazard, it does not provide enough flexibility for modelling PDs. It has a Markovian property, which means that the risk of default is the same at any point in time as it was when the loan was granted.

2.4.8.2 Weibull Distribution

By setting ϑ to 1, the density function of the Generalised Gamma distribution becomes:

$$f(t) = \frac{1}{\sigma t} (\lambda t)^{\frac{1}{\sigma}} \exp\left[-(\lambda t)^{\frac{1}{\sigma}}\right] = \frac{1}{\sigma} \lambda^{\frac{1}{\sigma}} t^{\frac{1}{\sigma}-1} \exp\left[-(\lambda t)^{\frac{1}{\sigma}}\right],$$

which is a density function of the two-parameter Weibull distribution. The corresponding survivor and hazard functions are given by $S(t) = \exp\left[-(\lambda t)^{\frac{1}{\sigma}}\right]$ and $h(t) = \frac{1}{\sigma} \lambda^{\frac{1}{\sigma}} t^{\frac{1}{\sigma}-1}$, respectively. The maximum likelihood estimates of λ and σ are obtained by maximising the log-likelihood function:

$$\ell(\lambda, \sigma; t_1, t_2, \ldots, t_n) = d \ln \frac{1}{\sigma} + d \frac{1}{\sigma} \ln \lambda + \left(\frac{1}{\sigma} - 1\right) \sum_{i=1}^{n} \sigma_i \ln t_i - \lambda^{\frac{1}{\sigma}} \sum_{i=1}^{n} t_i^{\frac{1}{\sigma}},$$

yielding an estimated baseline survivor function $\widehat{S}_0(t) = \exp\left[-(\widehat{\lambda}t)^{\frac{1}{\widehat{\sigma}}}\right]$. Covariates are incorporated by letting λ depend on a vector of covariates using a log-linear model $\ln\lambda = \boldsymbol{\beta}^\top \mathbf{x}$. Note that if the default time $T \sim \text{Exponential}(\lambda)$, then $T^{\frac{1}{\sigma}}$ has a Weibull distribution with scale and mean parameters σ and μ, respectively, where $\mu = \ln\lambda$.

2.4.8.3 Log-logistic Distribution

The survivor and hazard functions of the Log-logistic distribution are given by $S(t) = \frac{1}{1+\lambda t^{\frac{1}{\sigma}}}$ and $h(t) = \frac{\lambda^{\frac{1}{\sigma}} t^{\frac{1}{\sigma}-1}}{1+\lambda t^{\frac{1}{\sigma}}}$, respectively. By rewriting the survivor function as:

$$S(t) = \frac{1}{1 + (\lambda^\sigma t)^{\frac{1}{\sigma}}},$$

and applying the re-parameterisation $\frac{1}{\lambda^\sigma} = \boldsymbol{\beta}^\top \mathbf{x}$, the AFT Log-logistic model directly models the survival time as:

$$t = \left(\frac{1}{S(t)} - 1\right)^{\sigma} \cdot \boldsymbol{\beta}^\top \mathbf{x}.$$

2.4.8.4 Log-normal distribution

Let ϕ and Φ be the probability density function and the cumulative distribution of the Normal distribution, respectively. The survivor and hazard functions of the Log-normal distribution are given by $S(t) = 1 - \Phi\left(\frac{\ln t}{\sigma}\right)$ and $h(t) = \frac{\left(\frac{1}{t\sigma}\right)\phi\left(\frac{\ln t}{\sigma}\right)}{\Phi\left(\frac{-\ln t}{\sigma}\right)}$, respectively. Covariates can be taken into account by rewriting the survivor function as:

$$S(t|\mathbf{x}) = 1 - \Phi\left(\frac{\ln t - \boldsymbol{\beta}^\top \mathbf{x}}{\sigma}\right).$$

2.4.9 Probability of Default

An annual PD, also known as a marginal default probability, is calculated using the Cumulative Incidence Function (CIF). The CIF at time t is the probability of default at or before t, and is equivalent to the cumulative distribution function of the time

to default. Its complement, the survivor function, denoted by $S(t|\mathbf{x})$, measures the probability of not defaulting in that year. There is a one-to-one relationship between these two functions because defaulting is the only event of interest [Cao et al., 2009]. If the borrower did not default in year t, the probability that he/she will default in year $t+1$ is equal to the probability that he/she will not survive another year, and is calculated as:

$$\begin{aligned}
\text{PD}(t|\mathbf{x}) &= P(t < T \leq t+1 | T > t, \mathbf{x}) \\
&= \frac{P\left(T \leq t+1|\mathbf{x}\right) - P\left(T < t|\mathbf{x}\right)}{P\left(T > t|\mathbf{x}\right)} \\
&= \frac{F(t+1|\mathbf{x}) - F(t|\mathbf{x})}{1 - F(t|\mathbf{x})} \\
&= \frac{S(t|\mathbf{x}) - S(t+1|\mathbf{x})}{S(t|\mathbf{x})} \\
&= 1 - \frac{S(t+1|\mathbf{x})}{S(t|\mathbf{x})}.
\end{aligned} \tag{2.19}$$

In this study, once each borrower has been scored, survival probabilities are visualised using nomograms. The lifetime PD is obtained by computing the cumulative default probability:

$$\text{PD}(T_i|\mathbf{x}) = 1 - \prod_{t=1}^{T_i} \left\{1 - PD(t|\mathbf{x})\right\}, \tag{2.20}$$

where T_i denotes the loan's maturity for borrower i.

2.5 Imputation for Variables with Missing Data

Data imputation is a technique for filling in missing values in a dataset. Most imputation strategies are based on the assumption that data are missing at random (MAR). A MAR value occurs when, given the values of other available variables, respondents with missing values differ from other respondents in the dataset at random. Missing completely at random (MCAR) and informative missing (IM) data are two more types of missing data. Data is missing in MCAR for reasons unrelated to the subject's characteristics or responses, whereas IM occurs when the proclivity for a variable to have missing responses is a function of the missing data itself,

including the case where the missing data is related to the subject's actual response values being systematically higher or lower [Harrell et al., 2001].

Despite being straightforward, standard imputation approaches such as mean, median, and mode imputation underestimate variance since missing cases of a variable are replaced by its mean, median, or mode. Furthermore, these imputation approaches ignore the correlation between variables. The multivariate imputation by chained equations (MICE) method, proposed by Van Buuren and Oudshoorn [1999], addresses these challenges. In a nutshell, MICE consists of two primary steps, which are as follows:

1. The conditional regression model of each incomplete variable on other variables as predictors is specified. This step exploits the existing correlation between variables.

2. The algorithm generates multiple imputations based on the fitted conditional regression model by constructing a Gibbs sampler based on specified conditional distribution.

Let n be the number of independent realisations of the p-dimensional covariate vector $(y, \mathbf{x})$, where the $n \times (p-1)$-dimensional matrix, $\mathbf{x}$, is fully observed, while y is an $n \times 1$ vector containing missing data. Given n_{miss} missing values, $n_{\text{obs}} = n - n_{\text{miss}}$ is the number observed values for y. Missing data are replaced based on predictions from the model, thereby contributing to more variability than mean, median, or mode imputation. Variables with complete data, i.e., $\mathbf{x}$, are typically used as predictors, and this method allows greater flexibility in choosing the imputation model. In this dissertation, MICE is utilised for imputing missing records for each incomplete variable in order to obtain a complete dataset for further analysis. On the contrary, the EM algorithm, which is also an imputation method, performs imputation for latent values of ψ, as discussed in Section 2.4.7.2. In that particular case, the survival dataset is assumed to be complete; only $\psi = (\psi_1, \ldots, \psi_n)$ contains partially missing information.

2.5.1 Bayesian Linear Regression

MICE can be used to perform Bayesian Linear Regression (BLR) imputation, under the method `norm` in the MICE package in `R`, which is a parametric multiple imputation method. The missing values for the response variable y are assumed to come from a probability distribution given by:

$$\text{Prob}\{\text{y is missing}\} = \Phi(\upsilon^\top \mathbf{x} + \varrho),$$

where ϱ is independent white noise; υ is a $(p-1)$-dimensional parameter vector; and Φ is the Normal cumulative distribution function. The conditional regression model of the response is given by $y = \boldsymbol{\beta}^\top \mathbf{x} + \epsilon$, where the residual $\epsilon \sim N(0, \sigma_\epsilon^2)$ and $\boldsymbol{\beta} \in \mathbb{R}^{p-1}$. More specifically, the response $y \sim N(\boldsymbol{\beta}^\top \mathbf{x}, \sigma_\epsilon^2 1)$. According to Little and Rubin [2019], imputation using a Bayesian linear regression model repeats the following steps $M \geq 2$ times:

1. The posterior step: using the fully observed subset of the data, as well as the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$, the residual variance is sampled from the inverse gamma function $\Gamma^{-1}\left[\frac{n_{\text{obs}}}{2}, \frac{(y_i - \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_i)^\top (y_i - \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_i)}{2}\right]$, the posterior distribution of the residual variance $\tilde{\sigma}_\epsilon^2 | y_i, \mathbf{x}_i$. The second round of random draws is performed using $N_{p-1}\left(\widehat{\boldsymbol{\beta}}, \tilde{\sigma}_\epsilon^2 (\mathbf{x}^\top \mathbf{x})^{-1}\right)$, which is the posterior distribution of the intercept and slope parameters $\tilde{\boldsymbol{\beta}} | y_i, \mathbf{x}_i, \tilde{\sigma}_\epsilon^2$.
2. The imputation step: In this step, n_{miss} missing values of y are drawn independently from the imputation model, i.e., $\tilde{y}_j \sim N(\tilde{\boldsymbol{\beta}}^\top \mathbf{x}, \tilde{\sigma}_\epsilon^2)$.

2.5.2 Predictive Mean Matching

Predictive mean matching (PMM) imputation, which is the basis for the MICE framework, fills a missing value of a target variable with the observed value from a donor with a similar predictive mean [Hong and Lynn, 2020]. In other words, n_{miss} values are imputed using n_{obs} observations, which constitute the donor sample. The main appeal of PMM is its flexibility and semi-parametric nature. According to Little [1988], the primary steps of PMM are:

1. Compute the predictive mean for the n_{obs} records of y as $\widehat{y}_i = \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_i$.
2. Compute the predictive mean for the n_{miss} records of y as $\tilde{y}_j = \tilde{\boldsymbol{\beta}}^\top \mathbf{x}_j$.
3. Based on the absolute scalar distance $\tilde{d}_{i,j}$, match each element of $\tilde{y}_j$ to its respective closest element of $\widehat{y}_i$.
4. Impute missing observation based on observed y_i where $\tilde{d}_{i,j}$ is the smallest.

Iterations continue until the stated maximum number of iterations, M, is reached. The purpose of the regression model in PMM is to output predictive mean values used in the metric that matches records with missing values to similar records with observed values.

2.5.3 Random Forest

MICE can implement Random Forest (RF) imputation via `CALIBERrfimpute` in `R`, proposed by Shah et al. [2014]. The RF imputation method is a non-parametric machine learning-inspired method. `CALIBERrfimpute` imputes missing values of the response y based on draws from an independent Normal distribution centred on the conditional mean predicted by the RF. The variance of the RF prediction error is estimated using the out-of-bag mean square error, which is presumed to follow a Normal distribution. Essentially, the out-of-bag mean square error is computed as the mean squared difference between each observed value and the tree-based prediction for which that observation is excluded, i.e., left-out from the bootstrap sample [Hong and Lynn, 2020]. Left-out observations constitute the out-of-bag sample. After fitting decision trees to different bootstrap samples, the RF aggregates the results to reduce the risk of overfitting.

Chapter 3

Methodology

3.1 Data

3.1.1 Data Source

This study examined personal loans issued between October 2009 and October 2014 from Prosper, a P2P marketplace for retail loans. The data is publicly available at the Udacity online data repository [1]. Data cleaning and preparation are performed using Python software. The Python software supports the Scikit-learn packages in the data processing pipeline. In addition, coding and debugging are made easier because of simple syntax. The `R` software is used for subsequent statistical analysis because of its vast in-built statistical packages.

3.1.2 Initial Variable Analysis

Explanatory variables consist of credit application and behavioural variables. Credit behavioural variables are relevant for loans that are current at a given time, whereas credit application variables pertain to borrowers' characteristics relevant during the credit granting stage. Variables with more than 50% of their entries missing are

[1] https://s3.amazonaws.com/udacity-hosted-downloads/ud651/prosperLoanData.csv

discarded during data extraction. As a result, eleven variables from the data extract were removed. Those that remained were categorised based on:

1. Credit history: this category includes application variables that contain the borrowers' past information. Credit agencies gather these variables; for example, in South Africa, Experian or Transunion generally collect this information.
2. Current financial status: these variables contain information about the borrowers' current and active credit facilities.
3. Indebtedness: these variables measure the extent of the borrowers' debt burden.
4. Loan details: these variables contain loan-specific information, such as loan amount, funding date, interest rate, and monthly instalment amount.
5. Borrowers' details: provide demographic information of the borrowers.

There are some explanatory variables in the data that have missing cases. In real-world data, missing data is a common problem, and it dramatically reduces the statistical power and predictive abilities of statistical models. Table 3.1 shows the variable description and the percentage of missing data.

Table 3.1: Variable name, description and percentage of missing data.

Variable name	**Description**	**Missing**
DebtToIncomeRatio	The borrower's debt-to-income ratio at the time the credit profile was pulled.	9.59%
Occupation	The occupation chosen by the borrower when the listing was created.	1.83%
MonthlyDebtRate	Amount of added debt burden the loan adds to the borrower's monthly expenditure.	0.90%
ProsperScore	A custom risk score based on historical data.	0.18%
ProsperRatingCategory	Rating assigned assigned at the time the listing was created.	0.18%
EmploymentStatusDuration	The number of months the borrower was employed at the time the listing was created.	0.03%

3.1.3 Imputation Method Selection

Different MICE imputation methods were explored in our study, and their performance was assessed by comparing pre- and post-imputation distributions using simulated datasets. Two simulated datasets, A and B, containing correlated variables, each with 10,000 records, were generated from multivariate Normal and Log-Normal distributions, respectively. These distributions are typical in most retail consumer credit data. For each of the four simulated variables in each dataset, the `ampute()` function, proposed by Schouten et al. [2018], in the MICE package was used to generate missing data using the MAR mechanism. Figure 3.1 illustrates the process for generating missing data in a multivariate setting.

The missingness patterns are at the core of the technique. A missing data pattern describes a specific combination of variables containing missing values and variables with complete values. Figure 3.1 depicts how the complete dataset is randomly partitioned into k subsets, where k is the number of missing data patterns. The size of these subsets and the relative frequency of the occurrence of missing data patterns may vary depending on the frequency vector.

In summary, the `ampute()` function requires an a complete dataset to perform amputation. The user specifies the missingness proportion and can customise missing data patterns. Moreover, the function allows the user to select the data missing mechanism, and that can be MCAR, MAR, or IM. In this study, we chose MAR since it is the most commonly exploited missing data mechanism. Weighted sum scores for all data rows are typically computed for the MAR mechanism. Essentially, a weighted sum score is the output of linear regression in which the coefficients are weight matrix values. For instance, a candidate data row i for pattern k has a weighted sum score calculated as:

$$\text{wss}_i = w_{k,1}y_{1,i} + w_{k,2}y_{2,i} + \cdots + w_{k,m}y_{m,i},$$

where the variable set $\{y_{1,i}, y_{2,i}, \ldots, y_{m,i}\}$ comprises values for case i; and $\{w_{k,1}, w_{k,2}, \ldots, w_{k,m}\}$ are pre-specified weights on row k of the weights matrix.

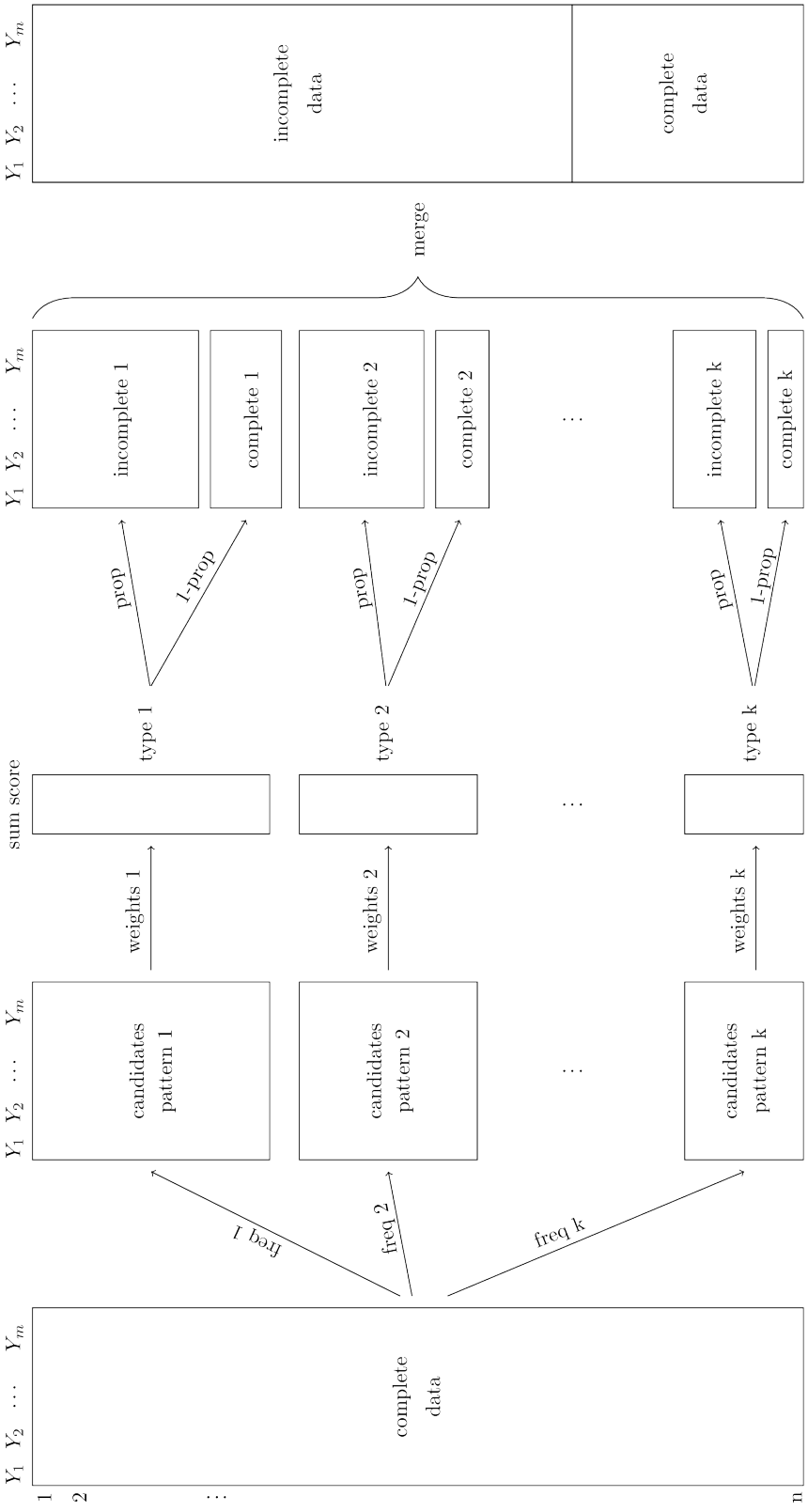
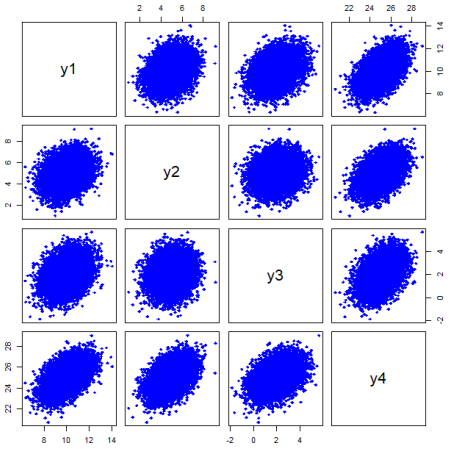
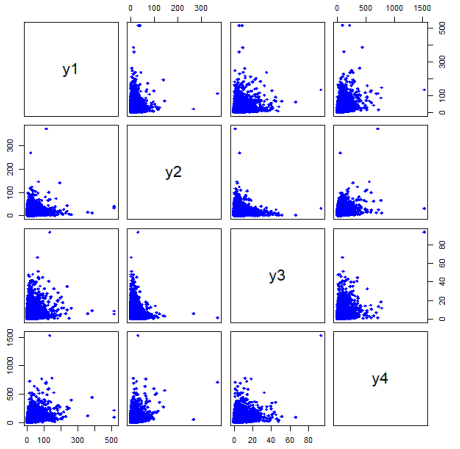


Figure 3.1: Amputation procedure

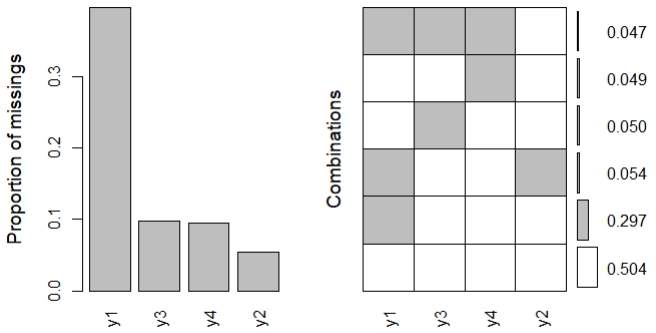
Simulated data sets A and B, shown in Figure 3.2, contain four correlated variables, thus $m = 4$ and since five data patterns were considered, then $k \in \{1, 2, \ldots, 5\}$.



(a) Simulated multivariate Normal data: A



(b) Simulated multivariate Log-Normal data: B



(c) Missing data pattern

Figure 3.2: Missing data patterns and simulated datasets

Figures 3.2(a) and 3.2(b) depict the correlation of variables from two simulated datasets, A and B, generated using multivariate Normal and Log-Normal distributions, respectively, whereas Figure 3.2(c) depicts the proportion of missing data for each variable as well as the corresponding missing pattern. Pattern 1, for example, contains the majority of missing values only in y_1 but not in other variables, while pattern 2 contains missing values for both y_1 and y_2 but not for y_3 and y_4, and so on.

Data imputation was performed for amputated datasets A and B using four different methods: BLR, PMM, RF, and mean, which was used as the performance benchmark. The performance of imputation methods was evaluated using the bias metric, which computes the average amount by which actual or ground truth values exceed estimated values using imputation. Bias should be closer to zero for better-performing imputation methods. In addition, the two-sample Kolmogorov-Smirnov (KS) test was conducted to determine if the imputed and ground truth samples were from the same distribution. The null hypothesis of the two-sample KS test states that both samples come from the same distribution. The D statistic from the KS test estimates the magnitude of the pairwise difference between empirical cumulative distributions. Van Buuren and Groothuis-Oudshoorn [2011] recommend a small number of iterations within the MICE framework, thus we conducted five iterations (i.e., $M = 5$) for each imputation.

3.1.4 Data Preparation for Survival Analysis

A unique loan key identifies each record within the dataset as belonging to a particular borrower. The variable *LoanStatus* indicates the actual classification of the loan, which is either completed, current, written off, defaulted, or cancelled. A loan is considered written off when the bank determines that there is no reasonable expectation that the borrower will repay the loan. The default status not only consists of loans that defaulted but also written off loans. Loans are classified as defaulted when they are 90+ days overdue. After 150 days, they are written off since there is no longer a reasonable expectation for further payments. Completed, current and cancelled loans are considered censored since they did not experience default during the study period. Cancelled loans form part of the analysis because they are scored and approved, therefore falling into the "score during normal day-to-day operations" [Siddiqi, 2012].

The variable *LoanOriginationDate* records when the loan was funded. It forms the entry point of an account, starting from October 2009 and ending in October 2014. The exit date, as specified by the variable *ClosedDate*, is the date when the last payment for the loan was received, thus effectively giving a time point when a borrower is absorbed

either into default or censorship. The number of months between the entry date and exit date is the observed survival time. A borrower with a "complete" loan status and observed survival time is less than the loan term has settled the loan early. Early-settled loans are part of the censored category. Typically, early settlement is modelled as a competing risk. However, the main focus of this dissertation is modelling dynamic PDs as part of IFRS 9 capital provisioning. Thus attrition rates or competing risks are secondary and not in the scope of the study. In addition, the default status is an absorbing state, preventing further defaults from occurring.

3.1.5 Exclusions

Loan-specific variables assessed at the credit granting stage, such as loan listing number, listing key, lender yield, and so on, are excluded from the analysis because they do not reflect the borrower's actual creditworthiness [Siddiqi, 2012]. In total, thirty variables fitting that classification were excluded. Examples include variables that fall into the loan details category, such as interest and fees charged, percentage funded, borrower rate, estimated return, and estimated effective yield. The variable *LoanStatus* can take the attributes listed in Table 3.2.

Table 3.2: Loan status categories.

Observations	**Count**	**Proportion (%)**
Current	33,442	63.10
Completed	14,784	27.89
Written off	4,017	7.58
Defaulted	757	1.43
Total	53,000	100.00

For this particular dataset, there were no cancelled loans. Under logistic regression, current loans are not taken into account since they have not yet matured, either through full repayment, written off, or default. Thus logistic regression is not considered in our study since it results in loss of information and gives a static view of PDs that is not very helpful under the current credit loss provisioning that requires dynamic PDs over the credit lifetime of borrowers. The final dataset comprises personal loans with a contractual loan repayment period of three years. In total, there

are 48,226 (91.02%) censored borrowers and 4,774 (8.97%) uncensored borrowers in the dataset.

3.1.6 Variable Binning

An overview of the variable binning/bucketing techniques used in this study is covered in this section.

3.1.6.1 Weight of Evidence

In credit scoring, it is recommended that numerical variables be converted to dummy variables to improve performance [Mays, 2001; Zeng, 2014]. As part of this study, numerical variables are encoded as dummy attributes and assessed for logic and ability to rank risks based on the weight of evidence (WOE). According to Siddiqi [2012], using WOEs and the resulting binned variables increases performance. Let δ be the binary response variable that takes a value of zero and one to indicate censored and uncensored borrowers, respectively. Assume a numerical variable x is discretised and has k factor levels as its attributes. WOE measures the predictive power of each attribute. For each factor level $j = 1, \ldots, k$, there is a corresponding response $\delta_j \in \{0, 1\}$, resulting in the event rate given by D_j. Let p_j and q_j be the relative frequencies of censored and uncensored borrowers belonging to attribute j, respectively. WOE for factor level j is given by:

$$\text{WOE}_j(x) = \ln\left[\frac{p_j}{q_j}\right] = \ln\left[\frac{1 - D_j}{D_j}\right] + \ln\left[\frac{N_T^e}{N_T^{ne}}\right] = -\text{logit}(D_j) + \ln\left[\frac{N_T^e}{N_T^{ne}}\right],$$

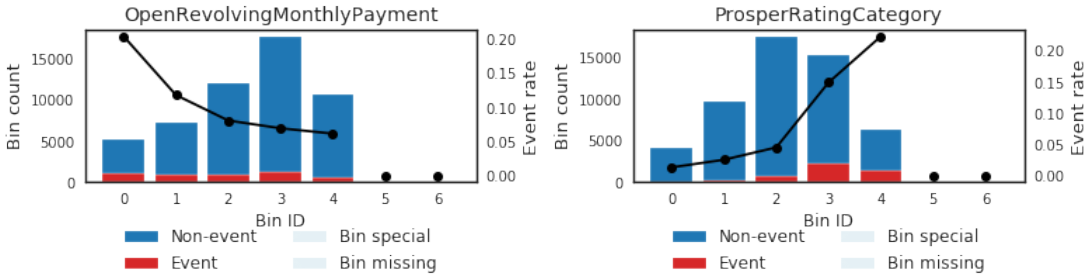
where N_T^e and N_T^{ne} is the total sum of events and non-events, respectively. Similarly, categorical variables with many classes are re-grouped into coarser bins by combining attributes with similar default rates. The binning process is set to ensure that each factor level has at least 5% of the sample volume. Where applicable, re-grouping is performed to ensure that the final binned covariate has WOE (and default rate) that is either monotonically increasing or decreasing to ensure that risk is ranked appropriately.

Table 3.3 depicts examples of the binning process for a numerical variable, *OpenRevolvingMonthlyPayment* and a categorical variable, *ProsperRatingCategory*. The bin attributes and the resulting monotonic trend are depicted in Figure 3.3.

Table 3.3: Binning process.

Bin ID	Bin	Count (%)	Non-event	Event	Event rate (%)	WoE
0	< 58.50	9.85	4,158	1,065	20.39	-0.95
1	$[58.50, 139.50)$	13.74	6,424	858	11.78	-0.30
2	$[139.50, 274.50)$	22.87	11,145	975	8.04	0.12
3	$[274.50, 630.50)$	33.38	16,468	1,222	6.91	0.28
4	≥ 630.50	20.16	10,031	654	6.12	0.42

Bin ID	Bin	Count (%)	Non-event	Event	Event rate (%)	WoE
0	[AA]	7.78	4,073	53	1.28	2.03
1	[A]	18.40	9,505	245	2.51	1.35
2	[B,C]	33.06	16,733	787	4.49	0.74
3	[D,E]	28.95	13,055	2,291	14.93	-0.57
4	[HR]	11.81	4,860	1,398	22.39	-1.07



(a) OpenRevolvingMonthlyPayment (b) ProsperRatingCategory

Figure 3.3: WOE binning process

The binning process for the rest of the potential predictors is depicted in Appendices A.1, and A.2, and the corresponding bin definitions shown in Appendix A.2.

3.1.6.2 Mantel-Haenszel Test

For logistic regression and other static regression methods, the use of WOE and the binning technique described in section 3.1.6.1 is usually all it takes to prepare

variables for model development. In survival analysis, each binned variable's factor level represents a separate survival group, thus one might be interested in pairwise comparisons of survival distributions and testing whether they are significantly different from each other or not. In this research, we implement a technique similar to Stepanova and Thomas [2002] that compares survivor functions of neighbouring bins using the Mantel-Haenszel test discussed in Section 2.4.2.4. Survivor curves that are not significantly different according to the Mantel-Haenszel test are merged.

3.1.7 Variable Selection

This study employs several variable selection methods to determine the most predictive variables for predicting the outcome (i.e., default) as well as to reduce redundancy by not allowing variables that convey the same information to be included in the model.

3.1.7.1 Variance Inflation Factor

Multicollinearity occurs when at least two explanatory variables are strongly correlated. Explanatory variables that are highly correlated provide similar information in the regression model, causing issues with fitting and interpreting the model. Thus, it is preferable to use only one of the correlated covariates. The variance inflation factor (VIF) measures how much of each variable's variance is explained by other explanatory variables. This metric is obtained by regressing each explanatory variable against other explanatory variables. Let R_j^2 be the determination coefficient of regressing explanatory variable x_j against all other explanatory variables. The VIF for variable x_j is given by:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}.$$

The reciprocal of VIF is known as torelance. A VIF value of one indicates the absence of multicollinearity for a given variable. More specifically, the tolerance value is equal to 1 when $R_j^2 = 0$, which implies orthogonality (independence) between x_j and all other predictors [Thompson et al., 2017]. As a rule of thumb, variables having VIF scores of more than ten (or tolerance values less than 0.1) are omitted from further analysis as multicollinearity is deemed severe [Alin, 2010; Daoud, 2017]. Initial VIF

scores for each of the variables are shown in the second column of Table 3.4; variables with VIF scores larger than ten are highlighted and constitute a set of variables that are excluded. VIF was recalculated to evaluate multicollinearity in the kept variables; the third column of Table 3.4 reveals that there are no variables with VIF larger than ten.

Table 3.4: VIF scores.

Variable name	Initial VIF	VIF: kept variables
MonthlyDebtRate	**221.02**	-
OpenRevolvingMonthlyPayment	6.72	4.19
TotalInquiries	3.59	2.93
AmountDelinquent	1.19	1.18
OpenRevolvingAccounts	**18.02**	-
AvgCreditScore	**19.38**	-
OpenCreditLines	**80.50**	-
EmploymentStatusDuration	2.23	1.85
CurrentDelinquencies	1.37	1.32
ProsperScore	**10.10**	-
TotalCreditLinespast7years	**47.93**	-
CurrentCreditLines	**75.14**	-
AvailableBankcardCredit	1.84	1.46
TotalTrades	**51.65**	-
DelinquenciesLast7Years	1.50	1.32
RevolvingCreditBalance	3.15	2.93
InquiriesLast6Months	2.58	2.41
DebtToIncomeRatio	**222.38**	-
TradesOpenedLast6Months	1.91	1.67

3.1.7.2 Variance Threshold

The variance threshold approach removes constant or quasi-constant variables whose variances do not meet a pre-defined threshold. A constant variable has no variability across all the samples. On the other hand, a quasi-constant variable contains the same values for very large subsets. Constant and quasi-constant variables are ineffective for predictive modeling since they are uninformative.

3.1.7.3 Information Value

Jeffrey's divergence or Information value (IV) is a univariate metric that measures the discriminatory and predictive power of an explanatory variable [Siddiqi, 2012]. Given two discrete probability distributions, P and Q for censored and uncensored borrowers, respectively, the Kullback-Leibler divergence is given by:

$$D_{\text{KL}}(P||Q) = \sum_{j=1}^{k} p_j \left(\ln p_j - \ln q_j\right) = \sum_{j=1}^{k} p_j \ln \left[\frac{p_j}{q_j}\right],$$

where p_j and q_j are as defined in Section 3.1.6.1. The IV of a binned variable, expressed in terms of the Kullback-Leibler divergence, is given by:

$$\begin{aligned}
\text{IV}(x) &= J(P||Q) \\
&= D_{\text{KL}}(P||Q) + D_{\text{KL}}(Q||P) \\
&= \sum_{j=1}^{k} p_j \ln \left[\frac{p_j}{q_j}\right] + \sum_{j=1}^{k} q_j \ln \left[\frac{q_j}{p_j}\right] \\
&= \sum_{j=1}^{k} (p_j - q_j) \ln \left[\frac{p_j}{q_j}\right] \\
&= \sum_{j=1}^{k} (p_j - q_j) \cdot \text{WOE}_j(x).
\end{aligned}$$

Covariates having weak discriminatory power in terms of IV are examined for elimination from the model development process. While the interpretation "weak" in the context of IV slightly varies, Siddiqi [2012] considers $\text{IV}(x) < 2\%$.

3.1.7.4 Gini Coefficient

The Gini coefficient or Accuracy Ratio, developed by Gini [1936], is a quantitative measure of discriminatory power given the distribution of events and non-events. Let $j = 1, 2, \ldots, k$ be attributes of variable x sorted monotonically in terms of default rate.

The Gini coefficient is given by:

$$\text{Gini}(x) = 1 - \left(\frac{2\sum_{j=2}^{k}\left(N_j^e \sum_{i=1}^{j-1} N_i^{ne}\right) + \sum_{m=1}^{k} N_m^e N_m^{ne}}{N_T^e N_T^{ne}} \right).$$

The Gini coefficient ranges from 0 for zero discriminatory power to 1 for maximal discriminatory power.

Table 3.5 shows the IV and Gini scores and highlights variables dropped from the analysis based on them having weak predictive and discriminatory power. All the variables in the table have been binned and assessed for group differences as discussed in Sections 3.1.6.1 and 3.1.6.2, respectively. In addition, all the potential variables have been assigned shortened names to facilitate exposition.

Table 3.5: IV and Gini scores.

Variable name	**Short name**	**IV(%)**	**Gini(%)**
DelinquenciesLast7Years	-	**1.06**	**4.81**
TotalInquiries	Tot_Inq_bn	2.27	7.15
AvailableBankcardCredit	AvlBnkCred_bn	15.81	19.89
EmploymentStatusDuration	EmpDur_bn	3.13	9.69
InquiriesLast6Months	InqLst6M_bn	12.11	18.46
RevolvingCreditBalance	RevCredBal_bn	19.43	21.55
TradesOpenedLast6Months	TrdsOplst6M_bn	3.70	7.84
AmountDelinquent	AmntDlnq_bn	5.39	8.72
CurrentDelinquencies	CurDlnq_bn	5.50	9.31
OpenRevolvingMonthlyPayment	OpnRevMnthPmt_bn	20.16	22.55
ListingCategory	ListCat	14.81	19.27
ProsperRatingCategory	PrpRatCat	80.52	45.91
EmploymentStatus	EmpStat	10.07	14.24
IncomeVerifiable	IncVerf	5.25	7.54
Occupation	OccCat	9.80	16.36
BorrowerState	BwrState	4.68	12.00
IsBorrowerHomeOwner	-	**1.32**	**5.73**
IncomeRange	IncRange	21.04	25.11

3.1.7.5 Variable Correlation

Pairwise-wise correlation, ρ, between variables is examined. Variables that are strongly correlated impart nearly the same information and weaken the regression model when they are both included as predictors. In other words, adding a strongly correlated variable to the model adds no incremental information, but only contributes noise to the model. One way to mitigate this problem is to retain only one variable from the pair. Cohen [2013] adopts the convention that assigns specific values of the absolute value of the correlation between covariates x and y, $|\rho|$, to the categories "small," "medium," and "high" to serve as operational definitions. According to the author, the range $0.1 \leq |\rho| < 0.3$ constitutes "small", while $0.3 \leq |\rho| < 0.5$ and $\rho \geq 0.5$ are "medium" and "high", respectively. Accordingly, in our analysis, predictors are strongly correlated when $|\rho| \geq 0.5$.

In this study, when two variables are strongly correlated, the one with the lower IV or Gini is discarded since a variable with a lower IV or Gini is less logically connected to what is being predicted. The methodology employed here is consistent with that proposed by Héberger and Rajkó [2002]. The authors computed and compared test statistics such as the McNemar's test, Fisher's exact test, and the Chi-square test. If a test reveals a statistically significant difference between correlated predictors, the "superior-inferior" or "winner-loser" classification is employed to eliminate subordinate variables. Table 3.6 shows strongly correlated variables; highlighted variables are excluded based on them having lower IV and Gini.

Table 3.6: Highly correlated variables.

Variable 1	Variable 2	Correlation
RevCredBal_bn	OpnRevMnthPmt_bn	0.75
Tot_Inq_bn	InqLst6M_bn	0.58

Figure 3.4 summarises the association between potential predictors as measured by Hoeffding's D.

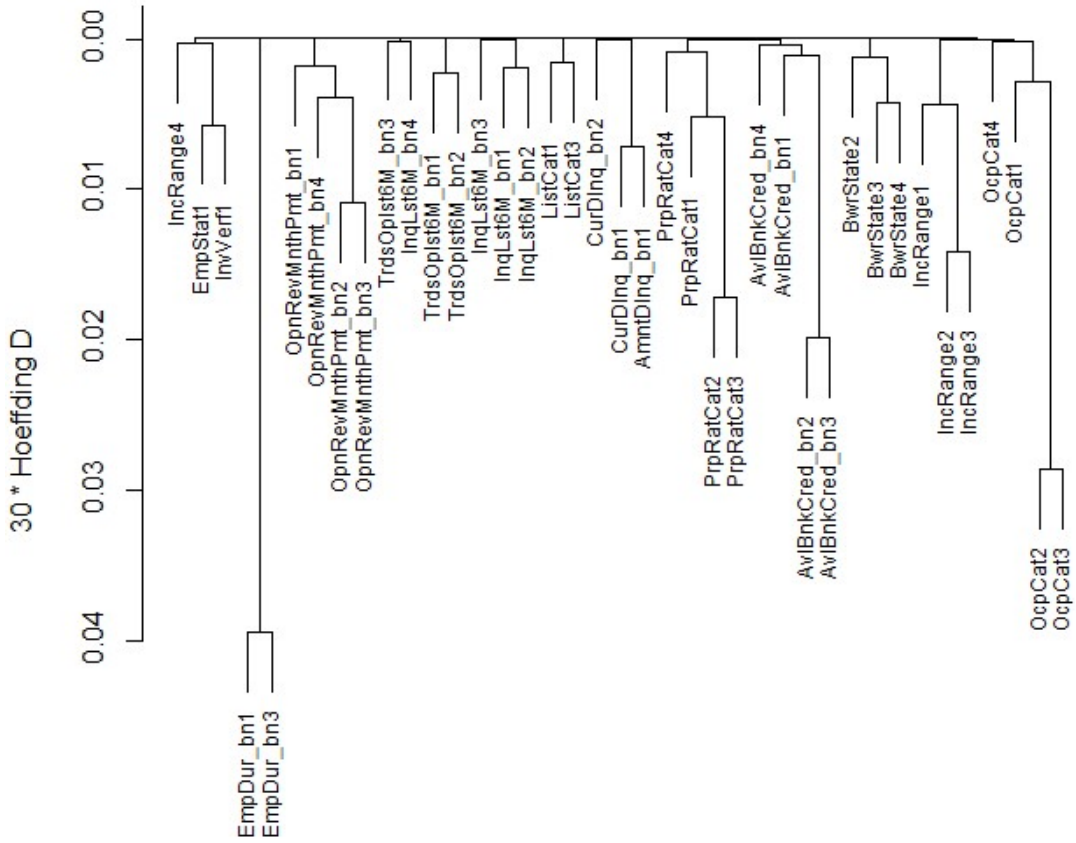


Figure 3.4: Hoeffding D measure

3.2 Assessment of the PH Assumption

Numerous diagnostic tools are employed to assess the PH assumption. This section provides a brief overview of the diagnostic tools used in this study.

3.2.1 Visual Inspection using Log Hazards Plots

The most widely used graphical technique for assessing the PH assumption involves comparing estimated log cumulative hazard functions over different categories of variables under investigation [Dabrowska et al., 1992]. The log cumulative hazard function is given by:

$$\ln H(t|\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x} + \ln H_0(t).$$

Considering two different specifications of the covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ corresponding to two subjects, the difference in the log cumulative hazard functions for these subjects is given by:

$$\ln H(t|\mathbf{x_2}) - \ln H(t|\mathbf{x_1}) = \boldsymbol{\beta}^\top(\mathbf{x}_2 - \mathbf{x}_1).$$

If the PH assumption holds, plotting the estimated log cumulative hazard curves for these subjects on the same graph shows parallel curves separated by a distance equal to the predictor values' differences.

3.2.2 Cox-Snell Residuals

The Cox-Snell residual, given by:

$$r_{C_i} = e^{\widehat{\boldsymbol{\beta}}^\top \mathbf{x}} \widehat{H}_0(t_i),$$

can be used to perform a graphical model validation for the PH assumption. Cox and Snell [1968] show that the Cox-Snell residuals have an exponential distribution when the PH assumption is met.

3.2.3 Schoenfeld Residuals

Given a covariate vector $\mathbf{x}$, the score function can be expressed as:

$$\ell'(\boldsymbol{\beta}) = \sum_{i=1}^{D} \left\{ \mathbf{x}_i - \sum_{k \in R(t_i)} \mathbf{x}_k \cdot p(\boldsymbol{\beta}, \mathbf{x}_k) \right\},$$

where $p(\boldsymbol{\beta}, \mathbf{x}_k) = \frac{e^{\boldsymbol{\beta}^\top \mathbf{x}_k}}{\sum_{j \in R(t_k)} e^{\boldsymbol{\beta}^\top \mathbf{x}_j}}$. Schoenfeld residuals for each covariate are the summed terms of the score function. Each term is essentially the difference between the observed value of the covariate for borrower i and its expected value, $\bar{x}(t_i)$ [Schoenfeld, 1982]. The residual for the i^{th} default time is given by:

$$\widehat{r}_i = x_i - \sum_{k \in R(t_i)} x_k \cdot p(\widehat{\boldsymbol{\beta}}, x_k) = x_i - \mathbb{E}[x_i] = x_i - \bar{x}(t_i).$$

If the PH assumption holds, then $\mathbb{E}[\widehat{r}_i] \approx 0$ and a plot of these residuals versus default time will be centred about zero. The GOF test proposed by Harrell et al. [1982] is used as an alternative method for evaluation of the PH assumption. The proposed GOF test investigates the correlation between Schoenfeld residuals and survival times. A correlation of zero indicates that the model met the PH assumption (the null hypothesis). Grambsch and Therneau [1994] showed that the expected value of the scaled residuals is given by:

$$\mathbb{E}[r_i^*] \approx \boldsymbol{\beta} + \boldsymbol{\beta}(t),$$

where $\boldsymbol{\beta}(t)$ is a time-dependent log hazard ratio. As a result, yet another GOF test is formulated as a hypothesis test for a constant $\boldsymbol{\beta}(t)$.

3.3 Model Validation

In this study, two model validation methods are employed to assess the performance of the survival analysis models under investigation in this dissertation. These methods are widely used and recommended in the fields of statistics and data science.

3.3.1 Validation using the Bootstrap

Given a dataset $\mathcal{D}$, a bootstrap dataset of the same size as $\mathcal{D}$ is created by sampling from $\mathcal{D}$ with replacement. The measure of error $\hat{e}_1$ (also known as an estimate of optimism) is derived using the first bootstrap dataset, Z_1. It is the difference between the accuracy index computed from the bootstrap sample and the one based on the original sample. According to Harrell et al. [2001], procedure can be repeated B times for some value of B between 200 and 300. Thus, for B different bootstrap datasets $\{Z_i\}_{i=1}^B$, B corresponding error estimates $\{\hat{e}_i\}_{i=1}^B$ are obtained and average optimism is computed based on these error estimates.

Let $\{\Upsilon_i\}_{i=1}^B$ be accuracy indexes obtained by fitting a model to $\{Z_i\}_{i=1}^B$ bootstrap samples. Efron [1983] proposes an improved bootstrap estimate of bias due to

overfitting or optimism given by:

$$\hat{e}_0 = \sum_{i=1}^{B} w_i \Upsilon_i,$$

where w_i are weights derived from bootstrap samples. The bias-corrected or overfitting adjusted estimate of accuracy index is computed by subtracting an updated estimate of optimism, $\hat{e}_0$, from the accuracy index derived from the original sample. This methodology was adopted in the study because the bias-corrected estimate of predictive accuracy is approximately equal to an unbiased estimate of future model performance. For more details on this method, the reader can refer to Efron [1983] and Efron and Tibshirani [1997]. In this dissertation, a total of 300 replications were performed to validate a model developed on the entire loan portfolio dataset.

3.3.2 Training and Test Dataset Split

In this validation technique, the personal loan dataset is partitioned into separate training and test datasets using random sampling. This method of validating the model's efficacy and efficiency is inspired by the data science methodology. Model fitting is implemented using the training dataset, while the test dataset is used for model evaluation. The dataset is partitioned based on the machine learning-inspired 80:20 ratio for the training and test datasets, respectively.

3.4 Evaluation Metrics

The performance of the models developed in this study was evaluated using model discrimination and model calibration evaluation metrics discussed in this section.

3.4.1 Validation for Model Discrimination

3.4.1.1 Area Under the Curve

A confusion matrix tabulates and compares predicted classes against the actual classes as depicted in Table 3.7.

Table 3.7: Confusion matrix.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

TP is the number of borrowers that defaulted on their loans and are correctly identified by the model. FP denotes the number of borrowers that are mistakenly labelled as defaulters on their loans. TN is the number of accurately categorised censored cases, whereas FN denotes the number of defaulted borrowers wrongly labelled as censored cases. Entries in Table 3.7 are used to calculate evaluation metrics shown in Table 3.8.

Table 3.8: Evaluation metrics.

Metric	**Formula**	**Description**
Specificity	$\frac{\text{FP}}{\text{TP+FP}}$	False positive rate (FPR) $= 1 - \frac{\text{TP}}{\text{TP+FP}}$.
Sensitivity	$\frac{\text{TP}}{\text{TP+FN}}$	True positive rate (TPR).

The receiver-operating characteristic (ROC) curve is a plot of Sensitivity (or TPR) versus (1 - Specificity) (or FPR) [Mandrekar, 2010]. A random model, also referred to as the baseline model, gives points along the main diagonal since FPR $=$ TPR on average. Models having ROC curves situated closer to the top-left corner are preferred. Area Under the Curve (AUC) measures the area below the ROC curve. It provides an aggregate measure of performance across all possible classification thresholds. AUC ranges in value from zero for a model whose predictions are completely wrong to one for a model whose predictions are completely correct.

3.4.1.2 Concordance Index

The Concordance Index, introduced by Harrell et al. [1982], also referred to as the c index, is the probability of concordance between predicted and observed survival time, where $c = 0.5$ for random predictions and $c = 1$ for perfectly discriminating models. Ideally, fitted models have c index values that lie between 0.55 and 0.7.

3.4.1.3 Somers' D

Somers' D is a measure of pure discrimination for survival models. It is the difference between concordance and discordance probabilities. Somers' D is given by:

$$D_{\text{xy}} = 2(c - 0.5),$$

where c is the concordance probability. This metric is equivalent to the Gini index.

3.4.1.4 Information Criteria

Given a set of competing models $\{M_K : k = 1, 2, \ldots, K\}$ indexed by $k = 1, 2, \ldots, K$, Alkaike's Information Criterion (AIC) is given by:

$$\text{AIC}(k) = -2 \ln \mathcal{L}_k(\widehat{\boldsymbol{\beta}}) + 2p_k,$$

where p_k are free parameters that maximise the likelihood function. The best-fit model is the one with the minimal AIC value. The Bayesian Information Criterion (BIC) is closely related to AIC. It finds the best model among a finite set of models. The BIC for model k is given by:

$$\text{BIC}(k) = -2 \ln \mathcal{L}_k(\widehat{\boldsymbol{\beta}}) + p_k \ln [n],$$

where n is the sample size. For both AIC and BIC, a trade-off exists between increasing the likelihood and overfitting. AIC and the BIC introduce a penalty term $2p_k$ and $p_k \ln [n]$, respectively, to counter-balance that trade-off. For the BIC, the penalty term is larger than in AIC, making it more stringent. The best-fit model is the one with

the lowest AIC or BIC value. Consistent AIC (CAIC), proposed by Bozdogan [1987], combines the benefits of both AIC and BIC via:

$$\text{CAIC}(k) = \alpha\text{AIC}(k) + (1-\alpha)\text{BIC}(k),$$

where α has a default value equal to 0.5. One of the most common criticisms of AIC is the number 2 attached to the penalty term, which some studies have found to be arbitrary.

3.4.2 Model Calibration

3.4.2.1 Calibration Plot

A calibration plot compares the predicted probability to the observed probability at a given point in time. Pairs of observed and predicted probabilities lie on the 45-degree angle line if the model is ideal, meaning that both probabilities correspond closely to each other.

3.4.2.2 Brier Score

The Brier score (BS) is a loss-type metric for prediction performance that evaluates model discrimination and calibration at the same time [Gerds et al., 2014]. Let n be the number of samples, $\widehat{S}(t|\mathbf{x})$ be the predicted survivor function and $\widehat{G}(t) = P(C > t)$ be the estimator of the conditional survivor function for censored borrowers. BS is computed as:

$$\text{BS}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\{0-\widehat{S}(t|\mathbf{x}_i)\}^2 \cdot 1_{\{T_i \le t,\ \delta_i=1\}}}{\widehat{G}(t_i)} + \frac{\{1-\widehat{S}(t|\mathbf{x}_i)\}^2 \cdot 1_{\{T_i>t\}}}{\widehat{G}(t)}\right].$$

A lower BS correspond to better prediction performance [Kronek and Reddy, 2008]. The Integrated Brier Score (IBS) provides an overal calculation of the model

performance at all available times:

$$\text{IBS}(t_{\text{max}}) = \frac{1}{t_{\text{max}}} \int_0^{t_{\text{max}}} \text{BS}(t) dt.$$

Another metric that will be used to evaluate performance is the Index of Prediction Accuracy (IPA), which is very much the same as the R^2 in standard linear regression model. The IPA is given by:

$$\text{IPA}(t) = 1 - \frac{\text{BS of Prediction Model}}{\text{BS of Null Model}}.$$

For the IPA metric, the higher the better.

Chapter 4

Results and Analysis

4.1 Introduction

The study was carried out to assess and estimate default times within a bank's loan portfolio under IFRS 9. This task was performed using survival analysis because of its versatility to model time to default on a forward-looking basis. Appendix B.1 illustrates the KM estimator of the survivor function for the loan portfolio dataset, as well as the smoothed survival estimate, indicating that the dataset is censored, so it makes sense to use survival analysis rather than traditional methods such as logistic regression. When working with real-world data, the problem of missing data is inevitable and quite prevalent. One of the goals of this study is to look into imputation approaches for dealing with missing data.

In this dissertation, several survival analysis models were examined. ROC, AUC, c-index, and Somers' D (Gini) were employed as discrimination performance metrics. Calibration plots, as well as BS, were used for validation for model calibration. In addition, Q-Q plots, Tukey-Anscombe plots, and model diagnostics such as Schoenfeld residuals were also employed to evaluate model fit. Finally, nomograms were employed to display covariate information and survival probabilities.

4.2 Data Imputation Results

The results shown in Table 4.1 reveal that mean imputation performed poorly for simulated datasets A and B discussed in Section 3.1.3. In both datasets, bias and the

Table 4.1: Simulated datasets: imputation methods evaluation.

Metric	Variable	Dataset A				Dataset B			
		Mean	BLR	PMM	RF	Mean	BLR	PMM	RF
|Bias|	y_1	0.114	**0.004**	0.011	0.014	1.841	0.106	**0.097**	0.296
	y_2	0.008	0.003	**0.002**	**0.002**	0.059	0.024	**0.010**	0.059
	y_3	0.013	**0.001**	0.006	0.007	0.043	0.019	0.017	**0.009**
	y_4	0.022	**0.001**	0.004	**0.001**	1.003	**0.025**	0.145	0.334
D statistic	y_1	0.244	**0.007**	0.012	0.013	0.227	0.074	**0.006**	0.011
	y_2	0.030	**0.002**	0.002	0.003	0.031	0.011	**0.003**	0.004
	y_3	0.053	**0.002**	0.004	0.004	0.062	0.021	0.003	**0.002**
	y_4	0.056	**0.003**	**0.003**	**0.003**	0.056	0.020	0.004	**0.003**
p-value	y_1	<.001	0.975	0.523	0.376	<.001	<.001	**0.995**	0.616
	y_2	<.001	1.000	1.000	1.000	<.001	0.569	**1.000**	**1.000**
	y_3	<.001	1.000	1.000	1.000	<.001	0.029	**1.000**	**1.000**
	y_4	<.001	1.000	1.000	1.000	<.001	0.043	**1.000**	**1.000**

D statistic are high compared to other imputation methods. The results from the two-sample KS test reveal a statistically significant difference between the imputed and ground truth samples for each variable. The probability density functions for multiple imputations are shown in Appendices A.4 and A.5 for datasets A and B, respectively. The results also highlight an important finding that appears in other studies. Notably, Hong and Lynn [2020] states that a particular imputation method "should not be recommended indiscriminately as a panacea for imputing missing data, especially when data are highly skewed". Despite the authors' focus on the missForest, an RF-based imputation, the same can be said of the methods we investigated. BLR imputation, a fully parametric model, performs best when the underlying sample is from a Normal distribution. The finding is not surprising given that the imputed values are drawn at random from a Normal distribution, as stated in Section 2.5.1. Furthermore, when the sample is from a Normal distribution, the performance difference between BLR, PMM, and RF is marginal. A marked difference is observed when the underlying distribution is non-Normal. More notably, Table 4.1 demonstrates that both PMM and RF outperform BLR imputation.

Moreover, the performance of PMM and RF is roughly comparable across the datasets.

In this study, BLR imputation is utilised for samples that are Normal or near Normal. For data samples from a skewed distribution, such as the Log-Normal distribution, we propose using either PMM or RF because their performance appears to be similar. Given the sheer number of attributes, imputations for missing categorical variables are based on the most frequent category.

4.3 Model Implementation Results

Table 4.2 provides a list and description of the models implemented in the study. The performance results are presented for each model category, which includes semi-parametric PH, parametric AFT, and mixture cure models.

Table 4.2: Models and description.

Model	**Model Description**
Model 1: Cox PH	Cox PH model.
Model 2: Stratified Cox PH	Stratified Cox Regression.
Model 3: Cox PH with Splines	Cox Regression with Penalised Splines.
Model 4: Weibull	Weibull survival model.
Model 5: Exponential	Exponential survival model.
Model 6: Log-logistic	Log-logistic survival model.
Model 7: Log-normal	Log-normal survival model.
Model 8: Logistic-Exponential	Exponential mixture cure model.
Model 9: Logistic-Weibull	Weibull mixture cure model.
Model 10: Logistic-Log-logistic	Log-logistic mixture cure.
Model 11: Logistic-Log-normal	Log-normal mixture cure model.
Model 12: Extended Cox - Step	Extended Cox regression with a step function.

4.3.1 Semi-parametric Models

Stepwise model selection using AIC as the model selection criterion was performed. The best-fitting model was assessed for the PH assumption using diagnostics such as the Schoenfeld residuals or the Cox-Snell residuals. Covariates that violated the PH

assumption were removed and considered in subsequent models. Appendix A.3 shows plots of log-hazards versus survival time for validating the PH assumption.

4.3.1.1 Validation using the Bootstrap

The bootstrapped model performance results for model discrimination based on Somers' D index are displayed in Table 4.3. There is no significant overfitting because the training and test values for Somers' D are comparable for all the semi-parametric models presented. In all models, optimism is marginal, showing that overfitting is minimal. The bias-corrected index, which indicates how well the model will discriminate prognosis in the future, shows that the Stratified Cox PH model outperforms all other presented.

Table 4.3: Semi-parametric models: Somers' D.

Model	Original sample	Training	Test	Optimism	Bias-corrected Index
Cox PH	0.300	0.301	0.298	0.003	0.297
Stratified Cox PH	0.483	0.483	0.478	0.004	**0.478**
Cox PH with Splines	0.300	0.301	0.299	0.002	0.298

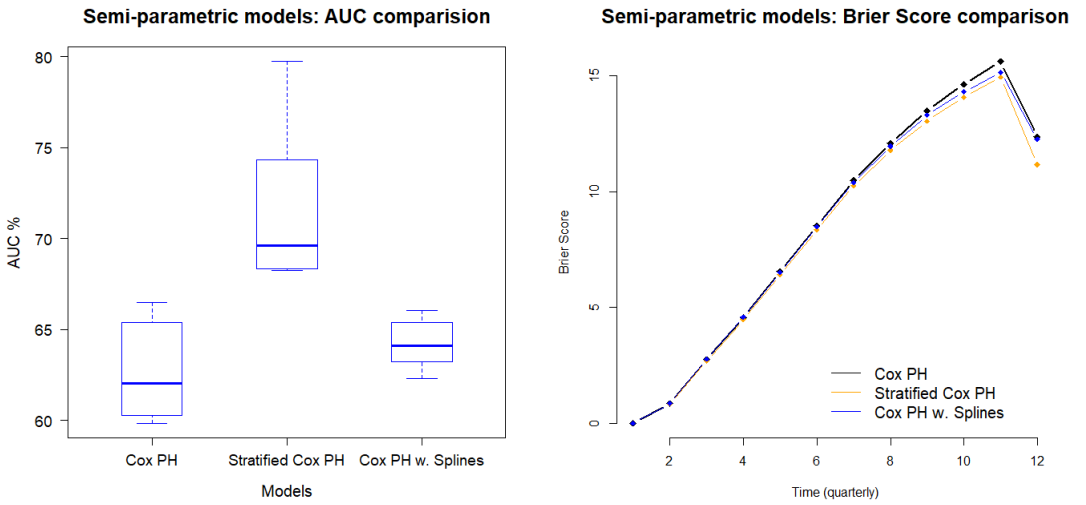
The results for the bootstrapped BS and IPA for calibration performance, as well as AUC for discrimination performance, are shown in Table 4.4. The results reveal that across the years, the stratified Cox PH model outperforms both the Cox PH model and the Cox PH with Splines in terms of annual calibration performance. Additionally, all the models presented performed better than the random model in terms of discrimination performance, i.e., AUC values are above 60% across all the models. Stratified Cox PH showed better discrimination performance than both Cox PH and Cox PH with Splines, as shown by the annual AUC values.

Figure 4.1(a), which shows the distribution of the average AUC over three years of prediction, demonstrates that the Stratified Cox PH model outperforms other semi-parametric models investigated across time, i.e., the Stratified Cox PH model yields higher average values. Furthermore, even though the Cox PH and Cox PH with Splines perform similarly, the boxplots indicate that the latter produces more stable average values. According to the calibration performance shown in Figure 4.1(b), the

Table 4.4: Semi-parametric models: BS, IPA, and AUC.

Period	Model	BS	IPA	AUC
Year 1	Cox PH	4.565	1.356	65.384
	Stratified Cox PH	**4.468**	**3.434**	**74.323**
	Cox PH with Splines	4.571	1.245	65.252
Year 2	Cox PH	12.089	1.424	61.090
	Stratified Cox PH	**11.782**	**3.933**	**68.482**
	Cox PH with Splines	12.004	2.115	62.702
Year 3	Cox PH	12.365	6.302	-
	Stratified Cox PH	**11.158**	**15.451**	-
	Cox PH with Splines	12.311	6.708	-

Stratified Cox PH outperformed other semi-parametric models consistently. Despite the marginal performance difference between the standard Cox PH model and the Cox PH with Splines during the first half of the prediction period, the Cox PH with Splines began to outperform during the second half.



(a) Semi-parametric: AUC comparison (b) Semi-parametric: Brier Score comparison

Figure 4.1: Semi-parametric Models: AUC and Brier Score comparison

4.3.1.2 Training / Test dataset Validation

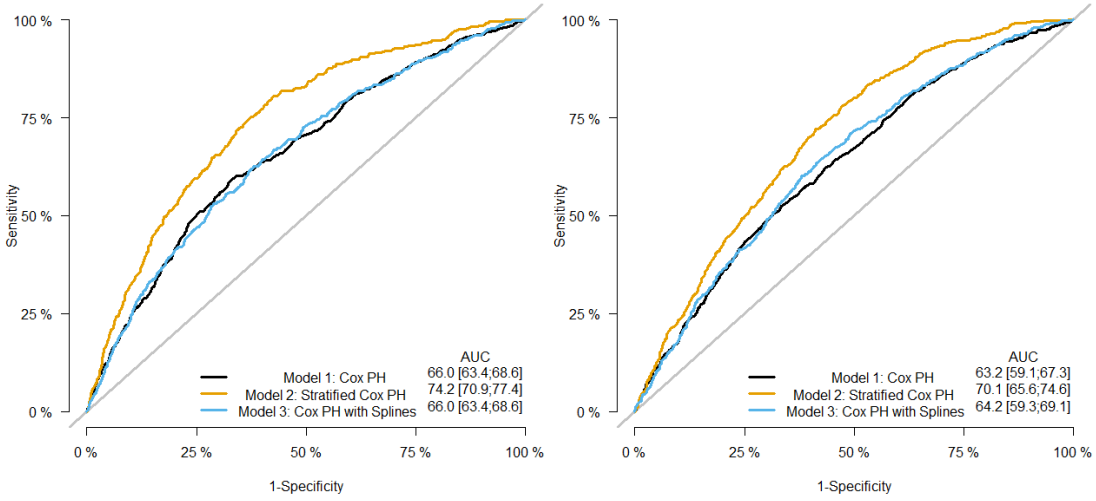
Based on the training c-index, AIC, BIC, and CAIC values in Table 4.5, the Stratified Cox PH model is the best-fit model among the semi-parametric models investigated.

Moreover, the Stratified Cox PH model outperforms both the Cox PH model and the Cox PH with Splines model in terms of calibration performance, as indicated by annual IBS values calculated using the test dataset.

Table 4.5: Semi-parametric models: c-index, AIC, BIC, CAIC, and IBS.

Model	Training Metrics				Test: IBS		
	c index	AIC	BIC	CAIC	Year 1	Year 2	Year 3
Cox PH	0.648	74,864	74,994	74,929	1.302	4.828	7.910
Stratified Cox PH	**0.744**	**53,015**	**53,145**	**53,080**	**1.274**	**4.722**	**7.644**
Cox PH with Splines	0.648	74,817	74,921	74,869	1.305	4.818	7.842

Annual ROC and AUC calculated using the test dataset are displayed in Figure 4.2. The ROC curves of the Stratified Cox PH model are positioned much further towards the top-left corner, indicating that this model outperforms the other two shown. Moreover, the test AUC values are very similar to the bootstrap AUC values shown in Table 4.4, as is the ranking of the performance of the models.



(a) Semi-parametric Models: Year 1 AUC (b) Semi-parametric Models: Year 2 AUC

Figure 4.2: Semi-parametric models: ROC and AUC for years 1 and 2

4.3.1.3 Model Selection and Analysis for Semi-parametric Models

According to the model validation performance metrics, the Stratified Cox Model outperformed other semi-parametric models presented. Appendix B.2 shows that the log-hazard ratios are significant at a 5% significance level according to the Wald test for this model. The Wald test is based on the Chi-square test for the null hypothesis that each regression coefficient is equal to zero. The p-value of the Wald test is less than 5% for each variable in the model, indicating that the log-hazards are statistically different from zero.

The hazard ratios and multi-level confidence intervals for the effects of explanatory variables in the fitted model are depicted by the forest plot in Figure 4.3. The forest plot displays the hazard ratios and the 95% multi-level confidence intervals, each relative to a reference level. For example, for the variable *TradesOpenedLast6Months*, borrowers who have opened at least one trade in the last six months are more likely to default than borrowers who do not have any trades over the same time period.

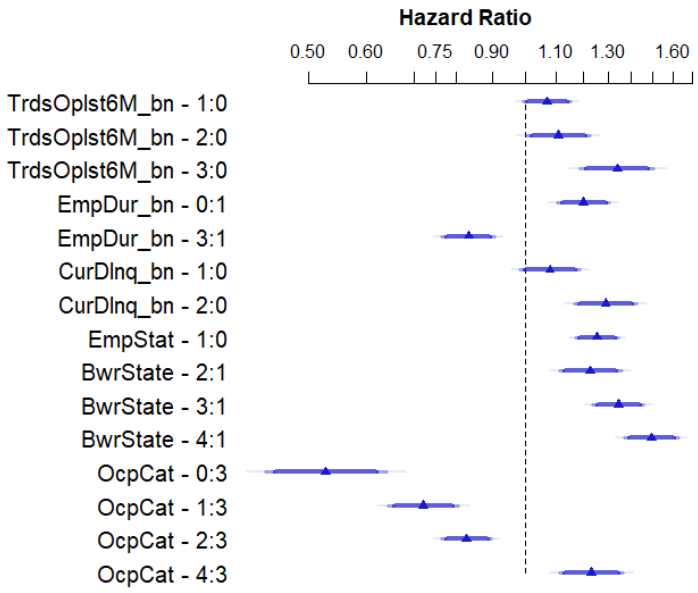


Figure 4.3: Stratified Cox PH: Forest plot

This is summarised better in Table 4.6, which provides the interpretation of the hazard ratios, as well as the risk ranking of each survivor group relative to the baseline or reference survivor groups. It is worthwhile to note that the risk ranking in terms of hazard ratios is *pari-passu* with the trends observed in the binning process, thus confirming the validity of the Stratified Cox PH model in this investigation.

Table 4.6: Interpretation of hazard ratios and risk ranking.

Predictor	**Effect relative to reference level**	**Hazards Ratio Explanation**
TrandesOplst6M	Positive	The default hazards for borrowers in bucket 1 are 24% percent more than for borrowers in bucket 0, which is a reference. Borrowers in buckets 2 and 3 face hazards that are 48% and 131% higher than those in bucket 0. This means that the default rate rises in proportion to the number of trades opened in the last six months.
EmpDur_bn	Negative	Using bucket 0 as a reference, the default hazards for borrowers in buckets 1 and 3 are 78% and 62% lower, respectively, indicating that the default rate decreases as employment duration increases.
CurDlnq_bn	Positive	Using bucket 0 as a reference, the default hazards for borrowers in buckets 1 and 2 are 37% and 76% higher, respectively, indicating that the default rate increases with the number of current delinquencies.
EmpStat	Positive	The default hazards of borrowers in bucket 1 is 35% higher than in bucket 0.
BwrState	Positive	Using bucket 1 as a reference, the default hazards in bucket 2 are 19% higher, while the default hazards in buckets 3 and 4 are 32% and 48% higher, respectively.
OcpCat	Positive	Using bucket 0 as a reference, the default hazards are higher for borrowers in lower-paying occupations. For example, bucket 0, which includes borrowers in the medical or judicial professions, has lower default hazards than bucket 4, which includes students and administrative assistants.

Schoenfeld residuals for each variable in the model are plotted against survival time in Figure 4.4. This serves as a basis for testing the correlation between the Schoenfeld residuals and the survival time, which is the GOF test for the PH assumption. The null

hypothesis of the test is that the correlation is zero. The p-value for the variable *OcpCat* is 4.73%, which is less than 5%. However, the global and the majority of the individual predictor p-values demonstrate that the correlation is not significantly different from zero at a 5% significance level: therefore, we fail to reject the null hypothesis and conclude that the PH assumption is met. Moreover, the residuals are centred around zero for all the predictors, i.e., $\mathbb{E}[\hat{r_i}] \approx 0$, which also is an additional confirmation that the PH assumption holds.

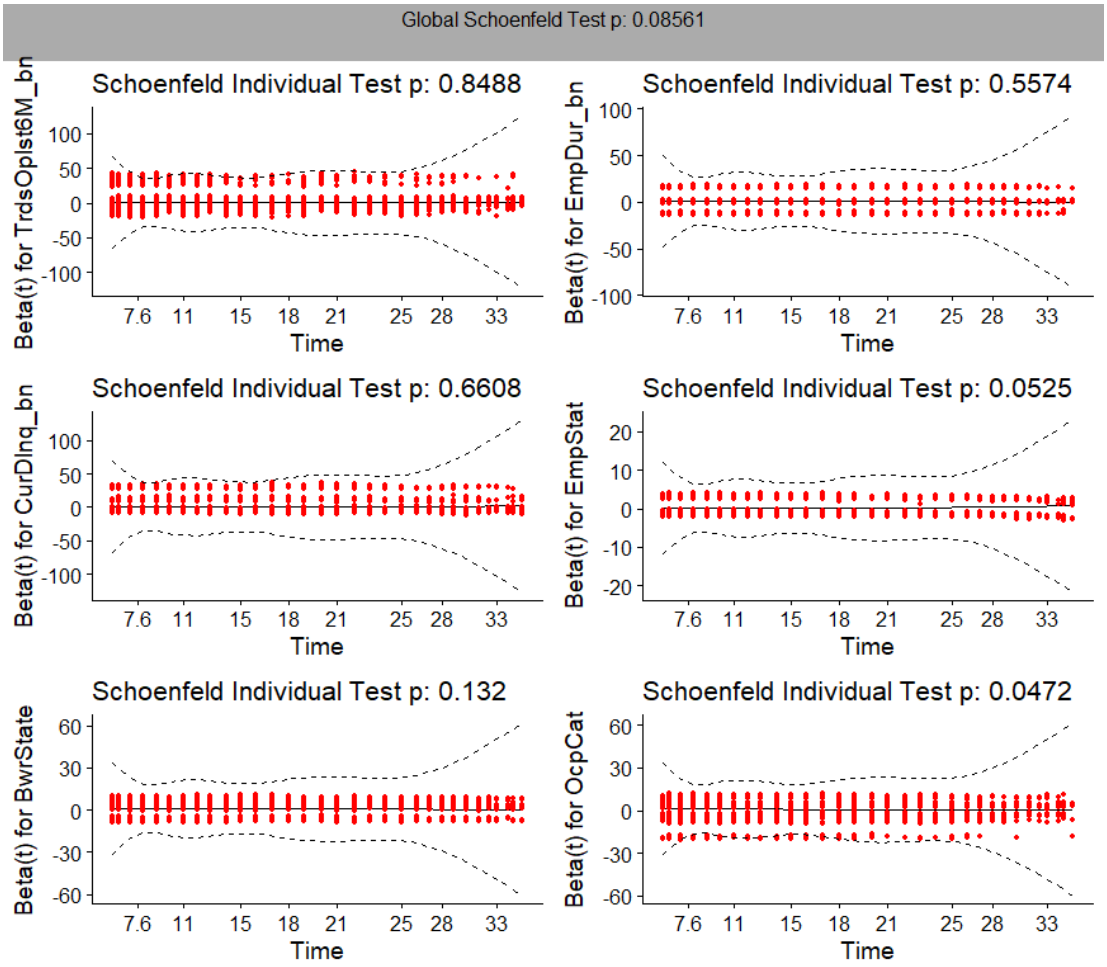


Figure 4.4: Stratified Cox PH: PH test using Schoenfeld residuals.

Plotting the Cox-Snell residuals versus the Exponential distribution quantiles is another diagnostic tool used to test for the PH assumption. Figure 4.5 depicts this graphical model validation. It demonstrates that the Cox-Snell residuals correlate

with the Exponential distribution quantiles for most of the points, thereby confirming that the PH assumption is met.

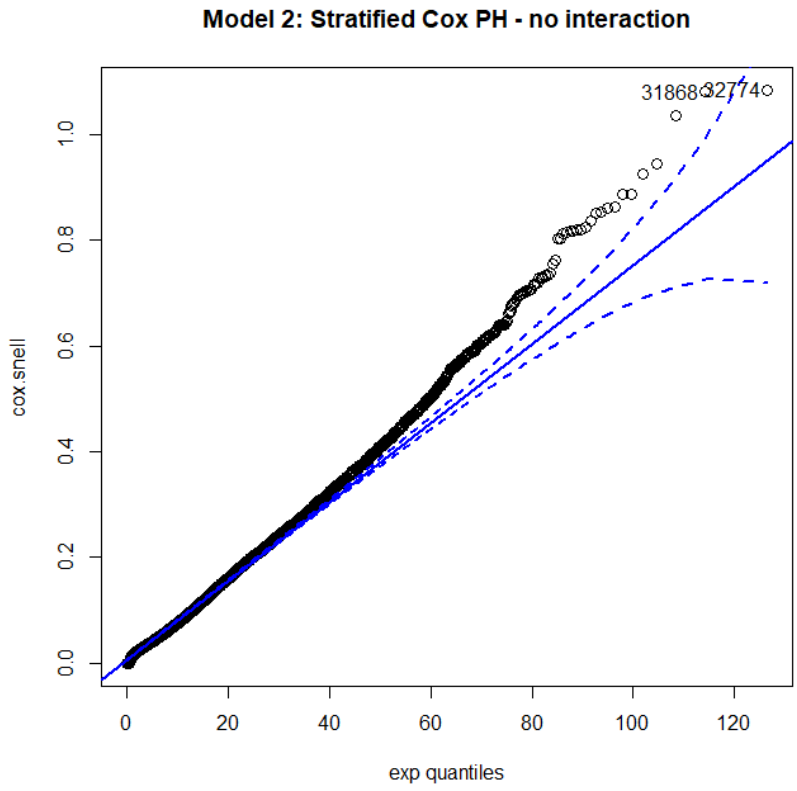


Figure 4.5: Stratified Cox PH: Cox-Snell residuals.

A nomogram for the Stratified Cox PH model, which displays multiple predicted survivor functions for the entire loan portfolio, is shown in Figure 4.6. It is observed that the predicted survival probabilities for 12 months are higher than those for 24 months and 36 months. This suggests that defaults are less probable during the first year of the loan. The risk of default gathers momentum during the second year of the loan term.

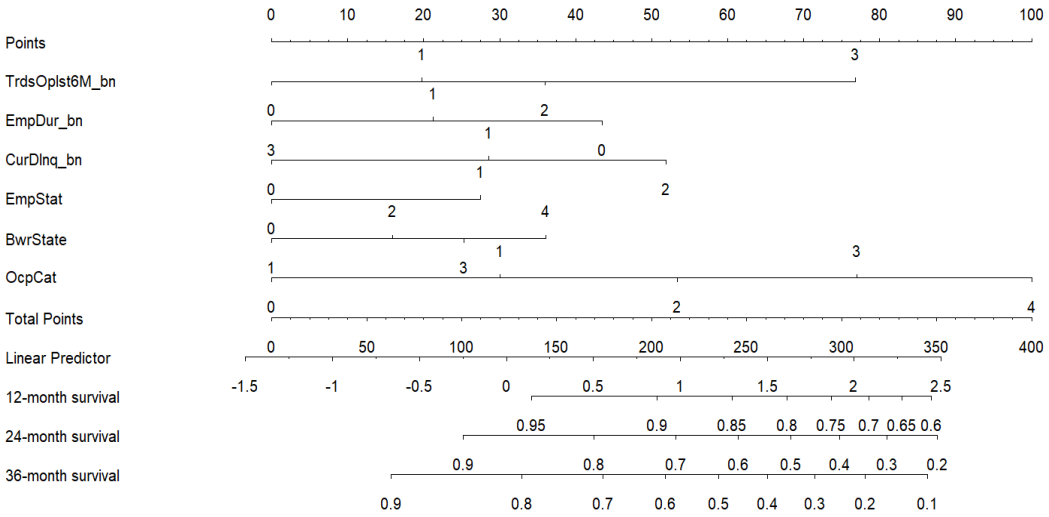


Figure 4.6: Stratified Cox PH: Nomogram.

4.3.2 Parametric Models

Variables that met the PH assumption were used to construct the Exponential and Weibull models. These two models are both PH and AFT models, thus justifying the inclusion of variables that satisfied the PH assumption. For both the Log-logistic and the Log-normal models, stepwise model selection was performed using the BIC to avoid overly complex models. This is because the BIC applies a larger penalty than AIC.

4.3.2.1 Validation using the Bootstrap

Table 4.7 shows the bootstrapped model performance results for model discrimination based on Somers' D index. Across all models, the optimism is marginal, meaning overfitting is negligible. Based on the bias-corrected Somers' D, the Log-logistic model and the Log-normal have comparable performance.

The results for the bootstrapped yearly BS and IPA for calibration performance, as well as the AUC for discrimination performance, are shown in Table 4.8. Based on

Table 4.7: Parametric models: Somers' D.

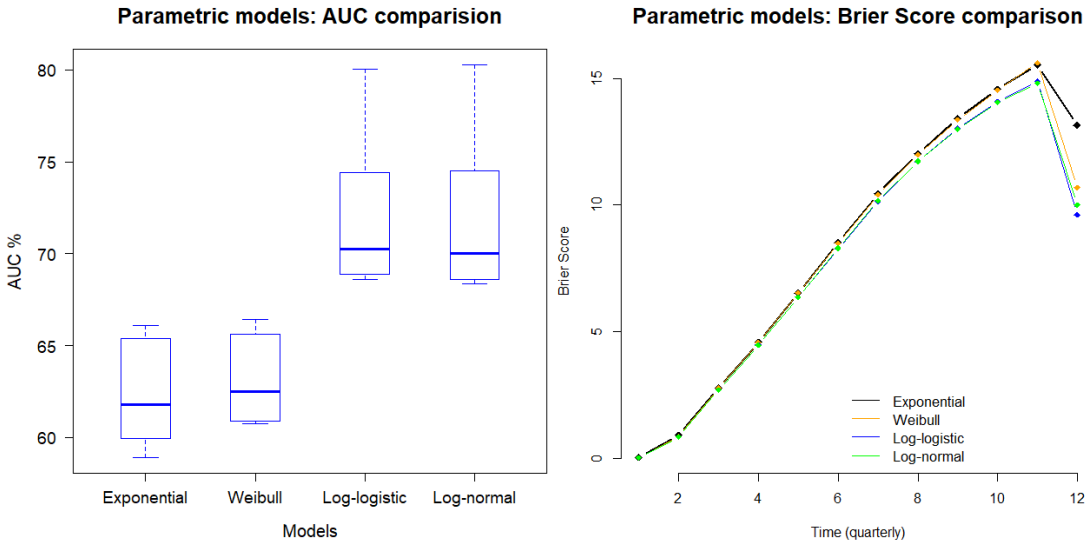
Model	Original sample	Training	Test	Optimism	Bias-corrected Index
Exponential	0.298	0.300	0.297	0.003	0.295
Weibull	0.300	0.300	0.297	0.003	0.297
Log-logistic	0.489	0.489	0.487	0.002	0.487
Log-normal	0.490	0.491	0.489	0.002	**0.488**

these evaluation metrics, the Log-Logistic and Log-Normal models performed better than the Exponential and Weibull models. Moreover, these two outperforming models perform similarly.

Table 4.8: Parametric models: BS, IPA, and AUC.

Period	Model	BS	IPA	AUC
Year 1	Exponential	4.592	0.825	65.403
	Weibull	4.573	1.202	65.451
	Log-logistic	**4.456**	**3.728**	74.360
	Log-normal	4.462	3.848	**74.494**
Year 2	Exponential	12.075	1.358	60.830
	Weibull	12.058	1.682	61.462
	Log-logistic	11.751	4.184	**68.718**
	Log-normal	**11.722**	**4.378**	68.545
Year 3	Exponential	13.178	-0.998	-
	Weibull	10.773	18.364	-
	Log-logistic	**9.692**	**26.555**	-
	Log-normal	9.993	24.000	-

The distribution of the AUC across three years of prediction, as well as the Brier Scores for the parametric models investigated, are displayed in Figure 4.7. From the results presented, the Log-Logistic and Log-Normal models outperformed the other parametric models investigated. These two models perform similarly in terms of discrimination and calibration (see Figures 4.7(a) and 4.7(b)).



(a) Parametric: AUC comparison (b) Parametric: Brier Score comparison

Figure 4.7: Parametric models: AUC and Brier Score comparison

4.3.2.2 Training / Test dataset Validation

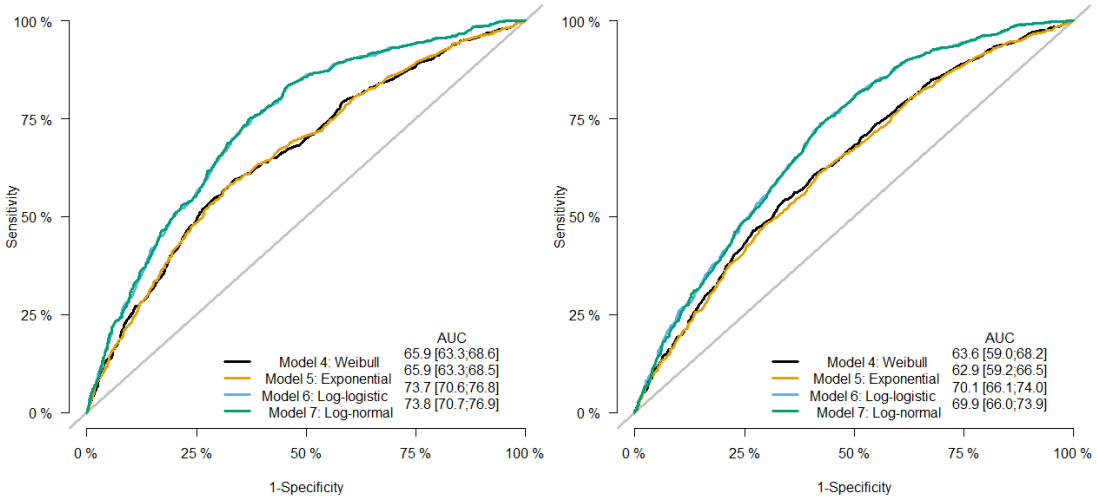
Based on the training c-index, AIC, BIC, and CAIC values displayed in Table 4.9, the best-fit model is the Log-normal model. In terms of calibration performance, AFT models outperform the Exponential model and the Weibull model, as evidenced by yearly IBS values based on the test dataset.

Table 4.9: Parametric models: c-index, AIC, BIC, CAIC, and IBS.

Model	Training Metrics				Test: IBS		
	c index	AIC	BIC	CAIC	Year 1	Year 2	Year 3
Exponential	0.647	46,523	46,670	46,597	1.333	4.853	7.945
Weibull	0.647	44,984	45,132	45,058	1.306	4.838	7.926
Log-logistic	0.744	43,333	43,497	43,415	1.280	4.719	**7.664**
Log-normal	**0.746**	**42,929**	**43,093**	**43,011**	**1.260**	**4.688**	7.675

The ROC curves of the Log-logistic and Log-normal models, depicted in Figure 4.8, are indistinguishable from one another and positioned much further to the top-left corner, indicating outperformance of these compared to the Exponential and Weibull

models. Moreover, the test AUC values are very similar to the bootstrap AUC values shown in Table 4.8, as is the ranking of the performance of the model.



(a) Parametric Models: Year 1 AUC (b) Parametric models: Year 2 AUC

Figure 4.8: Parametric models: ROC and AUC for years 1 and 2

4.3.2.3 Model Selection and Analysis for Parametric models

According to the model validation performance, the Log-logistic model and the Log-normal model outperformed other semi-parametric models presented. This section summarises the specific details about the fitted Log-logistic and Log-normal models. Firstly, the regression coefficients of these models are statistically significant at a 5% significance level according to the Wald test (see Appendices B.3 and B.6). Secondly, the forest plots for both these models logically rank time to default amongst different survival groups (see Appendices B.5 and B.8).

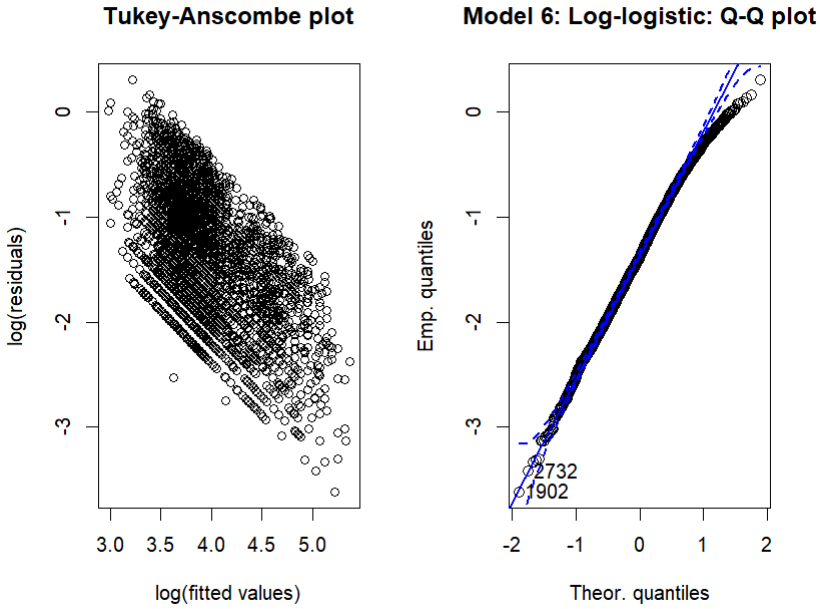
The risk ranking is summarised in Table 4.10, which provides an interpretation of the accelerated failure time constants for the Log-normal model. The survival ratios for the Log-logistic model are interpreted similarly since, based on the results, the difference between the regression coefficients of these two models is marginal. It is important to note that the risk ranking in terms of accelerated failure time constants or

Table 4.10: Interpretation of survival time ratios and risk ranking.

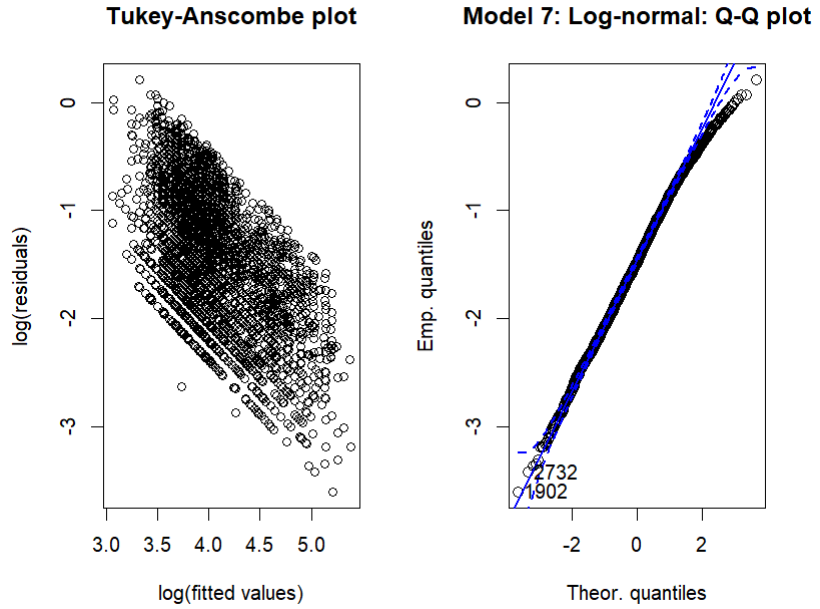
Predictor	**Effect relative to reference level**	**Survival Ratio Explanation**
AmntDlnq_bn	Negative	The borrowers in bucket 1 have a shorter time to default than borrowers in bucket 0, which is a reference the level, by a factor of 0.89.
InqLst6M_bn	Negative	Using bucket 0 as a reference, borrowers in buckets 1 have a shorter time to default by a factor of 0.85. Borrowers in bucket 2 have a shorter time to default by a factor of 0.80, while for the worst ranking buckets 3 and 4, the factor is 0.73 and 0.63, respectively.
BwrState	Negative	Using bucket 1 as a reference, borrowers in bucket 2 have a shorter time to default by a factor of 0.89, while borrowers in buckets 3 and 4 have a shorter time to default by a factor 0.83 and 0.78, respectively.
IncVerf	Negative	The borrowers in bucket 1 have a shorter time to default than borrowers in bucket 0, which is a reference the level, by a factor of 0.78.
PrpRatCat	Negative	Using bucket 0 as a reference, borrowers in buckets 1 have a shorter time to default by a factor of 0.79. Borrowers in bucket 2 have a shorter time to default by a factor of 0.61, while for the worst ranking buckets 3 and 4, the factor is 0.37 and 0.33, respectively.
OcpCat	Negative	Using bucket 0 as a reference, borrowers in buckets 1 have a shorter time to default by a factor of 0.83. Borrowers in bucket 2 have a shorter time to default by a factor of 0.78, while for the worst ranking buckets 3 and 4, the factor is 0.70 and 0.62, respectively.

survival time ratios described in Table 4.10 is consistent with the trends seen in the binning process, thus confirming the validity of the best performing parametric models investigated in the study. Plots of the empirical quantiles of the residuals versus theoretical residuals as well as the Tukey-Anscombe plots for the Log-logistic and the Log-normal models are displayed in Figures 4.9(a) and 4.9(b), respectively. These figures show a satisfactory fit, i.e., the empirical residual quantiles and the theoretical residuals are aligned along the 45-degree angle line for most of the data points, meaning that the quantiles correspond closely to each other. The Tukey-Anscombe plots for both models display residuals that have a similar distribution over all fitted values, thus further underlining the quality of the fitted

models.



(a) Log-logistic: Q-Q plot and Tukey-Anscombe plot



(b) Log-normal: Q-Q plot and Tukey-Anscombe plot

Figure 4.9: Q-Q plot and Tukey-Anscombe plot

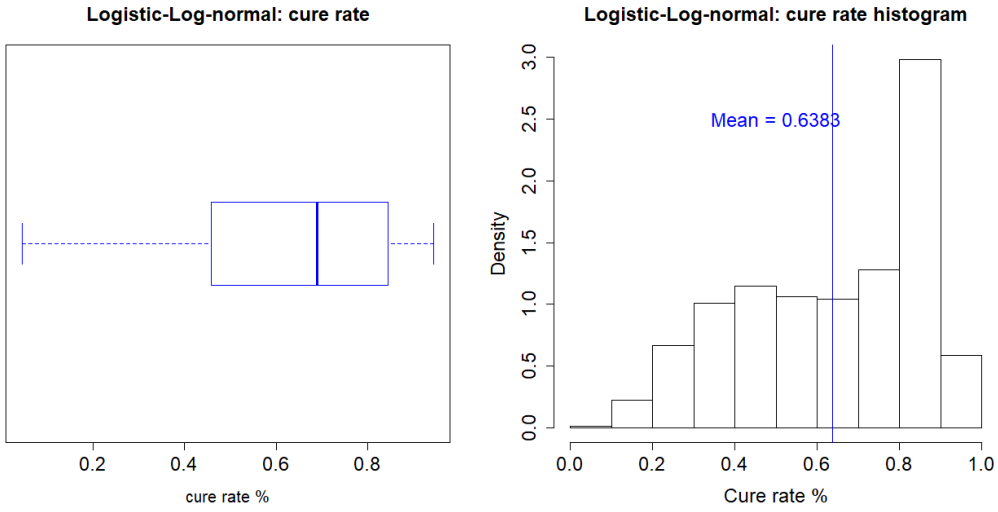
4.4 Mixture Cure Models

Comparison between parametric mixture models assuming different AFT models was assessed using AIB, BIC, and CAIC, where a lower value for each of these metrics indicates a better fit. For these models, evaluation metrics such as AUC, ROC, and BS are not readily available in standard `R` packages, so they must be developed, thus falling beyond the scope of this dissertation. The results in Figure 4.11 show that the Logistic-Log-normal mixture cure model is more efficient than the Logistic-Exponential, Logistic-Weibull, and Logistic-Log-logistic mixture cure models in the analysis of the Prosper dataset. Based on this top-performing model,

Table 4.11: Mixture Cure Parametric Models: AIC, BIC, and CAIC.

Mixture Cure Model	Evaluation Metrics		
	AIC	BIC	CAIC
Logistic-Exponential	45,257	45,422	45,339
Logistic-Weibull	43,232	43,405	43,318
Logistic-Log-logistic	43,127	43,300	43,214
Logistic-Log-normal	**42,943**	**43,116**	**43,029**

survivor functions were computed for all borrowers. Furthermore, the cure rates were estimated for the entire loan term, and are displayed in Figure 4.10.



(a) Boxplot: Cure probabilities (b) Histogram: Cure probabilities

Figure 4.10: Logistic-Log-normal cure rates

Using mixture cure models have an added benefit; these type of models make it relatively simple to identify distinct borrower cohorts and model the survival experience of borrowers who are more likely to default and those who are not. This is critical for IFRS 9 stage allocation or as part of a comprehensive credit risk management process that includes borrower segmentation and loss minimisation; non-susceptible borrowers are identified and assigned to lower risk buckets, freeing up capital that would otherwise have been held if no segmentation was performed. From Figure 4.10(b), which shows the histogram of the probability of cure, the average cure rate is 63.8%. The survival ratios and the corresponding risk ranking for uncured borrowers are summarised in Table 4.12.

Table 4.12: Interpretation of survival time ratios and risk ranking: Mixture cure model.

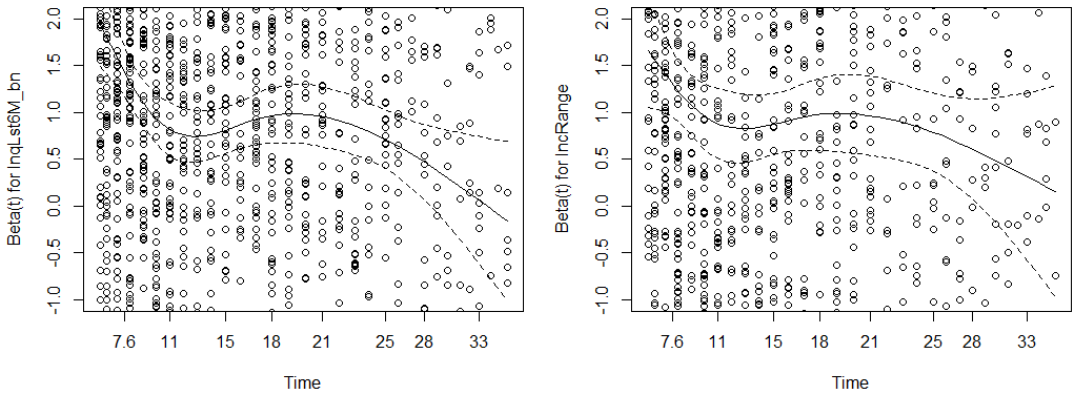
Predictor	**Effect relative to reference level**	**Survival Ratio Explanation**
AmntDlnq_bn	Negative	The borrowers in bucket 1 have a shorter time to default than borrowers in bucket 0, which is a reference the level, by a factor of 0.77.
InqLst6M_bn	Negative	Using bucket 0 as a reference, borrowers in buckets 1 have a shorter time to default by a factor of 0.70. Borrowers in bucket 2 have a shorter time to default by a factor of 0.60, while for the worst ranking buckets 3 and 4, the factor is 0.47 and 0.36, respectively.
BwrState	Negative	Using bucket 1 as a reference, borrowers in bucket 2 have a shorter time to default by a factor of 0.73, while borrowers in buckets 3 and 4 have a shorter time to default by a factor 0.64 and 0.55, respectively.
IncVerf	Negative	The borrowers in bucket 1 have a shorter time to default than borrowers in bucket 0, which is a reference the level, by a factor of 0.50.
PrpRatCat	Negative	Using bucket 0 as a reference, borrowers in buckets 1 have a shorter time to default by a factor of 0.55. Borrowers in bucket 2 have a shorter time to default by a factor of 0.33, while for the worst ranking buckets 3 and 4, the factor is 0.10 and 0.08, respectively.
OcpCat	Negative	Using bucket 0 as a reference, borrowers in buckets 1 have a shorter time to default by a factor of 0.67. Borrowers in bucket 2 have a shorter time to default by a factor of 0.57, while for the worst ranking buckets 3 and 4, the factor is 0.44 and 0.33, respectively.

According to Table 4.12, the survival time ratios are consistent with the trends observed in the binning, and hence the model ranks risk accurately. Higher-risk groups have a shorter time to default when compared to those estimated using standard parametric models, as evidenced by the magnitude of survival ratios. This makes sense given that the survivor function modelled using mixture cure models is only for borrowers who are at risk of default.

The AIC, BIC, and CAIC evaluation metrics show that mixture cure models outperform the standard models investigated. The top-performing mixture cure model, for instance, the Log-normal mixture cure model, differs marginally from the top-performing standard parametric AFT models. However, standard models are incapable of categorising borrowers into two distinct cohorts depending on their susceptibility to default.

4.5 Extended Cox Regression

Covariates that violated the PH assumption were incorporated into the Cox regression model by allowing their coefficients to vary with time. Figure 4.11 shows how the coefficients of the covariates *InqLst6M_bn* and *IncRange* vary with time.



(a) Schoenfeld residuals: binned InqLst6M

(b) Schoenfeld residuals: IncRange

Figure 4.11: Time varying log hazard ratios

The Extended Cox with time step function incorporates these variables by making use of a step function that partitions the time index according to the evolution of $\beta(t)$ shown in Figure 4.11. More specifically, the regression coefficient of *InqLst6M_bn* and *IncRange* appear to have a sharp decline in the first 9 months, then followed by a gradual increase up to 18 months and finally decreasing for the remainder of the period. Thus, this model divides the data into three epochs of the first 9 months, 9-18 months, and greater than 18 months.

Appendix B.1 shows that the log-hazard ratios, as well as the time interactions, are significant at a 5% significance level. The GOF test results of the test of the correlation between the Schoenfeld residual and the survival time are shown in Appendix B.2. Table 4.13 shows the bootstrapped model performance results for model discrimination.

Table 4.13: Extended Cox with Step function: Somers' D.

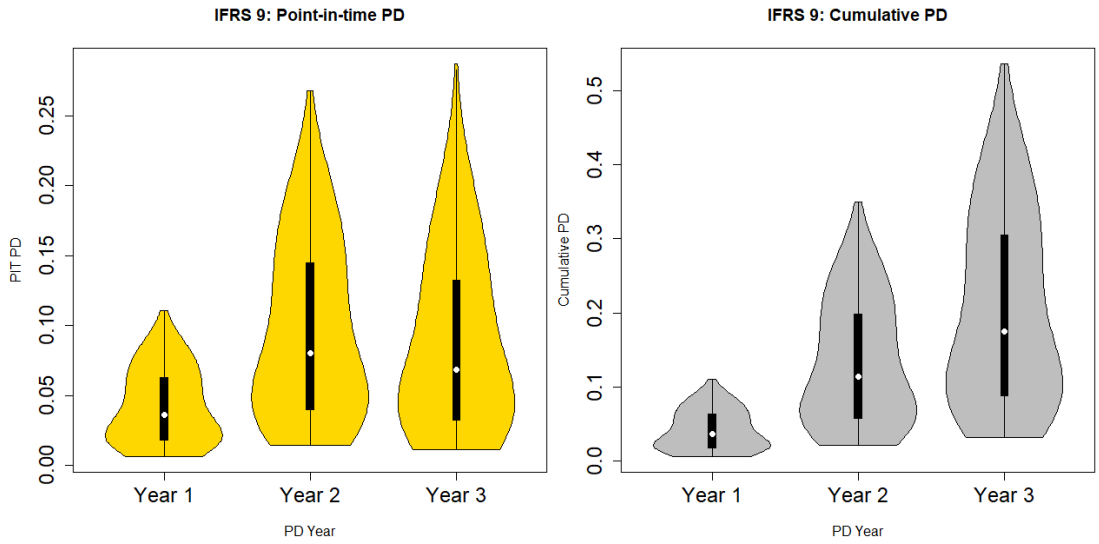
Original Sample	Training	Test	Optimism	Bias-corrected Index
0.412	0.415	0.407	0.005	0.407

The c-index and the CAIC is 71.21% and 73,976, respectively, which is a marked improvement compared to the standard Cox PH model. Unfortunately, due to a pre-processing step of the data into the start-stop format, computation of other performance measures such as bootstrap IB, AUC, and IPS are beyond the scope of our study.

4.6 Point-in-time PDs and Capital Impact

In this section, an overview of the PIT PDs and credit lifetime PDs calculated using Equations 2.19 and 2.20, respectively, is provided. Figure 4.12(a) shows violin plots of the PIT PDs for the entire loan portfolio. The PDs are computed using the Logistic-Log-normal mixture cure model, which is the best performing model among the mixture cure models investigated. The spike in the median PIT PD during the second year of the loan shows that the default experience has worsened. Moreover, the PDs are more dispersed in years two and three, signalling the need to increase

capital provisions as a loss buffer for those periods. In year three, the median PIT PD declines. This is commonly observed for term loans because, following a turbulent period in which defaults were more prevalent, the loan portfolio now includes borrowers who are less likely to default as their loans approach maturity.



(a) PIT PD (b) Cumulative PD

Figure 4.12: PIT PD and Cumulative PDs

In general, the shape of the distribution for PIT PDs in Figure 4.12(a) suggests that yearly PDs are largely concentrated around their respective lower quartile. The cumulative PDs shown in Figure 4.12(b) exhibit the same pattern. The lifetime ECL for the Prosper loan portfolio for loans originated between 2009 and 2014 is $7.91 million as shown in Figure 4.13.

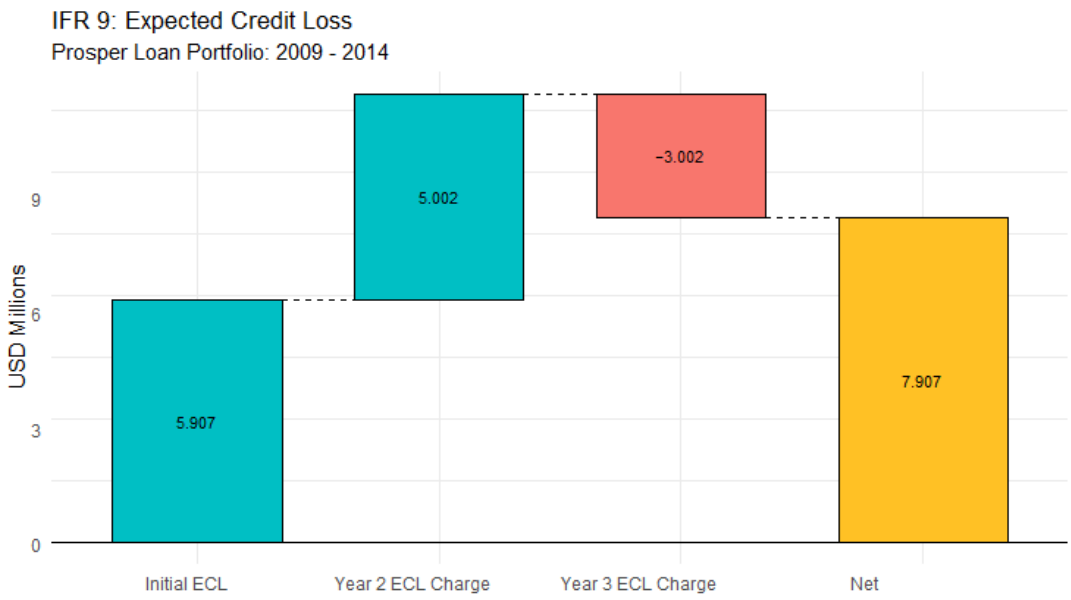


Figure 4.13: IFRS 9: expected credit loss

As there is no collateral variable in the data set, the loans in the portfolio are assumed to be unsecured. For unsecured transactions, the BCBS recommends an LGD of 50%. For this simplified calculation of ECL, the LGD was kept constant for the duration of the loan. The loan amount at origination plus three months' worth of interest is a proxy for EAD. This is consistent with the definition of default used in this study. Beginning with a 12-month ECL of \$5.91 million estimated using PIT PDs in year one, capital provisions increased by \$5 million during year two. This corresponds to the predicted PD profile, which peaks in the second year as more defaults are expected. In the final year of the loan, PDs are expected to decline such that net provisions are \$7.91 million, resulting from a reduction of \$3 million in capital provisions.

4.7 Summary

Several survival analysis models were employed to estimate PD and predict default times. The study investigated the entire loan portfolio containing three-year loans, and each account was monitored monthly from entry to exit. The dataset covered a

five-year period which aligns with the minimum historical data period required to estimate PDs. Multiple imputation techniques were employed where data was missing in order to obtain complete datasets for analysis. For PH models, potential covariates were investigated for the PH assumption using the GOF test of the correlation between the Schoenfeld residuals and time, as well as the plot of the log hazards versus time. Important covariates (based on IV and Gini) that failed the PH assumption were incorporated into the stratified Cox PH model. In addition, an extended Cox regression model was implemented to adjust for time-varying coefficients.

Model validation was performed using bootstrap and the one-time 80:20 training and test data split. Model performance and complexity were quantified based on probabilistic statistical measures such as AIC, BIC, and CAIC. Validation for model discrimination performance was measured using ROC, AUC, c-index, and Somers' D. Validation for model calibration performance was assessed using calibration plots, BS, IPA, and IBS. Nomograms were also constructed to visualise estimated survivor functions for the entire loan portfolio. The stratified Cox regression, the Logistic-Log-Normal mixture cure model, and two of the AFT models, namely the Log-Logistic and Log-Normal models, outperformed all the other models. The Logistic-Log-normal mixture model performed similarly to the standard Log-normal model in terms of AIC, BIC, and CAIC. However, mixture cure models have an advantage in that they allow for the estimation of the probability of cure, making it reasonably easier to identify borrowers that are more at risk of default and estimate their survival experience. In addition, the Cox PH model with Splines performed better than the standard Cox PH across most performance measures, confirming the results from earlier studies by Luo et al. [2016] and Dirick et al. [2017]. Moreover, incorporating time-varying effects to model time-varying regression coefficients or time-dependent covariance improves predictive performance significantly, which is confirmed by Bellotti and Crook [2009].

An estimate of lifetime ECL was provided for the entire loan portfolio. Calculating lifetime PDs as part of the credit risk management process allows intermittent capital provisions, which fluctuate in response to changes in the borrowers' risk profile, to be part of a broader risk budgeting plan. This is crucial because it helps credit providers to

design a solid capital budgeting plan that matches the expected credit risk experience of the entire portfolio, as required by IFRS 9.

Chapter 5

Summary and Conclusion

5.1 Introduction

Traditionally, banks have been developing models that estimate one-year PD models, a practice in line with the Basel II regulatory requirements. The 2007 – 2009 financial crisis paved a path for a revised approach to capital provisioning as part of the bank's credit risk management. The IASB and FASB worked together to update accounting standards in favour of an expected credit loss paradigm, which necessitates the use of forward-looking PDs. The IFRS 9 became a cornerstone of credit risk management in mid-2014 and has gained momentum since then. This study investigated the application of survival analysis in estimating forward-looking PDs as part of the bank's credit risk management under the IFRS 9.

5.2 Limitations and Recommendations

The execution of the study was impeded by a lack of powerful computational resources. This was notably true for the Logistic-Cox PH mixture cure model. As a result, this model was not explored further. Instead, various parametric mixture cure models were applied, and the findings showed that these models performed significantly better than

standard models.

According to the findings, for non-proportional hazards, the study recommends using the Logistic-Log-normal mixture cure model. Cox regression, on the other hand, is less constrained because it makes no parametric assumptions about the distribution of time to default. For situations where powerful covariates fail the PH assumption, the Cox regression model offers a way to incorporate them via stratification or adjusting for time-varying coefficients by splitting the time-axis according to Schoenfeld residuals. Moreover, time-varying covariates can be incorporated using other extended versions of the Cox regression, which is crucial for the situation where external longitudinal macroeconomic factors are required. This study recommends exploring a wide range of mixture cure and Cox regression methodologies for estimating PIT PDs.

5.3 Conclusion

According to the results of this study, survival analysis is an effective method for calculating PIT PDs, and hence plays a critical role in credit risk management. Furthermore, because survival analysis can handle censoring, it is ideal for calculating the IFRS 9 impairments for the entire loan portfolio of retail bank without having to exclude accounts with incomplete information. In the study, it was demonstrated that loss loan provisioning, which is commonly used by traditional banks, can be extended to P2P lending platforms. The survival models developed were based on a range of statistically valid methods supported by the literature. The findings from the top-performing models, namely, the Logistic-Log-normal, standard Log-normal, and stratified Cox regression models, were satisfactory in terms of model complexity, ability to discriminate prognosis in the future, and calibration performance. The study highlights further benefits associated with each model type. Cox regression methodologies offer flexibility to perform stratification and observe survival experience within each stratum. Moreover, Cox regression can handle non-proportional hazard via time-varying log hazard ratios in its extended iterations, which is especially relevant in credit risk management where external longitudinal macroeconomic factors are involved. On the other hand, mixture cure models make it

relatively simple to identify distinct borrower cohorts, which is critical in credit risk management.

5.4 Further Work

Further exploration may entail incorporating panel data consisting of covariates tied to the general economic fundamentals. Retail banks use this methodology to obtain macro-economic sensitive PIT PDs. In addition, PIT PDs can be estimated using machine learning-inspired survival analysis approaches such as Random Forest. The method can easily be adapted to include longitudinal and multivariate outcomes, which may be useful in assessing competing risks. Furthermore, because evaluation metrics for mixture cure models are not readily available, they can be developed in future work.

References

Aalen, O. (1976). Nonparametric inference in connection with multiple decrement models. *Scandinavian Journal of Statistics*, 3:15–27. 6, 19

Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3):370–374. 56

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609. 2

Altshuler, B. (1970). Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences*, 6:1–11. 6, 19

Amico, M. and Van Keilegom, I. (2018). Cure models in survival analysis. *Annual Review of Statistics and Its Application*, 5:311–342. 34, 36, 39

Ata, N. and Sözer, M. T. (2007). Cox regression models with nonproportional hazards applied to lung cancer survival data. *Hacettepe Journal of Mathematics and Statistics*, 36(2):157–167. 27, 28, 31

Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., Tiburtius, P., and Funk, B. (2011). Online peer-to-peer lending-a literature review. *Journal of Internet Banking and Commerce*, 16(2):1–18. 11

Baesens, B., Van Gestel, T., Stepanova, M., Van den Poel, D., and Vanthienen, J. (2005). Neural network survival analysis for personal loan data. *Journal of the Operational Research Society*, 56(9):1089–1098. 8

Bellini, T. (2019). *IFRS 9 and CECL Credit Risk Modelling and Validation: A Practical Guide with Examples Worked in R and SAS*. Academic Press, London, UK. 2, 3, 13, 14

Bellotti, T. and Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, 60(12):1699–1707. 9, 91

Beran, J. and Djaïdja, A.-Y. K. (2007). Credit risk modeling based on survival analysis with immunes. *Statistical Methodology*, 4(3):251–276. 8

Berkovich, E. (2011). Search and herding effects in peer-to-peer lending: evidence from prosper. com. *Annals of Finance*, 7(3):389–405. 11

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370. 67

Breslow, N. E. (1972). Discussion of Professor Cox's paper. *Journal of the Royal Statistical Society: Series B*, 34:216–217. 26, 38

Cai, C., Zou, Y., Peng, Y., and Zhang, J. (2012). smcure: An R-package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, 108(3):1255–1260. 37

Cao, R., Vilar, J. M., and Devia, A. (2009). Modelling consumer credit risk via survival analysis. *SORT: Statistics and Operations Research Transactions*, 33(1):3–30. 43

Carling, K., Jacobson, T., Lindé, J., and Roszbach, K. (2007). Corporate credit risk modeling and the macroeconomy. *Journal of Banking & Finance*, 31(3):845–868. 8

Chen, X., Zhou, L., and Wan, D. (2016). Group social capital and lending outcomes in the financial credit market: An empirical study of online peer-to-peer lending. *Electronic Commerce Research and Applications*, 15:1–13. 11

Chen, Y. Q. and Wang, M.-C. (2000). Analysis of accelerated hazards models. *Journal of the American Statistical Association*, 95(450):608–618. 35

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge, New York, USA. 60

Cox, C., Chu, H., Schneider, M. F., and Munoz, A. (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine*, 26(23):4352–4374. 39, 40

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202. 7, 21

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276. 24

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):248–265. 62

Dabrowska, D. M., Doksum, K. A., Feduska, N. J., Husing, R., and Neville, P. (1992). Methods for comparing cumulative hazard functions in a semi-proportional hazard model. *Statistics in Medicine*, 11(11):1465–1476. 61

Daoud, J. I. (2017). Multicollinearity and regression analysis. In *Journal of Physics: Conference Series*, volume 949, page 012009. IOP Publishing. 56

De Boor, C. and De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-verlag, New York, USA. 32

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22. 36

Dirick, L., Claeskens, G., and Baesens, B. (2017). Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society*, 68(6):652–665. 9, 10, 15, 31, 91

Djeundje, V. B. and Crook, J. (2019). Identifying hidden patterns in credit risk survival data using generalised additive models. *European Journal of Operational Research*, 277(1):366–376. 9

Edwards, G. (2014). The upcoming new era of expected loss provisioning. *SEACEN Financial Stability Journal*, 2:13–24. 13

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565. 27

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331. 63, 64

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560. 64

Emmert-Streib, F. and Dehmer, M. (2019). Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, 1(3):1013–1038. 20

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158. 20

Farewell, V. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38(4):1041–1046. 10

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, 14(3):257–262. 10

Fisher, L. D. and Lin, D. Y. (1999). Time-dependent covariates in the cox proportional-hazards regression model. *Annual Review of Public Health*, 20(1):145–157. 29

Fitzpatrick, T. and Mues, C. (2021). How can lenders prosper? comparing machine learning approaches to identify profitable peer-to-peer loan investments. *European Journal of Operational Research*, 294(2):711–722. 11

Fleming, T. R. and Harrington, D. P. (1984). Nonparametric estimation of the survival distribution in censored data. *Communications in Statistics-Theory and Methods*, 13(20):2469–2486. 21

Florez-Lopez, R. and Ramon-Jeronimo, J. M. (2014). Modelling credit risk with scarce default data: on the suitability of cooperative bootstrapped strategies for small low-default portfolios. *Journal of the Operational Research Society*, 65(3):416–434. 9

Frykström, N. and Li, J. (2018). IFRS 9–the new accounting standard for credit loss recognition. *Economic Commentaries*, 3:1–13. 3, 12

Gerds, T. A., Andersen, P. K., and Kattan, M. W. (2014). Calibration plots for risk prediction models in the presence of competing risks. *Statistics in Medicine*, 33(18):3191–3203. 67

Gini, C. (1936). On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series*, 208(1):73–79. 58

Glaser, R. E. (1980). Bathtub and related failure rate characterizations. *Journal of the American Statistical Association*, 75(371):667–672. 40

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526. 63

Greenwood, M. et al. (1926). A report on the natural duration of cancer. *A Report on the Natural Duration of Cancer.*, (33):1–26. 18

Hague, I. P. (2004). IAS 39: Underlying principles. *Accounting in Europe*, 1(1):21–26. 3

Hall, P. and Opsomer, J. D. (2005). Theory for penalised spline regression. *Biometrika*, 92(1):105–118. 32

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546. 63, 66

Harrell, F. E. et al. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, volume 608. Springer, New York, USA. 6, 17, 18, 20, 24, 25, 29, 44, 63

Hashim, N., Li, W., and O'Hanlon, J. (2016). Expected-loss-based accounting for impairment of financial instruments: The FASB and IASB proposals 2009–2016. *Accounting in Europe*, 13(2):229–267. 13

Héberger, K. and Rajkó, R. (2002). Variable selection using pair-correlation method. environmental applications. *SAR and QSAR in Environmental Research*, 13(5):541–554. 60

Hong, S. and Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Medical Research Methodology*, 20(1):1–12. 45, 46, 70

Jiang, C., Wang, Z., and Zhao, H. (2019). A prediction-driven mixture cure model and its application in credit scoring. *European Journal of Operational Research*, 277(1):20–31. 2, 32

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, New Jersey, USA. 19, 24, 29, 38

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481. 6

Keramati, A. and Yousefi, N. (2011). A proposed classification of data mining techniques in credit scoring. In *the Proceeding of 2011 International Conference of Industrial Engineering and Operations Management*, pages 22–24, Kuala Lumpur, Malaysia. Journal. 2

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer, New York, USA. 18, 30, 31

Kleinbaum, D. G. and Klein, M. (2010). *Survival analysis*. Springer, Atlanta, ATL. 6, 7, 14, 16, 23, 27, 40

Kronek, L.-P. and Reddy, A. (2008). Logical analysis of survival data: prognostic survival models by detecting high-degree interactions in right-censored data. *Bioinformatics*, 24(16):i248–i253. 67

Kuk, A. Y. and Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–541. 9, 34, 39

Leo, M. and Sharma, S. (2019). Machine learning in banking risk management: A literature review. *Risks*, 7(1):1–22. 1, 2

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287–296. 45

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons, Michigan, USA. 45

Luo, S., Kong, X., and Nie, T. (2016). Spline based survival model for credit risk modeling. *European Journal of Operational Research*, 253(3):869–879. 9, 31, 91

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316. 65

Mariotto, C. (2016). Competition for Lending in the Internet era: The case of Peer-to-Peer Lending Marketplaces in the USA. *Communications & Strategies*, 8(103):35. 11

Mays, E. (2001). *Handbook of credit scoring*. Global Professional Publishing, London, UK. 54

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2):449–470. 8

Moore, D. F. (2016). *Applied survival analysis using R*. Springer, New Jersey, USA. 16, 19, 20, 25, 26, 29

Narain, B. (1992). Survival analysis and the credit granting decision. *Credit Scoring and Credit Control*, 109:121. 7, 10

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966. 6, 19

Noh, H. J., Roh, T. H., and Han, I. (2005). Prognostic personal credit risk model considering censored information. *Expert Systems with Applications*, 28(4):753–762. 2, 7

O'Sullivan, F. (1986). A Statistical Perspective on lll-posed Inverse Problem. *Statistical Science*, 1(4):502–527. 32

Peng, Y. (2003). Fitting semiparametric cure models. *Computational Statistics & Data Analysis*, 41(3):481–490. 9, 33, 37

Peng, Y. and Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1):237–243. 39

Peng, Y., Dear, K. B., and Denham, J. (1998). A generalized F mixture model for cure rate estimation. *Statistics in Medicine*, 17(8):813–830. 10

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., and Schmid, M. (2019). A review of spline function procedures in R. *BMC Medical Research Methodology*, 19(1):1–16. 31

Prentice, R. L. (1974). A log gamma model and its maximum likelihood estimation. *Biometrika*, 61(3):539–544. 39

Qing Chen, Y. (2001). Accelerated hazards regression model and its adequacy for censored survival data. *Biometrics*, 57(3):853–860. 35

Renke, S., Roestoff, M., and Haupt, F. (2007). The National Credit Act: New parameters for the granting of credit in South Africa. *Obiter*, 28(2):229–270. 5, 14

Roestoff, M., Haupt, F., Coetzee, H., and Erasmus, M. (2009). The debt counselling process: Closing the loopholes in the National Credit Act 34 of 2005. *PER: Potchefstroomse Elektroniese Regsblad*, 12(4):247–305. 5

Sarlija, N., Bensic, M., and Zekic-Susac, M. (2009). Comparison procedure of predicting the time to default in behavioural scoring. *Expert Systems with Applications*, 36(5):8778–8788. 8

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241. 62

Schouten, R. M., Lugtig, P., and Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930. 49

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study. *American Journal of Epidemiology*, 179(6):764–774. 46

Siddiqi, N. (2012). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, volume 3. John Wiley & Sons, New Jersey, N.J. 52, 53, 54, 58

Stepanova, M. and Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research*, 50(2):277–289. 8, 56

Sy, J. P. and Taylor, J. M. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56(1):227–236. 38

Therneau, T., Crowson, C., and Atkinson, E. (2017). Using time dependent covariates and time dependent coefficients in the Cox model. *Survival Vignettes*, 2(3):1–25. 30

Thomas, L. C., Ho, J., and Scherer, W. T. (2001). Time will tell: behavioural scoring and the dynamics of consumer credit assessment. *IMA Journal of Management Mathematics*, 12(1):89–103. 7, 24

Thompson, C. G., Kim, R. S., Aloe, A. M., and Becker, B. J. (2017). Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. *Basic and Applied Social Psychology*, 39(2):81–90. 56

Tong, E. N., Mues, C., and Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research*, 218(1):132–139. 10, 32

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45:1–67. 52

Van Buuren, S. and Oudshoorn, K. (1999). *Flexible multivariate imputation by MICE*. Leiden: TNO, Wassenaarseweg, Netherlands. 44

Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of "permanent employment" in Japan. *Journal of the American Statistical Association*, 87(418):284–292. 10

Zeng, G. (2014). A necessary condition for a good binning algorithm in credit scoring. *Applied Mathematical Sciences*, 8(65):3229–3242. 54

Zhang, J. and Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine*, 26(16):3157–3171. 38

Zhang, J. and Peng, Y. (2009). Accelerated hazards mixture cure model. *Lifetime Data Analysis*, 15(4):455–467. 36, 38

Appendix A

Additional Results

Table A.1: Binned predictors and bin definitions.

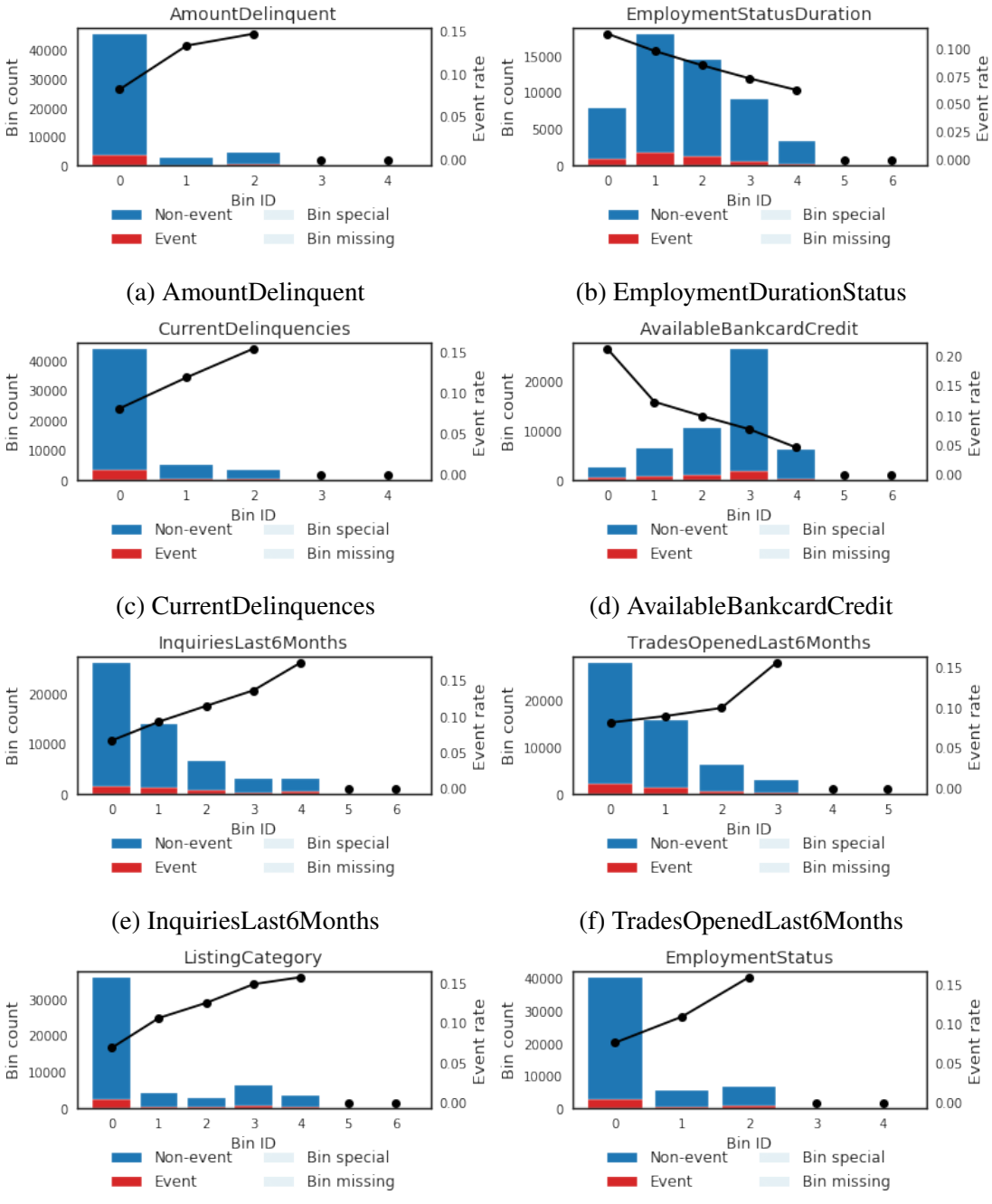
Year	Bin ID	Attribute	WOE
AmntDlnq_bn	0	< 3.50	0.105
	1	$[3.50, 438.50)$	-0.438
	2	≥ 438.50	-0.553
EmpDur_bn	0	< 15.50	-0.255
	1	$[15.50, 69.50)$	-0.090
	2	$[69.50, 150.50)$	0.065
	3	$[150.50, 285.50)$	0.228
	4	≥ 285.50	0.394
CurDlnq_bn	0	< 0.50	0.114
	1	$[0.50, 1.50)$	-0.315
	2	≥ 1.50	-0.617
AvlBnkCred_bn	0	< 47.50	-1.008
	1	$[47.50, 509.50)$	-0.348
	2	$[509.50, 2322.50)$	-0.105
	3	$[2322.50, 26326.50)$	0.176
	4	≥ 26326.50	0.732
InqLst6M_bn	0	< 0.50	0.330
	1	$[0.50, 1.50)$	-0.033
	2	$[1.50, 2.50)$	-0.270
	3	$[2.50, 3.50)$	-0.468
	4	≥ 3.50	-0.764

Table A.2: Binned predictors and bin definitions (continued).

Year	Bin ID	Attribute	WOE
TrdsOplst6M_bn	0	< 0.50	0.111
	1	$[0.50, 1.50)$	0.007
	2	$[1.50, 2.50)$	-0.109
	3	≥ 2.50	-0.623
ListCat	0	[0, 1, 8, 9, 10, 11, 14, 15, 16, 17, 20]	0.298
	1	[2]	-0.180
	2	[6, 13]	-0.366
	3	[7]	-0.565
	4	[3, 5, 12]	-0.633
EmpStat	0	[Employed]	0.190
	1	[Other, Self-employed]	-0.208
	2	[Retired, Not employed]	-0.644
InvVerf	0	[TRUE]	0.086
	1	[FALSE]	-0.610
OccCat	0	[Dentist, Judge, Doctor, . . .]	0.801
	1	[Computer Programmer, Electricia1 Engineer, . . .]	0.418
	2	[Teacher, Tradesman - Mechanic, . . .]	0.177
	3	[Construction, Postal Service, Bus Driver, . . .]	-0.199
	4	[Administrative Assistant, Religious, Student, . . .]	-0.454
BwrState	0	[MT, WY, NE, DC, CO, MA, AR, AZ, OR, . . .]	0.247
	1	[HI, NH, VT, DE, OK, IL, CT, MN, KS]	0.018
	2	[ID, VA, WA, NV, SC, FL, AK, PA, WV]	-0.094
	3	[WI, GA, MS, AL, OH, MO, NC]	-0.241
	4	[MD, LA, TN, RI, KY, SD]	-0.384
IncRange	0	[100,000+]	0.710
	1	[75,000 - 99,999]	0.467
	2	[50,000 - 74,999]	0.167
	3	[25,000 - 49,999]	-0.352
	4	[Not Employed, 0, 1 - 24,999]	-0.806

Note that Table A.2 uses shortened variable names. The actual long names are listed

in Table 3.5.



(a) AmountDelinquent

(b) EmploymentDurationStatus

(c) CurrentDelinquences

(d) AvailableBankcardCredit

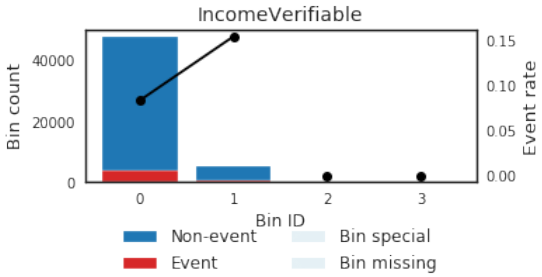
(e) InquiriesLast6Months

(f) TradesOpenedLast6Months

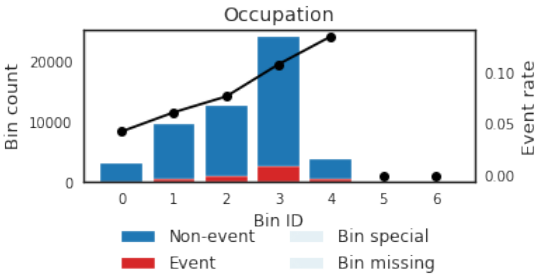
(g) ListingCategory

(h) EmploymentStatus

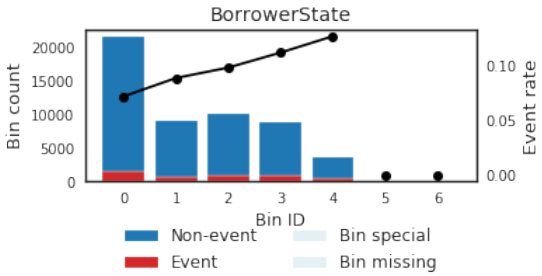
Figure A.1: Initial variable binning process



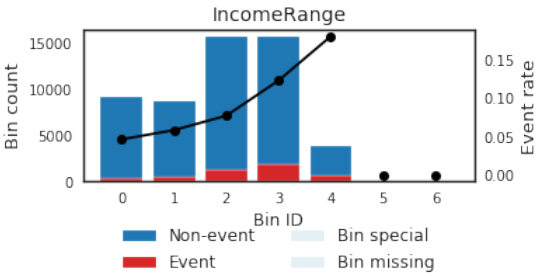
(a) IncomeVerifiable



(b) Occupation



(c) BorrowerState



(d) IncomeRange

Figure A.2: Initial variable binning process (continued)

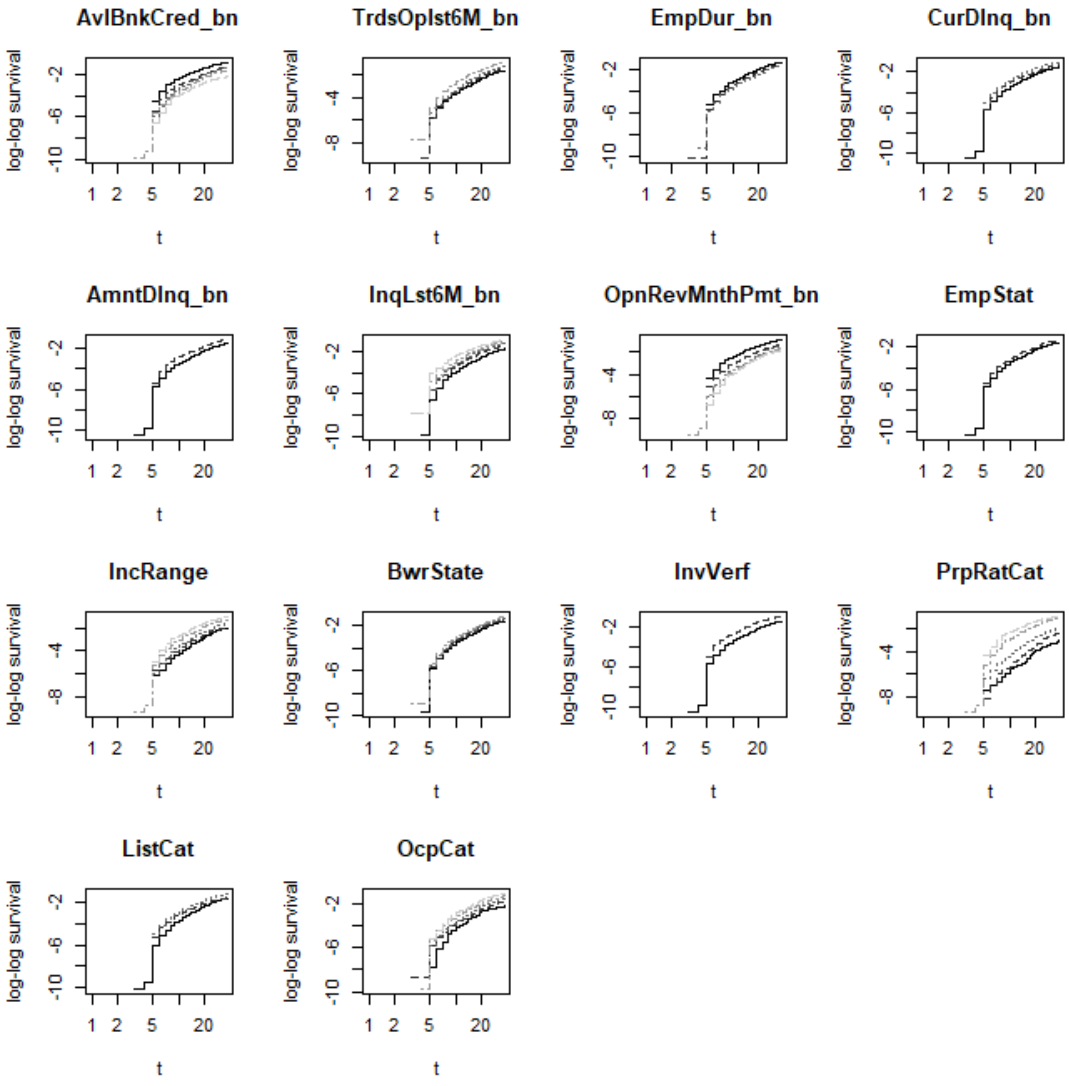
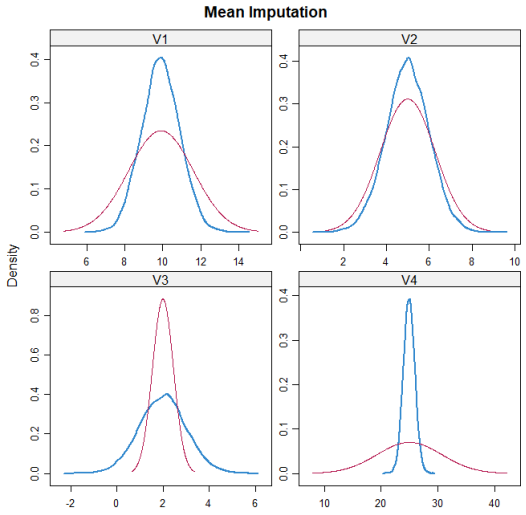
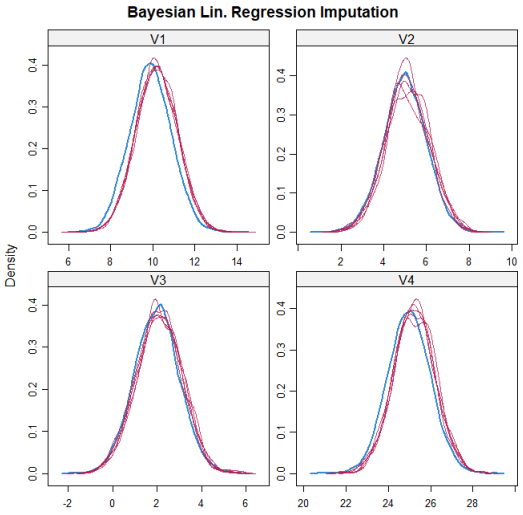


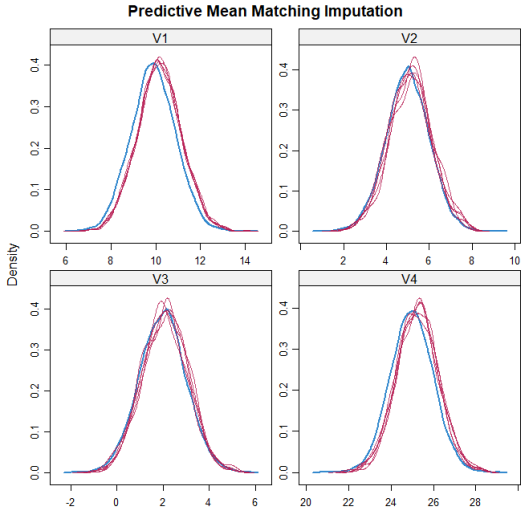
Figure A.3: Log hazard plots.



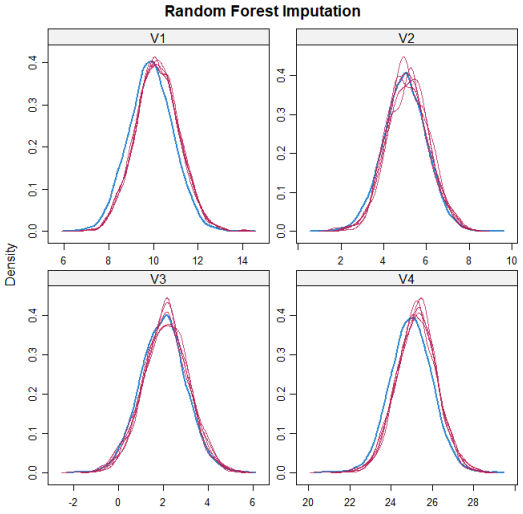
(a) Mean



(b) Bayesian Linear Regression

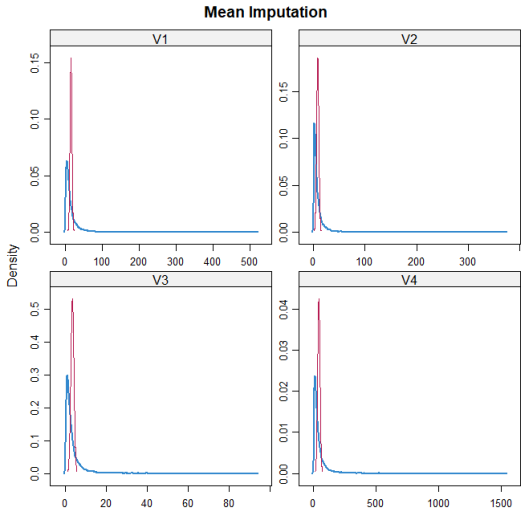


(c) Predictive Mean Matching

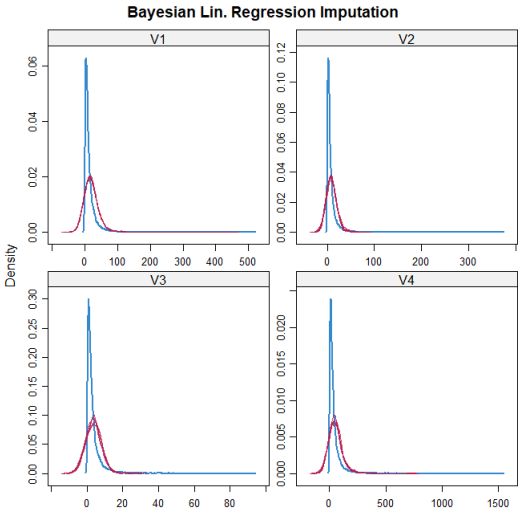


(d) Random Forest

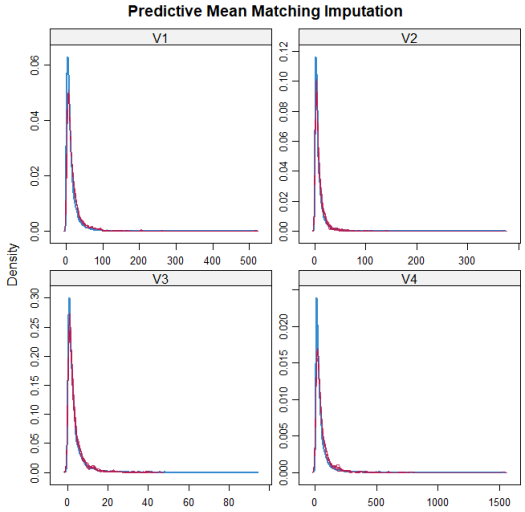
Figure A.4: MICE Imputation: A



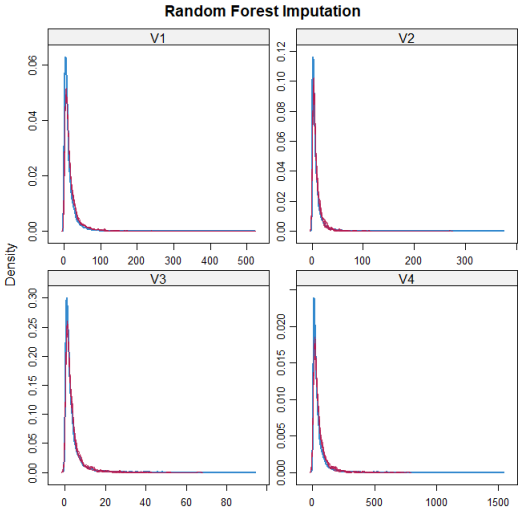
(a) Mean



(b) Bayesian Linear Regression



(c) Predictive Mean Matching



(d) Random Forest

Figure A.5: MICE Imputation: B

Appendix B

Extra Results

B.1 Kaplan-Meier Estimator

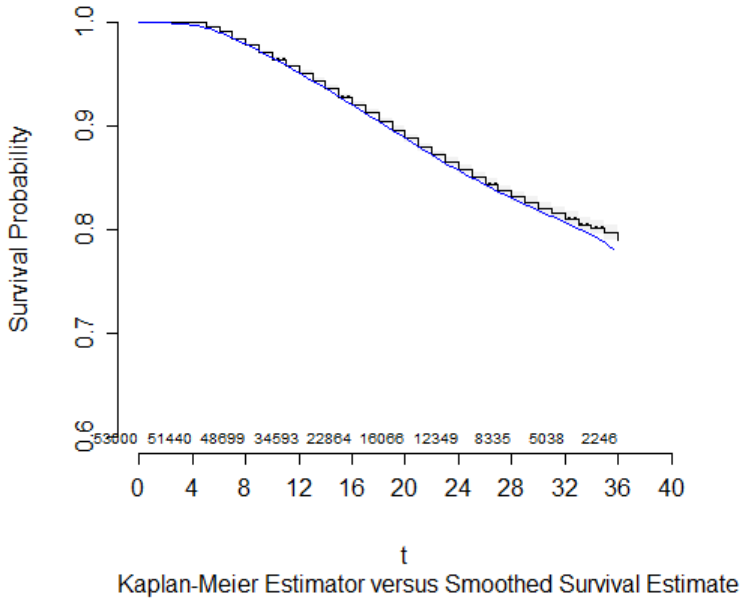


Figure B.1: KM estimator.

B.2 Model 2: Stratified Cox PH

$$\text{Prob}\{T \geq t|\mathbf{x}\} = S_0(t)^{e^{\boldsymbol{\beta}^\top \mathbf{x}}}, \text{ where}$$

$\hat{\boldsymbol{\beta}}^\top \mathbf{x} =$

-0.628

$+0.0671[\text{TrdsOplst6M\_bn1}] + 0.103[\text{TrdsOplst6M}_\text{b}\text{n2}] + 0.292[\text{TrdsOplst6M\_bn3}]$

$-0.184[\text{EmpDur\_bn1}] - 0.368[\text{EmpDur\_bn3}]$

$+0.0767[\text{CurDlnq\_bn1}] + 0.255[\text{CurDlnq\_bn2}]$

$+0.228[\text{EmpStat1}]$

$+0.206[\text{BwrState2}] + 0.295[\text{BwrState3}] + 0.401[\text{BwrState4}]$

$+0.312[\text{OcpCat1}] + 0.45[\text{OcpCat2}] + 0.64[\text{OcpCat3}] + 0.849[\text{OcpCat4}]$

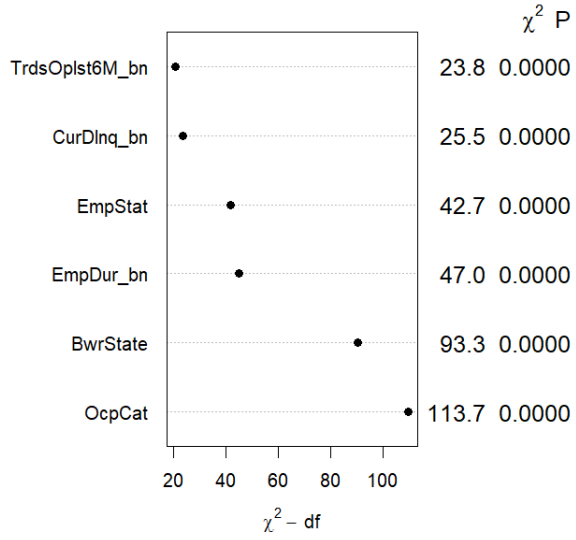


Figure B.2: Stratified Cox PH: Wald test

B.3 Model 6: Log-logistic

$$\text{Prob}\{T \geq t\} = [1 + \exp(\frac{\log(t) - \boldsymbol{\beta}^\top \mathbf{x}}{0.5209381})]^{-1} \text{ where}$$

$\hat{\boldsymbol{\beta}}^\top \mathbf{x} =$

5.63

$-0.107[\text{AmntDlnq\_bn1}]$

$-0.164[\text{InqLst6M\_bn1}] - 0.221[\text{InqLst6M\_bn2}] - 0.329[\text{InqLst6M\_bn3}]$

$-0.45[\text{InqLst6M\_bn4}] - 0.123[\text{BwrState2}] - 0.186[\text{BwrState3}] - 0.252[\text{BwrState4}]$

$-0.242[\text{InvVerf1}]$

$-0.3[\text{PrpRatCat1}] - 0.578[\text{PrpRatCat2}] - 1.11[\text{PrpRatCat3}] - 1.2[\text{PrpRatCat4}]$

$-0.183[\text{OcpCat1}] - 0.254[\text{OcpCat2}] - 0.378[\text{OcpCat3}] - 0.505[\text{OcpCat4}]$

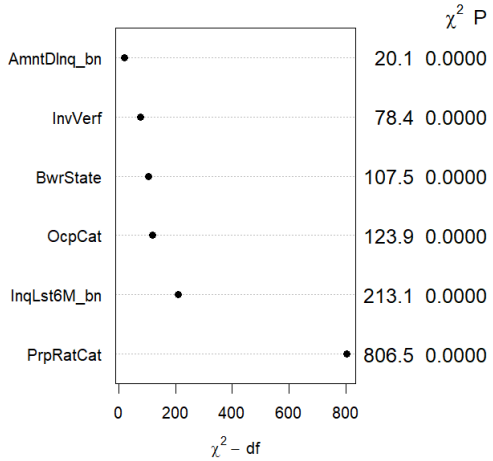


Figure B.3: Log-logistic: Wald test.

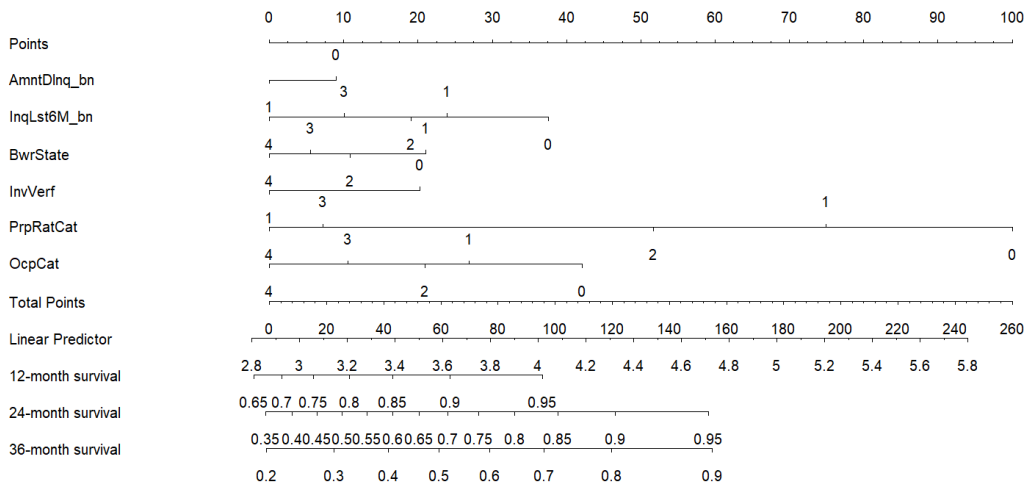


Figure B.4: Log-logistic: Nomogram.

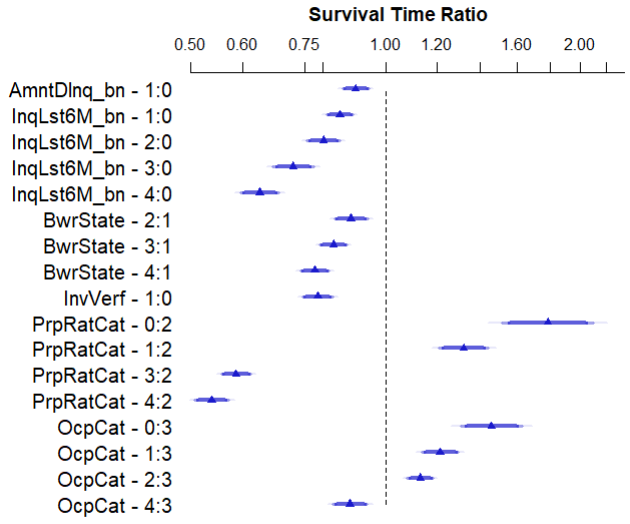


Figure B.5: Log-logistic: Forest plot.

B.4 Model 7: Log-normal

$$\text{Prob}\{T \geq t\} = 1 - \Phi(\frac{\log(t) - X\boldsymbol{\beta}}{1.00245}) \text{ where}$$

$\hat{\boldsymbol{\beta}}^{\top}\mathbf{x} =$

5.62

$-0.121[\text{AmntDlnq\_bn1}]$

$-0.158[\text{InqLst6M\_bn1}] - 0.223[\text{InqLst6M\_bn2}] - 0.321[\text{InqLst6M\_bn3}]$

$-0.457[\text{InqLst6M\_bn4}] - 0.122[\text{BwrState2}] - 0.183[\text{BwrState3}] - 0.246[\text{BwrState4}]$

$-0.253[\text{InvVerf1}]$

$-0.237[\text{PrpRatCat1}] - 0.492[\text{PrpRatCat2}] - 1[\text{PrpRatCat3}] - 1.12[\text{PrpRatCat4}]$

$-0.184[\text{OcpCat1}] - 0.244[\text{OcpCat2}] - 0.362[\text{OcpCat3}] - 0.478[\text{OcpCat4}]$

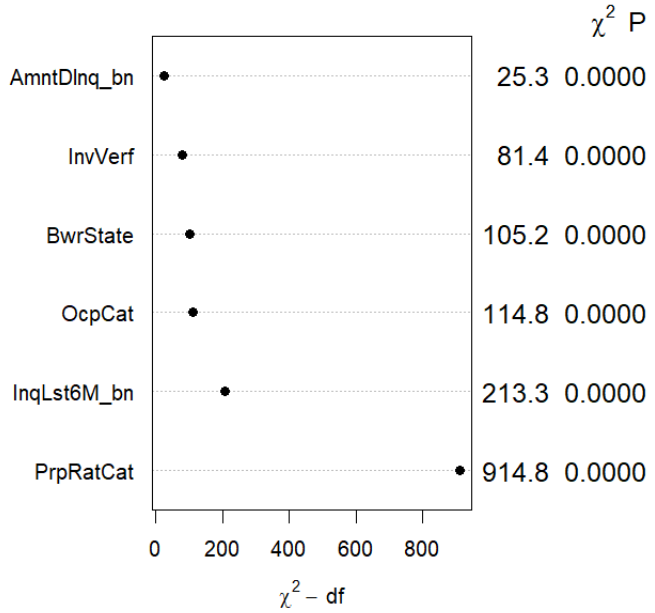


Figure B.6: Log-normal: Wald test.

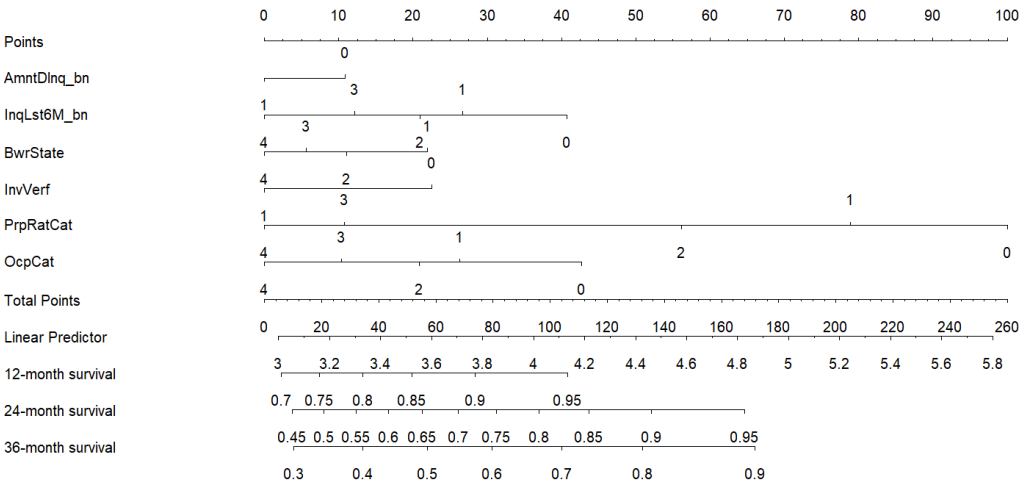


Figure B.7: Log-normal: Nomogram.

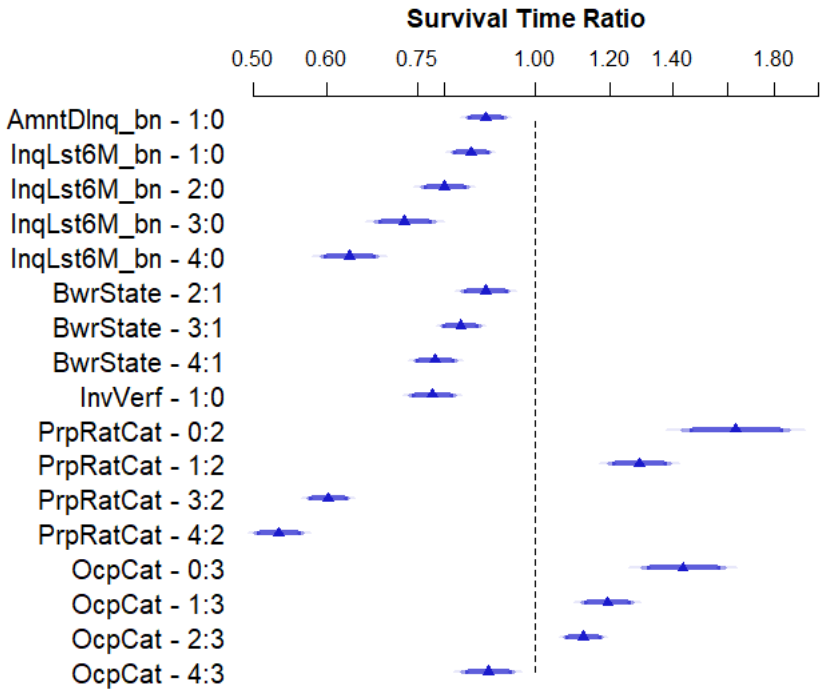


Figure B.8: Log-normal: Forest plot.

B.5 Extended Cox regression with a step function

Table B.1: Extended Cox: Wald Statistics.

Variable Name	ChiSquare	d.f	P-value
TrdsOplst6M_bn	71.89	3	<.001
EmpDur_bn	24.50	2	<.001
CurDlnq_bn	112.61	2	<.001
EmpStat	19.49	1	<.001
BwrState	109.06	3	<.001
OcpCat	67.13	4	<.001
InqLst6M_bn (Factor+Higher Order Factors)	483.25	12	<.001
All Interactions	47.57	8	<.001
ListCat (Factor+Higher Order Factors)	148.02	6	<.001
All Interactions	46.55	4	<.001
IncRange (Factor+Higher Order Factors)	318.23	12	<.001
All Interactions	47.41	8	<.001
InqLst6M_bn * timegroup (Factor+Higher Order Factors)	47.57	8	<.001
timegroup * ListCat (Factor+Higher Order Factors)	46.55	4	<.001
timegroup * IncRange (Factor+Higher Order Factors)	47.41	8	<.001
TOTAL INTERACTION	147.50	20	<.001
TOTAL	1800.86	45	<.001

Table B.2: Extended Cox regression: PH Test using Schoenfeld Residuals.

Variable Name	P-value
TrdsOplst6M_bn	0.593
EmpDur_bn	0.236
CurDlnq_bn	0.082
EmpStat	0.126
BwrState	0.521
OcpCat	0.425
InqLst6M_bn:strata(timegroup)	0.058
ListCat:strata(timegroup)	0.163
IncRange:strata(timegroup)	0.665
GLOBAL	0.125

Appendix C

Code

C.1 Python and R Code

The following functions are written as part of the research project.

C.1.1 Data Extraction and Preparation in Python

```
class Data_Prep(object):
    def __init__(self):
        self.get_data()

    def get_data(self):
        # Extract data, check for completely empty column and drop them
        #url = "https://s3.amazonaws.com/udacity-hosted-downloads/ud651/
            prosperLoanData.csv"
        raw = pd.read_csv('../data/prosperLoanData.csv')
        nan_value = float('NaN')
        raw.replace("", nan_value, inplace=True)
        raw.dropna(how='all', axis=1, inplace=True)
        lst = ['Completed', 'Current', 'Chargedoff', 'Defaulted', 'Cancelled']
        raw = raw.query('LoanStatus in @lst')
        '''
        Remove duplicate LoanKey id
        '''
```

```
        raw.drop_duplicates(subset ="LoanKey", keep = False, inplace = True)
        raw.index = raw['LoanKey']
        raw.drop('LoanKey', axis='columns', inplace=True)
        '''
        Formatting dates date_format
        '''
        raw['LoanOriginationDate'] = format_date(raw,'LoanOriginationDate')
        raw['ClosedDate'] = format_date(raw,'ClosedDate')
        raw = raw.loc[(raw['LoanOriginationDate']>='2009-11-02') & (raw['Term'
            ]==36)]
        self.data = raw
        return self.data.info()

    def default_flag(self):
        end = '2014-11-03'
        default_status = ['Chargedoff', 'Defaulted']
        self.data['default_flag'] = self.data['LoanStatus'].apply(lambda x: 1 if
             x in default_status else 0)
        self.data['ClosedDate'] = self.data['ClosedDate'].fillna(pd.Timestamp(
            end)) # adding final date to current status

    def survival_time(self):
        data = self.data.copy()
        '''
        Preliminary date formats and calculate obsersation time in months
        '''
        data['obstime'] = (data['ClosedDate'].dt.year - data['
            LoanOriginationDate'].dt.year) * 12 + (data['ClosedDate'].dt.month -
                            data['LoanOriginationDate'].dt.month)
        self.data = data.loc[(data['obstime'] > 0) & (data['obstime'] <=36)]

    def create_variables(self):
        data = self.data.copy()
        '''
        Create custom variables
        '''
        data['AvgCreditScore'] = (data['CreditScoreRangeUpper'] + data['
            CreditScoreRangeLower']) / 2

        data['MonthlyDebtRate'] = np.where((data['StatedMonthlyIncome']!=0) | (
            data['StatedMonthlyIncome'].isna()),
```

```
                                                    np.minimum(data['
        MonthlyLoanPayment'] / data['StatedMonthlyIncome'], 1.0), nan_value)

    cols = ['CreditScoreRangeLower', 'CreditScoreRangeUpper', '
        MonthlyLoanPayment', 'StatedMonthlyIncome']
    data.drop(cols, axis='columns', inplace=True)
    self.data = data

def calc_perc_missing(self):
    # Calculate percentage missing and display
    data = self.data.copy()
    perc_missing = data.isnull().sum() * 100 /len(data)
    self.missing_df = pd.DataFrame({'column_name': data.columns,  'percent_
        missing': np.round(perc_missing,4)})
    return print(self.missing_df.to_string(index=False))

def id_cols_empty_perc(self, empty_perc):
    # Specify the missing of empty
    self.empty_perc = empty_perc * 100
    idx_empty = np.where(self.missing_df['percent_missing'] >= self.empty_
        perc)
    self.empty_perc_df = self.missing_df.iloc[idx_empty]
    return print (self.empty_perc_df.to_string(index=False))

def del_empty_perc_cols(self):
    '''
    Delete nearly empty cols and print number of continuous & categorical
        variables
    '''
    cols = self.empty_perc_df.iloc[:,0].to_list()
    self.data.drop(cols, axis='columns', inplace=True)
    '''
    Calculate the number of continuous and categorial variables excluding id
    '''
    return print('# of continous variables:{}\n# of categorical variables:{}
        '                  .format(len(self.data.select_dtypes('float').count
        ()),
                           len(self.data.select_dtypes('object').count())))


def loan_status_counts_before_excl(self):
```

```
        data = self.data.copy()
        return print(data['LoanStatus'].value_counts())

    def loan_status_counts_after_excl(self):
        data = self.data.copy()
        return print(data['LoanStatus'].value_counts())

    def drop_post_loan_issuance_vars(self, drop_list):
        '''
        Drop variables whose values are not known at loan issuance
        '''
        self.data.drop(drop_list, axis='columns', inplace=True)
        return self.data

class Constant_QuasiConstant(BaseEstimator, TransformerMixin):
    def __init__(self, threshold):
        self.threshold = threshold
        self.selector = None
        self.init_selector()

    def init_selector(self):
        self.selector = VarianceThreshold(threshold=self.threshold)

    def fit(self, X, y=None):
        return self

    def transform (self, X):
        df = X.select_dtypes('float').copy()
        '''
        Removing Constant Variables Variance Threshold
        '''
        constant_filter = VarianceThreshold(threshold=0.0)
        constant_filter.fit(df)
        len(df.columns[constant_filter.get_support()])
        constant_columns = [column for column in df.columns
                            if column not in df.columns[constant_filter.get_
                                support()]]
        '''
        Dropping constant columns
        '''
        df.drop(labels=constant_columns, axis=1, inplace=True)
```

```
        '''
        Removing Quasi-Constant Variables Using Variance Threshold
        '''
        self.selector.fit(df)
        qconstant_columns = [column for column in df.columns
                             if column not in df.columns[self.selector.get_
                                 support()]]
        print ("None Quasi-constant variables: %3.1f; Quasi-constant features :
            %3.1f " %
               (len(df.columns[self.selector.get_support()]),len(qconstant_
                   columns)))
        for column in qconstant_columns:
            print("Quasi-constant variable at %3.1f percent threshold is %14s:"
                % (self.threshold *100, column))
        '''
        Now select only non-constant and non-quasi-constant variables
        '''
        df.drop(labels=qconstant_columns, axis='columns', inplace=True)
        return df

num_dict = {'ProsperScore':'norm',
            'EmploymentStatusDuration':'pmm',
            'DebtToIncomeRatio':'pmm',
            'MonthlyDebtRate':'pmm'}
dict_num_strategy = num_dict
variance_threshold  = Constant_QuasiConstant(threshold=0.99)
imp_numerical = MiceImputer(k=3,
                            n=10,
                            strategy=dict_num_strategy,
                            predictors='all',
                            imp_kwgs={'fill_value': 'random'},
                            seed=42,
                            visit='default',
                            return_list=True)
num_pipeline = Pipeline(steps=[('num_processor', numerical_processing()),
                               ('num_imputer', imp_numerical),
                               ('mice_average', mice_average()),
                               ('variance_threshold', variance_threshold),
                              ])
try:
    loans_imp_num = num_pipeline.fit_transform(loans)
```

```
except TypeError as te:
    print(f"{te.__class__.__name__}: {te}")
cat_dict = {'Occupation':'categorical',
            'ProsperRatingCategory':'categorical'}
dict_cat_strategy = cat_dict
imp_categorical = SingleImputer(strategy=dict_cat_strategy,
                                predictors='all',
                                imp_kwgs={'fill_value': 'random'},
                                seed=42)
cat_pipeline = Pipeline(steps=[('cat_processor',categorical_processing()), ('cat
    _imputer', imp_categorical),])
try:
    loans_imp_cat = cat_pipeline.fit_transform(loans)
except TypeError as te:
    print(f"{te.__class__.__name__}: {te}")
class VIF(BaseEstimator, TransformerMixin):
    def fit(self, X, y=None):
        self.X = X
        return self
    def transform(self, X):
        X = self.X.copy()
        df = pd.DataFrame()
        df['VIF Factor'] = [variance_inflation_factor(X.values, i) for i in
            range(X.shape[1])]
        df['features'] = X.columns
        idx = np.where((df['VIF Factor'] >= 10) | (df['VIF Factor'].isna()))
        df_vars_flag = df.iloc[idx]
        print(df)
        drop_vif_cols = []
        for i in list(range(df_vars_flag.shape[0])):
            drop_vif_cols.append(df_vars_flag.iloc[i-1,1])
        print(drop_vif_cols)
        X.drop(drop_vif_cols, axis='columns', inplace=True)
        return X

class CORR(BaseEstimator, TransformerMixin):
    def fit(self, X, y=None):
        return self
    def transform(self, X):
        data = X.copy()
        correlated_features = []
```

```
        correlation_matrix = data.corr()
        for i in range(len(correlation_matrix.columns)):
            for j in range(i):
                if abs(correlation_matrix.iloc[i, j]) > 0.5:
                    colname = correlation_matrix.columns[i]
                    correlated_features.append(colname)
        print(correlated_features)
        X.drop(correlated_features, axis='columns', inplace=True)
        return X

class Binning(BaseEstimator, TransformerMixin):
    def __init__(self, dtype, trend = True):
        self.trend = trend
        self.dtype = dtype
    def fit(self, X, y=True):
        return self
    def transform (self, X, y=True):
        X = X.copy()
        variables = X.columns.to_list()
        if self.trend:
            var_bin_list = self.trend
        else:
            var_bin_list = list(zip(variables,['auto'] * len(variables)))
        for i in var_bin_list:
            x = X[i[0]].values
            optb = OptimalBinning(name=i[0], dtype=self.dtype, solver='cp',
                monotonic_trend=i[1], max_n_bins=5)
            optb.fit(x, y)
            binning_table = optb.binning_table
            binning_table.build()
            rcParams['figure.figsize'] = 5, 2
            binning_table.plot(metric="event_rate")
            X[i[0] + '_bin'] = optb.transform(x, metric="indices")
            print(binning_table.build())
            print(binning_table.analysis(pvalue_test="chi2"))
        X = X.applymap(str) # convert bin id to strings
        X.drop(variables, axis='columns', inplace=True)
```

stats_code_doc_Prosper.py

C.1.2 Imputation Analysis in R

```
# Multivariate Normal: PMM
    ------------------------------------------------------------------------
# Specifying method for imputing
meth[c("y1", "y2", "y3", "y4")]  = "pmm"
set.seed(42)
mice.imputed <- mice(result$amp, method = meth, predictorMatrix = PredM, m=5)
#Plot density
windows()
densityplot(mice.imputed, main = "Predictive Mean Matching Imputation")
complete.pmm <- complete(mice.imputed)
# compute Bias and Perc Bias
for (i in 1:length(colnames(testdata))){
  actual = testdata[,i]
  predicted = complete.pmm[,i]
  print(paste("Bias", colnames(testdata)[i], bias(actual, predicted), sep = " :
      "))
  print(paste("% Bias", colnames(testdata)[i], percent_bias(actual, predicted),
      sep = " : "))
}
# Multivariate Normal: Norm
    ------------------------------------------------------------------------
meth[c("y1", "y2", "y3", "y4")]  = "norm"
set.seed(42)
mice.imputed <- mice(result$amp, method = meth, predictorMatrix = PredM, m=5)
#Plot density
windows()
densityplot(mice.imputed, main = "Bayesian Lin. Regression Imputation")
complete.norm <- complete(mice.imputed)
# compute Bias and Perc Bias
for (i in 1:length(colnames(testdata))){
  actual = testdata[,i]
  predicted = complete.norm[,i]
  print(paste("Bias", colnames(testdata)[i], bias(actual, predicted), sep = " :
      "))
  print(paste("% Bias", colnames(testdata)[i], percent_bias(actual, predicted),
      sep = " : "))
}
# Multivariate Normal: Mean
    ------------------------------------------------------------------------
meth[c("y1", "y2", "y3", "y4")]  = "mean"
```

```
set.seed(42)
mice.imputed <- mice(result$amp, method = meth, predictorMatrix = PredM, m=5)
#Plot density
windows()
densityplot(mice.imputed, main = "Mean Imputation")
complete.mean <- complete(mice.imputed)
# compute Bias and Perc Bias
for (i in 1:length(colnames(testdata))){
  actual = testdata[,i]
  predicted = complete.mean[,i]
  print(paste("Bias", colnames(testdata)[i], bias(actual, predicted), sep = " :
      "))
  print(paste("% Bias", colnames(testdata)[i], percent_bias(actual, predicted),
      sep = " : "))
}
# Multivariate Normal: RF
    ------------------------------------------------------------------------
meth[c("y1", "y2", "y3", "y4")]  = "rf"
set.seed(42)
mice.imputed <- mice(result$amp, method = meth, predictorMatrix = PredM, m=5)
#Plot density
windows()
densityplot(mice.imputed, main = "Random Forest Imputation")
complete.rf <- complete(mice.imputed)
# compute Bias and Perc Bias
for (i in 1:length(colnames(testdata))){
  actual = testdata[,i]
  predicted = complete.rf[,i]
  print(paste("Bias", colnames(testdata)[i], bias(actual, predicted), sep = " :
      "))
  print(paste("% Bias", colnames(testdata)[i], percent_bias(actual, predicted),
      sep = " : "))
}
```

Inputation_sim_doc.R

C.1.3 Model Building and Further Analysis in R

```
# Further explanatory data analysis part 1: log rank test -----------------
# create a function that outputs log rank test and plot survival survival curves
    for each variable's category
LogRnkFun <- function(data, features.select, i){
  cols <- colnames(data[, features.select])
  f.f <- as.formula(paste("Surv(time, status)",
                          paste(cols[i], collapse = "+"),
                          sep = "~"))
  pdiff <- survminer::pairwise_survdiff(f.f, data = data)
  print(pdiff)
  survminer::ggsurvplot(surv_fit(f.f, data = data), data = data)
}
# User Defined Fuctions ----------------------------------------------------
nom_plot <- function(s){
  sv = Survival(s)
  surv12 = function(x) sv(12, lp=x)
  surv24 = function(x) sv(24, lp=x)
  surv36 = function(x) sv(36, lp=x)
  plot(nomogram(s, fun=list(surv12, surv24, surv36),
                funlabel=c('12-month survival', '24-month survival', '36-month
                    survival')), cex.var=.75, cex.axis =.75, lmgp=.25)
}
AUC_overtime <- function(s){
  ggplot(data = s$AUC$score, aes(x=times,y=AUC))+
    geom_point()+
    geom_line()
}
smoothSEcurve <- function(yy, xx) {
  # use after a call to "plot"
  # fit a lowess curve and 95% confidence interval curve
  xx.list <- min(xx) + ((0:100)/100)*(max(xx) - min(xx)) # make list of x values

  # Then fit loess function through the points (xx, yy) at the listed values
  yy.xx <- predict(loess(yy ~ xx), se=T, newdata=data.frame(xx=xx.list))

  lines(yy.xx$fit ~ xx.list, lwd=2)
  lines(yy.xx$fit - qt(0.975, yy.xx$df)*yy.xx$se.fit ~ xx.list, lty=2)
  lines(yy.xx$fit + qt(0.975, yy.xx$df)*yy.xx$se.fit ~ xx.list, lty=2)
}
# Kaplan-Meier Estimator ---------------------------------------------------
par(mfrow=c(1,1))
```

```
limy <- c(0.6,1)
survplot(npsurv(S ~ 1), conf="none", n.risk=TRUE, ylim = limy,   xlab=expression
    (t))
survplot(npsurv(S ~ 1, type="fleming-harrington", conf.int=FALSE), add=TRUE, lty
    =3) # Altschuler-Nelson-Aalen-Fleming-Harrington estimator
title(sub="Kaplan-Meier Estimator", adj=0, cex=.7)

## median follow-up time using "reverse" Kaplan-Meier
median(loans$time)
status.followup <- 1 - status
## Calculate hazard function  - obtained by dividing time into equal intervals
    of width 6 months
pehaz.smooth <- muhaz(loans$time, loans$status, bw.method = "local", b.cor = "
    left", max.time = maxtime)
## plot smooth hazard function
haz <- pehaz.smooth$haz.est
times <- pehaz.smooth$est.grid
surv.smooth <- exp(-cumsum(haz[1:(length(haz)-1)]*diff(times)))
survplot(npsurv(S.f ~ 1), conf="bands", n.risk=TRUE, ylim = limy,  xlab=
    expression(t))
survplot(npsurv(S.f ~ 1, type="fleming-harrington", conf.int=FALSE), add=TRUE,
    lty=3) # Altschuler-Nelson-Aalen-Fleming-Harrington estimator
lines(surv.smooth ~ times[1:(length(times)-1)], lty=1, lw=0.75, col="blue")
title(sub="Kaplan-Meier Estimator versus Smoothed Survival Estimate", adj=0, cex
    =.7)

# Cox Proportiona Hazard Model (All binned variables)
    ----------------------------------------------
f.0 <- as.formula(
  paste(outcome,
        paste(predictors, collapse = " + "),
        sep = " ~ "))
print(f.0)
# # Observe how predictors varies over time
windows()
aa_fit <-with(loans.train, aareg(f.0))
autoplot(aa_fit)
# Fit Cox PH model
f.cph.0 <- with(loans.train, cph(f.0, x = TRUE, y = TRUE, method = "breslow"))
step.vars <- as.formula(
  paste("upper",
```

```
        paste(predictors, collapse = " + "),
        sep = " =~ "))
print(step.vars)
## Stepwise model selection
tic("Stepwise model selection")
f.cph.step.0 <- stats::step(f.cph.0, scope = list(upper=step.vars, lower=~
    PrpRatCat), direction = "both")
toc()
## Calculate model statistics
print(f.cph.step.0, latex = T, coefs = F) # Model fit diagnostics
### Testing PH assumption: Plot KM log hazard against time (visual check)
windows()
co <- gray(seq(0,0.8, 0.2))
par(mfrow=c(4,4), mar=c(4,4,4,2) + 0.1)
for (i in 1:n.vars.step) {
  f.i <- as.formula(
    paste(outcome,
          paste(predictors.step[i], collapse = " + "),
          sep = " ~ "))
  f.km.i <- with(loans.train, npsurv(f.i))
  plot(f.km.i, fun = "cloglog", xlab = expression(t), ylab = "log-log survival",
       col = co, lty =c(1:maxlevel), main=predictors.step[i], cex=.7)
  }
### Testing PH assumption: Correlation with time test (Schoenfeld residuals)
phtest <- cox.zph(f.cph.step.0, terms = T) # H0: rho = 0, rho is correlation of
    regression coefficients with time
phtest
windows()
par(mfrow=c(4,4), mar=c(3,4,4,2) + 0.1)
for (i in 1:n.vars.step) {
  plot(phtest, var = predictors.step[i], ylim = c(-5,5))
  abline(0,0, col=2, cex=.7)
}

# Application to IFRS 9: Waterfall and Violin Plots
    ------------------------------------------------
##  Calculate Cumulative Incidence
n <- nrow(loans)
cif.12 <- nricens::get.risk.coxph(f.cph.step.0.str.a, t0 = 12)
cpd1 <- cif.12[c(1:n)]
cif.24 <- nricens::get.risk.coxph(f.cph.step.0.str.a, t0 = 24)
```

```
cpd2 <- cif.24[c(1:n)]
cif.36 <- nricens::get.risk.coxph(f.cph.step.0.str.a, t0 = 36)
cpd3 <- cif.36[c(1:n)]

pd.df <- data.frame(matrix(ncol = 7, nrow = n))
pd.df <- pd.df %>%
  rename( LoanKey = X1, cpd1 = X2,  cpd2 = X3, cpd3 = X4, mpd1 = X5, mpd2 = X6,
      mpd3 = X7 )
pd.df$cpd1 <- cpd1
pd.df$cpd2 <- cpd2
pd.df$cpd3 <- cpd3

# Calculate PIT PDs
pd.df$mpd1 <- cpd1
pd.df$mpd2 <- (pd.df$cpd2 - pd.df$cpd1)/ (1 - pd.df$cpd1)
pd.df$mpd3 <- (pd.df$cpd3 - pd.df$cpd2)/ (1 - pd.df$cpd2)
# Import file  containing loan amount (proxy for EAD), interest charged (to
    compute discounted ECL) and assume LGD flat at 50\%
loans.ECL.PIT <- full_join(x = loans.ECL, y = pd.df[, c(1, 5:7)], by = "LoanKey"
    ) %>%
  arrange(LoanKey)
head(loans.ECL.PIT, 5)
# assume flat LGD of 15%
lgd <- 0.5
loans.ECL.PIT$LGD <- rep(lgd, nrow(loans.ECL.PIT))
loans.ECL.PIT$EAD <- loans.ECL.PIT$LoanOriginalAmount * (1 + loans.ECL.PIT$
    BorrowerRate / 12)^(0.25) # using pio method
# Calculate discount factors
loans.ECL.PIT$dfct1 <-   1 / (1 + loans.ECL.PIT$BorrowerAPR)
loans.ECL.PIT$dfct2 <-   1 / (1 + loans.ECL.PIT$BorrowerAPR)^2
loans.ECL.PIT$dfct3 <-   1 / (1 + loans.ECL.PIT$BorrowerAPR)^3
# compute ECL
loans.ECL.PIT$ECL1 <-  loans.ECL.PIT$dfct1 * loans.ECL.PIT$mpd1 * loans.ECL.PIT$
    EAD * loans.ECL.PIT$LGD
loans.ECL.PIT$ECL2 <-  loans.ECL.PIT$dfct2 * loans.ECL.PIT$mpd2 * loans.ECL.PIT$
    EAD * loans.ECL.PIT$LGD
loans.ECL.PIT$ECL3 <-  loans.ECL.PIT$dfct3 * loans.ECL.PIT$mpd3 * loans.ECL.PIT$
    EAD * loans.ECL.PIT$LGD
# Convert to millions
tot_ECL1 <- sum(loans.ECL.PIT[, 'ECL1']) / 1000000
tot_ECL2 <- sum(loans.ECL.PIT[, 'ECL2']) / 1000000
```

```
tot_ECL3 <- sum(loans.ECL.PIT[, 'ECL3']) / 1000000
print(c(tot_ECL1, tot_ECL2, tot_ECL3))
ECL_change_yr_2 <- tot_ECL2 - tot_ECL1
ECL_change_yr_3 <- tot_ECL3 - tot_ECL2
print(c(ECL_change_yr_2, ECL_change_yr_3))
# Draw waterfall for ECL
category <- c("Initial ECL", "Year 2 ECL Charge", "Year 3 ECL Charge")
amount <- c(8.92, 5.00, -7.03)
ecl.chart <- data.frame(category, amount)
waterfall(ecl.chart,
          calc_total=TRUE,
          total_axis_text = "Net",
          total_rect_text_color="black",
          total_rect_color="goldenrod1") +
  # scale_y_continuous(label=scales::dollar) +
  labs(title = "IFR 9: Expected Credit Loss",
       subtitle = "Prosper Loan Portfolio: 2009 - 2014",
       y="USD Millions",
       x="") +
  theme_minimal()
# Violin plots for PIT PDs and Cum. PDs
windows()
vioplot(pd.df$mpd1, pd.df$mpd2, pd.df$mpd3,
        names=c("PIT PD Year 1", "PIT PD Year 2", "PIT PD Year 3"),
        xlab = "PD Year",
        ylab = "PIT PD",
        col="gold")
title("IFRS 9: Point-in-time PD")
windows()
vioplot(cpd1, cpd2, cpd3,
        names=c("Cumulative PD Year 1", "Cumulative PD Year 2", "Cumulative PD
            Year 3"),
        xlab = "PD Year",
        ylab = "Cumulative PD",
        col="gray")
title("IFRS 9: Cumulative PD")
# Comparing  Models --------------------------
## Plot ROC over 12,24 months follow-up period : For Semi-parametric
tic("ROC curve")
windows()
t.vector.auc <- c(12,24)
```

```
for (t in 1:length(t.vector.auc)){
  score.roc <- Score(list("Model 1: Cox PH" =f.cph.step.0.ph ,
                          "Model 2: Stratified Cox PH"=f.cph.step.0.str,
                          "Model 3: Cox PH with Splines"=f.cph.step.1.ph),
                     formula = Hist(time, status) ~ 1,
                     data = loans.test,
                     times = t.vector.auc[t],
                     plots = "roc", metrics = c("auc","brier"))
windows()
plotROC(score.roc, brier.in.legend = T , cex = .75)

}
toc()

a <- seq(3,36,3)
length(a)
auc.semi <- data.frame(matrix(ncol = 3, nrow = length(a)-2))
auc.semi <- auc.semi %>%
  rename( mod1 = X1, mod2 = X2,  mod3 = X3)

windows()
boxplot(auc.semi$mod1, auc.semi$mod2, auc.semi$mod3,
        main = "Semi-parametric model: AUC comparision",
        at = c(1,2,3),
        names = c("Cox PH", "Stratified Cox PH", "Cox PH w. Splines"),
        las = 1,
        xlab = "Models",
        ylab = "AUC %",
        boxwex = 0.45,
        border = "blue",
        horizontal = F,
        notch = F
)
bs.semi <- data.frame(matrix(ncol = 3, nrow = length(a)))
bs.semi <- bs.semi %>%
  rename(mod1 = X1, mod2 = X2,  mod3 = X3)

windows()
plot(bs.semi$mod1, type="b", pch = 18, xlab="Time (quarterly)", ylab="Brier
    Score", main="Semi-parametric: Brier Score Comparison",
     col="black", lwd = 2, axes=F, cex.lab = 1, cex.axis = 0.5)
```

```
axis(1)
axis(2)
lines(bs.semi$mod2, type="b", pch = 18, col="orange")
lines(bs.semi$mod3, type="b", pch = 18, col="blue")
legend(8,2, legend = c("Cox PH", "Stratified Cox PH", "Cox PH w. Splines"), col=
    c("black", "orange", "blue"), lty = 1, cex=0.7, box.lty=0)
## Plot ROC over 12,24 months follow-up period : For Parametric
tic("ROC curve")
windows()
t.vector.auc <- c(12,24)
for (t in 1:length(t.vector.auc)){
  score.roc <- Score(list("Model 4: Weibull"=f.weib,
                          "Model 5: Exponential"=f.exp,
                          "Model 6: Log-logistic"=f.llogis,
                          "Model 7: Log-normal"=f.lnorm),
                     formula = Hist(time, status) ~ 1,
                     data = loans.test,
                     times = t.vector.auc[t],
                     plots = "roc", metrics = c("auc","brier"))

  windows()
  plotROC(score.roc, brier.in.legend = T , cex = .75)

}
toc()
windows()
boxplot(auc.para$mod1, auc.para$mod2, auc.para$mod3, auc.para$mod4,
        main = "Parametric model: AUC comparision",
        at = c(1,2,3,4),
        names = c("Exponential", "Weibull", "Log-logistic", "Log-normal"),
        las = 1,
        xlab = "Models",
        ylab = "AUC %",
        boxwex = 0.45,
        # col = c("gray", "orange","blue"),
        border = "blue",
        horizontal = F,
        notch = F
)

windows()
```

```
plot(bs.para$mod1, type="b", pch = 18, xlab="Time (quarterly)", ylab="Brier
    Score",
     main="Parametric: Brier Score Comparison",
     col="black", lwd = 2, axes=F, cex.lab = 1, cex.axis = 0.5)
axis(1)
axis(2)
lines(bs.para$mod2, type="b", pch = 18, col="orange")
lines(bs.para$mod3, type="b", pch = 18, col="blue")
lines(bs.para$mod4, type="b", pch = 18, col="green")
legend(8,2, legend = c("Exponential", "Weibull", "Log-logistic", "Log-normal"),
       col=c("black", "orange", "blue", "green"),
       lty = 1, cex=0.7, box.lty=0)
```

Msc_Dis_Code_doc_v2.R