Statistical Modelling for Home Loans and Regulatory Credit Risk Capital Forecast

Department of Mathematical Statistics and Actuarial Sciences

Faculty of Natural and Agricultural Sciences

Bloemfontein

Submitted in fulfilment of the degree of Master of Actuarial Sciences

BY

Paulosi Lucky Mazibuko

2006054882

BC480010

January 2019

Supervisor: Prof Maksim Finkelstein

DECLARATION

I, Paulosi Lucky Mazibuko, declare that the master's dissertation that I herewith submit for the master's qualification in Actuarial Sciences at the University of the Free State, is my own independent work, and that I have not previously submitted it for a qualification at another institution of higher education. I have used information from other sources, and I have given credit by proper and complete references of the source material, so that my research can be distinguished from what was quoted from other sources. I acknowledge that failure to comply with the instructions regarding referencing will be regarded as plagiarism.

I furthermore cede copyright of the thesis in favour of the University of the Free State.

23/01/2019

Paulosi Lucky Mazibuko
(Author)

DEDICATION

My paper is dedicated to my wife, Ntaoleng Motaung, and my daughter, Wathandwa Mazibuko, for all the support through the difficult journey to finish my dissertation successfully. All thanks to my supportive supervisor, Professor Maksim Finkelstein, who believed in me, that I could take this challenge to build my model and complete my master's degree.

ACKNOWLEDGEMENTS

My research was completed successfully, due chiefly to the ongoing support from my wife, mother and colleagues. I would like to thank my wife, Ntaoleng Motaung, for all the support, patience and love throughout the year, while completing this research study. Also, a special thanks to my supervisor, Prof. Maksim Finkelstein, who provided me with support and guidance, and believed in me. Moreover, I would like to thank the School of Mathematical Statistics and Actuarial Sciences for giving me this great opportunity to prove my capabilities and strength in the Actuarial Sciences and Mathematical Statistics field. Lastly, I hereby extend my gratitude to the financial institution – one of the largest in South Africa, for the data I used, the time and platform to work on, and resources and software provided to be used for this research study.

ABSTRACT

In commercial credit institutions, valuation of default is useful for moneylenders such as banks and other companies that make a practice of credit scoring as quantitative research to determine the creditworthiness of an individual or borrower. There are several statistical models that are used in the bank for credit scoring. Logistic and Survival Analysis models are the most-utilised scoring models by lenders, among others. The main intention of this paper is to model and predict the likelihood of non-payment for a mortgage loans in financial institutions. To range these objectives, two statistical approaches, namely Logistic Regression and Survival Analysis, are used to a large dataset of mortgage loans by one of the financial institutions. In this paper, it has been shown that the Survival model is a good method on likelihood of non-payment, in contrast with Logistic Regression. The results of the final modelling for both approaches shows parallel fit in Receiver Operator Characteristic (ROC) with the Logistic Regression model outperforming the Survival model in both training and testing dataset. In prediction of defaulted and non-defaulted results on mortgage loans, Logistic Regression still has better performance than Survival Analysis in both training and testing datasets. In general, the results show that the Survival Analysis method is competitive with the Logistic Regression method traditionally utilised in the financial institutions. Moreover, by methods for a vast, genuine dataset, time reliance was notable that made accessible more precise credit risk scoring and imperative insight into self-motivated market impacts that can educate and upgrade related decision-making.

Keywords Credit Score; Logistic Regression; Survival Analysis; Probability of Default; decision-making; Receiver Operator Characteristic; Market Impacts

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviations	Description
AFT	Accelerated Failure Time
AIC	Akaike Information Criterion
AUC	Area under Curve
BCBS	Basel Committee on Bank System
BIC	Bayesian Information Criterion
CPH	Cox Proportional Hazard
CR	Credit Risk
CS	Credit Scoring
ECOA	Equal Credit Opportunity Act
IRB	Internal Ratings Based
IV	Information Value
K-M	Kaplan Meier
LR	Logistic Regression
L-T	Life Tables or Actuarial Estimator
N-A	Nelso Aalen
OR	Odds Ratio
ROC	Receiver Operating Characteristics
SA	Survival Analysis
WoE	Weight of Evidence

CHAPTER 1: INTRODUCTION TO RISK MODELS

1.1. Introduction

The framework of agreement strategies, namely Basel 2 and Basel 3, and the consequent increased essential for more precise credit risk controls, shows that the investigation of survival has become more necessary as time goes on. Factually, the survival model is mostly utilised within the engineering and life insurance contexts, where the period until an occurrence is analysed – e.g. the period until decease or engine failure (Dirick et al., 2017).

Survival Analysis has been made known by Narain (1992) as different from Logistic Regression on the credit context (Dirick et al., 2017). The benefits of the exploitation Survival model during this setting as point to non-payment, are often modelled, and not simply whether a borrower can default or not. It offers a transparent method of assessing the seeming profitableness of a borrower, and non-payments of loan of Survival Analysis method match and combine things once a case has payment within the observation period (Dirick et al., 2017). The non-parametric approach is utilised to give the likelihood of default in the conditional supply purpose of the period to non-payment (Đurović, 2017).

The Survival Analysis model will embrace shortened and censored data within the progression analysis as associated to the Logistic model. The right, left and internal censoring are three kinds of censoring in Survival Analysis models. The foremost common kind of censoring come upon in SA data is right censored (Survival). The right censored defines as the event that is not discovered in the study. In a credit setting, borrowers do not default; thus, a great deal of data in the study is right-censored (Jaber, 2017).

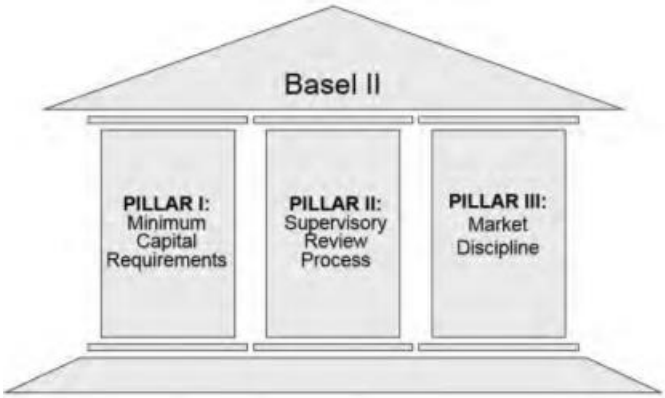
Logistic Regression (LR) is another model which can be used to develop chance of default. It is utilised to describe a possible future event, a binary result such as (1 or 0, Yes or No, True or False), given a set of things that one control, that causes other things to change (such as test scores) – for example, forecasting chances of winning Lotto. To signify binary/categorical result, one uses numbers that change. The Logistic model is categorical, where one uses log of odds as a dependent variable. In other words, it forecasts the likelihood of existence of occurrence by appropriate numbers to a logit function (Memić, 2015).

The next sections give more detailed information about The New Basel Capital Accord, along with 3 Pillars of Basel 2. Sections 1.3 and 1.4 discuss credit scoring and risk profile review, and how it is used in a bank.

1.2. The New Basel Capital Agreement

In 1974, the Basel Committee on Bank management (BCBS) was based by way of a setting for systematic support between its participant nations on banking superior materials. The BCBS describes its original goal as improvement of financial stability by rising superior apprehend however and therefore the quality of banking management worldwide. Thereafter, it had better quality to monitor and make sure the capital competence of banks and the banking industry. Basel Committee in Bank management (BCBS) introduced the Basel Accords that have three banking laws, particularly Basel 1, Basel 2 and Basel 3, that can be further explained in the next sub-sections. The BCBS offers approvals on banking guidelines in relation to operational, wealth and market risk. The main aim of the accords is to make sure that the monetary institutions have satisfactory capital on account to encounter necessities and absorb unforeseen losses.

On 26 June 2004, The BCB management unconfined International Convergence of Capital Measurement and Capital Standards: A revised Framework, which is commonly recognised as the Basel 2 Agreement. In Basel 2, separately from Credit and Market Risk; Operational Risk was carefully considered in Capital Adequacy Ratio Control (Roy et al., 2013). The Basel 2 Agreement focuses on three aspects/Pillars of the Basel Capital Agreement namely on the following figure:



Source:https://www.researchgate.net/figure/Metaphorical-Representation-of-the-Pillars-Supporting-Basel-II_fig1_5144280

1.2.1. Pillar 1: Least Principal Necessities

The design of Least Regulatory Capital is a continuation of the 1988 Basel Agreement. Basel II also studies the following:

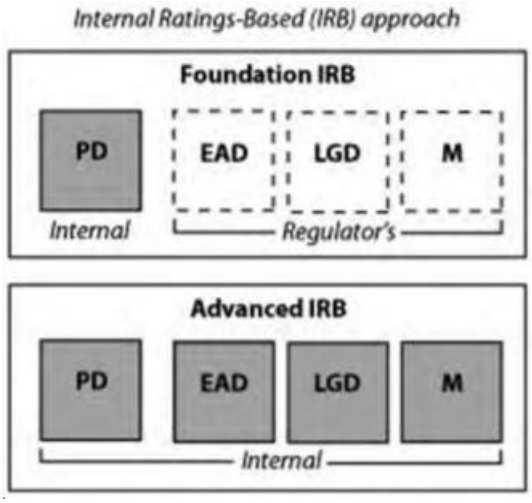
- Hazard management incentives
- Innovative operational risk capital trust
- Risk weighted assets (RWA)
- Market risk mostly unaffected

The following formula shows how to calculate the Capital of Risk Weighted Asset Ratio (CRAR):

$$CRAR = \frac{Tier\ 1\ Capital + Tier\ 2\ Capital}{Risk\ Weighted\ Assets}$$

Pillar 1 of Basel Accord II enables institutions to calculate their own credit risk capital internally in either two ways:

1. The consistent method
2. The Internal Ratings Based (IRB) method (namely Foundation and Advanced method): permits banks to shape and utilise their individual inside risk evaluations, to changing degrees. The IRB method is constructed on the following four main limitations:
    a. Likelihood of non-payment (PD): the likelihood that an advance will not be reimbursed and will in this way fall into non-payment in the following year;
    b. Loss Given Default: the evaluated financial misfortune, communicated as a percentage of exposure, which will be brought about if an obligor goes into default;
    c. Exposure at Default: a proportion of the fiscal introduction, should an obligor go into default;
    d. Maturity: is the period to the document instalment date of a loan or other budgetary instrument.



Source:https://support.sas.com/content/dam/SAS/support/en/books/developing-credit-risk-models-using-sas-enterprise-miner-and-sas-stat/66220_excerpt.pdf

However, in this study, only Probability of Default will be measured for advanced Internal Ratings-Based and Foundation approach.

1.2.2. Pillar 2: Supervisory Review Process

Basel 2 had offered forces to the controllers to administer and check the bank's risk administration framework and capital appraisal arrangement. The controllers can likewise request for cradle capital separated from least capital necessity by BCBS. RBI has requested for 9% CRAR, which is over 8%

recommended by BCBS. Controllers are enabled to supervise the internal risk assessment routines proposed in Pillar 1.

1.2.3. Pillar 3: Market Discipline

This market discipline had made exposure of a bank's risk-taking positions and capital, required. This progression was focused to present market discipline through discovery.

1.3. Credit Scoring

Credit scoring (CS) is the construction that bolsters the creditors either to give an advance to an applicant, or not (Mageto et al., 2015). It denotes the likelihood that an applicant will not pay on a dedication by neglecting to make fundamental instalments (Basel, 2000). Usually, this was done utilising judgemental scoring frameworks, using a credit/loan assessment. Variables proportionate to instalment history, bank and exchange references, age, size and kind of business, country of inception, and spending plan, are scored and weighted to give a general FICO assessment. In any case, be that as it may, improvement of this framework is extremely tedious and costly (Capon, 1982).

There are several statistical models that are used among banks to determine the credit score for a person who requires any form of financial credit. Logistic Regression and Survival Analysis are absolute, most regularly utilised scoring models by loan specialists, among others, that estimate likelihood of default on obligations.

Likelihood of non-payment is a term unfolding the likelihood of a non-payment over a time horizon. It gives an estimation of the likelihood that a debtor cannot make its debt obligations. Probability of default is a key credit risk parameter, and a credit risk model aims to distinguish between good and bad customers (Dirick et al., 2017).

To determine probability of default, the scoring system will be built. Logistic Regression is commonly used to build/develop a probability of default model (Ferreira et al., 2015). The disadvantages of Logistic Regression are that the impact of changes in macroeconomic variables is typically not considered. Significance of changes in time is not considered in Logistic Regression models, but survival analysis has been deployed in credit scoring which will address those issues (Chmielewska, 2016).

A proportion made had been investigated in growing non-payment models to pact with credit risk. In many conditions, the non-payment high-quality of exploration inclines to depend on Survival Analysis model regression based on its suitability.

The purpose of my paper is to use the Cox Proportional Hazard (CPH) model as another method for modelling credit risk, relate it to Logistic Regression, and address some limitations of Logistic Regression.

1.4. Review of risk profiles

It is an assessment of an individual or organisation's readiness to take risks, as well as the dangers to which an association is uncovered. A risk profile is vital for deciding an appropriate investment asset distribution for a portfolio. Associations utilise a risk profile to moderate potential risks and threats. Many factors impact the default rates, such as loan to value, risk credit score, and month on book, etc. Financial Institutions and academic researchers suggest that the mortgage default rate relies on credit scores such as TransUnion, Experian, etc. The usage of statistical models, the scoring companies consider several features, separately, of these five parts to regulate credit risk: expenses history, up-to-date level of obligation to debts, forms of credit utilised, duration of credit past, and new debts.

The key risk drivers within mortgage can best be analysed by examining the relationship among the following variables:

1. Current credit score and score at the time the account was reserved.
2. House price index associated with the property's location.
3. The loan to value (LTV) based on original, or on a derived, adjustment, considering the house price appreciation over the years from the origination LTV.

In addition to the features described above, credit risk can depend on macroeconomic variables and influences. In economic recessions, the default likelihood rises and risk ratings decline. The macroeconomic factors that are considered in this paper include interest rate, inflation, prime rate and house price index, as described above, sourced primarily from a South African financial institution.

1.5. Goals and Objectives

The aim of this paper is to build PD estimation using the Survival model approach, and compare it to Logistic Regression. In this institution, Survival Analysis has not been used to estimate PD; however, it will be modelled to improve PD estimation and reduce risk by forecasting. The objectives of this paper are to do the following:

1. Find factors which affect default rates.
2. Apply Survival models: show good- and bad-risk customers, calculate the probability of surviving to a specified duration, and calculate default rates on bank's mortgage loans.
3. Forecast/project default rates, using a Survival model.

4. Conduct a univariate analysis for every customer covariate, and select factors fit for separating risks.
5. Fit the Cox regression model to build a dataset for the default occasion, estimating default as the censored data.
6. Hazards assumption has been assessed by each model.
7. Fit a Logistic Regression model for each event.
8. Do a comparison between Logistic and Cox regression, based on predicting loans which are likely to default.

1.6. Research Data

This study explores a dataset obtained in the consumer credit context. The analysis looks at facility level information, rather than at customer level. That means that if a customer holds more than one account, this study treats each account separately. The dataset will consist of all active accounts between Jan 2017 and Dec 2017 (1-year data). Application and behavioural variables are provided per account in the dataset. Datasets with variables such as income amount, age and credit bureau risk score, will be reserved at the period of request. For this purpose, an account will be taken as having not been paid if it reaches three months or extra in the opening twelve months. Mortgage loans that are not being paid are declared as bad, and a payment is mentioned as a good account. The repayment status is given per account per month under observation. A fixed workout/outcome period will be determined, and used in the calculation of forward looking probabilities. A workout period is the sum of periods it takes for the bulk of accounts to be absorbed into the events of attention.

1.7. Sources of Data

This paper uses consumer credit data retrieved from one of the leading South African commercial institutions. The institution approved the criteria outlined in the Basel Accord. This indicates that the data complies with international standards, and that the data is trustworthy for study dedication.

1.8. Research Hypothesis

Traditionally, the problem is addressed using statistical models such as lasso, logistic regression, and decision trees models. These techniques are not appropriate to handle censored data. If the data is missing, it is considered as censoring in the survival analysis model. Logistic regression limitations are explained as follows:

Limitations of Logistic Regression:

1. Impact of changes in Macro Economic variables is not considered in Logistic Regression based likelihood of default models (PD).
2. Predicted Probability of Default is assumed to remain constant across the Outcome Period.

3. Prediction of time to default is not calculated.

1.9. Skeleton of Chapters

Chapter One: *"Introduction"* explains the new Basel Capital Accord, background of credit scoring, problem statement, purpose of the study, and limitations of logistic regression which were addressed in survival analysis.

Chapter Two: *"Literature Review"* looks at the history and progression of statistical models. It highlights the names of authors, titles of the journal articles, year of journal article, papers used, volume and version.

Chapter Three: "Methodology" looks at model building of Logistic Regression and the Cox Proportional Hazard model, model development by fitting Logistic and Cox models using Rstudio, and, lastly, checking model performance such as area under the ROC curve, confusion matrix, Gains Table and Lift Chart, and as well as model performance comparison.

Chapter Four: *"Data and Preliminary Analysis"* gives more description of data used, variables obtained, data bucketing (univariate and bivariate analysis), multivariate analysis, and stratified random sampling.

Chapter Five: *"Estimation and Results".* This chapter gives the detailed results of the Logistic Regression and Survival Analysis model. It has versions of progression, high-quality, performance and contrast, and the outcomes are presented graphically and numerically. These results were carried out to decide the methods to strive to perform better for some customer credit unit statistics, inside the existence of opposing risks and long-term non-defaulters. The SAS and Rstudio were utilised to analyse the mortgage portfolio.

Chapter Six: *"Discussion and recommendations".* This chapter gives conclusion to the Logistic Regression and Survival Analysis model. Recommendations are given for future research, in this chapter.

1.10. Conclusion

This chapter introduces the methods suggested for use in this paper, the background of the study, aims and objectives, sources of data, research hypothesis and outline of chapters. In the past, survival analysis was utilised in the engineering framework, and health, since the time extent until an occasion is analysed. It is being used for consumer mortgage loans data which is like lifetime data as it alarms a follow-up on the behaviour of events over time. Survival Analysis regression, which involves time dependent, handle censored and truncated data, addresses the limitations of Logistic Regression. This study is aimed at comparing two approaches, namely the Survival Analysis and Logistic Regression models, in the existence of opposing risks – which is non-payment.

CHAPTER 2: REVIEW OF THE LITERATURE

2.1. Introduction

This section introduces the history, progression and improvement of CS structures, Logistic Regression, SA and probability of default in financial institutions which are associated with the topic of study. It gives the list of the author(s), area of study, year of journal and the papers used. Statistical methods will be applied to model credit risk, chances, pitfalls and limitations of certain methods. The improvement on the credit scoring helps to change the business world over time.

2.2. Credit Scoring

It is an approach for characterising the hazard/risk of a loan applicant (Abdou and Pointon, 2011). Leasers can make choices utilising a credit score whether to grant a client credit or not. A moneylender commonly makes two sorts of choices: to begin with, whether to give credit to a new modern application or not, and secondly, how to deal with existing applications, including whether to increase their credit limits or not (Thomas, et al., 2002). In the 1980s, it was utilised. In agreement with Thomas et al. (2002), the accomplishment of CS in credit cards implied that the institutions must begin utilising a rating system for other products such as mortgage loans and personal loans, whereas within the final limited years, rating was utilised for domestic credits and little trade advance.

Sometime recently, in the computer age, credit-permitting choices depended on subjective human evaluation in a method called judgmental procedures. There was no sanctioning set up to manage and control choices made (Capon, 1982). Agreeing with Capon (1982), before dispatch of the Equal Credit Opportunity Act (ECOA), passed in 1974, credit frameworks separated giving of credit based on sexual orientation and conjugal status. ECOA actualised equal opportunities in getting to credits by customers in any case of sexual introduction and conjugal status. Judgemental methodologies for giving credit, which included person ruling by a credit officer on a case preface, were supplanted by means of a robotised strategy for settling and utilising credit choices, insinuated to as credit scoring; it is not only banks with credit scoring; retailers and others utilised the credit scoring system (Capon, 1982).

Numerical scoring frameworks were first created in the postal order trade in the 1930s, and advanced by utilising the substantial private financial businesses. In a typical framework, various indicator qualities were decided for their capacity to segregate between the individuals who keep to their credit agreement and the ones who did not make repayments, and points were granted to distinctive levels of every characteristic. An applicant was arbitrated on affiliation amid his/her summated score, crosswise over qualities, and freely set acknowledge/dismiss cut-offs. Initial frameworks presented such attributes as job, length of business, credit bureau clearance, individual reference, conjugal position, financial balance, neighbourhood, life insurance, sex and race. Numerical scoring structures assume an essential job in progression, when contrasted with judgemental techniques;

however, dissemination of quantitative strategies did not happen until the advancement of the vital computer innovation in the mid-1960s (Capon, 1982).

Nowadays, a credit scoring system needs less data to decide, because CS models have been evaluated to incorporate just those variables which are factually as well as altogether related with reimbursement execution through judgemental choices, have no measurable essentialness, and along these lines no factor decrease strategies are accessible. Credit scoring models endeavour to address the inclination that would come about because of considering the reimbursement pasts of just acknowledged requests, and not all requests. They do this expecting how disallowed applications would have performed if they had been acknowledged. An extra fundamental advantage of credit scoring is that the equivalent can be examined effectively by various credit experts or analysts and given similar weights.

Measurable models, for example Logistic Regression and Survival Analysis, have been deployed in the credit scoring frameworks. As stated by Dirick et al. (2017), Survival Analysis needs to be utilised within the medical setting and concluded manufacturing, where the time length is until the point that an event is investigated, for instance, the time through until the point that demise on the other hand machine dissatisfaction (Kalbfleisch and Prentice, 2002).

As indicated by Gupta (2017), Survival Analysis as an option to Logistic Regression, was introduced by Narain (1992). The principle benefit of utilising SA regression in credit risk setting is that an opportunity to non-payment can be displayed, and not simply whether an individual would or would not make payment (Thomas et al., 2002). Numerous specialists reviewed the case of Narain (1992), and started to utilise further developed procedures, when contrasted with the parametric accelerated failure time survival systems. With its adaptable, nonparametric standard risk, the Cox PH model remained the primary option in contrast to the accelerated failure time model according to Banasik et al. (1999), and further created by Stepanova and Thomas (2002) to broaden together Cox PH and AFT models by utilising, amid the remains, granular grouping as well as period-shift covariates further developed by Bellotti and Crook (2009).

In this paper, we will be adding to the current study by examining contract credits informational collections from one of the banks in South Africa, utilising the Cox PH model, and utilising measurable default time forecasts and monetary appraisal techniques, by foreseeing the future estimation of the credit, fitting to every model sort considered: the ‘‘plain’’ Survival Analysis (SA) models.

2.3. Logistic Regression

LR is a statistical method for studying a dataset in which there are one or more independent variables that decide a result. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). Cox (1958) was the first to develop Logistic Regression in 1958. The twofold

Logistic model was utilised to predict the likelihood of a paired reaction dependent on at least one indicator factor. It enables one to state that the nearness of a hazard factor expands the chances of a given result by an explicit factor. The model is an alternate likelihood display and not a classifier (Cox, 1958). The following journals or papers are being studied for this paper:

1. **Tri-Dung Nguyen, Shi-Wei Shen & Udechukwu Ojiako (2013) Modelling the predictive performance of credit scoring. ACTA COMMERCII. Independent Research Journal in the Management Sciences 13(1).**

The aim of their investigation was to study the projecting execution of credit-scoring frameworks in Taiwan.

Research configuration, style and technique: utilised an information test of 10,349 records drawn somewhere in the range of 1992 and 2010; LR models were utilised to think about the prescient execution of CS frameworks.

Results: A trial of Goodness-of-fit checked that CS models that consolidated the "Taiwan Corporate Credit Risk Index", microeconomic factors and macroeconomic factors had more projecting power. This prescribes macroeconomic factors do have informative influential for non-payment loan probability.

Applied consequences: the uniqueness in the examination was 3 credit risk regression were built up to anticipate commercial company's non-payments dependent on various microeconomic and macroeconomic variables, for example, the Taiwan Corporate Credit Risk Index, resource development taxes, stock record (SI) and total national output (GDP).

Role: the investigation utilises distinctive Goodness-Of-Fits and ROC amid the study of the strength of the prescient intensity of these aspects.

2. **Deni Memić* (2015) Evaluating Credit Default Using Logistic Regression and Multiple Discriminant Analysis: Empirical Evidence from Bosnia and Herzegovian. ACTA COMMERCII. Interdisciplinary Description of Complex Systems 13(1):128-153.**

The aim of their paper was to survey the likelihood of non-payment presence of the lending marketplace in Bosnia and Herzegovina. As such, the primary motivation behind the paper was to anticipate loan non-payment, or to make an expectation display that recognises non-payments and non-defaulters' institutions, in view on the fiscal statistics acquired from the monetary accounts, utilising many technique methods. The techniques utilised in the paper are LR and multiple discriminant analysis.

Data and methods

Financial institutions of Bosnia and Herzegovina were dissected, as the example for the study was formed. Essential examples incorporate organisations from both B&H articles, as they are assumed as relatively distinct managing account markets. Information for the investigation was gathered from a few information bases, freely, for non-payment information and comparing money related information, as no essential information occurs in Bosnia and Herzegovina. Non-payment organisations were distinguished by means of comparing time of non-payment event. Information was utilised from a few financial institutions in Bosnia and Herzegovina to guarantee that the example speaks to the greater part of the financial institutions. Money-related proportions, as fundamental non-payment indicators, were picked in view of the pertinent writing. The study incorporates 31 budgetary and two (2) dummy indicator factors, which were gathered for all tested organisations, up to four times prior to non-payment. They were coordinated with non-payment information, showing whether an organisation is failing to pay, or solid.

Results

The outcomes displayed the framed representations consuming more prescient limit. Aimed at "logit models", a few factors are extra powerful on the non-payment expectation, more than others. Profit for resources is measurably huge in every one of the four-time frame preceding failing to pay, consuming more relapse quantities, or more effect on the model's capacity to anticipate non-payment. Comparative outcomes are gotten for MDA models. It is likewise discovered that prescient capacity contrasts between LR model and numerous discriminant examination.

3. Taha Zaghdoudi (2013) "Bank Failure Prediction with Logistic Regression". International Journal of Economics and Financial Issues 3(2):537-543.

The aim of their paper was to create the microeconomic factors which can forecast the banking shortcoming.

Data and methods:

The information utilised in their work was gathered from the yearly reports of the Central Bank of Tunisia and the Tunisian association of banks and monetary organisations. Their exploration depends on yearly information crosswise over eight (8) years, from 2002 to 2010 for the 14 widespread Tunisian banks.

The act of gathered monetary proportions from the Tunisian banks' accounting reports, shapes their battery of pointers supported by the CAMEL typology, from which they need to choose the proportions that have a solid prescient capacity to develop a prevision model of bank imperfection from it. The act of a vector of proportions chose from development by a stepwise regression choice, like a vector of logical factors in their logistic model, have furnished them with sensible results with

anticipated signs and meanings. In like manner, the greatest relevant proportions in the clarification of financial institutions' imperfection at the Tunisian banks, are the decline of keeping money productivity, and the capacity of financial institutions to refund their obligations – which have all the earmarks of being a highly odd proportion.

Results:

The input accomplished utilising their temporary model, demonstrated that a bank's capacity to refund its dedication, the coefficient of banking tasks, bank benefit per labourer and impact monetary proportion, has a negative effect on the likelihood of disappointment.

4. **Arindam Bandyopadhyay (2006) "Predicting Probability of Default of Indian Corporate Bonds: Logistic and Z-score Model Approaches". Journal of Risk Finance 7(3):255-272.**

Aim

- Their study intention purposed at rising a first cautioning sign model for foreseeing commercial non-payment in creating a marketplace economy indistinguishable to India. In the meantime, it additionally plans to exhibit techniques for specifically evaluating corporate likelihood of non-payment utilising money-related and, in addition, non-financial factors.

Methods:

- They utilised the "Multiple Discriminate Analysis" for emerging "Z-score models" for forecasting commercial share non-payment in India. Statistical modelling used in their study was the LR model aimed at forecasting the likelihood of failing to pay.

Results

- Z-score model established in their exploration appeared just a high arranging force on the projected example, yet, in addition, demonstrated a high prescient power as far as its capacity to identify bad organisations in the test dataset. The model plainly performs better than another two challenging models containing "Altman's" unique and creating marketplace of proportions separately in the Indian setting. For logit study, the exact outcomes uncover that incorporation of money related and non-monetary restrictions would be helpful in further precisely portraying non-payment risk.

5. **Clemma J Muller & Richard F MacLehose (2014) "Estimating Predicted Probabilities from Logistic Regression: Different Methods Correspond to Different Target Populations". International Journal of Epidemiology 43(3):962-970.**

Background: They inspected three regular advanced to forecast projected likelihoods following confounder-balanced LR: peripheral institutionalization (projected likelihoods added to a balanced mean mirroring the confounder dispersal in the objective populace) expectation at the models (restrictive anticipated likelihoods considered through balancing each confounder to its modular worth), and forecast that every strategy that compares to an alternate target populace is overlooked in preparation. Expectation at the methods is regularly mistakenly translated as evaluating normal likelihoods for the general investigation populace, and besides, yields illogical evaluations within the sight of dichotomous confounders. Non-payment directions in famous measurable programming packages frequently lead to inadvertent misuse of forecast in the methods.

Approaches: They identify errors in evaluated probability on these methods, namely marginal standardization forecast at the methods that ascertain the anticipated likelihood of the outcomes for every presentation level, accepting that everybody in the populace had the most widely recognized estimations of the confounders, and expectation at the means that figure the anticipated likelihood of the outcomes by presentation outcome, expecting that each individual in the dataset has the average estimation of one another's confounder, and talk about suggestions for translation, and give syntax for SAS and Strata.

Outcomes: Peripheral institutionalization grants induction to the aggregate individuals from which information is drawn. Forecast at the methodologies or methods permits deduction just to the related stratum of perceptions. With dichotomous confounders, forecast at the methods matches to a stratum that does exclude any genuine remarks.

2.4. Survival Analysis

1. **Jamil J. Jaber (2017) "Credit Risk Assessment using Survival Analysis for Progressive Right-Censored Data". Journal of Internet Banking and Commerce 22(1):2-18.**

The motivation behind the subject field was to utilise different non-parametric and parametric models to assume the likelihood of non-payment, that were utilised for checking the execution of an example of conceded instalment risk book.

Data and Methods

The sampling information of credit book acquired for their investigation was gathered from a financial institution in Jordan and covers classified data using a loan of advances. The month-on-month information of the credit book was gathered from 2010-January to 2014-December. The span of book is 4,393, whereas the aggregate value of non-instalments all through the 5-year time is 495. For the example information, an applicant is announced non-payment when his/her hard money portion is not settled within 3 logbook months.

The best parametric and non-parametric models are cautiously picked utilising a few "goodness-of-fit" principles; to be specific for parametric are Mean Squared Error, Akaike information criterion and Bayesian information criterion, and for non-parametric are Standard error and Mean Absolute Deviation. The anticipated non-payment likelihood is connected to assess the credit danger of a business book at 99.9% confidence interval (CI) and a few period limits (3, 6, 9, and 12 months).

Outcomes

In their study, the assessed Probability of Default Gompertz model was utilised for anticipating the most pessimistic scenario non-payment rate of a credit book at 99.9% CI and a few time prospects. The most pessimistic scenario non-payment rate is another component compulsory to figure the Risk Weighted Assets, that is the equation for ascertaining the principal prerequisites in Basel II Internal Rating Based (IRB). The outcomes demonstrate that the assessments of Probability of Default and most pessimistic scenario non-payment rate increment amid the one-year time, while the appraisals of copula connection decline amid a similar time. The outcomes are normal, since the Probability of Default (PD) and most pessimistic scenario non-payment rate has encouraging association though the PD, and copula connection has a bad association.

Recommendations

For further investigation, they intended to join the macroeconomic impacts in the forecast of Probability of Default. What's more, the idea of hazard exchange through protection approaches for lessening the credit danger of book can be well thought out, and examines on the expectation of Probability of default which considers protection strategies for diminishing credit dangers will be completed in their future research.

2. **Dyana Kwamboka Mageto, Samuel Musili Mwalili & Anthony Gichuhi Waititu (2015) "Modelling of Credit Risk using Random Forests versus Cox Proportional Hazard Regression". American Journal of Theoretical and Applied Statistics 4(4):247-253.**

Aim

The aim of their paper was to present Random Survival Forests (RSF) as another technique for modelling loan hazard, and to relate it to the CPH model.

Data and Methods

The data applied in their trial was optional information. It was acquired from driving business banks in Kenya. The credit candidates in the investigation were arbitrarily selected from the financial

institutions record including of seventy divisions. The trial acquired depended on arrangement of individual credits whose development was 45 months. The investigation therefore involved credits reserved from the long stretch of January-2004 to September-2008. The example acquired comprised 250 male candidates and 250 female candidates.

- *Random Survival Forest (RSF)*: 500 trials information booked 108 non-payment accounts. The family "surv" forest has constructed the model with 2000 trees with 3 factors tied at each split. In their study they utilised non-payment divided criteria – i.e. the log rank test measurement. The mistake rate on doing the execution assessment of the out-of-sack (OOS) appraisals of blunder recommended that when the subsequent model was connected the mistake was acquired as is littler than 0.5, thus inferring that they do not have enough proof to reason that the indicators are not imperative in foreseeing the likelihood of non-payment. In conclusion, this indicates to be a good model. The factors vital, as indicated by RSF, are Marital Status, Employment, Home Ownership and Educational level, while sex and age were the slightest critical.
- *Cox Proportional Hazard Model (CPH)*: With progression with their investigation of CPH model, period and position were relapsed against alternate factors.

Results

The Cox PH and RSF models were utilised in their paper. The "Harell's concordance index" (C-index) for Random Survival Forest model stood at 0.4378, whereas of the Cox PH model acquired stood at 0.3376. From their study, it demonstrates that the Cox proportional hazard model has a littler Harell's concordance index value to that of Random Survival Forest. It is apparent that the Cox model outperformed RSF according to Harell's concordance index.

Discussion and Prescriptions

Cox Proportional Hazard model was observed to be a superior model for evaluating the likelihood of non-payment, as coordinated to Random Survival Forest. In the two models, variables such as marital status, employment and home ownership were observed to be the regular vital factors. Be that as it may, the Random Survival Forest model showed highest education level as an imperative variable also. It was likewise discovered that gender and age do not influence, and were not vital in anticipating the likelihood of non-payment.

They prescribed the utilisation of different strategies to display credit hazard, such as the AFT and Kaplan-Meier models, to see how the models would carry on.

3. **Denis V. Rylov, Dmitry V. Shkurkin Anna A. Borisova (2016) "Estimation of the Probability of Default of Corporate Borrowers". International Journal of Economics and Financial Issues 6(S1):63-67.**

Objectives

Their study was to demonstrate the likelihood of non-payment of development organisations with the utilisation of logit-models of twofold decision dependent on monetary announcing information, institutional attributes, and additionally macroeconomic pointers, as a device for bookkeeping impact of cyclic economy.

Data and Methods

The premise of the database of study assisted different sources: data investigative framework FIRA PRO, information posted on the sites: Bank of Russia, Federal State Statistics Service, the Supreme Arbitration Court of the Russian Federation, the International Monetary Fund, and Bank for International Settlements. Logical position of the exploration depended on being crafted by remote (Altman, Beaver, Merton and others) and Russian (Karminsky, Peresetsky, Pomozanov and others) creators. The study utilised strategies – for example, a survey of the logical examination, blend, characterisation, virtual investigation and sorting in the practical portion of the utilised approaches of techniques for measurable investigation and econometric displaying.

The utilisation of these methodologies, because of freely accessible information on Russian organisations, was carried out to choose the most prevailing risk pointers (monetary, macroeconomic and institutional) and executed multifaceted displaying non-payment likelihood dependent on the chose issues. Systematisation and organising of different procedural parts of the assessment of the PD, permitted to shape a comprehensive perspective of the current strategies for assessing the Probability of Default, considering the benefits and shortcomings of these techniques and the degree of their usability to the Russian training. The outcomes of this investigation were the premises, and were utilised in choosing methods and demonstrating devices as a feature of developing their personal models to forecast the Probability of Default for the Russian organisations.

Results

In the beginning of 2014, the nature of loaning to non-money associations added up to about 56% book and 39% of the estimation of Russian banks' properties. As per these scientists, the dimension of extraordinary obligation of the corporate credit book will, in general, grow. More growth in the offer of corporate non-payments in the books of banks may cause unsteadiness in the managing of an

account division and the money-related framework. A substantial extent of loaning in the Russian market signified loaning development businesses. The disasters of 2007-2009 and 2015-2016 presented that business in this trade was mostly influenced by macro-economic stuns, that prompts to curiosity in the development of a model-estimation of non-payment likelihood for development business.

4. **Lore Dirick, Gerda Claeskens & Bart Baesens (2017) "Time to default in credit scoring using survival analysis: a benchmark study". Journal of the Operational Research Society 68(6):652-665.**

Aim

The aim of their paper was to determine period until non-payment in CS model using SA.

Data and Methods

Ten (10 genuine informational collections were utilised, and they utilised three primary assessment ways to deal with model performance: Area under curve, non-payment period expectation contrasts and future credit number approximation. They demonstrated that Cox Proportional Hazard models are all especially good, particularly a Cox Proportional Hazard model in mixture with penalised keys for the constant covariates.

Results

They found that the Cox Proportional Hazard display is superior to the multiple event mixture cure model, yet the mixture cure model does not accomplish fundamentally unique in the greater part of the cases, and is one of` the best models utilising financial assessment. It has the benefit of not demanding the survival capacity to go to 0 when period goes to boundlessness, which frequently is the most proper for CS information.

Recommendations

They expressed that, from their discoveries, it would be more intriguing to additionally broaden the mixture cure model and concentrate the execution of the subsequent model in correlation with a Cox Proportional Hazard regression with punished projections. They state that it should be possible by taking into consideration projections in the constant covariates. Moreover, it is intriguing to execute every one of the models once more over information that has been coarse-ordered, and contrasts its outcomes with the outcomes in this investigation. It is intriguing to contrast the outcomes of coarse

order with the spline-based strategies in this investigation, which can be an option for taking care of nonlinearity in the information.

5. T. Bellotti & J. Crook (2009) "Credit Scoring with Macroeconomic Variables Using Survival Analysis". Journal of the Operational Research Society 60(12):1699-1707.

Purpose

The aim of their paper was as follows:

- to prove that the Survival Analysis model is modest for forecast of non-payment as compared to the LR model.
- to also investigate the theory that likelihood of non-payment is influenced by general conditions in the economy after some time – i.e. incorporation of the macroeconomic factors gives a measurably huge enhancement in forecasts of non-payment.

Data

- Credit card requests and month-on-month performance information from a UK bank was utilised. The card accounts opened somewhere in the range of 1997 and 2001 were utilised as a training informational collection, and those opened somewhere in the range of 2002 and 2005 were utilised as a test informational collection. Every dataset contained more than 100k records with application factors – for example, salary, age, house and work status alongside a FICO score reserved at the period of request of a loan.
- A record is in default state on the chance that it went three months or more inside the initial year of their investigation. A record that defaults is alluded to as a bad account and a non-defaulting account is alluded to as a good account. The informational collection, utilising this definition, showed that the extent of awful cases in the information was little.
- The following macro-economic factors were utilised: Interest Rates (IR), Earnings, FTSE, Unemployment (Unemp), Production (Prod), House Price Index (House) and Consumer Confidence Index (CC). These macro-economic factors were chosen as the highest expected to affect non-payment. A positive value implied that as the estimation of the macro-economic factors increased, this was connected to an increase in danger of default, and the other way around – e.g. interest rate had a positive value, implying that expansion in financing cost is relied upon to put further worry into the economy, resulting in rise in non-payment, while production that has a negative value is a pointer of enhancing the economy, giving conditions to diminished danger of default.

Methods

- Subsequently the information was skewed regarding good to bad accounts; more noteworthy weight was given to the bad accounts. This is feasible for both Cox Proportional Hazard and Logistic Regression models, which meanwhile use Maximum Likelihood Estimation for which bad accounts can be incorporated into the probability function multiple times. Training data was demonstrated utilising Cox Proportional Hazard model to show time to default with each macro-economic factor. Cox Proportional Hazard model was utilised, since it takes into consideration incorporation of macro-economic factors as Time Varying Covariates (TVCs). This appeared differently in relation to the LR model which is a standard model for scoring. A Cox Proportional Hazard model without macro-economic factors was additionally worked, to decide if any elevate in execution was because of the utilisation of Cox Proportional Hazard model or the incorporation of macro-economic factors.
- Each macro-economic factor was then cooperated with an application variable and added to the essential model. It was normal that a few classes of credit buyers would be more inclined to changes in financial conditions, than others. The inspire of the model was then estimated utilising the Log Likelihood Ratio (LLR) got from the Maximum Likelihood strategy utilised to appraise the model. The connection giving the most reduced p-value for its LLR is incorporated into the ideal macro-economic Cox Proportional Hazard model.

Assessment:

- The ideal model was surveyed as far as the two of its logical power on the training data and its predictive power on the autonomous test set.
- The Cox model was evaluated as a logical model by announcing its fit to the training data with and without macro-economic factors, utilising Log Likelihood Ratio. The importance of every coefficient in the model is resolved utilising a Wald statistic resultant from MLE. The Wald statistic pursues chi-square statistics; thus a p value can be figured for the null hypothesis that the coefficient value is 0.

Results and Conclusion:

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

6. **Precious Mdlongwa, Hausitoe Nare, Thandekile Hlongwane & Isabel L. Moyo (2014) "Censored Regression Techniques for Credit Scoring (CS): A Case Study for the Commercial Bank of Zimbabwe. International Journal of Economics and Finance 6(10).**

Purpose:

The purpose of their article was to calculate the risk associated with CS in the Commercial Bank of Zimbabwe.

Data and Methods:

Data

The informational index utilised secured individual advances as of 2010-01-01 until 2012-01-01. Linear and Buckley James regression tests were utilised to locate the informative factors impacting period to non-payment and reimbursement. In their investigation of client grouping, statistical procedure (i.e. Discriminant Analysis) was employed.

Results

Time of life, conjugal status, credit reason and time at present place of employment were observed to be directly identified with period to non-payment. Time to refund was observed to be directly identified with age, conjugal position and credit reason. The 67.51% of the first accounts were observed to be accurately ordered. Buckley James regression did better than linear regression; subsequently, it was observed to be the greatest reasonable strategy in deciding factors influencing dangers in credit offering.

Recommendations

These researchers suggested that the business bank of Zimbabwe should attempt and watch the credit execution of every client and go about when the advance goes bad. It is recommended that the bank ought to set up a loan risk supervisory crew that ought to oversee the accompanying activities that will help in limiting credit chance:

- Rebuilding the FICO rating sheet and reallocate scores to every one of the factors that influence defaulting and reimbursement.
- Employing the Buckley James technique, as it ended up being better performing.
- Reviewing the base age for a credit candidate, since examination demonstrated that twenty-one years is not legitimate for a credit request.

- Studying the clients that are below lone and wedded in the Fico assessment piece, since they have widows and single men.
- Carefully checking the credit execution of every client thinking about survival analysis also.

7. José Pereira (2014) "Survival Analysis Employed in Predicting Corporate Failure: A Forecasting Model Proposal". International Business Research 7(5).

Aim

The principle motivation behind this paper was to display the expectation of company economic failure dependent on survival analysis (SA), a methodology that remains its benefits.

Methods

The model created in their paper applies persistence period, risk proportion as the reliant factor, and goes up against bad and good organisations originated as of a similar populace, thinking about the next cases as censored data. The principle favourable position of the model utilised depends on the extra data it gives. This methodology, gives an alternate point of view, since the survival curve of examination of an organisation permits them to express the probability of an organisation survival past a given time, and henceforth the danger of sinking into economic failure. In any case, correspondingly to what occurs with different techniques, the precision of the model created in their study depended completely on the nature of the information which bolsters the reason for demonstration.

Their study depended on the proportionality of hazards, which may not always be the case. Another pertinent restriction is the trouble of getting the survival periods – i.e. when the marvel that is being examined happens.

Results

In view of the outcomes found from the example utilised, they cannot help suspecting that this technique suggests great points of view when utilised for the advancement of determining models in the liquidation investigation. They are persuaded that utilising a progressively huge example of businesses, together with inspected accounts, as well as consolidating subjective variables, it might be conceivable to build up a model with a higher prescient power, which might be of incredible helpfulness for decision-making.

2.5. Conclusion

In this chapter, studies of modelling credit risk were debated, from before the advent of computers to more developed methodologies. Recently, before the computer age, credit-permitting choices depended on human evaluation in an approach called the Judgmental approach. According to Capon (1982), before dispatch of the ECOA, in 1974, credit frameworks separated giving of credit based on sex and marital status. The present procedures are led by Basel, and all people are given equal chances because of credit scoring that is employed during credit application. Statistical models, for example LR and SA, have been deployed in the CS frameworks. According to Gupta (2017), SA is another option to LR, which was first introduced by Narain (1992). The benefit of utilising Survival Analysis regression in credit risk is that an opportunity to non-payment can be displayed, and not simply whether a customer would pay or fail to make payment. With its adaptable non-parametric standard risk, the Cox PH model remained a primary option, in contrast to the accelerated failure time model according to Banasik et al. (1999) and further created by Stepanova and Thomas (2002) to broadened together Cox PH and AFT models by utilising, amid the remains, granular grouping as well as period-shift covariates further developed by Bellotti and Crook (2009).

CHAPTER 3: METHODOLOGY

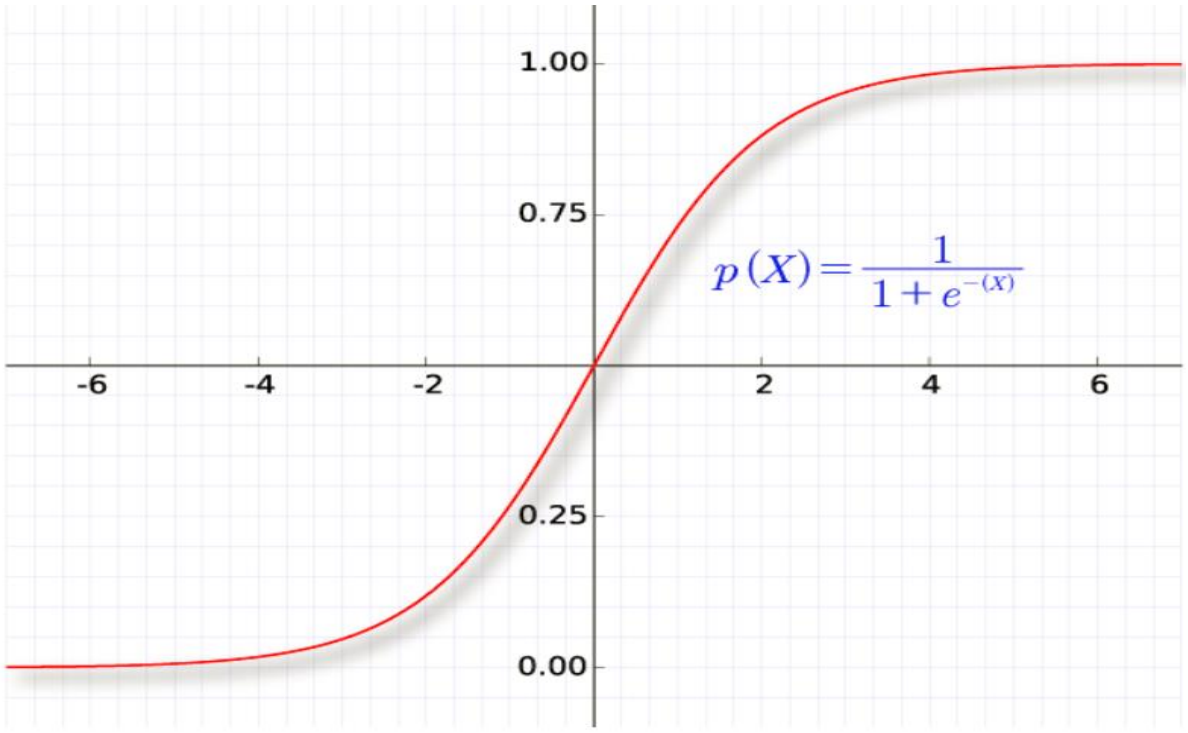
3.1. Introduction

The importance of this thesis is to analyse the likelihood of non-payment in mortgage loans. Data from one of the financial institution was used to model probability of default (PD) in a credit risk context. The bucketing of variables is necessary, to obtain the variables which are predictable (this is fully explained in Chapter 4). Univariate and multivariate data analysis shows trends over time and weight of evidence, as well as information value. Logistic and Cox models were built, and the best model was chosen by checking the performance of these methods, by financial institution, for credit scoring. Next sections fully explain the methods used to build PD models with different statistical procedures such as LR and SA regressions.

3.2. Logistic Regression

LR is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). As with all regression analyses, the Logistic Regression may be prognosticative analysis. It is employed to explain information data, and to clarify the link on the dependent binary variable and one, or a lot of, nominal, ordinal, interval or ratio-level freelance variables.

Form of the Standard Logistic Function:



Source: http://www.thefactmachine.com/wp-content/uploads/2015/03/13-Sigmoid.gif

The odds ratio proves the important part in understanding the results found from logit analysis. The OR is measured as the family member of the probabilities that state occurs to the likelihood that would not occur. The logit model such as random character of the sample, collinearities of variable independency of observation, are assumptions that must be met.

The LR is not the same as systematic regression due to dependent variable is binary. An amount of the likelihood of the consequence is specified by the odds of existence of an event. The odds of default are given by:

$$odds\ of\ default = \frac{p}{1-p} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p.$$

To solve this equation for p, one first applies the exponential function to both sides of the equation:

$$\exp\left(log\left(\frac{p}{1-p}\right)\right) = \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p).$$

Recall that $\exp(z) = e^z$ so that the right-hand side of the previous equation is

$$\exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) = e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}$$

Hence with log, $\exp(log(z)) = z$

Thus, the left-hand side is

$$\exp\left(log\left(\frac{p}{1-p}\right)\right) = \frac{p}{1-p}$$

Thus, after putting exponent on both sides, logistic regression equation becomes:

$$\frac{p}{1-p} = \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

$$= e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}$$

After multiplying both sides by $1-p$,

$$p = (1-p)\, e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}$$

$$p = e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p} - pe^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}$$

$$p + pe^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p} = e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}$$

Next, factor out the p,

$$p\left(1 + e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}\right) = e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}$$

$$p = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}{1 + e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}$$

Multiply both sides with $e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_{p)}}$. Since $e^{-z}e^{z} = 1$, this gives

$$p = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}{1+ e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}} * \frac{e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_{p)}}}{e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_{p)}}}$$

$$p = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p} * e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_{p)}}}{e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_{p)}} + e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p} * e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_{p)}}}$$

$$= \frac{1}{e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_{p)}} + 1}$$

The likelihood that a loan predicted to default is modelled by the LR model, and is shown below:

$$P(Default|x_i) = \pi(x) = \frac{e^{\alpha+\sum\beta_i x_i}}{1+ e^{\alpha+\sum\beta_i x_i}} = \frac{1}{1+ e^{-(\alpha+\sum\beta_i x_i)}}$$

, where: $P(Default|x_i) - dependent\ variables$ usually determines the probability of default

$\alpha - intercept$

$\beta_i,\ i = 1, 2, \dots, k - weights\ (coefficients)$

$x_i, \quad i = 1, 2, \dots, k\ - explanatory\ variables - financial\ ratios$

$P(Default|x_i)$ / Probability of Default assumes the value between 0 and 1, where 0 – "good" and 1 – "bad" enterprise (Chmielewska, 2016).

3.2.1. Model Development

The LR approach is made up of choosing an example of borrowers, and categorising customers from good or bad in a model. The sorting is constructed on the reimbursement outcome completed in an agreed time. The purpose of this LR model is to forecast credit risk, and to get covariates that are significant in CR forecast.

The records in the book are ticked as good or bad, constructed on specific standards. These standards are grounded on recompense performance. For instance, a distinctive standard for a bad account is a payment delay of more than 90 days. Payments made on time is classified as a good account. Lastly, accounts for which it difficult to be classified either as good or bad are indicated as vague, and are not included in this paper. Steps to build LR credit risk model are as follows:

1. Fitting Logistic Regression

Final Model Selection:

To fit the LR model, Proc-Logistic was utilised in SAS statistical software program, and is carried out and brings the Analysis of Maximum Likelihood Estimation. The SAS program will execute the logistic regression with all the preliminary set of factors based on univariate and multivariate evaluation distinctive in Chapter 4.

Model Fit Statistics:

Assessing the "Global Null Hypothesis: Beta = 0" table contains the results of 3 test(s) of the global null hypothesis that all coefficients (apart from the intercept) are equal to 0, often thought of as a test of the utility of the whole model:

$$\beta_1 = \beta_2 = \cdots = \beta_k = 0$$

These tests are asymptotically equivalent (equivalent in very large samples), so will usually agree. In cases where they do not, many prefer the likelihood ratio test (Hosmer and Lemeshow, 2013).

The Model Fit Statistics table displays two relative fit statistics, AIC and SC (BIC), in addition minus two multiply the log likelihood of the model. Generally, you will want the values in the Intercept and Covariates column, as these reflect the full model just fit. Lower values on AIC and SC signify better fit (both penalise adding more parameters to the model).

2. Assess Model Fit

Assessment of model fit is intended to measure how well model expected values match observed values. Within the case of Logistic Regression models, these measures quantify how well the predicted likelihood of the result matches observed likelihood across all divided groups of subjects.

One of the most-used test statistics is the Deviance χ^2, and its purpose is to measure how well the current model is from a perfectly fitting saturated model. A saturated model produces expected chances that completely match the determined chances by estimating a separate parameter for every distinct covariate pattern. Pearson χ^2 on other hand measures lack of fit in a slight way. Large (and significant) χ^2 values for both signify poor fit:

$$Deviance\ \chi^2 = 2\sum_{j=1}^{J} \log(\frac{O_j}{E_j})$$

$$Pearson\ \chi^2 = \sum_{j=1}^{J} \frac{(\ O_j - E_j\ )\text{^}2}{E_j}$$

3. Influence Diagnostics

Observations whose outcome values do not match the expected pattern predicted by the model (have large residuals) and who have extreme predictor values (high leverage) can have a disproportionate influence on the model coefficients and fit. I will be assessing how influential observations are affecting the model estimates, before concluding that the model fit is satisfactory.

I will use the DFBETAS and C diagnostic measures to help my model to identify influential observations. The DFBETAS statistic is calculated for each covariate pattern, and measures the change in all regression coefficients if the model were to be fit without that covariate pattern. DFBETAS, for observation $j, is\ a\ vector$ of standardised differences aimed at each coefficient i :

$$DFBETASi_j = \frac{\Delta_i \widehat{\beta_j}}{\sigma_i}$$

$\Delta\widehat{\beta_j} = vector\ or\ difference\ in\ parameter\ estimates\ when\ covariate\ pattern\ j.$

$\sigma_i = standard\ error\ of\ the\ ith\ coefficient.$

The C diagnostic is also calculated for each covariate pattern, so will be the same for observations with the same covariate values. It measures how much the overall model fit changes with deletion of that pattern. For observation j, C is calculated as:

$$C_j = \frac{\chi^2 h_j}{\left(1 - h_j\right)^2}$$

$\chi^2 = square\ of\ the\ Pearson\ residual, measure\ of\ bad\ fit$

$h_j = diagonal\ element\ of the\ hat\ matrix\ for\ covariate\ pattern, measure\ of\ leverage$

3.2.2. Model Performance

I will be evaluating our best model based on both its descriptive influence on the numbers and its projecting influence on the test data set. By hand-picking a very good model from the pool of experimental statistical models, we compared the ROCs, AUROC, KS, Gini etc.

3.2.2.1. Area under the ROC curve: AUC

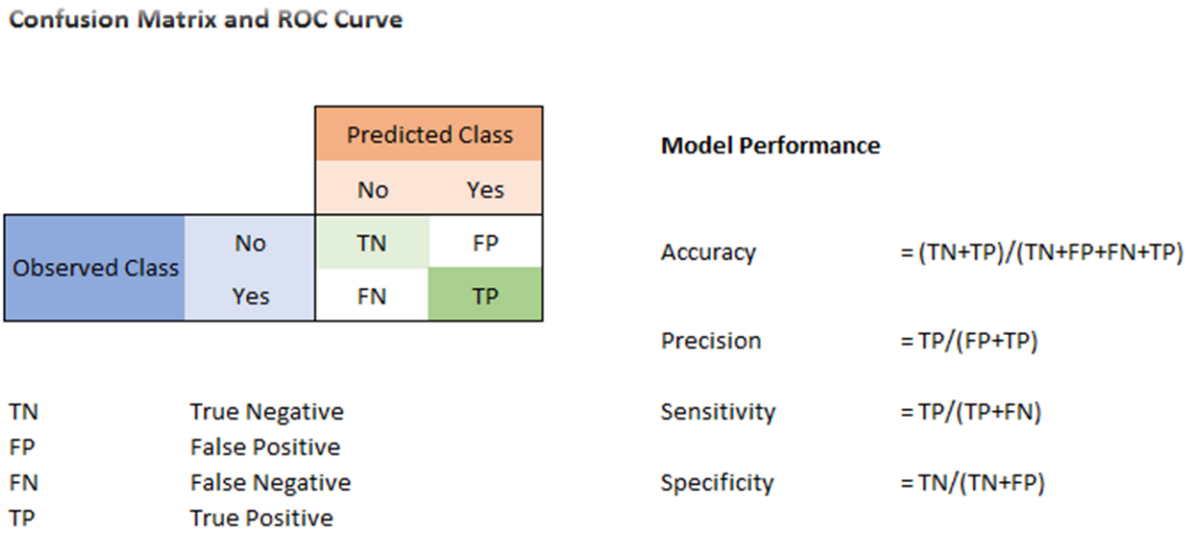
This remains the most common quantity of accuracy for machine learning techniques for binary classification problems. It is best to use as **AUROC** or **AUROCC** rather than using **AUC** as 'AUC' can be an area under any curve and we are interested in only "Area under ROC curve".

The greatest imaginable assessment (100 percent sensitive and 100 percent specific) would have an area under the curve of 1.0 (100%), since the entire graph would fall under the curve. While it is doable that the "perfectly bad test" would have an area of zero, I would simply redefine what is positive and negative and it would become a very good test.

Receiver operator characteristic (ROC) curves display the give-and-take of sensitivity and specificity across predicted probability cut-offs. The y-axis measures sensitivity. The x-axis measures (1-specificity). Therefore, points toward the top maximise sensitivity while points toward the left maximise specificity. A point at the very top left of the graph represents a predicted probability cut-off where sensitive and specificity are both maximised, where all Predicted=0 and Predicted=1 have been correctly classified. In that scenario, all observations with Observed=0 have a projected likelihood lower than the cut off, and all observations with Observed=1 have a projected likelihood advanced than the cut off. That is an unlikely situation. Producing an ROC curve for a fitted regression model is quite simple in proc logistic. Simply specify the plots=roc on the proc logistic statement to request the plot. I added the option (only) to prevent the influence plots from being generated.

3.2.2.2. Confusion Matrix

Confusion matrix is another method used to check performance measurement for a machine learning classification problem where output can be two or more classes. It is a table with four (4) different combinations of predicted and actual values. It is used to measure Recall, Precision, Specificity, Accuracy and AUROC curve.



Source: http://dni-institute.in/blogs/wp-content/uploads/2015/02/ConfusionMatrix.png

Confusion matrix can interpret the results as follows:

Predictive Power	TP & TN rate
Acceptable	>60%
Good	>70%
Very Good	>85%

3.2.2.3. Kolmogorov-Smirnov Goodness-of-Fit

In this paper, KS Goodness-of-Fit checks if two classes vary significantly. It measured whether the model can classify “not default” class from “default” class very well. If they are completely separated (i.e. AUROCC=100%) then the value of **KS** will be 1 (100%), and if they are same then the value of **KS** will be 0 (0%). This means the higher value is better classification power. In simple term KS = Maximum distance between distributions of two classes.

3.2.2.4. Gini Statistics

Gini is another commonly used measure to evaluate goodness of fit for binary classification problems. Which has a straight relationship with **AUROCC** (Gini = 2*AUROCC - 1) or “Somer's D = (Concordant Percent - Discordant Percent) / 100. It should be greater than 0.4”.

3.2.2.5. Cross Validation

For model prediction, I would like an estimation method with low bias and low variance. There are many reasons for the bias and variances, such as model misspecification, data scarcity, over fitting, etc. Cross-validation is one of the testing methods to check for the bias and variance. Cross-validation may be a computer concentrated procedure, utilising all accessible cases as training and test illustrations. It imitates the utilisation of training and test datasets by recurrently training the algorithm K times with a division 1/K of training instances cleared out for testing purposes.

This kind of hold-out assessment of execution lacks computational proficiency because of the recurring training, but the last-mentioned are intended to lower the modification of the estimate (Stone, 1974). There are types of cross-validation, including “leave-p-out”, “leave-one-out”, “k-fold”, and “repeated random sub-sampling validation” which the approach used in this paper. Individual round of such cross-validation includes subdividing an example of data into balancing subgroups, executing the analysis on the subgroup, and confirming the study on the additional. To follow the sampling selection, 70% of random sample is selected as the training dataset, and the rest 30% for the testing dataset for each round.

As alternative to LR model, the Survival model is built and explained in the next section. It addresses the advantages of Logistic Regression.

3.3. Survival Analysis

SA approach is a division of statistics for analysing the projected time until one or more occurrences occur. There are three basic areas that are considered in the survival model, which are not addressed in the logistic regression model which are, namely (Chmielewska, 2016):

- Period that the event of the notable proceedings is predictable

- The power of changes amid notable proceedings, and
- Amount and order of proceedings

In this paper, Cox Proportional Hazard is used. The main reason is that it includes macroeconomic variables as well as time-varying covariate. The Cox regression model approach is the combination of the parametric and non-parametric approach. The non-parametric approach gives data about variations of each behaviour's systems in period. In the parametric technique, the period among proceedings is implicit to be a random variable with specific distribution.

Censoring is exceptionally characteristic for occasion history information. In case data is not accessible at that point, it is censored. The foremost normal is right censoring when the time until occasion is not acknowledged, but it is longer than perception time (Chmielewska, 2016). The fundamentals of SA are specified, such as symbolisation and mutual difficulties.

This paper considers strategies for the investigation of data, which demonstrate the time when an occasion happens. It focuses on an occasion called default, a term normally utilised as a part of association with infringement of obligation contract conditions, e.g. absence of will or incapacity to pay obligation. Because a survival analysis concept is utilised, I expect that every customer will confront default sooner or later, conceivably exceptionally far off.

The default may occur long after the clients' death or withdrawal from the bank, hence the nature of this study approach implies heavy censoring.

Let X be positive random variables representing the time of default of a client, suppose that X is F and density function f. The distribution of X use mainly hazard function, conventionally denoted by $h(.)$, which is defined as

$$h(t) = \lim_{h \to 0+} \frac{1}{h} P(t \leq X < t + h | X \geq t) \ . \qquad (1)$$

A hazard function identifies the distribution of X.

If survival function is $(t)\ by\ S(t) = 1 - F(t)$, therefore hazard function is

$$h(t) = \lim_{h \to 0+} \frac{F(t+h) - F(t)}{h} \frac{1}{S(t)} = \frac{f(t)}{S(t)} = -\frac{d}{dt} log S(t) \qquad (2)$$

and therefore

$$S(t) = 1 - F(t) = \exp[-\int_0^t h(s)ds] \qquad (3)$$

Cumulative hazard function, denoted $\Lambda(.)$ is as follows

$$\Lambda(\mathrm{t}) = \int_0^t h(u)\,du = -logS(t)$$

The above of cumulative hazard function is address being at risk at time t. This explains that a client had not defaulted before time t and will be able to trace him in a next period.

The next subsections explain different types of estimation of Survival Function, Measures of Central tendency, Test of Equality over strata, Model Selection and Development and Model Assessment.

3.3.1. Estimation of Survival Function

There are two alternatives for assessing the survival, namely:

a. Non-parametric estimation: includes Kaplan Meier, Life-Tables and Nelson Aalen estimators.
b. A parametric model $\lambda(\mathrm{t})$ created on a density function $f(t)$.

The following sub-sections will cover the class of non-parametric estimators which includes the Kaplan–Meier, Life-table, and Nelson-Aalen estimators.

3.3.1.1. Kaplan Meier Estimator

Non-parametric statistics utilised for forecasting the survival function from the lifetime data is called a K-M estimator, and is recognised as product limit estimator. For instance, it is utilised in a health context to calculate the segment of patients living for a certain period after treatment. It can be utilised to calculate the span of time individuals continue to be jobless after a work loss, the period to let-down engine parts. K-M survival function estimator is measured as:

$$\hat{S}(t) = \prod_{t_i \le t} \frac{n_i - d_i}{n_i},$$

Where n_i: $number\ of\ subjects\ at\ risk$

d_i : $number\ of\ subjects\ who\ fail, both\ at\ time\ t_i$

Because of this, each period is the conditional likelihood above time t_i, so the likelihood of surviving above time t_i, specified focus has lasted until time t_i. The survival function forecast of the absolute and total/without any limits or restrictions probability of survival above time t is received by multiplying all conditional likelihoods until time t.

3.3.1.2. Lifetable or actuarial estimator

The lifetable method is very common in actuary and demography. It is particularly suitable for grouped data. Life-tables are frequently utilised within the insurance industry to appraise life expectancy and to set premiums. Life-tables are utilised broadly in biostatistical examination called a **cohort life table**. The cohort life table reviews the encounters of members over a pre-defined follow-up time in a cohort study or in a clinical trial, until the period of the occasion of intrigue or the conclusion of the study, whichever comes to begin with.

$$\hat{S}(t_j) = \prod_{l \le j} \hat{p}l$$

With $\widehat{p_j} = 1 - \widehat{q_j}$ (conditional probability of surviving), $\widehat{q_j} = \frac{d_j}{r'_j}$ (conditional probability of defaulting), and $r'_j = r_j - \frac{c_j}{2}$

3.3.1.3. Nelson-Aalen estimator

The N-A estimator is a non-parametric estimator of the total risk rate function in occasion of half-completed data. It is utilised in survival hypothesis, unwavering quality and life insurance, to conjecture the expanding aggregate of foreseen preliminaries. An occasion can be the default of a non-repairable part, the demise of an individual, or any occurrence for which the exploratory unit stays in the default state from the time when it changed on. The main aim of the Nelson-Aalen estimator is on the cumulative hazard at time t $\left(\widehat{\Lambda}_{NA}(t)\right)$.

$$\left(\widehat{\Lambda}_{NA}(t)\right) = \sum_{j.\tau_j \le t} \frac{d_j}{r_j}$$

After we calculated $\left(\widehat{\Lambda}_{NA}(t)\right)$, we can develop the Fleming-Harrington estimator of $S(t)$

$$\hat{S}_{FH} = \exp(\widehat{\Lambda}_{NA}(t))$$

3.3.2. Measures of Central tendency

Central tendency describes the central point of the distribution, and variability describes how the scores are scattered around that central point. The three common measures of central tendency are the mean, median, and mode. This paper is focused on forecasts of survival period at which 50% or 25% of the population have defaulted. Due to the slope frequently shown with follow-up times, medians are regularly distant better markers of a normal survival time.

3.3.3. Test of Equality over strata

Test of Equality over strata in this paper calculates the measurement for the non-parametric "Log-Rank" and "Wilcoxon" tests that are shown below:

$$Q = \frac{[\sum_{i=1}^{m} w_j(d_{ij} - \hat{e}_{ij})]^2}{\sum_{i=1}^{m} w_j^2 \hat{v}_{ij}},\ where$$

$d_{ij}\ is\ observed\ number\ of\ failures\ in\ stratum\ i\ at\ time\ t_j, \hat{e}_{ij}\ is\ the\ expected\ number\ of\ failures\ in\ stratum\ i$ at time

$t_j, \hat{v}_{ij} - estimaor\ of\ the\ variance\ of\ d_{ij}$

$w_i - weight\ of\ the\ differnce\ at\ time\ t_j$

These statistics total the weighted contrasts between the experiential number of defaulters and the projected number of defaulters for each stratum at each timepoint, expecting the same survival function of each stratum. In other words, if all strata have the same survival function, at that point we anticipate the same extent to default in each interim. In case these extents methodically contrast among strata over time, at that point the Q measurement will be expansive and the null hypothesis of no contrast among strata is more likely to be rejected. The log-rank and Wilcoxon tests in the results table vary in the weights w_j used. The log-rank test utilises $w_j = 1$, so contrasts at all time interims are weighted similarly. The Wilcoxon test employs $w_j = n_j$, so that contrasts are weighted by the number at hazard at time t_j, in this way open-handed more weight to contrasts that happen prior in follow-up time. Other non-parametric tests utilising other weighting plans are accessible through the test=option on the strata statement.

3.3.4. Model Selection and Development

The incorporation concerning collaborations of utilisation yet macroeconomic factors may likewise primary to higher measurable models, since, generally, more groups of credit customers will stay one-sided to changes inside monetary stipulations, than others. Collett (1994) depicted the methodologies that help for choosing a model methodology connected to administer which associations with grasp. The model fit will be estimated utilising the log-likelihood ratio (LLR) coming from maximum likelihood technique utilised to assess the model. Each covariate, the associations giving the smallest p-value for its LLR, is checked in the ideal CPH model. Variables of this paper will be fitted in the CPH model, and in this way the variables are noteworthy at 5% dimension of significance. In addition, the Cox proportional hazards are clarified further, below:

The Cox proportional hazards

Suppose reliance of the hazard rate on period is evacuated by communication the hazard rate because of $h_0(t)$, a benchmark risk rate which portrays the risk rates reliance on time alone, and $r(x, \beta_x)$, which depicts the risk rates reliance on the other x covariates:

$h(t) = h_0(t) r(x, \beta_x)$

In this parameterisation, $h(t)$ will measure up to $h_0(t)$ when $r(x, \beta_x) = 1$. It is naturally engaging to let $r(x, \beta_x) = 1 \; when \; all \; x = 0$, along these lines making the pattern hazard rate, $h_0(t)$, comparable to a regression intercept. Above, risk rate was talked about, as reliance on its covariates as an exponential function helpfully enables the regression coefficients to go on any value while still constraining the hazard rate to be positive. The exponential capacity is likewise equivalent to one when its contention is equivalent to zero. Let $r(x, \beta_x) = \exp(x\beta_x)$, and the risk capacity will be given by

$h(t) = h_0(t)\exp(x\beta_x)$

This parameterisation shapes the CPH model. It is called a risk model in view of the fact that risk rates between two sets with settled covariates will remain consistent after some time in this model. For instance, the risk rate when time t when $x = x_1$ would then be $h(t|x_1) = h_0(t)\exp(x_1\beta_x)$, and at time t when $x = x_2$ would be $h(t|x_2) = h_0(t)\exp(x_2\beta_x)$. The covariate impact of x, at that point is the proportion between these two risk rates, or a hazard ratio (HR):

$$HR = \frac{h(t|x_2)}{h(t|x_1)} = \frac{h_0(t)\exp(x_2\beta_x)}{h_0(t)\exp(x_1\beta_x)}$$

Notice that the standard risk rate, $h_0(t)$ is offset, and that the risk proportion does not rely upon time t:

$$HR = e^{\beta_x(x_2 - x_1)}$$

The risk proportion (HR) will in this manner remain consistent after some time, with settled covariates. In view of this parameterisation, covariate impacts are multiplicative as opposed to added additive, and are communicated as risk ratios, as opposed to risk contrasts. As we see over, one of the extraordinary favourable circumstances of the Cox model is that assessing indicator impacts does not rely upon making presumptions about the type of the benchmark risk function, $h_0(t)$, which can be left unspecified. Rather, we require just accepting that whatever the pattern risk function is, covariate impacts multiplicatively move risk function, and these multiplicative movements are steady after some time.

Thus, after model build it is always good practice to assess our model. The next subsection will explain ways to assess our perfect model.

3.3.5. Model Assessment

This subsection explains ways to assess our perfect model in relation to both descriptive power on the data and its prescient power on the autonomous test informational collection.

3.3.5.1. Predictive Performance

To decide its values as a credit scoring framework, the Cox PH demonstrate is verified as an indicator of default. Forecasts are made utilising survival likelihoods for various survival times registered utilising the Cox PH model. Given a cut-off limit, the survival likelihoods are utilised as scores to foresee great or awful cases. That is, if a case has survival likelihood at a year that is more prominent than the cut-off, then it is anticipated as good (i.e. it is anticipated as enduring default), else it is anticipated as a bad case. Forecasts are made with LR comparably utilising a cut-off on PDs registered by the LR display.

3.3.5.2. Descriptive Model

The Cox Proportional Hazard model is estimated as a descriptive model by uncovering its fit to the readiness data with and without macroeconomic variables, utilising log-probability proportion (LLR). The centrality of each coefficient in the model is settled utilising a Wald measurement got from the most probability estimation. The Wald measurement takes after a chi-square measurement, so a p-value can be figured for the invalid hypothesis that the coefficient value is zero (Hosmer and Lemeshow , 1999).

At the point when a covariate x collaborates with at least one covariate $y_1, \dots . y_n$, it is hard to quickly decide the impact of x on the Probability of Default (PD). Be that as it may, it is conceivable to decide the minimal impact on log-risk, $\gamma_x \; of \; x$, restrictive on the collaboration's terms (Brambor et al., 2006) , as

$$\gamma_x = \beta_x + \sum_{i=1}^{n} \beta_{xy_i} y_i \qquad (6)$$

$where \; \beta_x and \; \beta_{xy_i} - coefficient \; estimates \; for \; x$ and each collaboration $xy_i \; for \; i = 1,2,..,n.$

In this research I will report a solitary figure for minor impact by substituting the mean estimations of every collaboration term $y_1, y_2, \dots \_ y_n$ in eq (6), along these lines giving an estimation of minimal impact of x for the mean perception. LR, Cox PH with time-shifting covariates, and Cox PH with macroeconomic factors and distinction between Cox PH with and without macroeconomic components are the rundown earlier desires for the impacts of each macroeconomic factor, and the indication of the watched marginal impact can be tried against this desire. In testing the model, we will necessitate that general coordinate.

As per Bellotti and Crook (2009), the essentialness of a macroeconomic impact might be evaluated by the amount of the unvarying marginal impact – i.e. the absolute value of the marginal impact duplicated by standard deviation of the macroeconomic factors over the period of the information. This will give an expected sign of the overall centrality of each macroeconomic factor in the model.

3.4. Model Performance Comparison

As discussed, choosing a very good model from the other models we must relate the ROCs, AUROC, KS, Gini etc. The Receiver Operating Characteristic graphs the sensitivity versus 1-specificity of the models at several cut-off points. For the fail outcome, sensitivity discusses to an element of financial records that are not paying, that the model properly recognises as they went bad or failed to pay, whereas specificity is a portion of financial records that are paying that the model properly recognises as paid accounts.

3.5. Conclusion

In this section, the theory of survival analysis methods was discussed, such as non-parametric SA approaches, as well as LR. The usage of the parametric SA approach requires an accurate specification of the model before usage. This is often difficult to determine, hence the use of a semi-parametric survival analysis approach – the CPH model methodology. The CPH regression approach is known to be a robust technique which can closely approximate parametric regression estimates. In the existence of challenging risks, the CPH model method is applied for each cause of failure in a cause-specific regression.

CHAPTER 4: DATA AND PRELIMINARY ANALYSIS

4.1. Introduction

In this paper, time until a certain default is very important. This chapter gives further description of data used, variables obtained, data grouping, univariate and multivariate analysis, and stratified random sampling. This chapter covers the variable selection to be used in the final model.

4.2. Description of data

This study explores a dataset obtained in the consumer credit context. The analysis looks at facility-level information, rather than at customer level. This means that if a customer holds more than one account, this study treats each account separately. The dataset will consist of all active accounts between Jan 2017 and Dec 2017 (1-year data). Application and behavioural variables are provided per account in the dataset. Datasets with variables such as income amount, macroeconomic variables, age and credit bureau risk score, will be reserved at the period of request. For this purpose, an account will be taken as having not being paid if it goes three months down or more within the first 12 months. An account that is not being paid is declared as a bad account, and a payment is mentioned as a good account. The repayment status is given per account per month under observation. A fixed workout/outcome period will be determined and used in the calculation of forward-looking probabilities. A workout period is the amount of time it takes for the bulk of accounts to be absorbed into the events of interest.

4.3. Variables

The following variables were chosen as those expected to affect non-payment of mortgage loans, including macroeconomic variables, and were made available for the dataset used:

1. Client ID
2. Client age
3. Client income
4. Client number of children
5. Client race
6. Client’s occupation
7. High education level
8. Loan type
9. Bond amount
10. Credit Risk Score – a credit score, between 301 and 850, that indicates the creditworthiness of the borrower will timely repay future obligations.
11. Date defaulted

12. Outstanding terms – the period it would take to recompense off the remaining of repayment loan as planned
13. Gender
14. Date of reimbursement
15. Debit Interest Rate – the existing debit interest rate on the mortgage credit, taking into consideration any loan fluctuations
16. Loan-to-Value ratio – is a borrowing risk valuation ratio that monetary organisations and other creditors observe before approving a home loan.

4.4. Good- Bad and Univariate Analysis

Univariate Analysis explores variables (attributes) one by one. Variables could be either categorical or numerical. There are different statistical and visualization techniques of investigation for each type of variable. Numerical variables can be transformed into categorical counterparts by a process called binning or discretization. It is also possible to transform a categorical variable into its numerical counterpart by a process called encoding.

4.4.1. Analyse Good-Bad (0-good, 1-bad)

In this data, default is our response/target column where one is declared as bad/event and zero declared as good/non-event. Table 4.1 and Figure 4.1, below, illustrate default and non-defaulters where we see ~35% default (#6,996) and non-defaulters are ~65% (#13,051).

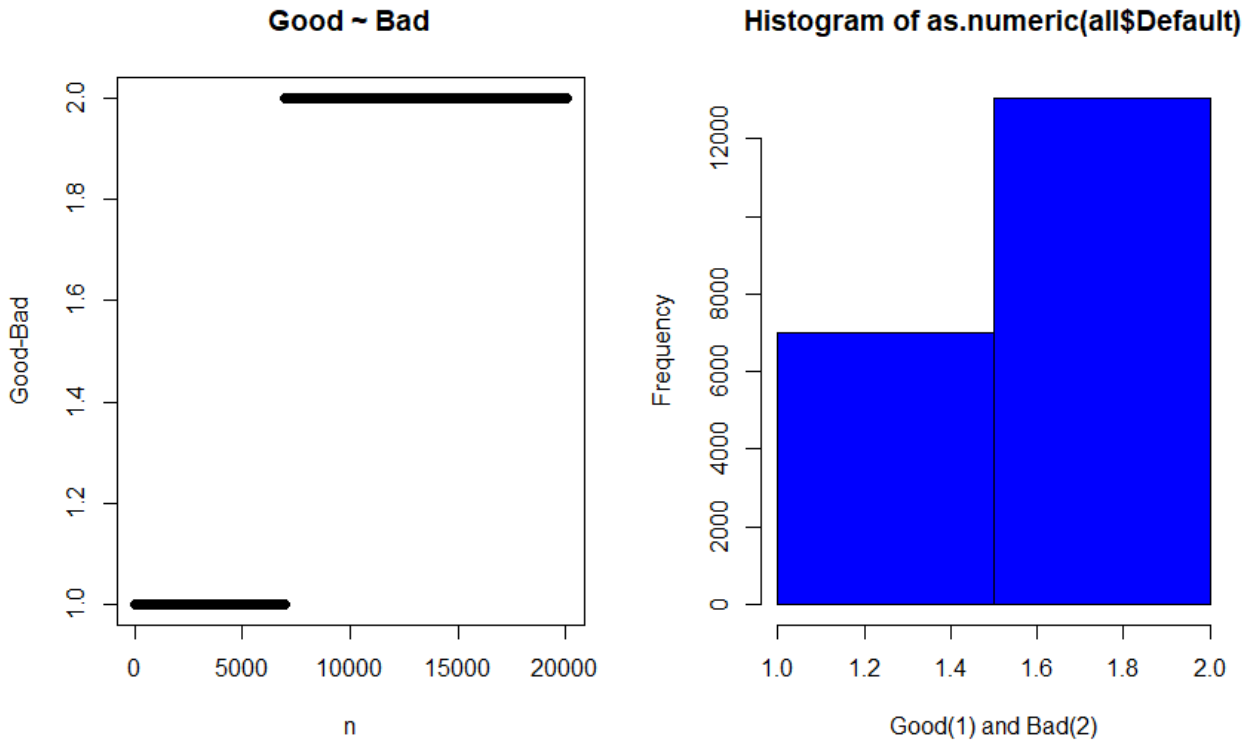


Figure 4. 1: Descriptive analysis of mortgage loans

	Count	Percentage
Bad	6996	34.9
Good	13051	65.1

Table 4. 1.: Frequency table for Good-Bad status

4.4.2. Univariate and bivariate analysis

Bivariate analysis is the simultaneous analysis of two variables (attributes). It explores the concept of relationship between two variables, whether there exists an association and the strength of this association, or whether there are differences between two variables and the significance of these differences. Bivariate analyses are directed to determine non-linear effects of the different risk drivers on the dependent variable, most notably prepayment rates. In this way, it can be assessed whether risk drivers can directly be included in the models, or should be included with a transformation. The figures for the continuous variables are constructed by combining observations in bins. The bin size determines the degree of smoothing visible in the figures. Decreasing the number of bins can aid in revealing a trend.

4.4.2.0. Weight of Evidence (WOE), Efficiency, and Information Value (IV)

WOE:

Woe shows predictive power of an independent variable in relation to a dependent variable. It evolved with credit scoring to magnify separation power between a good customer and a bad customer. Here it is one of the measures of separation between two classes (good/bad, yes/no, 0/1, response/no-response, default/no default). It is defined as:

$$WOE = In\left(\frac{Distribution\ of\ Non-Events(Good)}{Distribution\ of\ Events\ (Bad)}\right)$$

It is computed from the basic odds ratio:

$(Distribution\ of\ Good\ Credit\ Outcomes)/(Distribution\ of\ Bad\ Credit\ Outcomes)$

Information Value (IV)

IV helps to select variables by using their order of importance to information value after grouping.

$$IV = \sum(\%Non-Events - \%Events) * WoE$$

Efficiency:

$$Efficiency = \frac{Abs(\%Non\ Events - Events)}{2}$$

Siddiqi (2006) recommends that variables with tremendously more information value offer suspicion. He delivers the subsequent rule of thumb for forthcoming variables according to their Information Value:

$less\ than\ 0.02 - unpredictive$
$between\ 0.02\ and\ 0.1 - weak$
$between\ 0.1\ and\ 0.3 - medium$
$between\ 0.3\ and\ 0.5 - strong$
$greater\ than\ 0.5 - suspicious$

I. The following variables do not have prediction power – very weak predictor (IV < 2%), therefore we exclude them from modelling:

Position	Variable	IV	IV Rank
1	Mortgage Interest rate	1.93	39
8	Client number of children	1.59	40
11	Marital status id	1.24	41
30	Marital Status	1.24	42
43	Marital indicator	0.56	43
22	Current loan term	0.17	44
37	Prime rate level	0.15	45
7	Cust sex id	0	46
31	Sex	0	47
42	Gender	0	48

II. The following variables are very weak predictors (2%<=IV< 10%), hence we may or may not include them while modelling:

Position	Variable	IV	IV Rank
41	Income estimate monthly	9.87	21
28	Income estimate annually	9.87	22
18	Mortgage interest rate	9.86	23
34	Loan to value	9.42	24
2	Income band	8.37	25
5	Client age	6.91	26
26	Account balance	6.91	27
20	Out loan balance amount	6.89	28
40	Age	6.56	29
45	Age band	5.89	30
15	Loan type	5.58	31
3	Client ID	4.24	32
35	Year	3.98	33
38	HPI Growth YoY perc	3.96	34
36	Inflation Growth YoY	3.88	35
33	newdate_num	2.16	36

39	year_numeric	2.16	37
44	time	2.16	38

III. The following variables have medium prediction power (10%<=IV< 30%), hence we will include them in modelling as we have less number of variables:

Position	Variable	IV	IV Rank
25	Previous amount paid	27.14	5
32	Term paid	25.24	6
4	Account ID	23.92	7
24	Remaining term	22.41	8
48	Survival time	22.41	9
19	Bond amount	18.88	10
9	Race id	17.56	11
29	Race	17.56	12
6	Income amount	17.29	13
16	Property Type	15.44	14
12	Highest education level	14.19	15
46	UNKEY	13.22	16
47	Education level	13.18	17
17	Purchase Price	11.06	18
10	Customer Occupation id	10.3	19
23	Monthly repayment value	10.03	20

IV. The variable below has a strong predictor with IV between 30% and 50%

Position	Variable	IV	IV Rank
13	Status	39.94	4

V. The following variables have very high prediction power (IV > 50%); it could be suspicious and require further investigation.

Position	Variable	IV	IV Rank
21	Historical amount due	309.66	1
14	Sub-status	152.39	2
27	Client bureau score	101.42	3

Table 4. 2**.:** Tables of Information Value for all variables

These tables offer more direct interpretation for relating IV across all variables. Variables of the same IV will be tied in IV Rank.

Variables of related landscapes usually perform similarly, and do have very high correlation with one another. The flow table of Information Value displays how these variables cluster together. In the above table, *Remaining Term* offers a very similar Information Value as *Term Paid*. As our model favours variables of lower necessity from one another, only one variable from each cluster needs to be chosen to enter the regression.

In the table above, historical amount due has the highest information value of 309.66%, which means customers have skipped payment recently, and are more likely to default. Moreover, client credit bureau score has an IV of 101.42%, which also means that clients with a lower credit risk score are much more likely to default. The next subsections will be illustrating the *information value* graphically.

4.4.2.1. Checking historical amount due

Historical amount due is the amount of arrears accumulated to the specific loan amount account, due to missed payments. The two figures below show how frequency of customers who defaulted is associated with account:

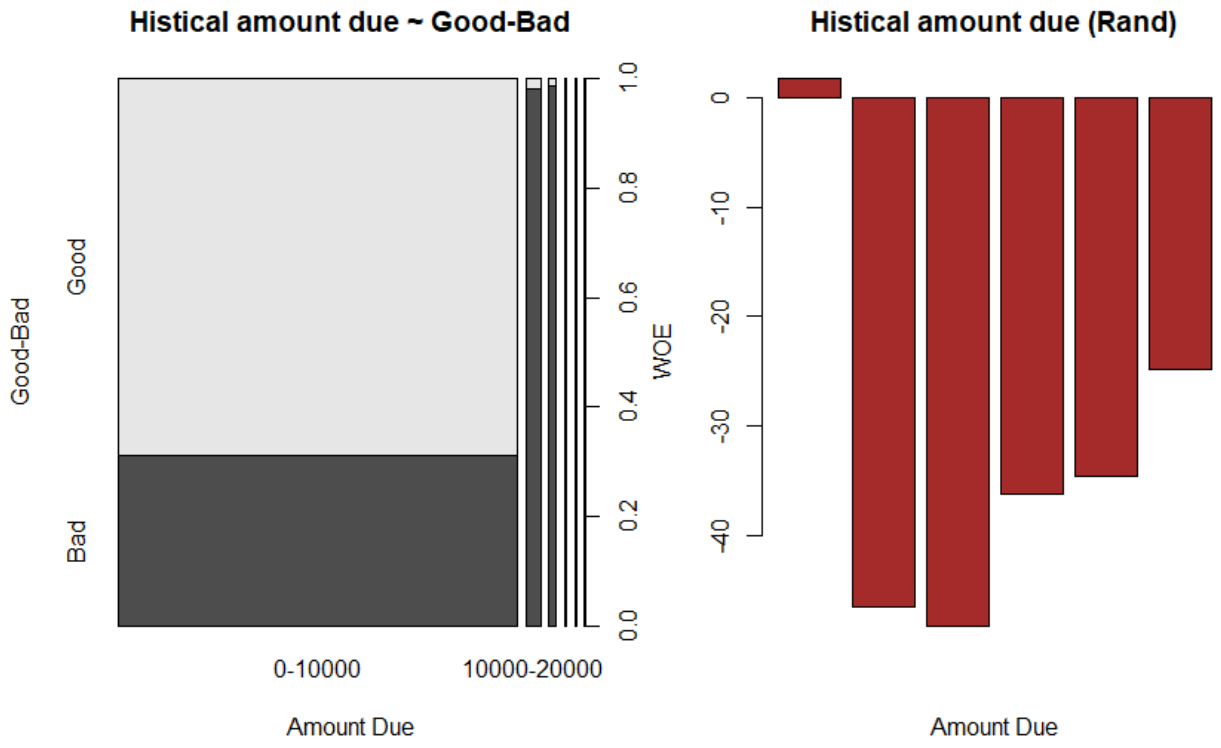


Figure 4. 2.**: (Left)** Account Distribution and default, and **(Right)** WoE for Each Account

The graph on the left shows a distribution of accounts by historical amount due and the default rate. The graph on the right shows the Weight of Evidence for each account. The following table is the numerical representation of good-bad accounts:

Names	1	0	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
0-10000	5846	13025	83.56	99.8	18871	94.13	69.02	4.56	-1.78	2.89072	8.12
10000-20000	659	12	9.42	0.09	671	3.35	1.79	9.91	46.51	43.39383	4.665
100000+	17	2	0.24	0.02	19	0.09	10.53	9.23	24.85	0.5467	0.11
20000-40000	350	5	5	0.04	355	1.77	1.41	9.92	48.28	23.94688	2.48
40000-60000	79	4	1.13	0.03	83	0.41	4.82	9.74	36.29	3.9919	0.55
60000-100000	45	3	0.64	0.02	48	0.24	6.25	9.7	34.66	2.14892	0.31

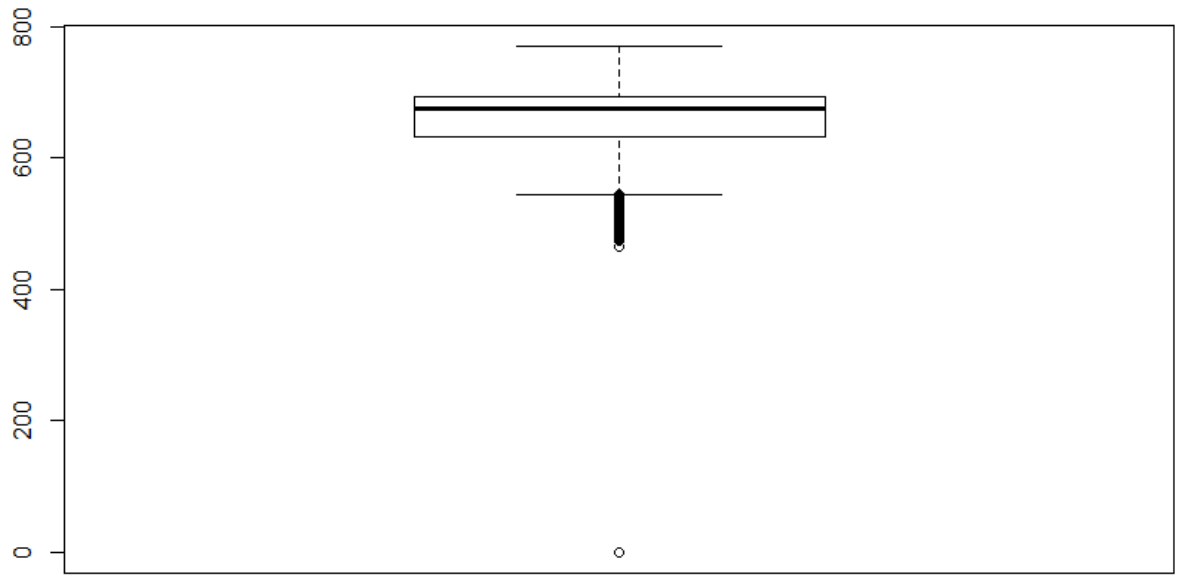
Table 4. 3.**:** Weight of Evidence ~ Checking Account Historical Amount Due ~ Good-Bad

Information Value is 76.92 and Efficiency is 16,235.

Table 4.3 is self-explanatory. Historical overdue amount is a key factor in the bank's monitoring process. It has high information value, which all means that the default rate is high when due amount for customers is high, as shown in Figure 4.2.

4.4.2.2. Client Bureau Score

The following two graphs shows the bivariate analyses for the variable behavioural risk score on number of customers (red) and percentage of customers defaulted (blue).



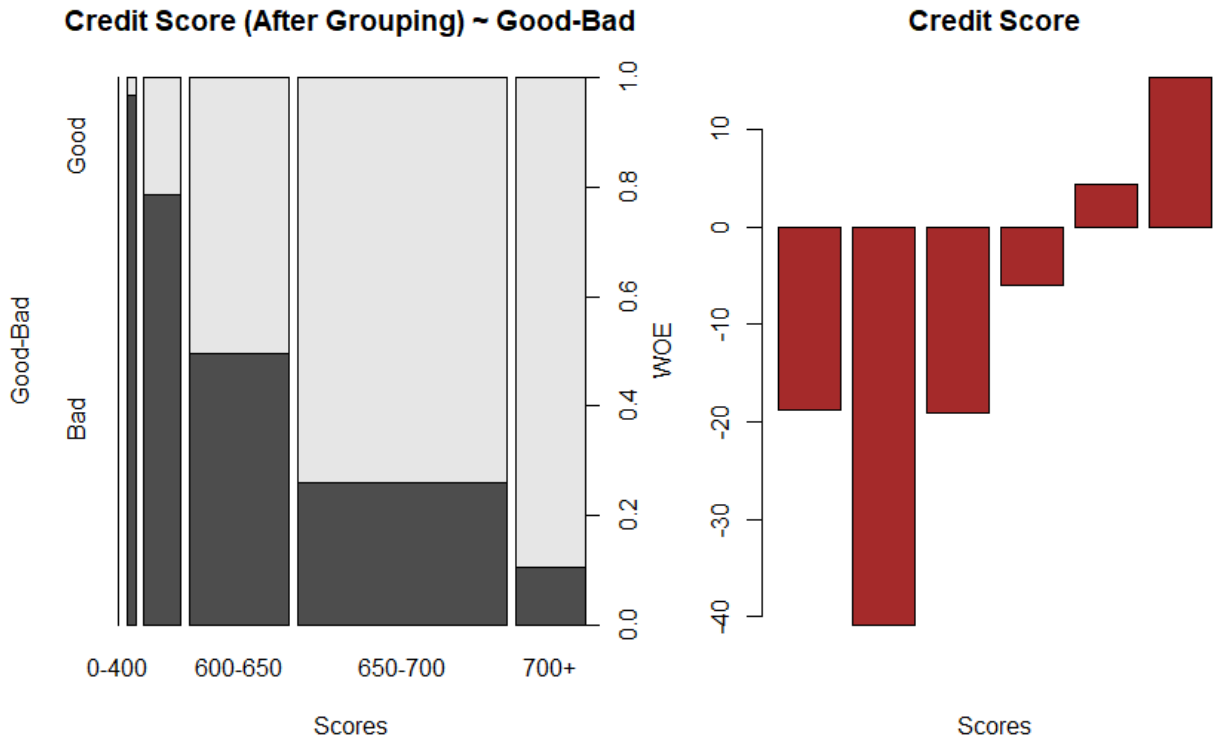


Figure 4. 3.**: (Left)** Risk score Distribution and default rate, and **(Right)** WoE for Each Risk Score

The following table shows numerical representation of good-bad accounts binned by client bureau risk score:

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
0-400	3	9	0.02	0.13	12	0.06	75	1.33	-18.72	0.20592	0.055
400-550	12	373	0.09	5.33	385	1.92	96.88	0.17	-40.81	21.38444	2.62
550-600	377	1375	2.89	19.65	1752	8.74	78.48	1.28	-19.17	32.12892	8.38
600-650	2387	2342	18.29	33.48	4729	23.59	49.52	3.53	-6.05	9.18995	7.595
650-700	7326	2555	56.13	36.52	9881	49.29	25.86	6.06	4.3	8.4323	9.805
700+	2946	342	22.57	4.89	3288	16.4	10.4	8.22	15.29	27.03272	8.84

Table 4. 4**.:** Weight of Evidence ~ Checking Account Client Bureau Score ~ Good-Bad

Information Value is 98.37 and Efficiency is 37.30.

As was discussed in Chapter 2, credit scoring determines your creditworthiness. A higher bureau risk score is associated with more creditworthiness. Therefore, it is expected that this score is inversely associated with the default rate. This is confirmed in Table 4.4. The bureau scores between 400 and 550 have the highest default rate of 96.88%. The default rate decreased significantly as associated with bureau risk sore, thus the higher bureau scores, the lower the default rates. This is due to not paying on time, skipping payments, or paying your credit card late – which can negatively impact your credit score. Paying your bills on time is a key way to improve your credit score, and it will automatically reduce default rate of loans significantly, as shown in the above graph.

4.4.2.3. Previous Amount Paid

Figure 4.4 shows the bivariate analyses for the previous amount paid variable on number of customers (red) and default rates (blue).

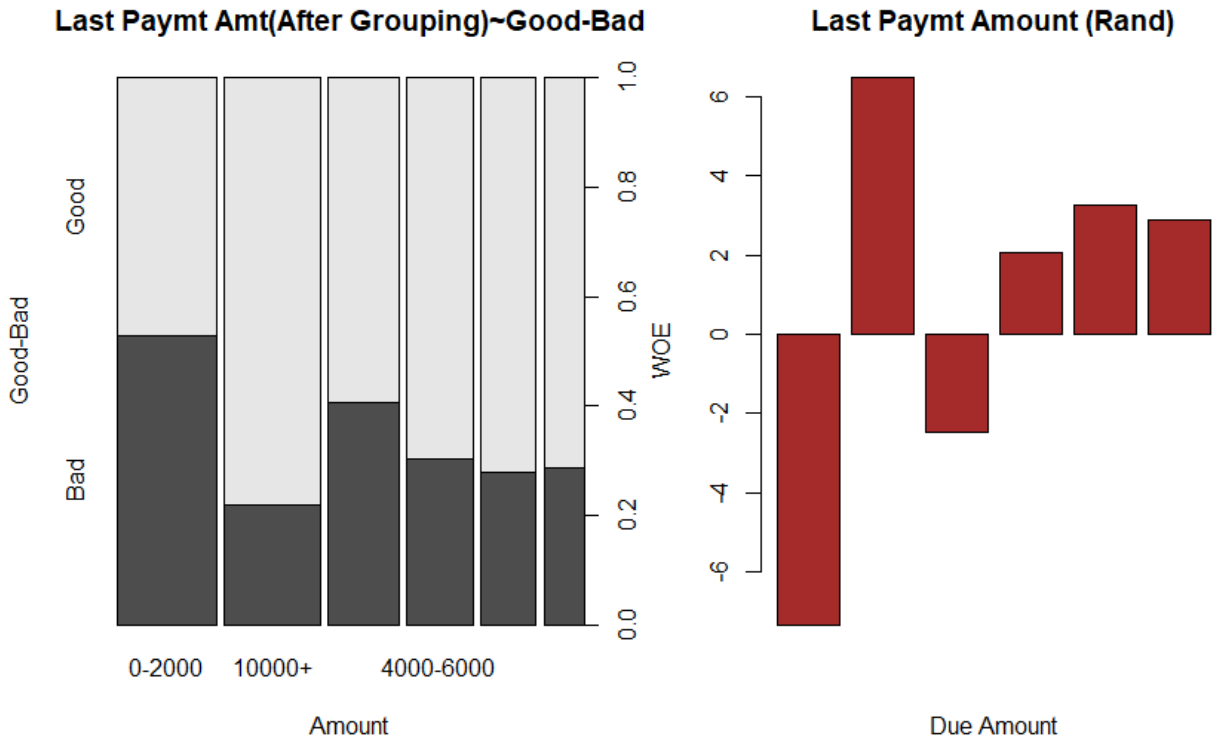


Figure 4. 4.: Previous Amount Paid Distribution and default rate, and WoE for Each Previous Paid

The following table shows numerical representation of good-bad accounts binned by last payment amount:

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
0-2000	2205	2460	16.9	35.16	4665	23.27	52.73	3.25	-7.33	13.38458	9.13
10000+	3518	987	26.96	14.11	4505	22.47	21.91	6.56	6.47	8.31395	6.425
2000-4000	1962	1345	15.03	19.23	3307	16.5	40.67	4.39	-2.46	1.0332	2.1
4000-6000	2162	944	16.57	13.49	3106	15.49	30.39	5.51	2.06	0.63448	1.54
6000-8000	1872	724	14.34	10.35	2596	12.95	27.89	5.81	3.26	1.30074	1.995
8000-10000	1332	536	10.21	7.66	1868	9.32	28.69	5.71	2.87	0.73185	1.275

Table 4. 5.: Weight of Evidence ~ Checking Account Previous amount paid ~ Good-Bad

Information Value is 25.40 and Efficiency is 22.47.

In Table 4.5, previous amount paid has information value of 25.40, which means customers have struggled to make payment recently, and will be likely to miss payments, or go bad. The graph in Figure 4.4 shows trends of defaults by previous amount paid. A higher previous amount paid is associated with a lower risk involved in ability to make payments. Therefore, its relationship with the default rate is expected to decrease significantly. A higher previous amount paid also means that customers have fewer resources to make unscheduled excess payments. In Figure 4.4, this is clear

from the decreasing trend. Since this effect is linear for most of the observations, the previous amount paid is included with no changes.

4.4.2.4. Term Paid

Term paid is traced monthly, and differs from 1 month to 240 months. Figure 4.5 (Left) shows the seasoning of the mortgages for the full sample of number of customers (red) and defaults (blue). Total number of customers is at the highest level for paid-off periods about eight (8) years, after which they remain stable to peak again at term paid of 20 years. Default rates increase steadily during the first few years after loan origination. After a small decrease in three (3) years in defaults, it rises again significantly to reach its peak after nine (9) years and decreased until term paid of 12 years, and rise again towards end of the loan.

The following two graphs show term paid trends with the default rate:

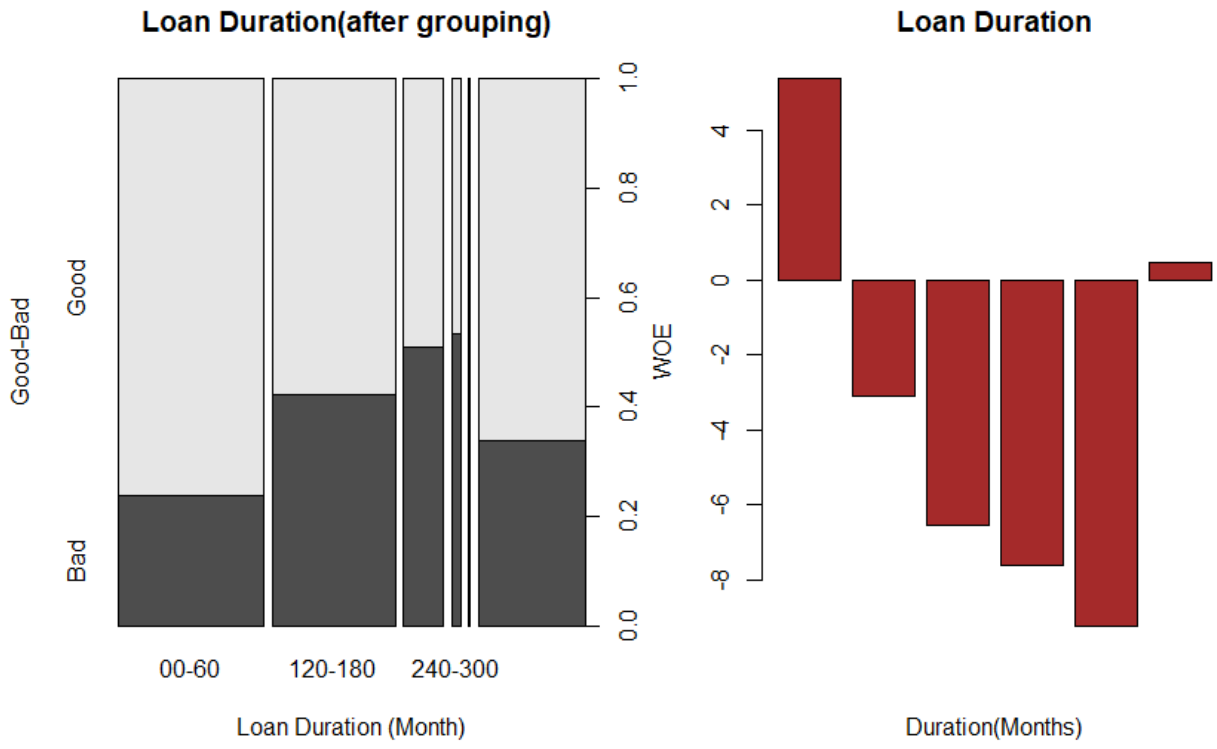


Figure 4. 5**.: (Left)**Term paid Distribution and default rate, and **(Right)** WoE for term paid

The following table shows numerical presentation of good-bad accounts binned by term paid:

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
00-60	5247	1642	40.2	23.47	6889	34.36	23.84	6.31	5.38	9.00074	8.365
120-180	3338	2438	25.58	34.85	5776	28.81	42.21	4.23	-3.09	2.86443	4.635
180-240	925	956	7.09	13.66	1881	9.38	50.82	3.42	-6.56	4.30992	3.285
240-300	186	214	1.43	3.06	400	2	53.5	3.18	-7.61	1.24043	0.815
300+	38	51	0.29	0.73	89	0.44	57.3	2.84	-9.23	0.40612	0.22
60-120	3317	1695	25.42	24.23	5012	25	33.82	5.12	0.48	0.05712	0.595

Table 4. 6**.:** Weight of Evidence ~ Checking Account Term paid of loan Good-Bad

Information Value is 17.88 and Efficiency is 17.92.

Since the effect of term paid on repayment rate is not linear over the terms of the mortgage, this variable is included as categorical variable by buckets. The bucket size is determined by confirming the occurrence of a similar amount of observations in each bucket.

4.4.2.5. Remaining Term of a loan

Remaining term is tracked each month, and a borrower can remain with months between 1 and 240 months. Figure 4.6 on the left and Figure 4.6 on the right show the bivariate analyses for the remaining terms on number of customers (red) and percentage of customers defaulted (blue).

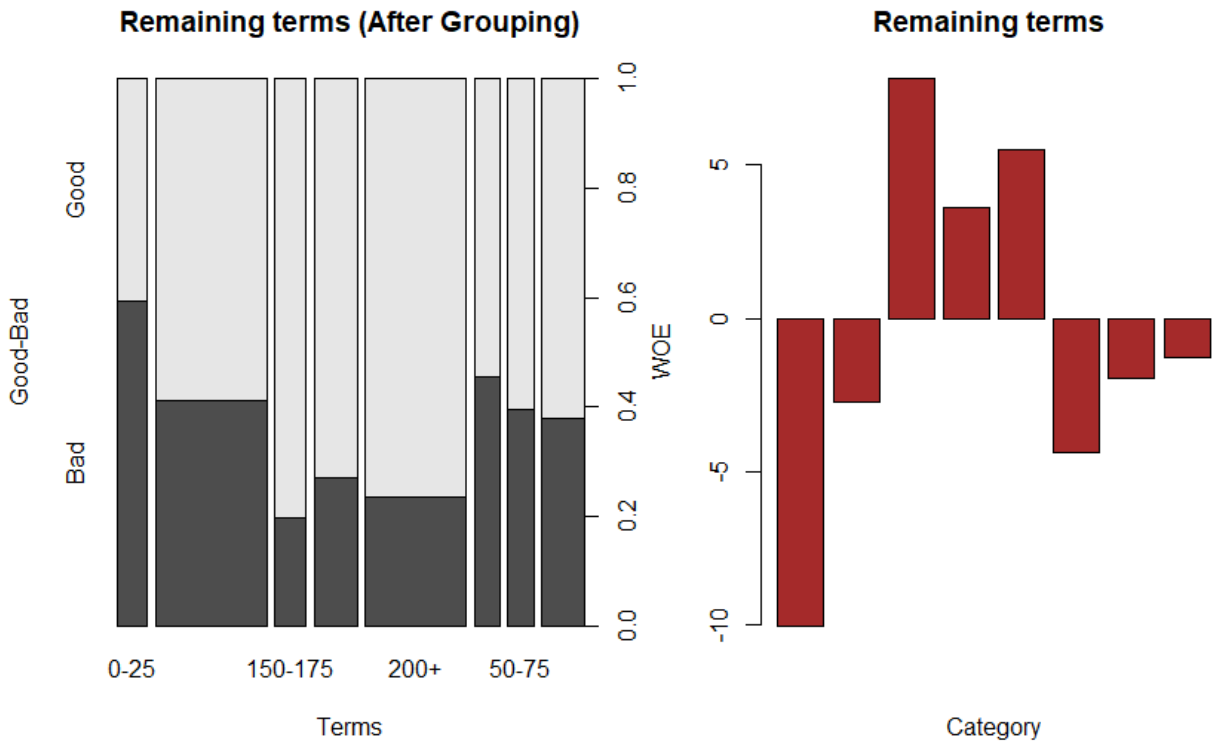


Figure 4. 6**: (Left)** Remain term Distribution and default rate, and **(Right)** WoE for Each Remain term

The following table shows numerical presentation of good-bad accounts binned by remaining term:

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
0-25	599	874	4.59	12.49	1473	7.35	59.33	2.69	-10.01	7.9079	3.95
100-150	3193	2243	24.47	32.06	5436	27.12	41.26	4.33	-2.7	2.0493	3.795
150-175	1213	297	9.29	4.25	1510	7.53	19.67	6.86	7.82	3.94128	2.52
175-200	1543	575	11.82	8.22	2118	10.57	27.15	5.9	3.63	1.3068	1.8
200+	3780	1166	28.96	16.67	4946	24.67	23.57	6.35	5.52	6.78408	6.145
25-50	668	555	5.12	7.93	1223	6.1	45.38	3.92	-4.37	1.22797	1.405
50-75	761	497	5.83	7.1	1258	6.28	39.51	4.51	-1.97	0.25019	0.635
75-100	1294	789	9.91	11.28	2083	10.39	37.88	4.68	-1.29	0.17673	0.685

Table 4. 7**.:** Checking Account Remaining term of loan Good-Bad

Information Value is 23.64 and Efficiency is 20.94.

Default rates decreased significantly from 14 months to 92 months, and then rise again to 44.4%. The increase in default rate may be due to borrowers settling mortgage loan accounts or skipping payments, same applies to the period of 17 years (default rate increased to ~28%). This is clear in Figure 4.6 by the decreasing trend in the blue line. Since this effect is linear for most of the observations, the remaining term is included with no changes.

4.4.2.6. Bond Amount

This only applies to mortgage loans (MLS): the amount of the bond which is registered and which covers the amount of the original loan. Figure 4.7 illustrates the bivariate analyses of the bond amount:

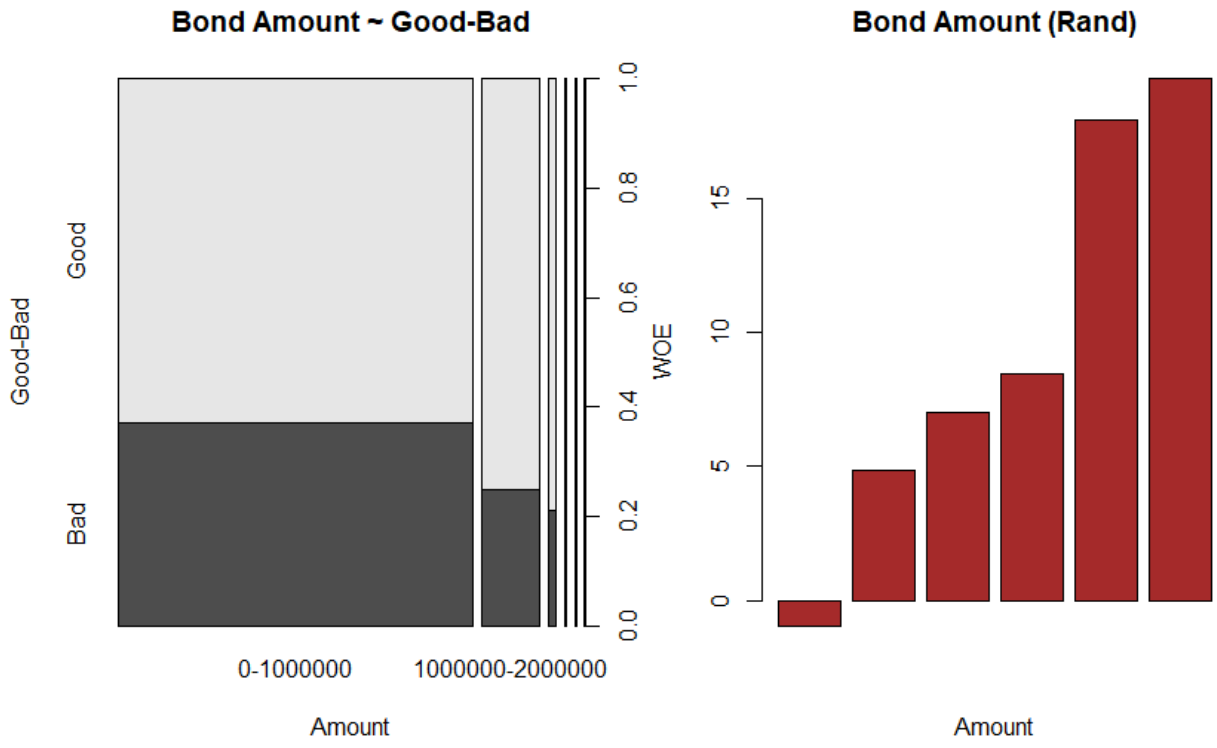


Figure 4. 7**:** Bond amount Distribution and default rate, and WoE for Each bond amount

The following table shows numerical presentation of good-bad accounts binned by bond amount:

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
0-1000000	10553	6212	80.86	88.79	16765	83.63	37.05	4.77	-0.94	0.74542	3.965
1000000-2000000	2076	685	15.91	9.79	2761	13.77	24.81	6.19	4.86	2.97432	3.06
2000000-3000000	290	77	2.22	1.1	367	1.83	20.98	6.69	7.02	0.78624	0.56
3000000-4000000	73	17	0.56	0.24	90	0.45	18.89	7	8.47	0.27104	0.16
4000000-5000000	31	3	0.24	0.04	34	0.17	8.82	8.57	17.92	0.3584	0.1
5000000+	28	2	0.21	0.03	30	0.15	6.67	8.75	19.46	0.35028	0.09

Table 4. 8**.:** Weight of Evidence ~ Checking Account bond amount Good-Bad

Information Value is 5.49 and Efficiency is 7.94.

The above table shows that the customers with bond amount less than one million rand are more likely to default. This is due to customers with or without Grade 12 education who are increasingly defaulting due to losing their respective jobs. In Table 4.8, on average, default rates are lower when bond amounts have increased, apart from the sharp peak between 0 and 1 million (and defaults).

4.4.2.7. Highest Education Level

Figure 4.8 illustrates the bivariate analyses of the highest education level:

Note: Educational level = 1 is Doctor and MBA, 2 is Masters, 3 is Honours, 4 is graduate/Bachelors, 5 is diploma, 6 is Grade 12, and 7 is without any formal education.

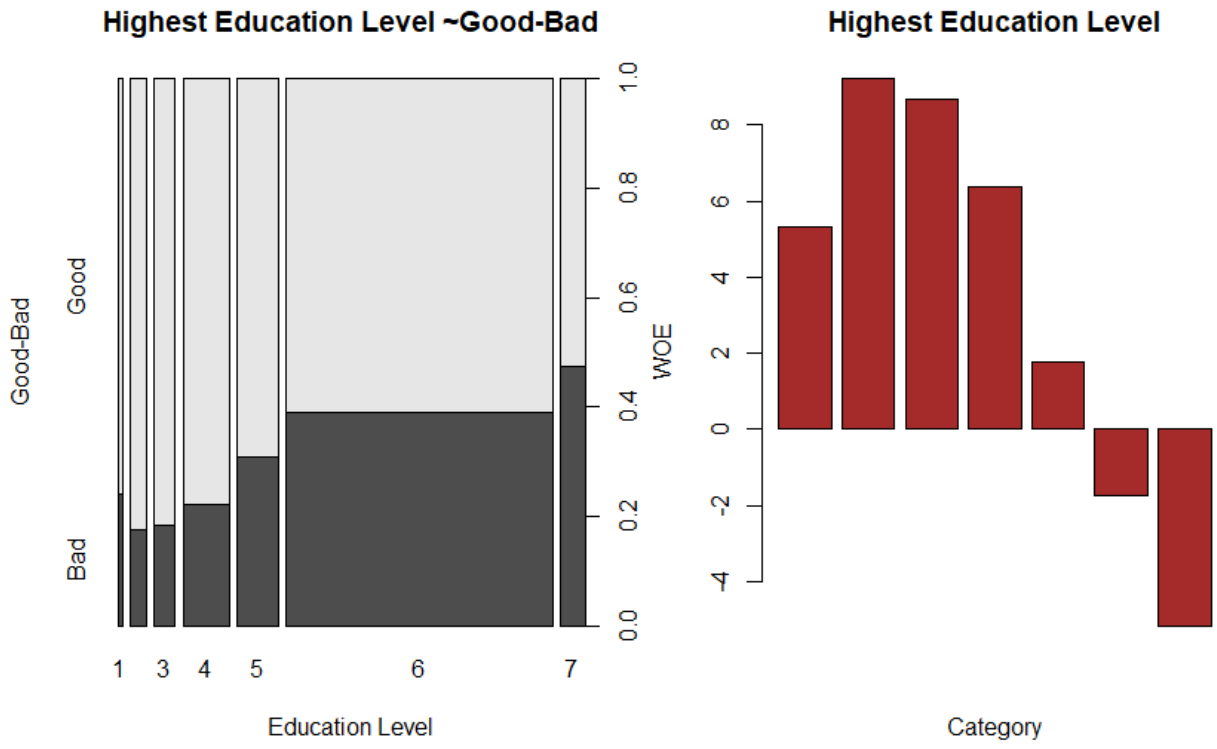


Figure 4. 8.: Education level Distribution and default rate, and WoE for Each Education level

The following table gives a numerical presentation of good-bad accounts binned by highest education level:

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
1	126	40	0.97	0.57	166	0.83	24.1	6.3	5.32	0.2128	0.2
2	613	131	4.7	1.87	744	3.71	17.61	7.15	9.22	2.60926	1.415
3	814	183	6.24	2.62	997	4.97	18.36	7.04	8.68	3.14216	1.81
4	1730	490	13.26	7	2220	11.07	22.07	6.54	6.39	4.00014	3.13
5	1342	602	10.28	8.6	1944	9.7	30.97	5.44	1.78	0.29904	0.84
6	7804	4991	59.8	71.34	12795	63.83	39.01	4.56	-1.76	2.03104	5.77
7	622	559	4.77	7.99	1181	5.89	47.33	3.74	-5.16	1.66152	1.61

Table 4. 9.**:** Weight of Evidence ~ Checking Education Level Good-Bad

Information Value is 13.96 and Efficiency is 14.78.

Figure 4.8 shows that number of borrowers and default rate differ per education level. In education 7, without formal education, default rates are higher (with 47.3%), and number of borrowers are lower compared to the rest of the bank mortgage loan customers. Number of borrowers are the highest in Grade 12 (education level 6), with the second highest default rates (~39%).

4.4.2.8. Purchase Price

Purchase price is the price that has been agreed upon by owner and customer for the transferring of an asset. Figure 4.9 shows the bivariate analyses of the purchase price:

Figure 4. 9.: Purchase Price Distribution and default, and WoE for Purchase Price

The following table gives a numerical presentation of good-bad accounts binned by purchase price:

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
0-500k	6527	4529	50.01	64.74	11056	55.15	40.96	4.36	-2.58	3.80034	7.365
1000k-2000k	2071	636	15.87	9.09	2707	13.5	23.49	6.36	5.57	3.77646	3.39
2000k-3000k	297	85	2.28	1.21	382	1.91	22.25	6.53	6.34	0.67838	0.535
3000k-4000k	93	13	0.71	0.19	106	0.53	12.26	7.89	13.18	0.68536	0.26
4000k+	40	6	0.31	0.09	46	0.23	13.04	7.75	12.37	0.27214	0.11
500k-1000k	4023	1727	30.83	24.69	5750	28.68	30.03	5.55	2.22	1.36308	3.07

Table 4. 10.: Weight of Evidence ~ Checking Purchase Price Good-Bad

Information Value is 10.58 and Efficiency is 14.73.

Purchase price is a scaled variable indicating the size of the loan, relative to the average loan size in the sample. The number of borrowers are higher for lesser loans. For mortgages with a below average size, the number of borrowers are flat, and default rates are decreasing significantly while purchase price is rising.

4.4.2.9. Mortgage Interest Rate

The following histogram shows distribution of interest rate by population:

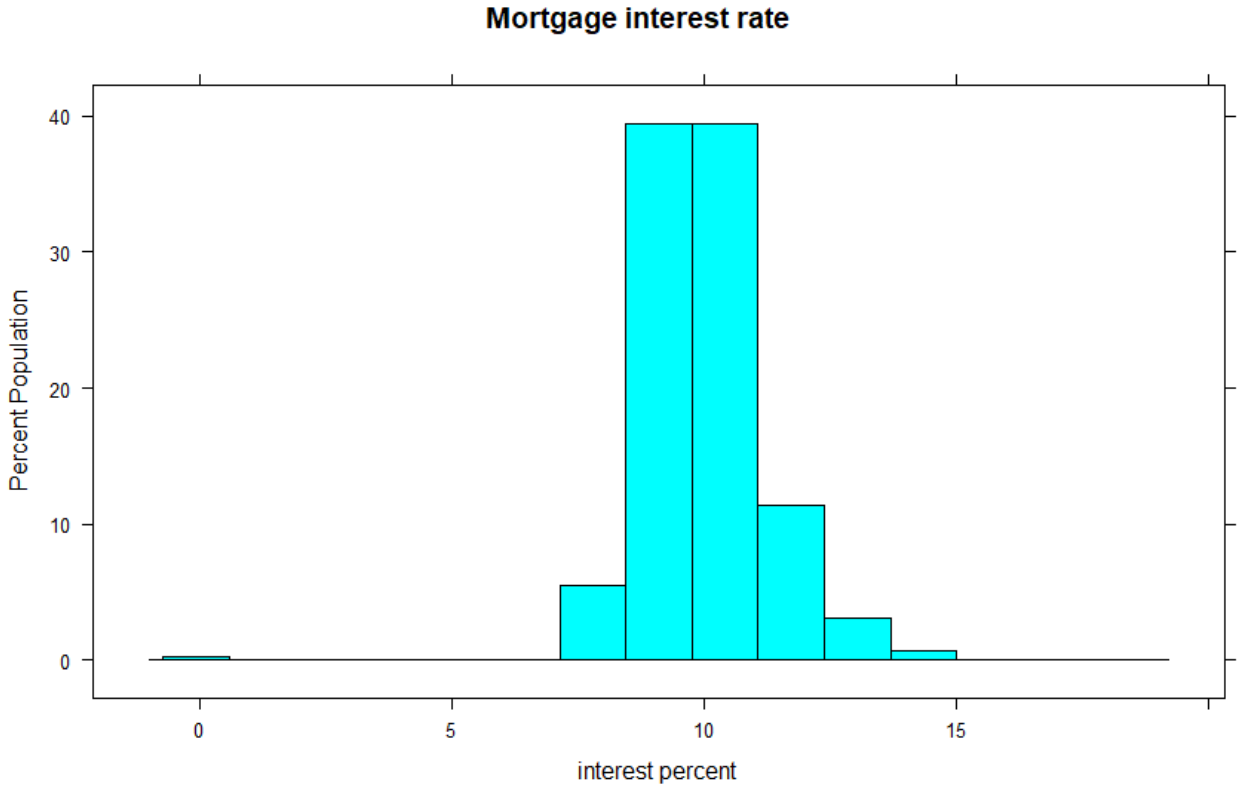


Figure 4.10 illustrates the bivariate analyses of the debit interest rate:

Figure 4. 10.: Mortgage interest rate Distribution and default rate, and WoE for Mortgage interest rate

The following table gives a numerical presentation of good-bad accounts binned by debit interest rate:

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
0-9	3530	1462	27.05	20.9	4992	24.9	29.29	5.64	2.58	1.5867	3.075
10 - 11	3816	2386	29.24	34.11	6202	30.94	38.47	4.62	-1.54	0.74998	2.435
11 – 12	1310	788	10.04	11.26	2098	10.47	37.56	4.71	-1.15	0.1403	0.61
12 – 13	370	374	2.84	5.35	744	3.71	50.27	3.47	-6.33	1.58883	1.255
13+	178	204	1.36	2.92	382	1.91	53.4	3.18	-7.64	1.19184	0.78
9 – 10	3847	1782	29.48	25.47	5629	28.08	31.66	5.36	1.46	0.58546	2.005

Table 4. 11.: Weight of Evidence ~ Checking Mortgage interest rate Good-Bad

Information Value is 5.84 and Efficiency is 10.16.

Table above shows that customers with higher interest rate are more likely to default. The customers with 12.2% interest rate has the highest default rate of 50.27% with the third smallest number of borrowers.

4.4.2.10. Loan to Value Ratio

It is defined as a loaning risk valuation ratio that financial administrations and other creditors inspect before offering a home. Figure 4.11 (Left) and 4.11 (Right) illustrate the bivariate analyses of the loan-to-value ratio (LTV).

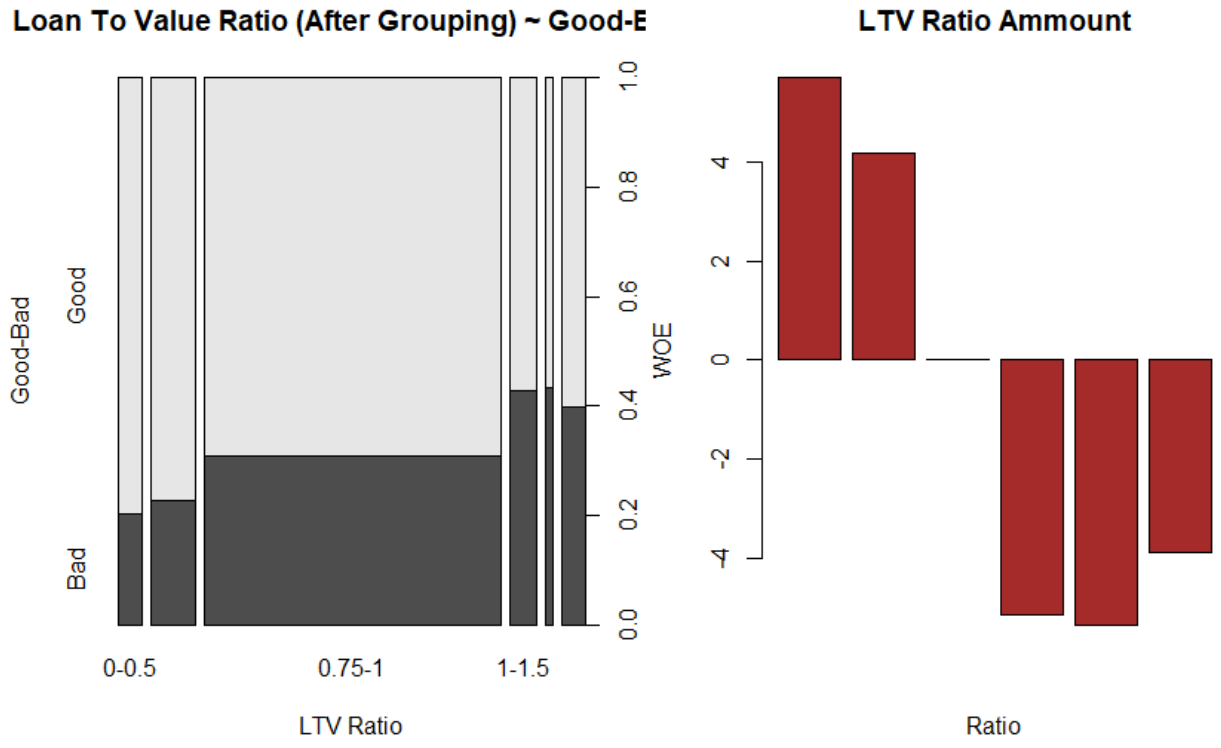


Figure 4.11.: **(Left)** Bivariate analyses of Loan to value Distribution and default rate, and **(Right)** WoE for Each Loan to value

The following table gives a numerical presentation of good-bad accounts binned by debit interest rate:

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
0-0.5	561	142	6.32	3.57	703	5.47	20.2	6.39	5.71	1.57025	1.375
0.5-0.75	1064	314	11.99	7.89	1378	10.72	22.79	6.03	4.18	1.7138	2.05
0.75-1	6224	2789	70.15	70.11	9013	70.14	30.94	5	0.01	0.00004	0.02
1-1.25	469	352	5.29	8.85	821	6.39	42.87	3.74	-5.15	1.8334	1.78
1.25-1.5	133	102	1.5	2.56	235	1.83	43.4	3.69	-5.35	0.5671	0.53
1.5+	421	279	4.75	7.01	700	5.45	39.86	4.04	-3.89	0.87914	1.13

Table 4. 12.: Weight of Evidence ~ Checking Loan to value ratio Good-Bad

Valuations with high loan to value ratios are generally seen as higher risk, and therefore, if the mortgage loan is approved, the loan normally costs the applicant more to borrow. A loan with high loan to value ratio needs the applicant to buy mortgage insurance to counterbalance the risk to the creditor. The result in Table 4.12 proves that the higher the LTV ratio, the higher the risk of customers defaulting.

4.4.2.11. Monthly repayment value

The static monthly payment for a static rate mortgage loan is the sum paid by the applicant every month, that certifies that the loan is paid off in full of interest at the end of its term. Figure 4.12 (Left) and 4.12 (Right) illustrate the bivariate analyses of the monthly repayment amount.

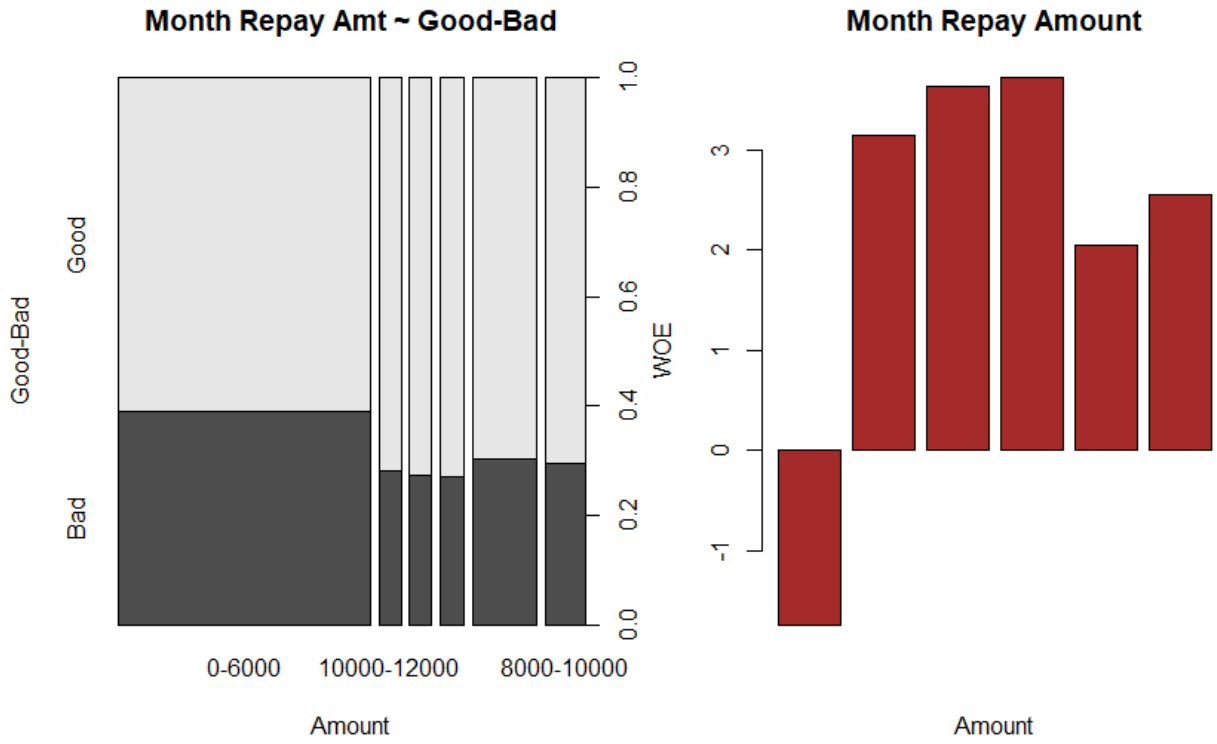


Figure 4.12**.: (Left)** Repay Amount Distribution and default rate, and **(Right)** WoE for Each Repay amount

The following table gives a numerical presentation of good-bad accounts binned by repayment amount on the monthly basis:

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
0-6000	7276	4641	55.75	66.34	11917	59.45	38.94	4.57	-1.74	1.84266	5.295
10000-12000	750	294	5.75	4.2	1044	5.21	28.16	5.78	3.14	0.4867	0.775
12000-15000	737	275	5.65	3.93	1012	5.05	27.17	5.9	3.63	0.62436	0.86
15000+	861	318	6.6	4.55	1179	5.88	26.97	5.92	3.72	0.7626	1.025
6000-8000	2103	918	16.11	13.12	3021	15.07	30.39	5.51	2.05	0.61295	1.495
8000-10000	1324	550	10.14	7.86	1874	9.35	29.35	5.63	2.55	0.5814	1.14

Table 4. 13.: Weight of Evidence ~ checking monthly repayment account Good-Bad

Information Value is 4.91 and Efficiency is 10.59.

The default rate increased significantly, due to missing monthly payments, as can be seen in table 13.

4.5. Multivariate Analysis

Multivariate analysis is utilised to study more complex sets of information than what univariate analysis strategies can handle. This type of analysis is almost always achieved with software (i.e. SPSS or SAS or Rstudio), as working with even the least of datasets can be devastating by hand.

4.5.1. Correlation Analysis

Correlation Analysis:

Variables satisfying all the univariate assessment tests were thought about for variable analysis. The figures below summarise results from the variable assessment on the idea of the subsequent criteria.

Proportions of Correlation	Interpretation
0.90 to 1.00 (−0.90 to −1.00)	Very high positive (negative) correlation
0.70 to 0.90 (−0.70 to −0.90)	High positive (negative) correlation
0.50 to 0.70 (−0.50 to −0.70)	Moderate positive (negative) correlation
0.30 to 0.50 (−0.30 to −0.50)	Low positive (negative) correlation
0.00 to 0.30 (0.00 to −0.30)	negligible correlation

The rule of thumb for interpreting the size of a correlation coefficient

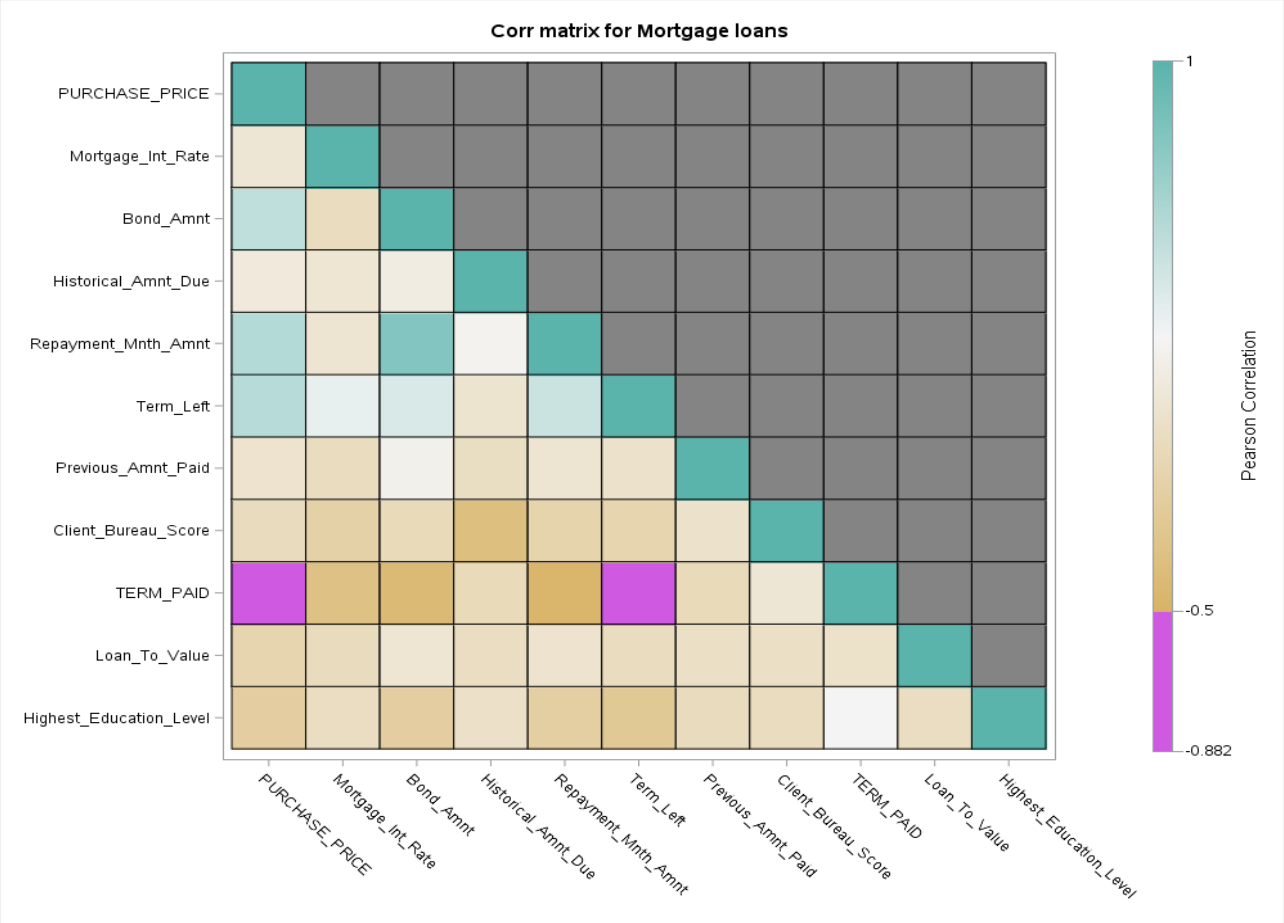


Figure 4.13.: Correlation Matrix for Mortgage loans

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	Historical Amount Due	Client Bureau Score	Previous Amount Paid	Term Paid	Remaining Term	Bond Amount	Highest Education Level	Purchase Price	Mortgage Interest Rate	Loan to value	Repayment Value
Historical Amount Due	1.00000	-0.36253	-0.01088	-0.05679	0.06193	0.15135	0.02070	0.12225	0.07421	-0.01257	0.21762
		<.0001	0.1995	<.0001	<.0001	<.0001	0.0147	<.0001	<.0001	0.2339	<.0001
	13895	13895	13895	13895	13895	13895	13895	13895	13895	8971	13895
Historical Amount Due											
Client Bureau Score	-0.36253	1.00000	0.02946	0.08627	-0.10435	-0.05573	-0.03115	-0.03936	-0.15462	0.00405	-0.12136
	<.0001		0.0005	<.0001	<.0001	<.0001	0.0002	<.0001	<.0001	0.7011	<.0001
	13895	13895	13895	13895	13895	13895	13895	13895	13895	8971	13895
Client Bureau Score											
Previous Amount Paid	-0.01088	0.02946	1.00000	-0.05392	0.02622	0.19334	-0.03456	0.05293	-0.02716	0.00518	0.06381
	0.1995	0.0005		<.0001	0.0020	<.0001	<.0001	<.0001	0.0014	0.6240	<.0001
	13895	13895	13895	13895	13895	13895	13895	13895	13895	8971	13895
Previous Amount Paid											
Term Paid	-0.05679	0.08627	-0.05392	1.00000	-0.88207	-0.41118	0.24631	-0.54769	-0.33616	0.02542	-0.46439
	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	0.0161	<.0001
	13895	13895	13895	13895	13895	13895	13895	13895	13895	8971	13895
Remaining Term	0.06193	-0.10435	0.02622	-0.88207	1.00000	0.39038	-0.24610	0.55609	0.31666	-0.02706	0.46056
	<.0001	<.0001	0.0020	<.0001		<.0001	<.0001	<.0001	<.0001	0.0104	<.0001
	13895	13895	13895	13895	13895	13895	13895	13895	13895	8971	13895
Remaining Term											

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	Historical Amount Due	Client Bureau Score	Previous Amount Paid	Term Paid	Remaining Term	Bond Amount	Highest Education Level	Purchase Price	Mortgage Interest Rate	Loan to value	Repayment Value
Bond Amount	0.15135	-0.05573	0.19334	-0.41118	0.39038	1.00000	-0.18427	0.51362	-0.02798	0.07090	0.80417
	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	0.0010	<.0001	<.0001
Bond Amount	13895	13895	13895	13895	13895	13895	13895	13895	13895	8971	13895
Highest Education Level	0.02070	-0.03115	-0.03456	0.24631	-0.24610	-0.18427	1.00000	-0.18560	-0.01184	-0.01267	-0.17600
	0.0147	0.0002	<.0001	<.0001	<.0001	<.0001		<.0001	0.1627	0.2303	<.0001
Highest Education level	13895	13895	13895	13895	13895	13895	13895	13895	13895	8971	13895
Purchase Price	0.12225	-0.03936	0.05293	-0.54769	0.55609	0.51362	-0.18560	1.00000	0.08306	-0.10934	0.56964
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
Purchase Price	13895	13895	13895	13895	13895	13895	13895	13895	13895	8971	13895
Mortgage Interest Rate	0.07421	-0.15462	-0.02716	-0.33616	0.31666	-0.02798	-0.01184	0.08306	1.00000	-0.03968	0.06690
	<.0001	<.0001	0.0014	<.0001	<.0001	0.0010	0.1627	<.0001		0.0002	<.0001
Mortgage interest rate	13895	13895	13895	13895	13895	13895	13895	13895	13895	8971	13895
Loan to value	-0.01257	0.00405	0.00518	0.02542	-0.02706	0.07090	-0.01267	-0.10934	-0.03968	1.00000	0.05640
	0.2339	0.7011	0.6240	0.0161	0.0104	<.0001	0.2303	<.0001	0.0002		<.0001
	8971	8971	8971	8971	8971	8971	8971	8971	8971	8971	8971
Repayment Value	0.21762	-0.12136	0.06381	-0.46439	0.46056	0.80417	-0.17600	0.56964	0.06690	0.05640	1.00000
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
Repayment Value	13895	13895	13895	13895	13895	13895	13895	13895	13895	8971	13895

Table 4. 14.: Correlation matrix of the key variables for the home loans portfolio

The correlation matrix suggests that some of the predictors are at least moderately marginally correlated. For example, term paid (TERM_PAID) and remaining term (RMNG_TERM) are strongly correlated (r = 0.882), and bond amount and monthly repayment amount are strongly correlated (r = 0.804). On the other hand, none of the pairwise correlations among past due amount, credit risk score, last payment amount, education level, debit interest rate and loan to value are particularly strong (r < 0.40 in each case).

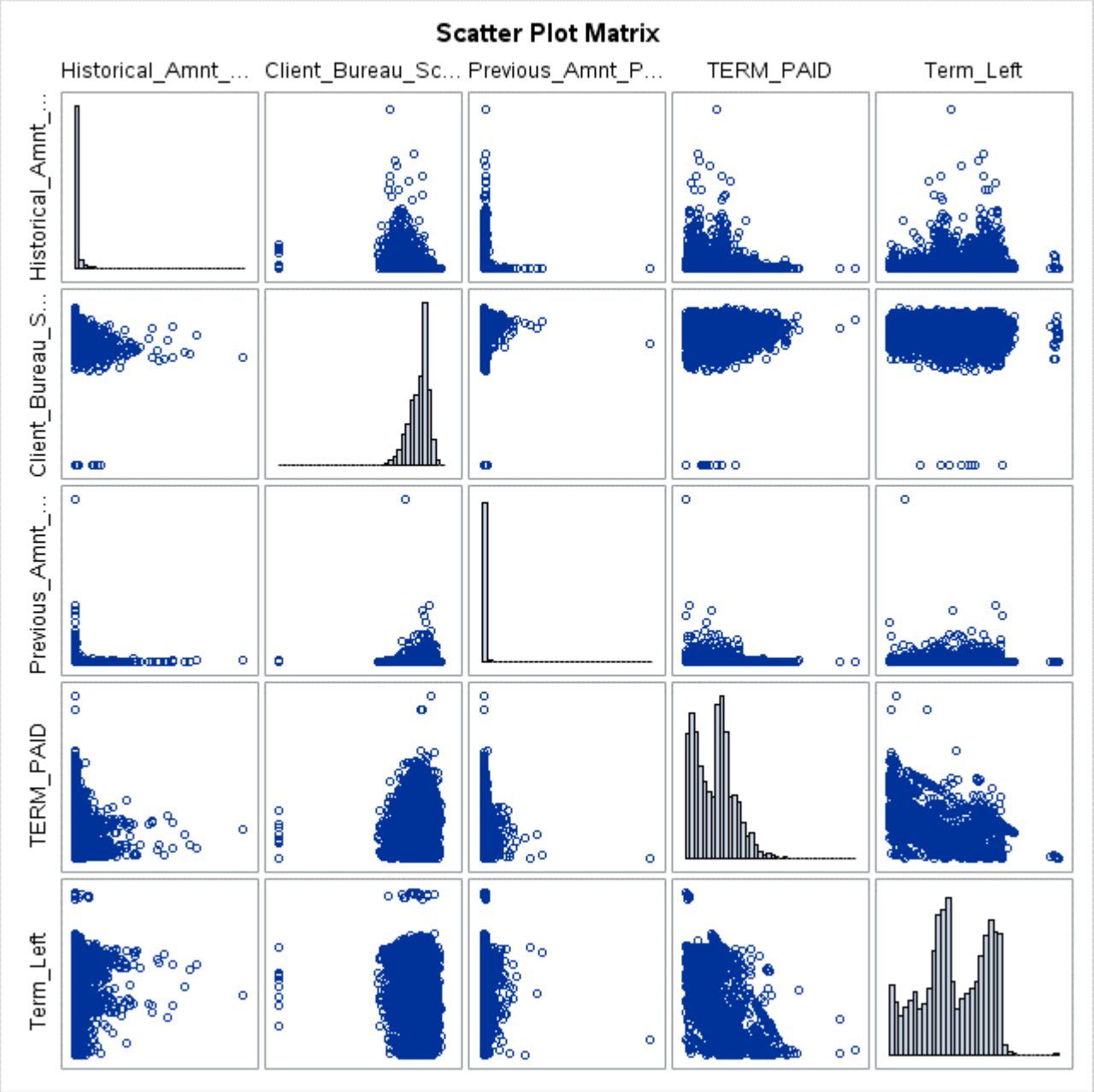


Figure 4. 14.: Scatter plot matrix of variables for the home loans

4.5.2. Variance Inflation factors (VIF)

It defines as ratio of variance in a model with many terms, divided by the variance of a model on term. It calculates the strictness of multicollinearity in an ordinary least squares regression analysis. It gives a guide that deals with how considerable the variance of a projected regression coefficient is enlarged, due to collinearity.

Regressing on eleven of the predictors, we obtain the following:

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	1.86765	0.088	21.22	<.0001	.	0
Previous Amount Paid	Previous Payment	1	-1.60E-07	5.97E-08	-2.68	0.0073	0.98578	1.01442
Highest Education Level		1	0.02442	0.00312	7.82	<.0001	0.9499	1.05275
Loan to value		1	-7.18E-08	4.40E-08	-1.63	0.1027	0.89049	1.12297
Client Bureau Score	Client Bureau Score	1	-0.00339	0.0000873 5	-38.75	<.0001	0.80245	1.24618
Historical Amount Due	Historical Amount Due	1	0.00001447	5.12E-07	28.26	<.0001	0.7955	1.25707
Mortgage interest rate	Mortgage interest rate	1	0.04329	0.00387	11.18	<.0001	0.58574	1.70724
Purchase Price	Purchase Price	1	-3.47E-08	1.26E-08	-2.75	0.0059	0.25011	3.9983
Term Left	Remaining Term	1	-0.00016226	0.0001600 6	-1.01	0.3107	0.23143	4.32101
Term Paid		1	0.00202	0.0001851 1	10.89	<.0001	0.20402	4.90154
Monthly Repayment Value	Monthly Repayment Value	1	-0.00000453	0.0000022 1	-2.05	0.0407	0.13151	7.60395
Bond Amount	Bond Amount	1	2.70E-08	2.63E-08	1.03	0.3039	0.09408	10.6291

Table 4. 15.: VIF Parameter Estimates

As can be seen, four of the variance inflation factors — 10.63, 7.60, 4.90 and 4.32 — are large. The VIF for the predictor *Bond Amount*, for instance, states that the variance of the estimated factor of *Bond Amount* is inflated by an aspect of 10.63 due to *Bond Amount* is extremely correlated with one of the other predictors in the model. The same applies to monthly repayment amount, term paid and remaining term. For instance, it shows that the variance of the estimated coefficient of monthly repayment amount, term paid and remaining term are inflated by a factor of 7.60, 4.90 and 4.32, respectively, because each is extremely correlated with one of the other predictors in the model. Remaining term, term paid, monthly repayment amount and bond amount were removed from the model, due to high VIF which is greater than four (4).

4.6. Stratified Random Sampling

It is a method of sampling data that contains the partition of people into lesser groups identified as strata. In this method, the strata are intended created on participants' common features or characteristics. We can split the data (given population) into random samples with 50-50, 60-40 or 70-30 ratios for *Training* (development sample on which model will be developed or trained) and **Test** (validation/holdout sample on which model will be tested) based on population size. In this exercise we will split the sample into 70-30. You may perform this step even before univariate analysis.

For modelling, a random sample of 70% of observations is selected for validation/training; the following split it by good or bad as strata.

	Count	Percentage
Bad	4898	34.9
Good	9136	65.1

The remaining 30% random sample observations are used for testing; the following table shows a split of good or bad events.

	Count	Percentage
Bad	2098	34.89
Good	3915	65.11

4.7. Conclusion

A description and explanation of the data has been given, where sample records of 20,047 loans were taken for the study. Variable reduction was done using the Information Value instrument to relate the predictive power along with other unlike variables. Weight of Evidence decorates additionally how each variable behaves. The eleven (11) best variables had higher information value, and were selected to be utilised to model Survival Analysis and Logistic Regression in the next chapter.

CHAPTER 5: ESTIMATION AND ANALYSIS

5.1. Introduction

This chapter offers the detailed analysis of Logistic Regression and Survival analysis model. It has a version improvement, choice, performance and contrast, and the outcomes have been presented graphically and numerically. These analyses were carried out to decide the approaches to striving to perform better for some customer credit unit statistics inside the existence of challenging risks and extended time survivors. The statistical applications, namely SAS and Rstudio, were used to analyse the mortgage loan data in credit risk context.

5.2. Model Selection and Development

5.2.1. Logistic Regression

5.2.1.1. Final Model Selection

The training dataset of mortgage loans is large at 13,895 loans with a not bad skewed split alongside binary output defaulted at 4,855 loans and 9,040 loans who did not default. Consequently, we have a large enough data sample size. The following table illustrates default = 1 and none defaulters = 0.

Ordered Value	Default	Total Frequency
1	Yes	4855
2	No	9040

Table 5. 1.: Response View

The LR model was fitted using Proc Logistic in SAS software. Table 5.2 illustrates the final model selected according to the stepwise regression.

MLE						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	15.3590	0.4348	1247.9027	<.0001	4681111
Client Bureau score	1	-0.0270	0.000572	2234.8246	<.0001	0.973
Previous amount Paid	1	-5.71E-6	9.499E-7	36.1458	<.0001	1.000
Highest Education level	1	0.3098	0.0202	235.4578	<.0001	1.363
Monthly repayment value	1	-0.00007	6.05E-6	119.2309	<.0001	1.000
Mortgage Interest Rate	1	0.0711	0.0159	20.0220	<.0001	1.074
Purchase Price	1	-4.04E-7	5.143E-8	61.7265	<.0001	1.000

Table 5. 2.: LR model MLE

Looking at Table 5.2, the highest education level and mortgage interest rate are having a positive sign, which means that the default rate is higher if the highest education level is Grade12 or without formal education and the mortgage interest rates are higher. The client bureau score is a negative sign which means that the default rate is lower when the bureau score is increasing. The previous monthly repayment value is negative, which also means that the default rate is lower when the

repayment value is higher (customers making payments monthly), as referred to in Chapter 4, section 4.4.2.11. The previous amount paid is negative, which also means that the default rate is lower when the previous amount paid is higher, and purchase price is negative – which also means that the default rate is lower when the mortgage purchase price is higher. All selected variables are significant with p-value < .05, and it is making business sense looking at the credit risk context.

Model Fit Statistics

Assessing the "Global Null Hypothesis: Beta = 0", the table contains the results of three (3) tests of the global null hypothesis that all coefficients (apart from the intercept) are equal to 0, often thought of as a test of the utility of the whole model:

$\beta_1 = \beta_2 = \cdots = \beta_k = 0$

These tests are asymptotically equivalent (equivalent in very large samples), so will usually agree. In cases where they do not, many prefer the likelihood ratio test (Hosmer and Lemeshow, 2013).

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4022.1526	6	<.0001
Score	3220.1578	6	<.0001
Wald	2585.9494	6	<.0001

Table 5. 3.: Testing null hypothesis that the beta = 0 for logistic regression model

The Model Fit Statistics table, below, displays two relative fit statistics, AIC and SC (BIC), in addition minus two multiply the log likelihood of the model. Generally, you will want the values in the Intercept and Covariates column, as these reflect the full model just fit. Lower values on AIC and SC signify better fit (both penalise adding more parameters to the model).

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	17984.306	13974.154
SC	17991.846	14026.929
-2 Log L	17982.306	13960.154

Table 5. 4.: Model Fit Statistics for logistic regression model

Assessing Model Fit

The following table shows Deviance and Pearson Goodness-of-fit statistics. It has been explained that Hosmer and Lemeshow (1980) built a goodness-of-fit statistic designed in the direction of more appropriate to use when J = n, i.e. number is n of observations utilised in the dataset for logistic.

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	13949.0634	14E3	1.0121	0.1572
Pearson	33616816.8	14E3	2439.183	<.0001

Table 5. 5.:" Deviance and Pearson Goodness-of-fit statistics"

Table 5.6, below, shows the Hosmer and Lemeshow $\hat{C}$ statistic, labelled Chi-Squared, with a non-significant p-value at p-value = 0.5068 > 0.05, suggesting good model fit.

Chi-Square	DF	Pr > ChiSq
7.2797	8	0.5068

Table 5. 6.: Hosmer and Lemeshow Goodness-of-fit test for Logistic Regression

Table 5.7 shows that proc logistic split the dataset into 10 groups, and then shows the observed and expected number of default = Yes and default = No in each group. The logistic regression accuracy plots are presented in Figure 5.1 with line of best fit. There are small discrepancies between observed and expected throughout the table.

		default = Yes		default = No	
Group	Total	Observed	Expected	Observed	Expected
1	1390	54	55.78	1336	1334.22
2	1389	131	137.01	1258	1251.99
3	1390	216	210.79	1174	1179.21
4	1389	273	285.99	1116	1103.01
5	1390	378	363.00	1012	1027.00
6	1389	426	444.32	963	944.68
7	1389	570	552.57	819	836.43
8	1390	702	715.31	688	674.69
9	1389	908	914.60	481	474.40
10	1390	1197	1175.62	193	214.38

Table 5. 7.: Hosmer and Lemeshow Partition – Logistic Regression model

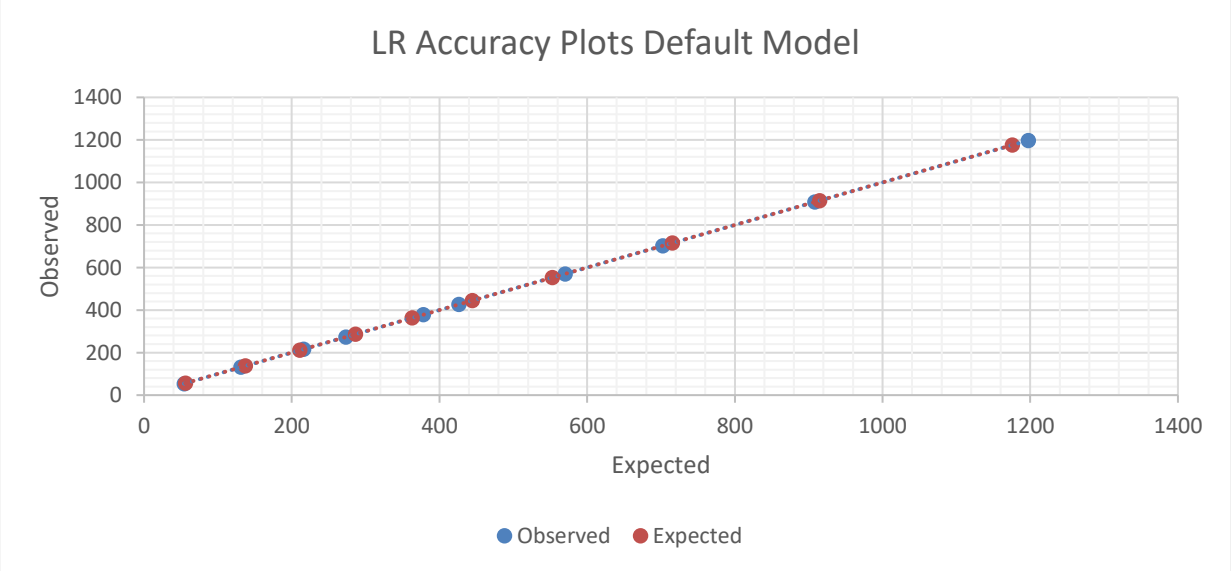


Figure 5. 1.: Accuracy Plots for Logistic regression

5.2.1.2. Influence Diagnostics

a. Identifying influential observations

The two plots below are formed by influence, and are plots of Pearson residuals ***Figure 5.2*** on the left, and plots of leverage ***Figure 5.3*** on the right-hand side. The influential observations usually have both large residuals and high leverage. From these two graphs, we can see the observations with the largest residual, and those observations are **937** and **10620** with the highest leverages.

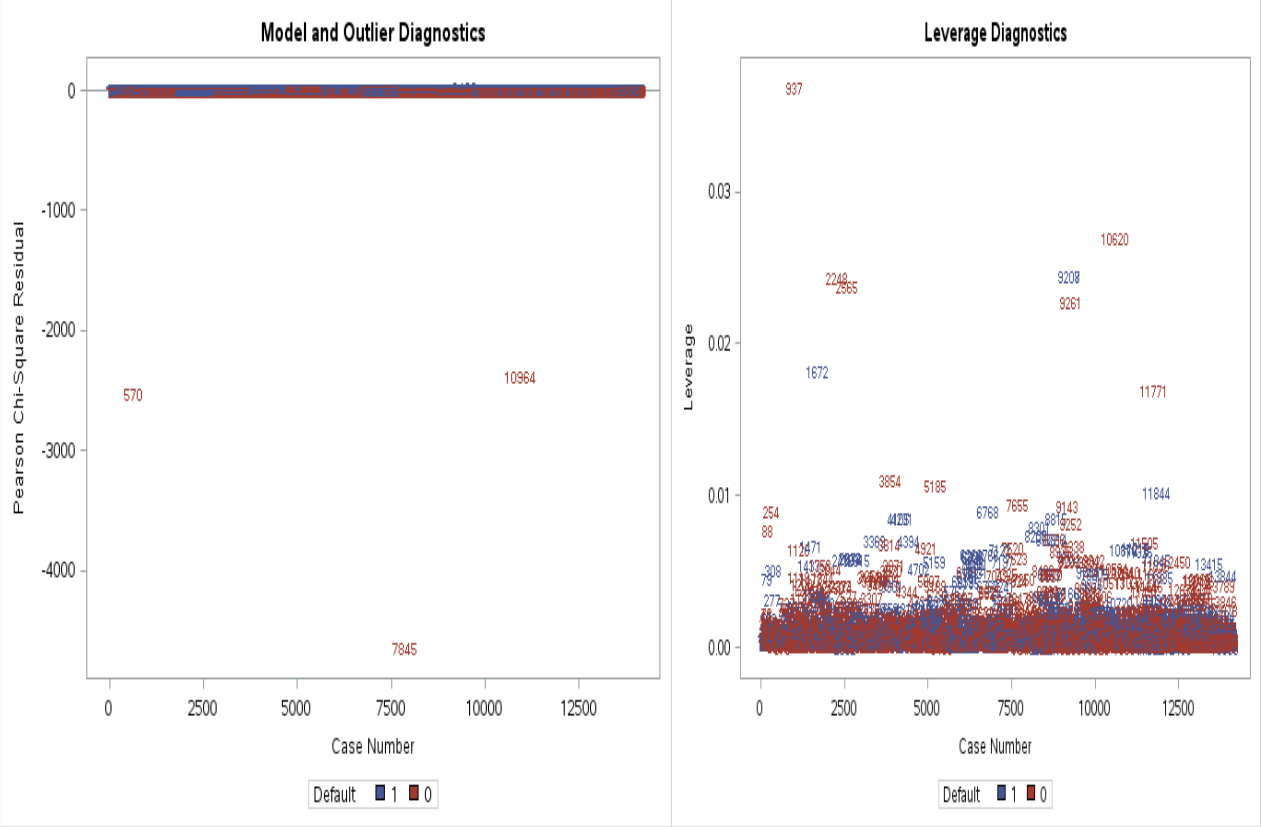


Figure 5. 2.: Left - Model and Outlier Diagnostics for LR Right - Leverage Diagnostics for LR

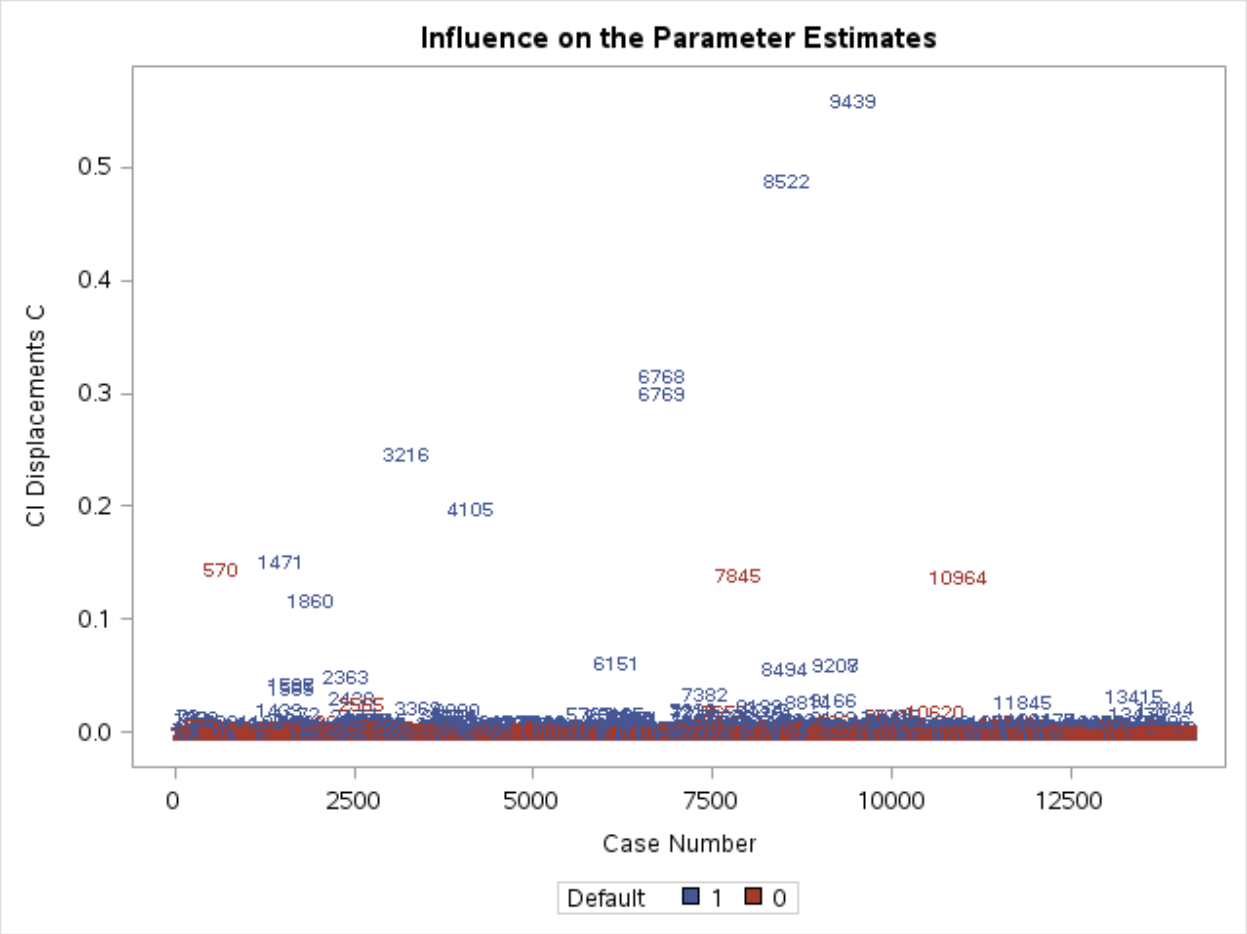


Figure 5. 3.: Influence on the Parameter Estimates for Logistic regression

Above is the plot of C diagnostic. Since C diagnostic is calculated using the Pearson residual and leverage, we can see that observations 9439, 8522, and 6768 have the highest C diagnostic values. Regarding cut-offs for deeming an observation "influential", Hosmer et al. (2013) do not recommend any cut-offs, but instead recommend identifying points that fall far away from other observations, which describes observations 9439, 8522, 6769 and 6768 on the graph of C diagnostic.

The next few graphs examine **DFBETAS** graphs. Model is checking for observations that are outliers. The two plots below are Dfbetas, and are plots of Influence on the Estimate of Bureau Score on the left, and plots of Influence on the Estimate of Previous Amount Paid on the right-hand side. In the DFBETAS plot for Bureau risk score, observations with the largest influence are 570, 7845 and 10964. In the graph on the right (DFBETAS of Historical amount paid), the plot shows 9439 and 8522 as the observations with the largest influence.

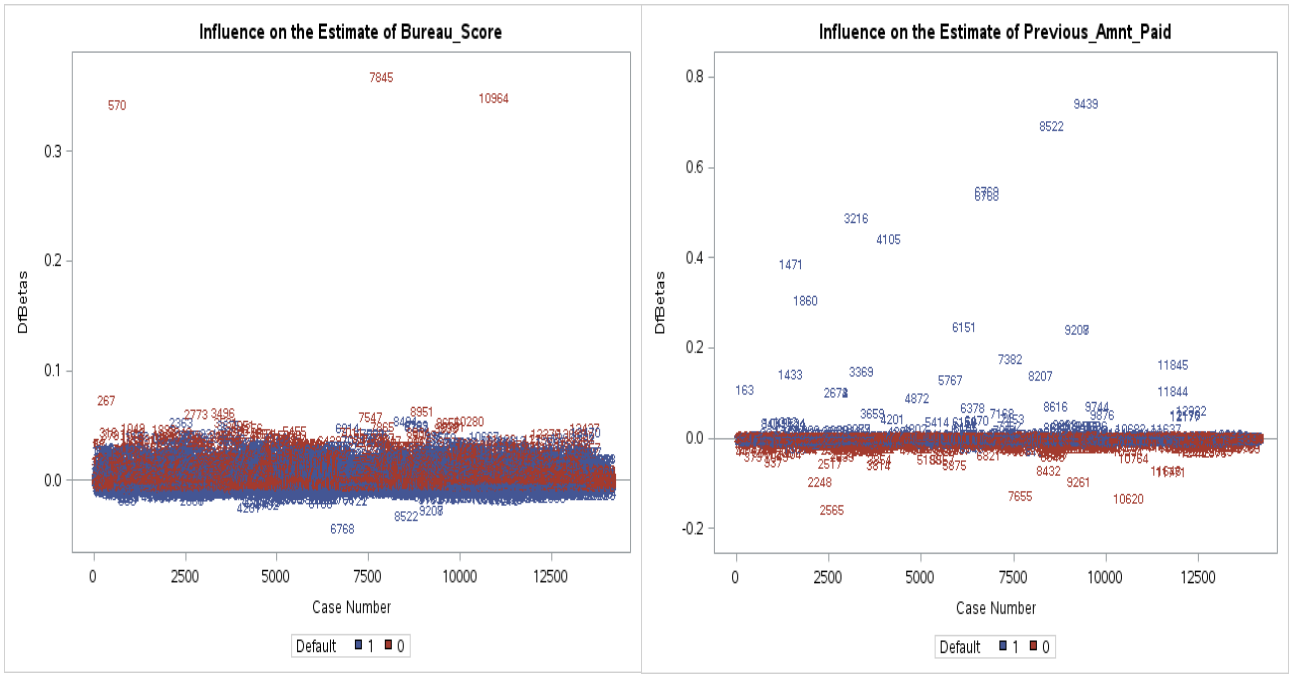


Figure 5. 4.: Left - influence on the Estimate of Bureau risk score. Right - influence on the Estimate of Historical amount paid.

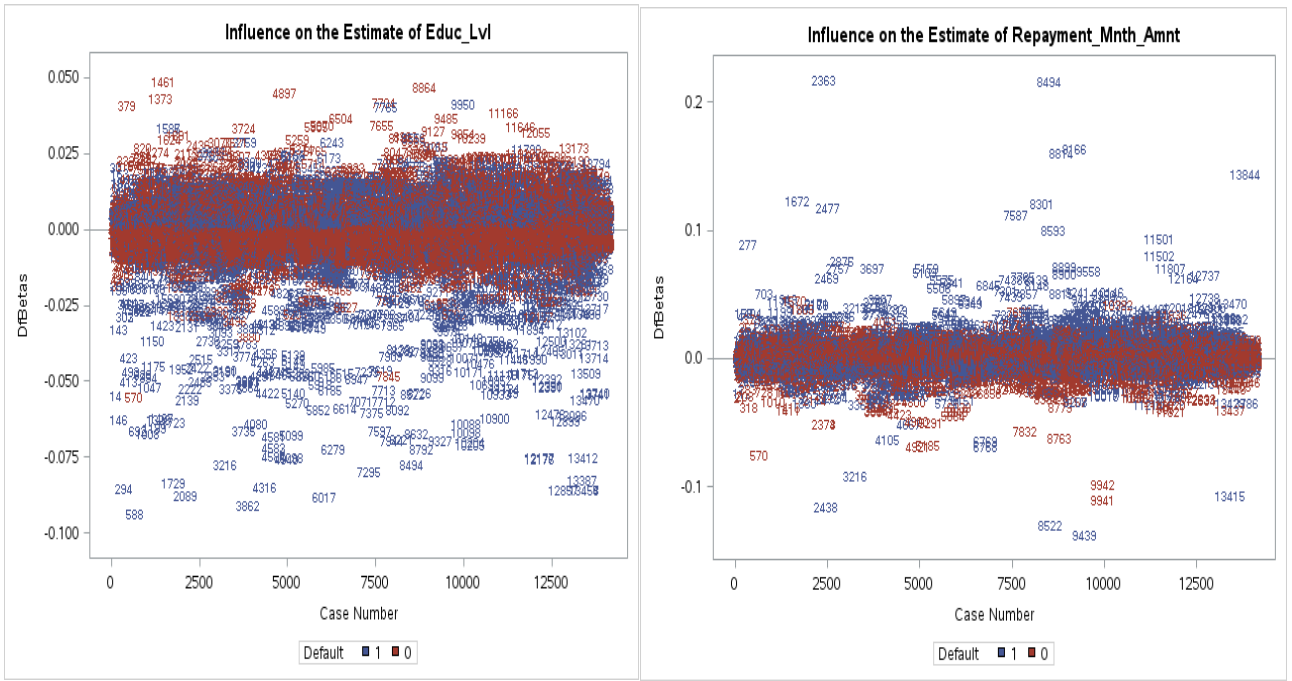


Figure 5. 5.: Left - influence on the Estimate of Education level. Right – influence on the Estimate of Repayment month amount.

In the DFBETAS plot for education level, the large influences are 588, 3862, and for repayment monthly amount are 2363 and 8494.

In the next two plots below, in the DFBETAS plot for debit interest rate, the large influences are 7845, 10964.

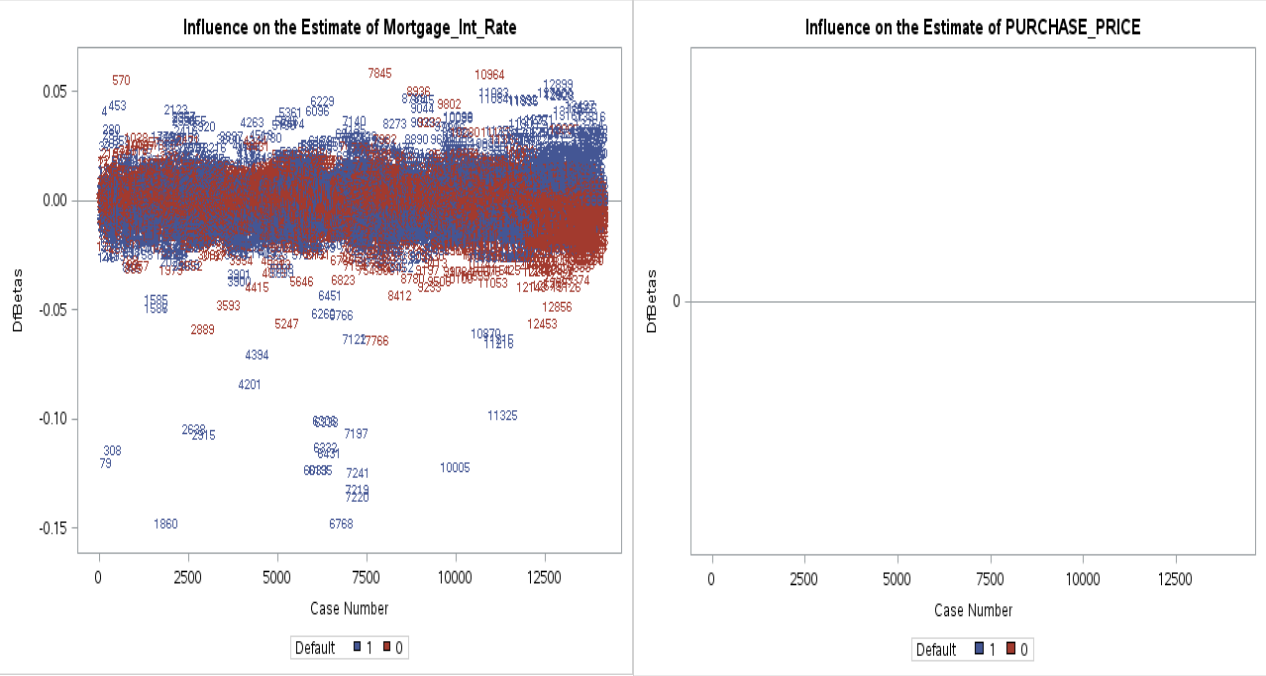


Figure 5. 6.: Left - influence on the Estimate of mortgage interest rate. Right – influence on the Estimate of Purchase Price.

b. Inspecting influential observations

We can conclude that we have identified observations 1860, 6768, 588, 3862, 2363, 8494, 570, 7845, 10964, 9439, 8522 and 6769 as unduly influential, and have checked whether we must include them in the model.

Obs	Client Id	Client nr of Children	Highest education level	Purchase price	Mortgage interest rate	Monthly repayment value	Previous amount paid	Client Bureau score	Default
1	772678	2	G12	0.00	8.750000	30239.29	30296.29	0	0
2	797116	2	DOC	0.00	9.750000	0.28	10861.41	707	1
3	2557816	0	G12	500000.00	0.000001	4345.03	340131.00	684	1
4	3510408	1	BCH	0.00	8.150000	40198.42	20000.00	665	1
5	7233305	2	DOC	0.00	9.500000	0.22	5.75	696	1
6	12109642	0	G12	810000.00	0.000001	6105.60	590175.82	617	1
7	12109642	0	G12	810000.00	10.100000	6105.60	590175.82	671	1
8	13372014	0	G12	0.00	8.500000	13040.27	13097.27	0	0
9	14381618	0	DOC	0.00	9.550000	41115.46	41172.46	650	1
10	14412282	0	HNR	900000.00	10.700000	0.00	740701.26	617	1
11	15631287	0		1450000.00	8.450000	0.00	789112.35	687	1
12	18938619	2	GRD	0.00	8.600000	23148.17	23205.17	0	0

Table 5. 8.: Influential observations on logistic regression model

Client ID 772678, 13372014 with Grade 12/matric qualification, with a high credit risk score would be expected to survive or not default by our model. However, they did not, so they have a large residual. Other observations, for example a client who paid his/her mortgage loans at a price of R740 701.26, had large leverage value. He/she in fact did not survive or default, so the fact that he/she is

influential suggests that the model expected him/her not to default. These observations will be included from the model.

5.2.2. Survival Analysis

5.2.2.1. Likelihood of Density Function (PDF)

$f(t)$ is the probability of detecting $Time$ at time t comparative to entirely survival times. If we integrate $f(t)$ on a variety of survival periods, then it provides the likelihood of defining a survival period with that interval $[a, b]$. The formula is given below.

$\Pr(a \leq Time \leq b) = \int_a^b f(t)dt = \int_a^b \lambda e^{-\lambda t} dt$, where

λ: $rate\ parameter\ of\ the\ exponential\ distribution\ and\ is\ equal\ to\ the\ reciprocal\ of\ the\ mean\ survival\ time$.

PDFs are fundamentally the histograms encompassed of bins of vanishingly small widths. As shown in Figure 5.7, the shorter survival times between 64 months and 150 months are extra-credible, representing that the risk of credit default in these periods is high.

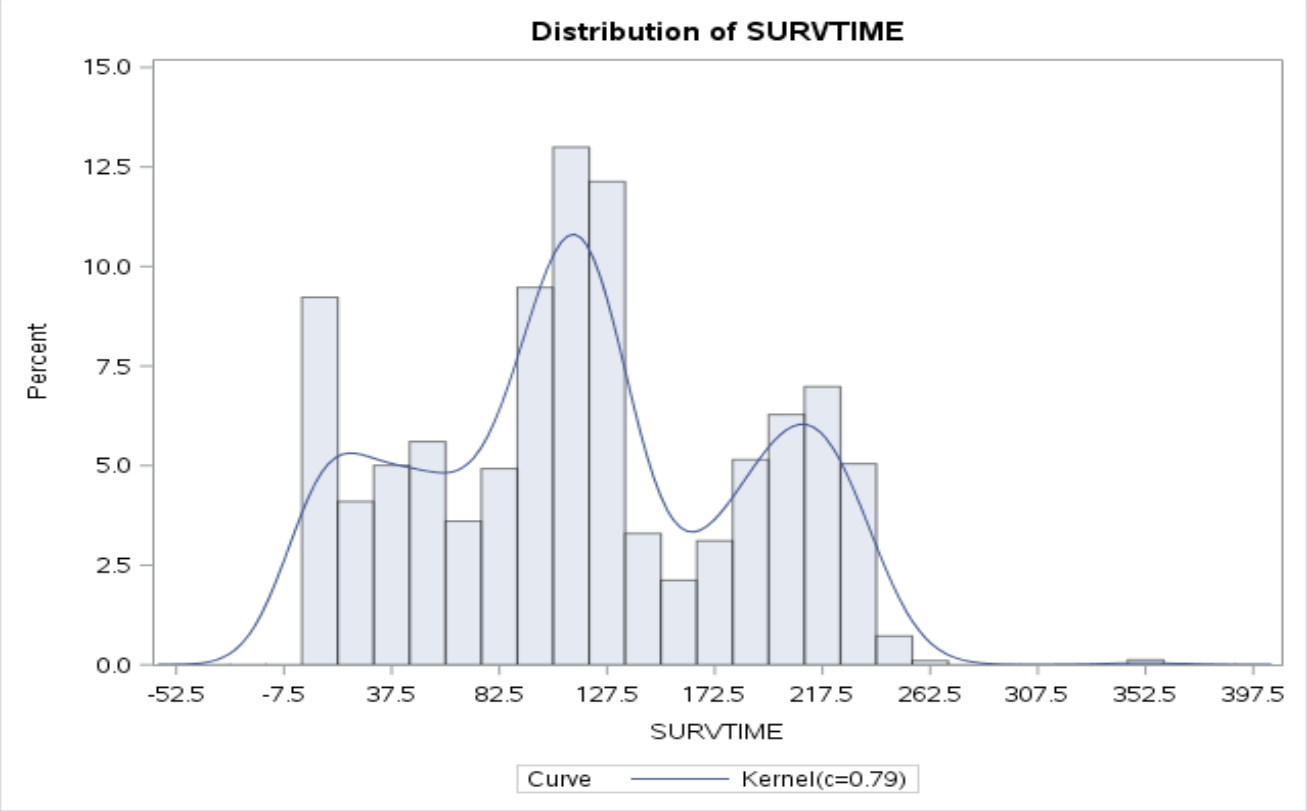


Figure 5. 7.: Distribution of the time to default for defaulted customers

Figure 5.7 is the probability of density function for all the defaulted loans, whereas Figure 5.8 is the probability of density function for all the loans in this paper. Figure 5.8 shows that many loans

censored around 84 and 228 months. The main reason for this is because many of the loans have not defaulted before censoring.

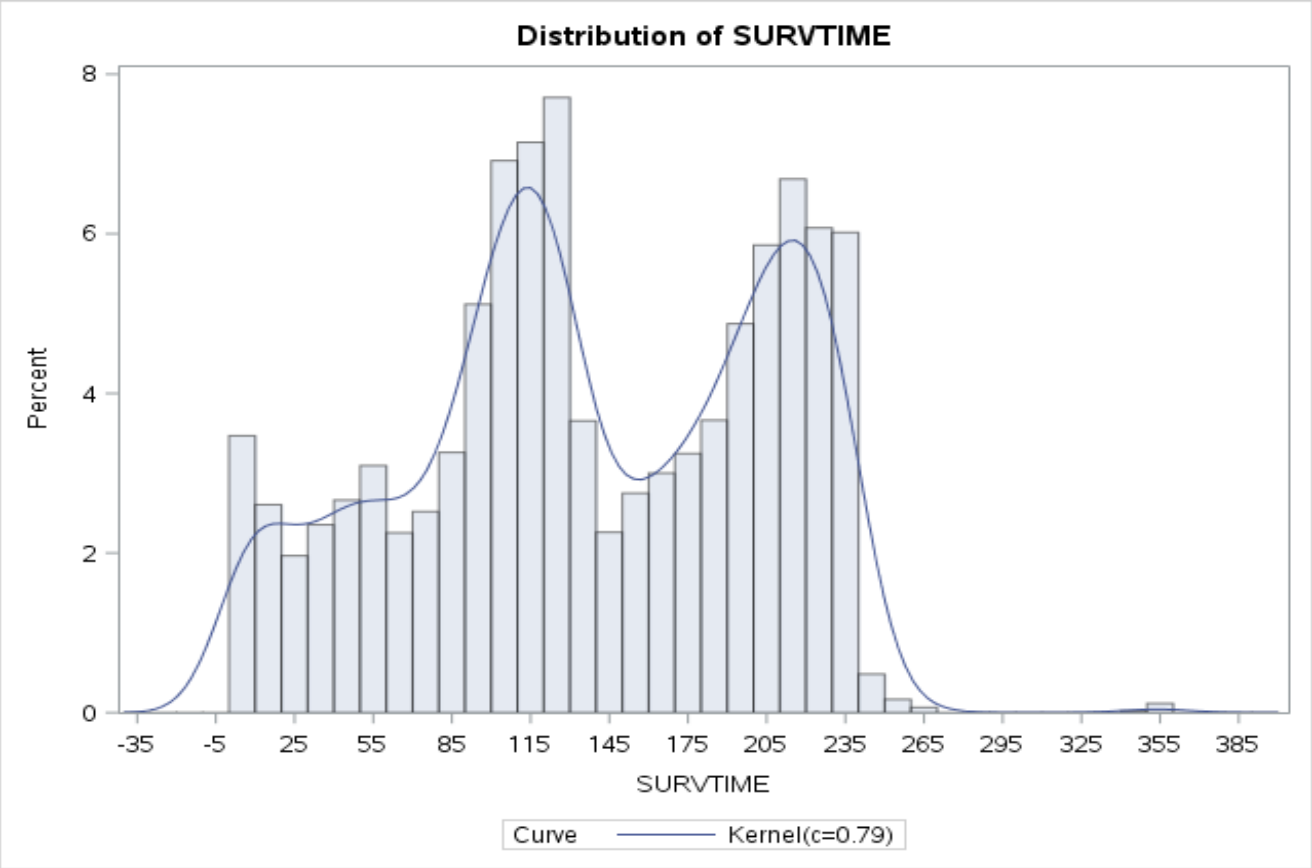


Figure 5. 8.: Spreading of the time to default for whole population

5.2.1.2. Cumulative Distribution Function, F(t)

CDF, $F(t)$, defines the likelihood of detecting $Time$ fewer than or equivalent to $some\ time\ t$, otherwise $\Pr(Timelet)$. Thus, it could be illustrated as shown below: $F(t) = \int_0^t f(t)dt$. The relationship between cdf and pdf also implies $f(t) = \frac{dF(t)}{dt}$. **Proc univariate** is used to display the estimation of the CDF in the SAS guide.

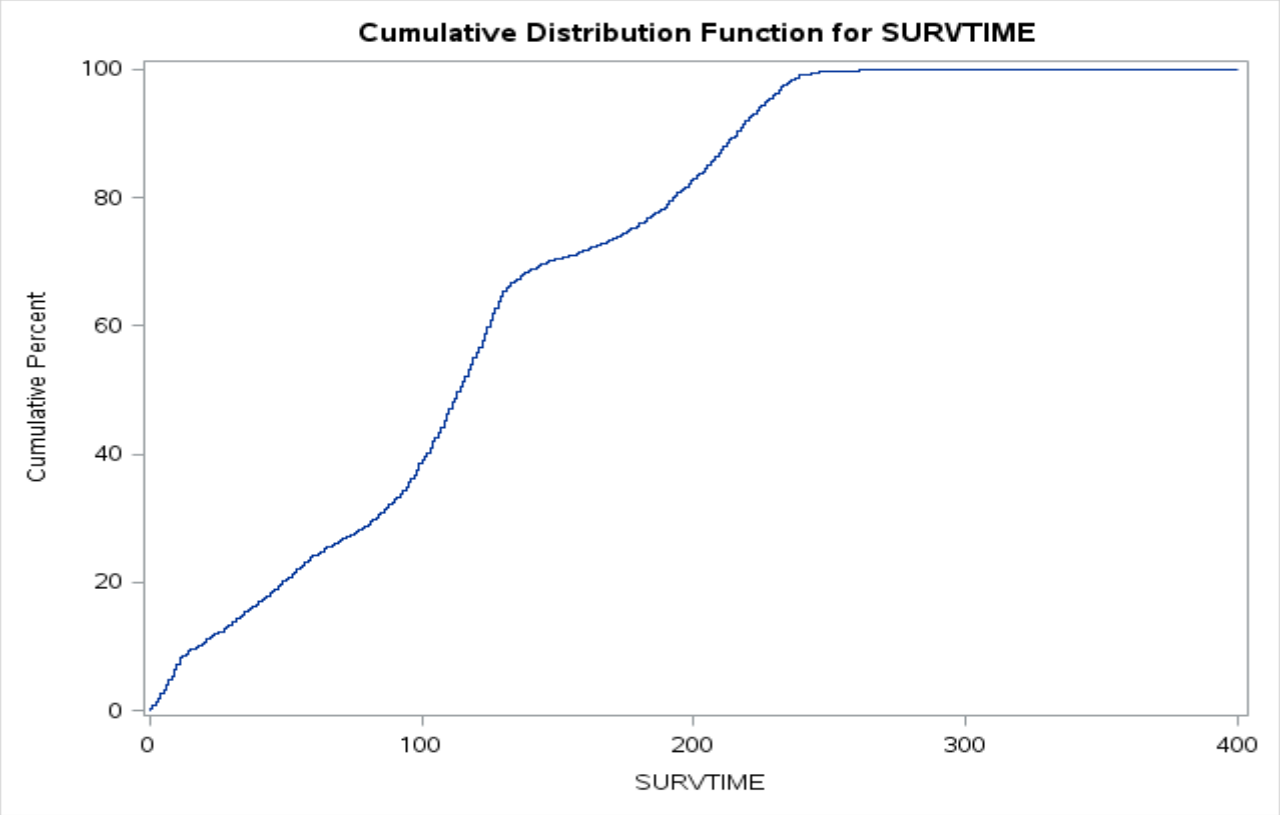


Figure 5. 9.: CDF of survival period

The above graph shows that the likelihood of surviving 114 months or less is close to 50%. As a result, through 114 months, a customer has collected a fair bit of risk, which gathers extra gradually after this point. In intervals where even periods are more plausible, the CDF will rise more quickly.

5.2.1.3. Descriptive SA

To determine some failure time random variables, the following are required:

1. A time origin
2. A time scale (years after going default)
3. Definition of the event

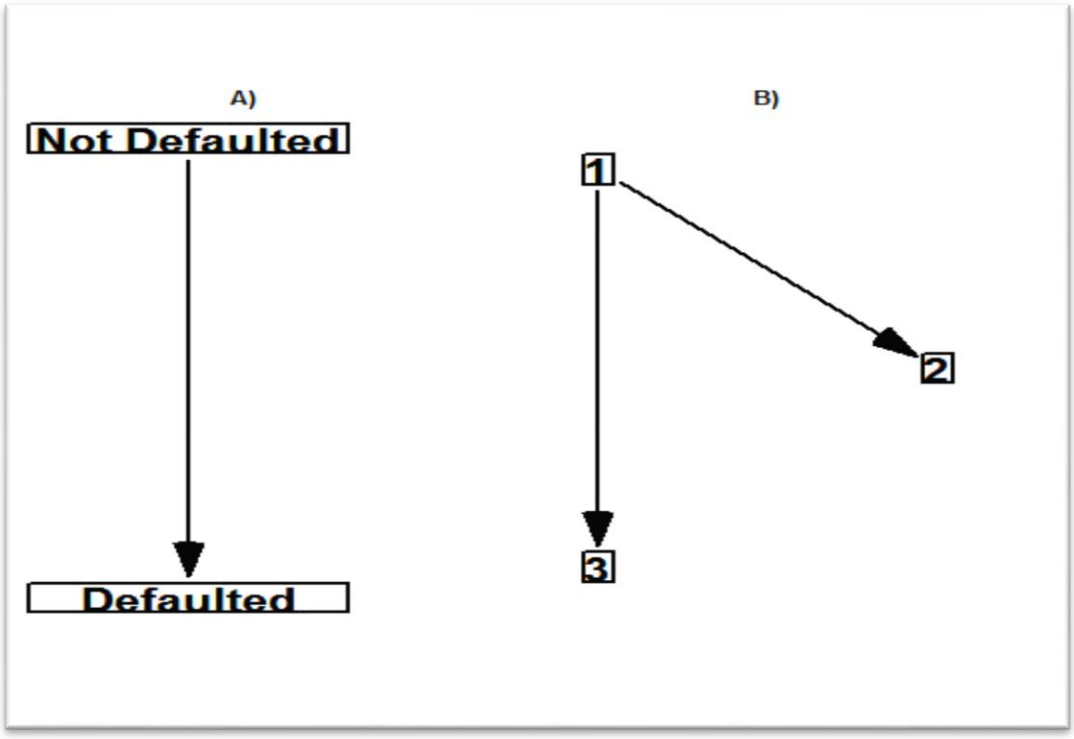


Figure 5. 10.: Box diagram for transitions/events

Not Defaulted	Defaulted
9040	4855

The results above show a split of our total observations; we have 9,040 (~65%) none defaulted and 4,855 (~35%) defaulted. The diagram of SA statistics in Figure 5.10 displays numerous structures which are naturally come upon in analysis of survival data.

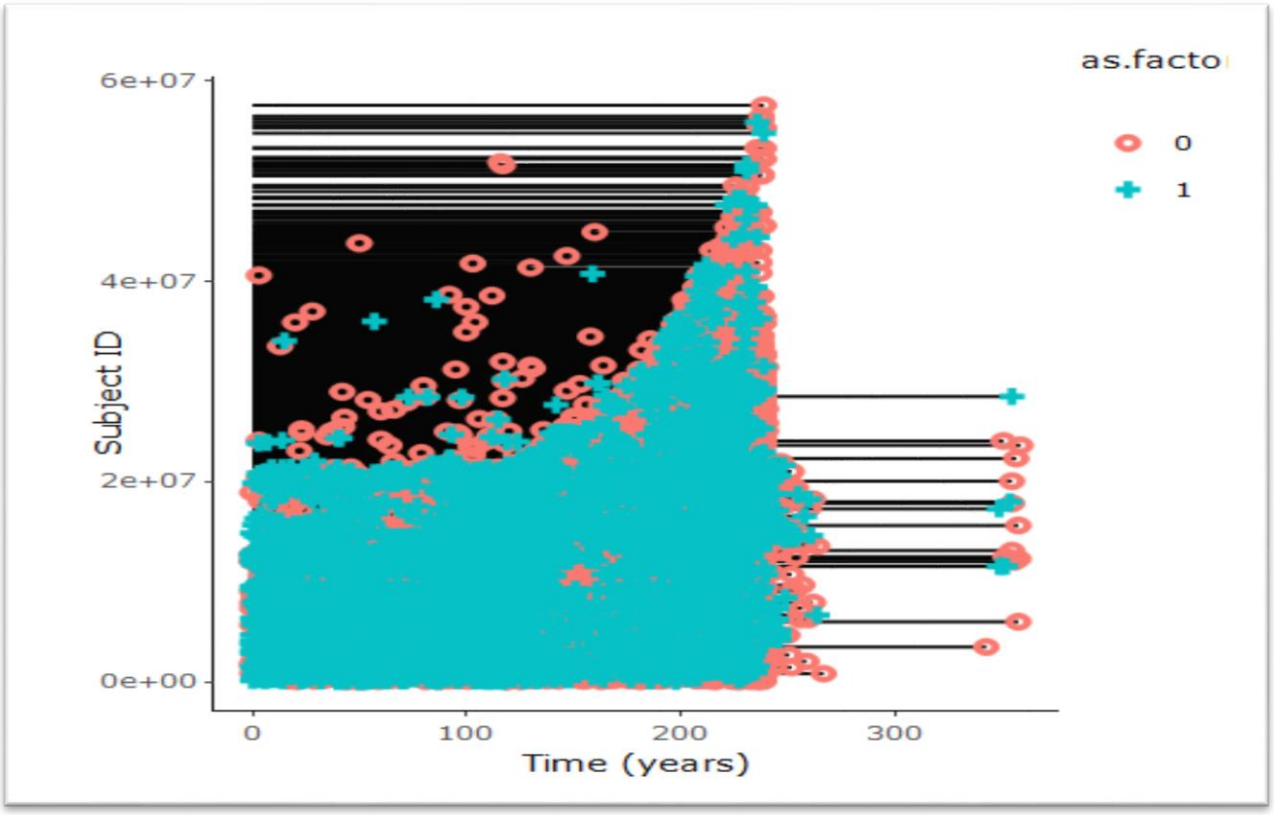


Figure 5. 11.: Possible representations of follow-up time. 0 None Defaulters and 1 Defaulter

Figure 5.11, above, illustrates follow-up time. We can see that default (blue) from mortgage loans is more likely to occur early after analysis, as opposed to default from other factors.

a. Estimation of Survival Function

Below are the results of non-parametric estimators. R programming was utilised to produce these results.

1. Kaplan Meier Estimator

The following table shows the Lifetable for Product-Limit survival estimates:

No of obs	*events*	*nmean	*se(nmean)	*median*	*0.95LCL*	*0.95UCL*
13895.00	4855.00	205.62	2.41	222.00	220.00	224.00

* restricted mean with upper limit = 359

Months	no.risk	no.event	Survival	std.err	lower 95% CI	upper 95% CI
0	13895	17	0.999	0.000297	0.9982	0.999
5	13683	145	0.988	0.000913	0.9865	0.99
10	13413	187	0.975	0.001336	0.9721	0.977
60	11651	823	0.913	0.002439	0.9079	0.917
110	8863	1119	0.817	0.003482	0.8103	0.824
160	5595	1195	0.693	0.004445	0.6844	0.702
210	2727	763	0.57	0.005511	0.5592	0.581
260	28	598	0.185	0.01999	0.1493	0.228
359	1	8	0.109	0.024102	0.0707	0.168

Table 5. 9.: Life-table for Product-Limit Survival Estimates

Table 5.9 is the illustration of K-M estimates of the survival function formed by ***fortify in Rstudio***. From the table above, we can see that the month interval showed in the first row is from zero month to just earlier than one month. At this point, 13,895 people at risk and 17 customers defaulted on the mortgage loan, since “Observed Events” is 17 and the forecast of the “Survival function” is 0.999. From 5 months to before 10 months, 145 customers defaulted, indicated by two rows of “remaining term” = 5 months and “Observed Events” = 145 in the final row wherever “remaining term” = 5.

During the interval [260,359), 1 out of 28 at risk defaulted, conditional probability of survival in this point of $\frac{28-1}{28} = 0.9643$. The table shows that the unconditional probability of surviving beyond 260 months is 0.1846, since $\hat{S}(260) = 0.1846 = p(surviving\ up\ to\ 260\ months) * 0.9643, therefore\ p(surviving\ up\ to\ 355 month) = \frac{0.1846}{0.9643} = 0.1914$. In table 5.9, the likelihood of surviving outside 260 months = 0.1914, the similar likelihood as is designed for remaining up to 260 months, so it means that the censored observations of survival estimates remain unchanged when we leave out of the study, only the number at risk.

The following graph illustrates the K-M estimate of the survival function, and it also shows how the survival function fluctuates over time; it was generated in **Rstudio**.

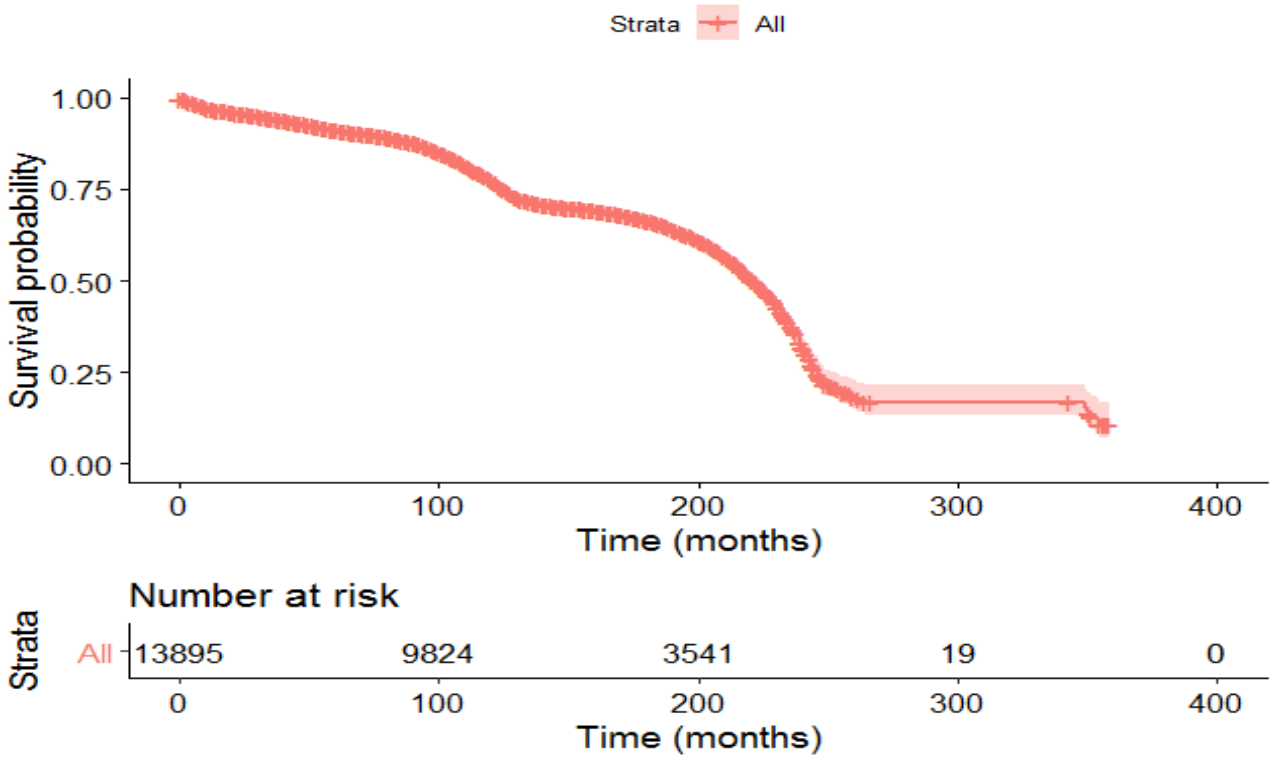


Figure 5. 12.: Product-Limit Survival Estimate

Figure 5.12 shows that the proportion of survival went from 1 (or 100%) (no events) at zero time up to 5 months later to approximately 98.8% at the 5-month point, at which time an event occurred. At 10 months another event reduced the cumulative probability of survival to approximately 97.5%, or, worded another way, at 10 months approximately 97.5% of subjects are estimated to be non-defaulters. At 160 months or 13.33 years an event reduced the cumulative probability of survival further to approximately 69.3% or 69.3% – customers are estimated to be non-defaulters.

2. Lifetable or actuarial estimator

The table below shows the life table or actuarial estimates. The interpretation of the table is as follows:

1st Interval 0-25 months: At month 0, the beginning of the 1st point (0-25 months), we have 13608.5 customers not defaulted yet, 399 customers defaulted in that point, and 573 customers are not defaulted. The modification was useful for the subject of customers censored during that point to show 13608.5 - (399/573) = 13607.8. The proportion of defaulting during interval 399/13608.5 = 0.0293. Among those at risk, proportion surviving interval 1-0.0293 = 0.9707. The likelihood that a customer lasts past 25 months, or survives the first point (utilising the upper limit of the point to define the month/period) is S25 = p25 = 1(0. 9707) =0.9707.

For the 2nd point, 25-50 months: The subject at risk is the subject at risk in the earlier point (0-25 months), and fewer for those who defaulted and are censored. The probability that a participant survives past 50 months/4.2 years is 0.9707.

Looking at the last interval, 275-300 months: The number at risk is 36.5 customers and number of event/defaulters is 25 customers. The probability of surviving in this interval is 2.5%. This is due to customers finishing mortgage loans.

Time	Nsubs	Nlost	Nrisk	Nevent	Surv	Pdf	Hazard	se.surv	se.pdf	se.hazard
0-25	13895	573	13608.5	399	1	0.0012	0.0012	0	1.00E-04	0.0001
25-50	12923	402	12722	451	0.9707	0.0014	0.0014	0.0014	1.00E-04	0.0001
50-75	12070	356	11892	534	0.9363	0.0017	0.0018	0.0021	1.00E-04	0.0001
75-100	11180	550	10905	869	0.8942	0.0029	0.0033	0.0027	1.00E-04	0.0001
100-125	9761	1060	9231	1459	0.823	0.0052	0.0069	0.0034	1.00E-04	0.0002
125-150	7242	464	7010	804	0.6929	0.0032	0.0049	0.0042	1.00E-04	0.0002
150-175	5974	207	5870.5	812	0.6134	0.0034	0.0059	0.0046	1.00E-04	0.0002
175-200	4955	402	4754	1061	0.5286	0.0047	0.01	0.0048	1.00E-04	0.0003
200-225	3492	555	3214.5	1616	0.4106	0.0083	0.0269	0.0049	2.00E-04	0.0006
225-250	1321	255	1193.5	988	0.2042	0.0068	0.0565	0.0044	2.00E-04	0.0013
250-275	78	8	74	22	0.0352	0.0004	0.014	0.0024	1.00E-04	0.0029
275-300	48	23	36.5	25	0.0247	NA	NA	0.0025	NA	NA

Table 5. 10.: Life table survival estimates

3. Nelson-Aalen Estimator

The following table illustrates the N-A estimator, which can be interpreted as follows: projected cumulative hazard at month three: $\hat{H}(3) = \frac{23}{13866} + \frac{21}{13833} + \frac{26}{13803} = 0.001884$. Therefore, we assume .001884 let-downs (everyone) at the last day of three months. The forecast of survival above three months is $\hat{S}(3) = \exp(-0.001884) = 0.998118$. This ties carefully with the K-M estimate of survival above three months of 0.99. In conclusion, 95% confidence interval of the true population survival rate 3 months after the several payments is between 99.23% and 99.5% for customers identified as survived for the first 3 months.

Time	n.risk	n.event	Survival	std.err	lower 95% CI	upper 95% CI
0	13895	17	0.999	0.000297	0.998	0.999
1	13866	23	0.997	0.000455	0.9961	0.998
2	13833	21	0.996	0.000561	0.9944	0.997
3	13803	26	0.994	0.00067	0.9923	0.995
4	13755	47	0.99	0.000831	0.9886	0.992
5	13683	28	0.988	0.000913	0.9864	0.99
10	13413	187	0.975	0.001336	0.972	0.977
60	11651	823	0.913	0.002439	0.9079	0.917
110	8863	1119	0.817	0.003483	0.8103	0.824
160	5595	1195	0.694	0.004448	0.6847	0.702
210	2727	763	0.571	0.005518	0.5597	0.581
260	28	598	0.188	0.020408	0.1503	0.23
359	1	8	0.114	0.025136	0.0704	0.168

Table 5. 11.: Nelson-Aalen estimator

b. Graphical comparison

The graph below shows the different estimates of the survival functions to evaluate potential differences.

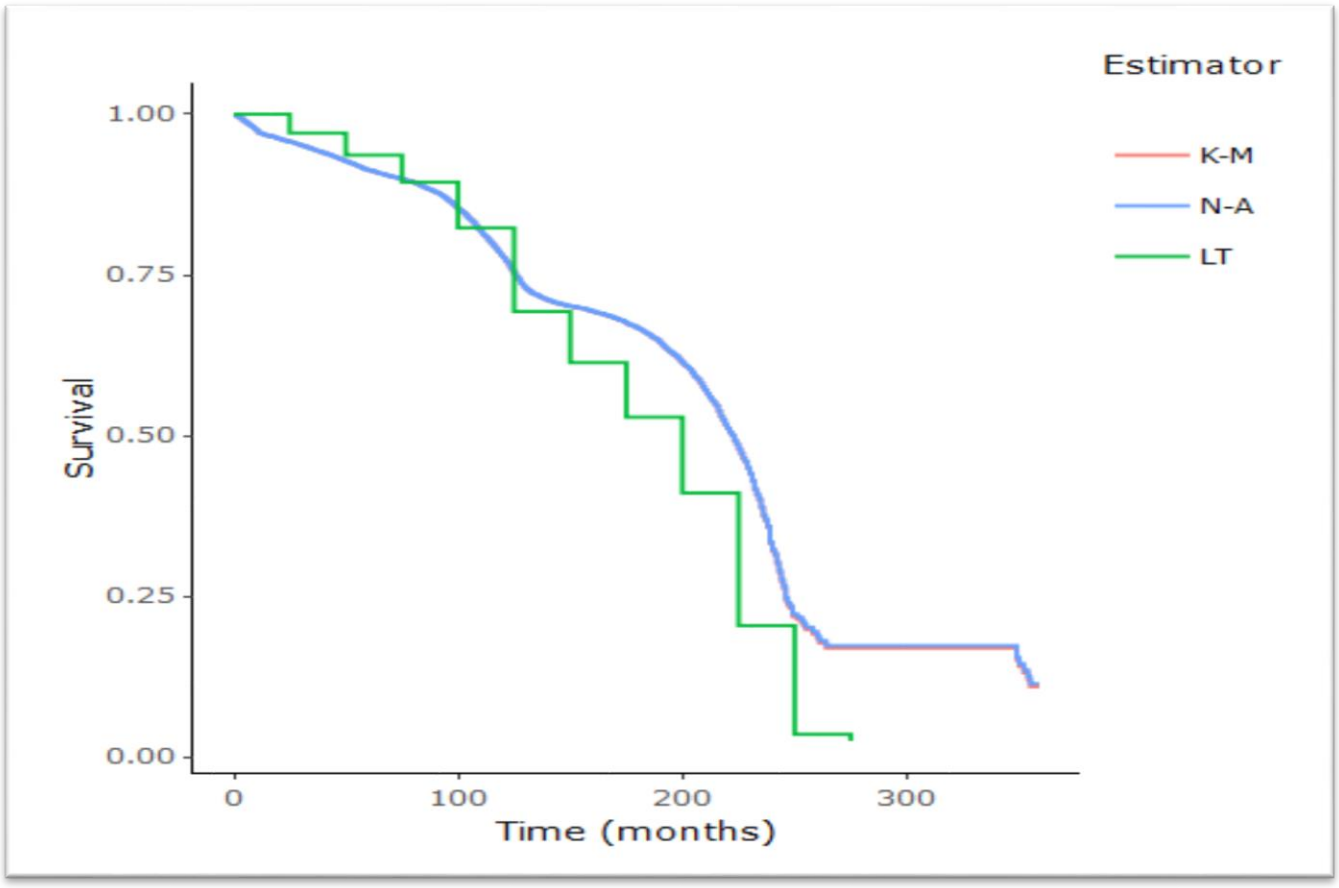


Figure 5. 13.: Comparison of Survival Estimates

K-M and N-A show that they are superior to LT. It suggests that it predicts survival times better than LT.

c. Measures of central tendency

The results in the table below show that in 222 months, 50% of the population is projected to have failed/defaulted in a mortgage loan. In 126 months, 25% of the population is projected to default, [0,127). The second 25% population is expected to default, [127,223).

Q	km.quantile	km.lower	km.upper	fh.quantile	fh.lower	fh.upper
25 0.25	126	124	128	126	124	128
50 0.5	222	220	224	223	220	224
75 0.75	246	244	253	246	244	254

Table 5. 12.: Quartile Estimates

The following figure is a graphical presentation of the estimated quantities (based on the survival curve using K-M):

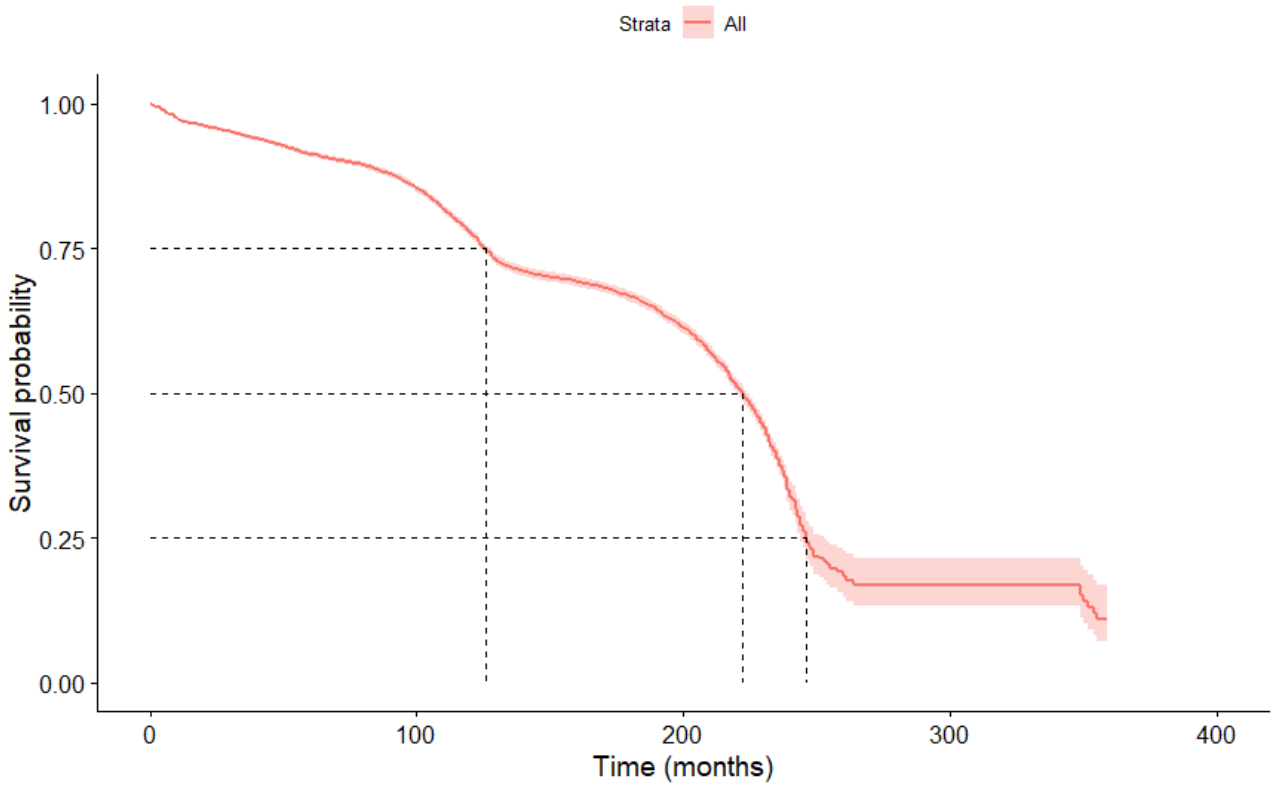


Figure 5. 14.: Survival probability of estimated quantities

d. Comparison of Survival Curves

Assume that you unsure that the survival function is different among numerous sets in your study (other groups seem to default a a great deal faster than others). The following plot shows Product-Limit Survival Estimates (with number of subjects at risk and 95% Hall-Wellner Bands):

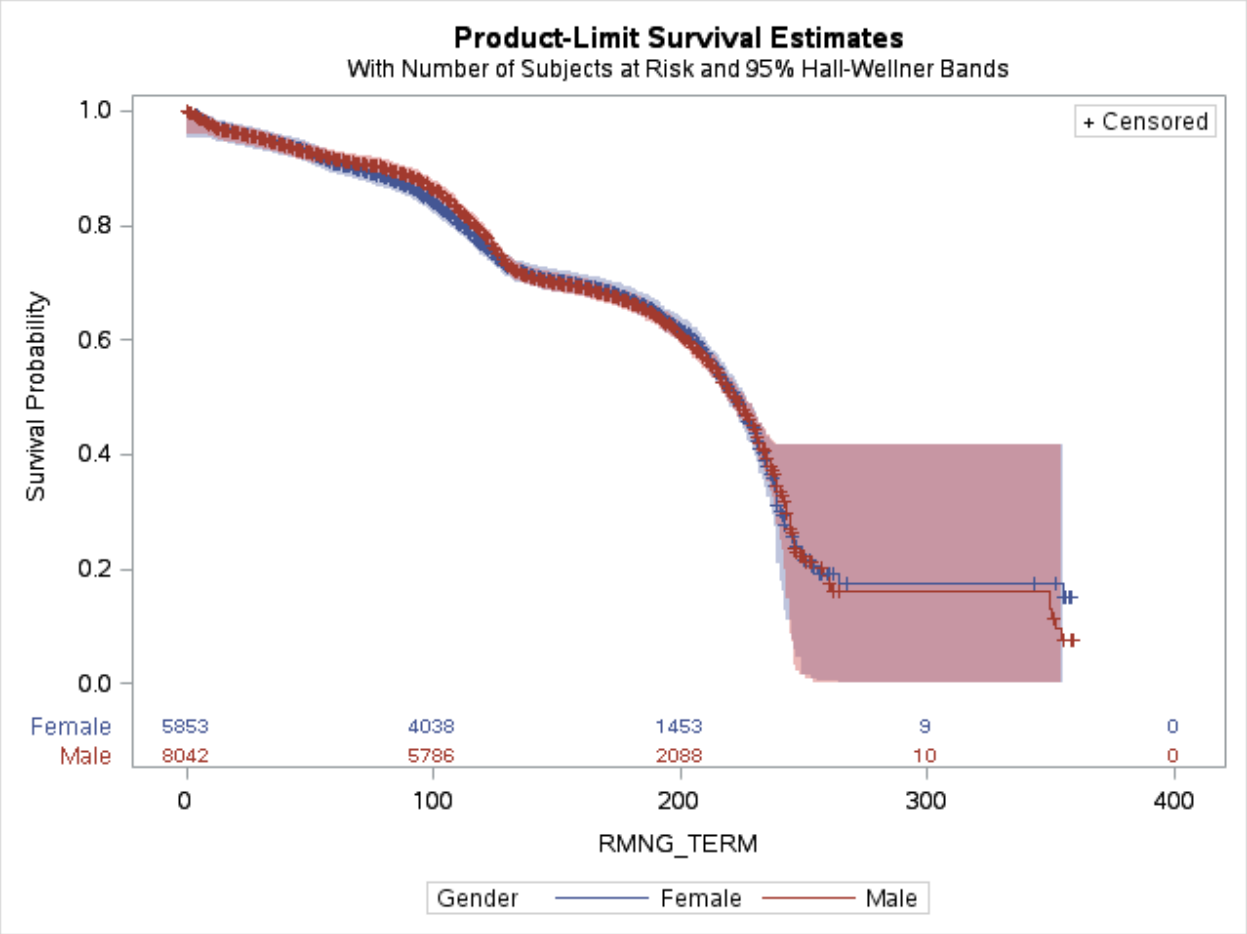


Figure 5. 15.: Estimated Survivor Functions of Genders

The above figure shows the K-M estimator stratified by sex; it shows that each gender had a poorer survival experience. This can be strengthened by the following 3 important assessments of equality shown in Table 5.13, below.

In the results, three Chi-square were discovered grounded by Test of Equality over Strata, whereas it supports our thought that survival varies among sex. The table below illustrates the Test of Equality over Strata:

Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.3689	1	0.5436
Wilcoxon	1.7602	1	0.1846
-2Log(LR)	0.0590	1	0.8081

Table 5. 13.: Test of Equality over Strata

Interpretation: Do NOT reject. Assumption of the null hypothesis has NOT led to an unlikely result (p-value = .5436 for the “Log-Rank test” and p-value = .1846 for the “Wilcoxon test”). It does not have statistically significant evidence that the survival distributions are not the same.

Cumulative hazard function

It is the integration of the hazard function. It can be understood as the likelihood of default at period x given survival until time x:

$$H(x) = \int_{-\infty}^{x} h(t)dt$$

Thus $H(t) = -In(1 - F(x))$

The following graph illustrates cumulative hazard rate by income:

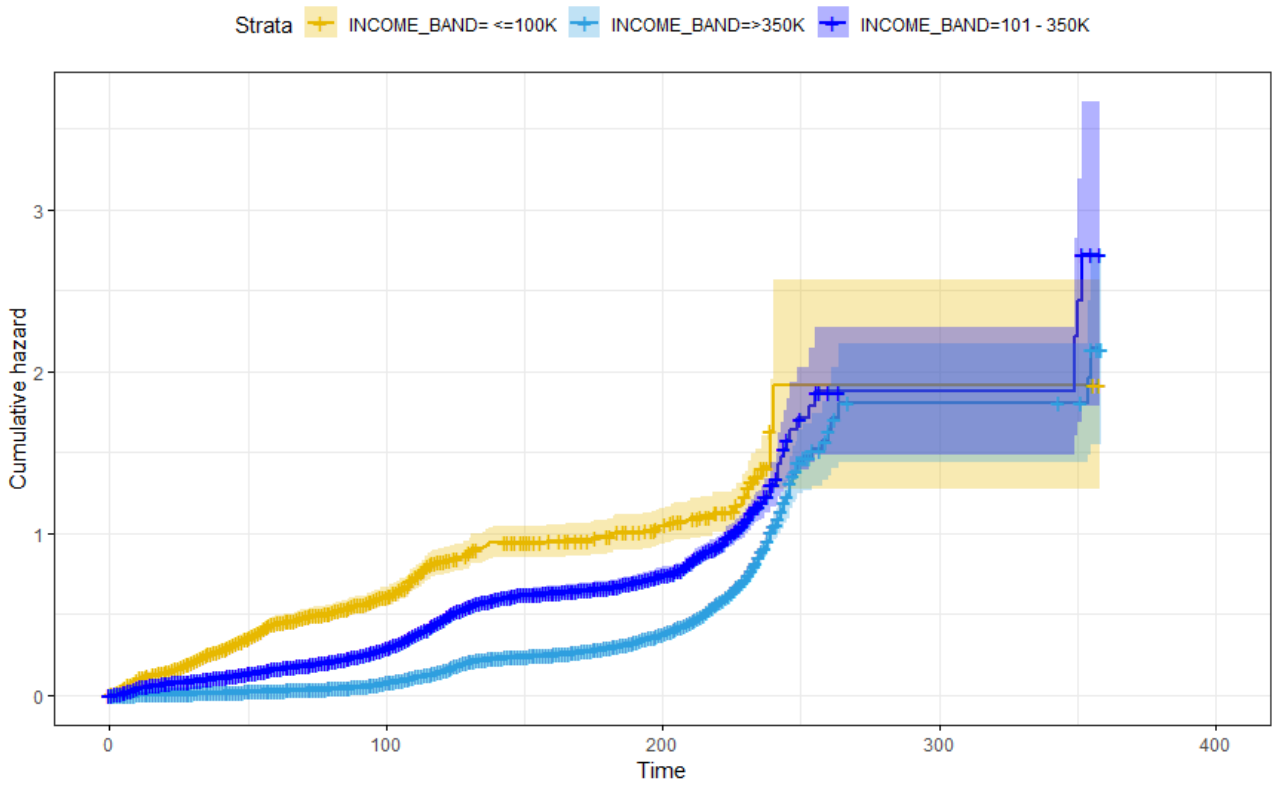


Figure 5. 16.: Estimation of Hazard Rate by Income band

The shapes in the plot are labelled by the income band in each group. From the graph, the hazard function looks flat at the start of period and increases until it is constant at around 250 months and stays same over time.

	N	Actual	Predicted	(A-P)^2/P	(A-P)^2/V
INCOME_BAND= <=100K	1267	583	212	649	690
INCOME_BAND=>350K	9483	2880	3749	201	899
INCOME_BAND=101 - 350K	3145	1392	894	278	343

Chi-squared= 1152 on two degrees of freedom, p= <2e-16

The “log-rank test” for change in survival shows p-value of p = <2e-16, proving that the income band groups differ significantly in survival. The following graph shows cumulative hazard by gender:

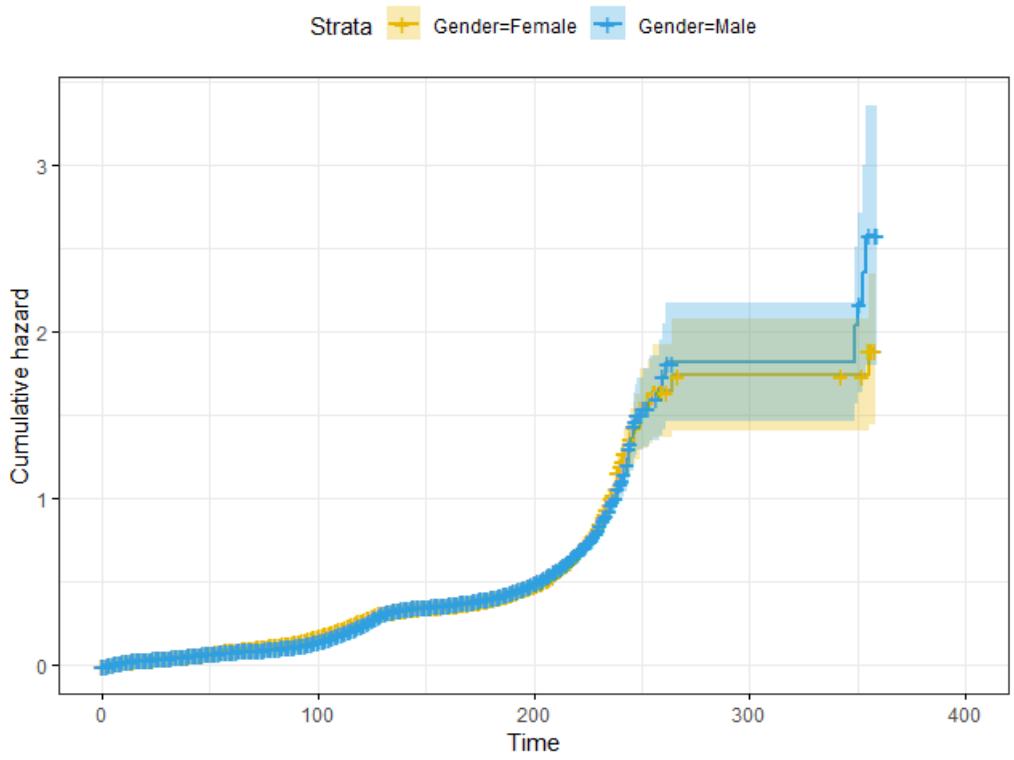


Figure 5. 17.: Estimation of Hazard Rate by Gender

	N	Actual	Predicted	(A-P)^2/P	(A-P)^2/V
Gender=Female	5853	2034	2013	0.215	0.369
Gender=Male	8042	2821	2842	0.152	0.369

Chi-squared= 0.4 on one degrees of freedom, p-value= 0.5

The “log rank test” for change in survival shows p-value = 0.5, proving that the gender group does not vary significantly in survival.

5.2.1.4. Outcomes of the Cox Proportional Hazard model

The feedback of the univariate results of each covariate in relation to survival time (in months) following the default, was calculated for the discrete covariates. The categorical value is differentiated into different indicator levels. We fit the multivariable Cox PH model that contains education level, customer credit risk score, past due amount, debit interest rate, term paid and monthly repayment amount, and default status. These covariates are significant at 5% level of significance.

```
  n= 13895, number of events= 4855

                              coef           exp(coef)      se(coef)         z       Pr(>|z|)
Highest_Highest_Education_Level             1.769e-01     1.193e+00     1.532e-02     11.542
<2e-16 ***
Client_Bureau_score          -5.142e-03     9.949e-01     1.474e-04    -34.884     <2e-16 ***
Previous_Amount_Due           2.469e-05     1.000e+00     8.607e-07     28.685     <2e-16 ***
Mortgage_Interest_Rate        1.054e-01     1.111e+00     1.186e-02      8.888     <2e-16 ***
Term_Paid                     1.563e-02     1.016e+00     1.840e-04     84.969     <2e-16 ***
Month_Repayment_Value        -9.329e-05     9.999e-01     4.334e-06    -21.528     <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1

                            exp(coef)         exp(-coef)   lower .95   upper .95
Highest_Highest_Education_Level             1.1935        0.8379       1.1582      1.2298
Client_Bureau_score           0.9949        1.0052        0.9946      0.9952
Previous_Amount_Due           1.0000        1.0000        1.0000      1.0000
Mortgage_Interest_Rate        1.1112        0.9000        1.0856      1.1373
Term_Paid                     1.0158        0.9845        1.0154      1.0161
Month_Repayment_Value         0.9999        1.0001        0.9999      0.9999

Concordance= 0.903 (se = 0.005)
Rsquare= 0.441   (max possible= 0.998 )
Likelihood ratio test= 8080   on 6 df,   p=<2e-16
Wald test            = 12074  on 6 df,   p=<2e-16
Score (logrank) test = 10863  on 6 df,   p=<2e-16
```

Table 5. 14.: Results of the univariable proportional hazards Cox regression model of mortgage loans

The asymptotic significance of likelihood ratio test (LRT), Wald test, and Score (logrank) test are less than 0.05, demonstrating that the model is significant. The analysis shows that the assessment statistics are in nearby agreement, and the null hypothesis is fast disallowed.

The variables education level, customer credit risk score, past due amount, debit interest rate, term paid and monthly repayment amount are significant (p-value <0.05).

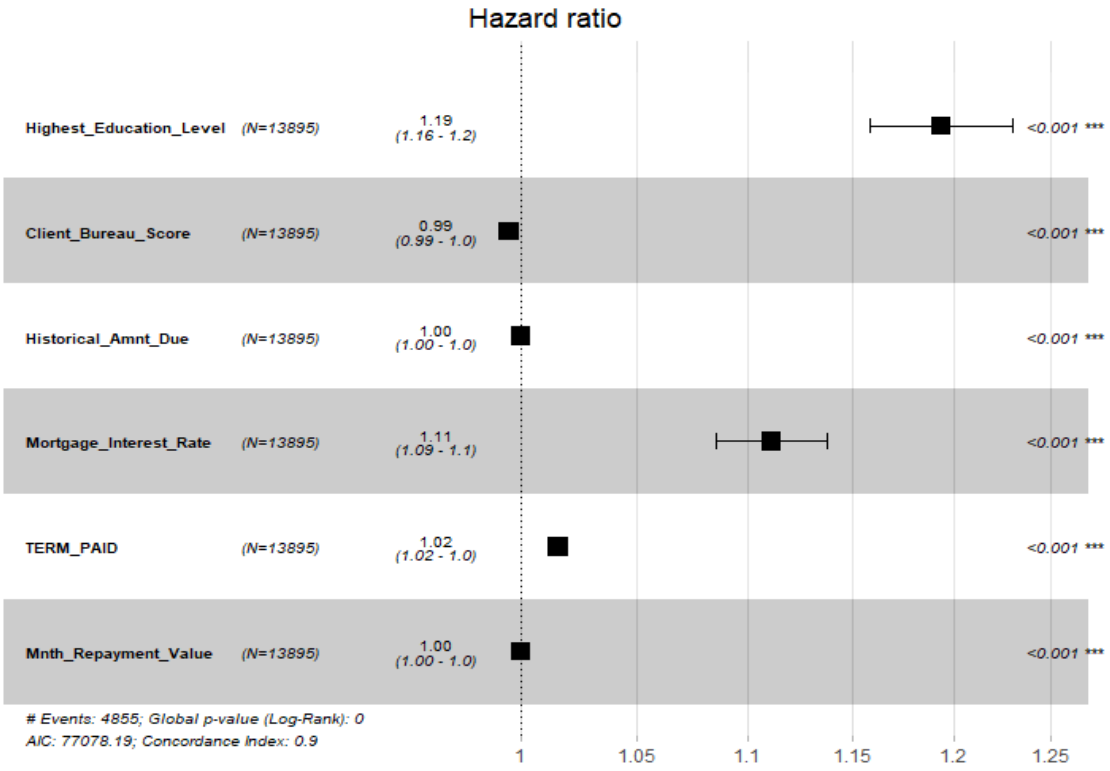


Figure 5. 18.: Hazard ratio of multivariate Cox PH

The p-value for debit interest rate is less than 0.05, by means of a hazard ratio HR = “exp (coef)” = 1.11, signifying a solid connection among mortgage interest rate and raised risk of mortgage loan default. Bring to a standstill the additional covariates the same, an advanced value of debit interest rate is related with a weak survival.

The asymptotic significance for highest education level is less than 0.001, with “hazard ratio HR” = “exp (coef)” = 1.19, 95% CI from 1.16 to 1.12. The hazard rate increased, a significant change (p-value <0.05). Customers with or without matric (Grade 12) are subject to danger of defaulting. Bring to a standstill the remaining covariates the same, an advanced education level is related to a weak survival.

The client bureau score has less asymptotic significance of 0.05. The Hazard Ratio is 0.99, with a 95% CI from 0.99 to 1. Since the CI for Hazard Ratio contains one, our analysis shows that client bureau score contributes less to the change in the Hazard Ratio. For instance, holding the other covariates the same, other credit risk score makes risk of default by a factor Exp (beta) = 0.99, or 1%, which is not contributing significantly.

The asymptotic significance for term paid is now 0.23. Whereas “Hazard Ratio = exp (coef)” = 1.02, with a 95% CI of 1.02 to 1.0. Due to the CI for Hazard Ratio take in 1, our findings show that term paid play a little role, and drift near significance. For instance, bring to a standstill the remaining covariates the same, a further term paid bring everyday risk of default by an aspect of “exp (beta)” = 1.02, or 2%, which is not an important influence.

The following section checks whether the data are sufficiently reliable with the hypothesis of proportional hazards regarding each of the covariates separately as well as globally.

5.2.1.5. Diagnoses of the model

1. ***Assessment of the PH***

The findings of tests of other time-dependent covariates in Table 5.15 were not significant (highlighted in yellow), thus no sufficient proof to reject proportionality assumption for some of the covariates at 5% level of significance.

	RHO	CHISQ	P
Highest_Education_Level	-0.0203	1.018	0.313
Client_Bureau_Score	0.0366	2.4899	0.115
Historical_Amnt_Due	0.0019	0.0431	0.836
Mnth_Repayment_Value	0.3152	534.801	2.55E-118
Mortgage_Interest_Rate	0.0753	11.3636	0.0007
Term_Paid	0.1353	40.47	2.00E-10
Highest_Education_Level: Client_Bureau_Score	0.0087	0.1829	0.669
Historical_Amnt_Due: Mnth_Repayment_Value	-0.0197	5.132	0.0235
Client_Bureau_Score: Mortgage_Interest_Rate	-0.059	6.9573	0.0084
Mnth_Repayment_Value: Term_Paid	-0.1474	104.3876	1.66E-24
GLOBAL	NA	921.3968	1.58E-191

Table 5. 15.: Result of test of proportionality assumption containing the variables in Table 5.14 and their interaction

"The **Schoenfeld residual** is the covariate value for the individual that failed, minus its expected value (Yields residuals for everyone who failed, for each covariate)."

The figures of the Scaled Schoenfeld Residuals and the lowness smooth curves are shown below in Figures 5.19(a) – 5.19(e) support the assumption of proportional hazards for individual covariates. That is, each subplot in the figure is random, smooth and approximates a horizontal through zero or slope approximately equal to zero. The plots show the variables that support the proportional hazards assumption.

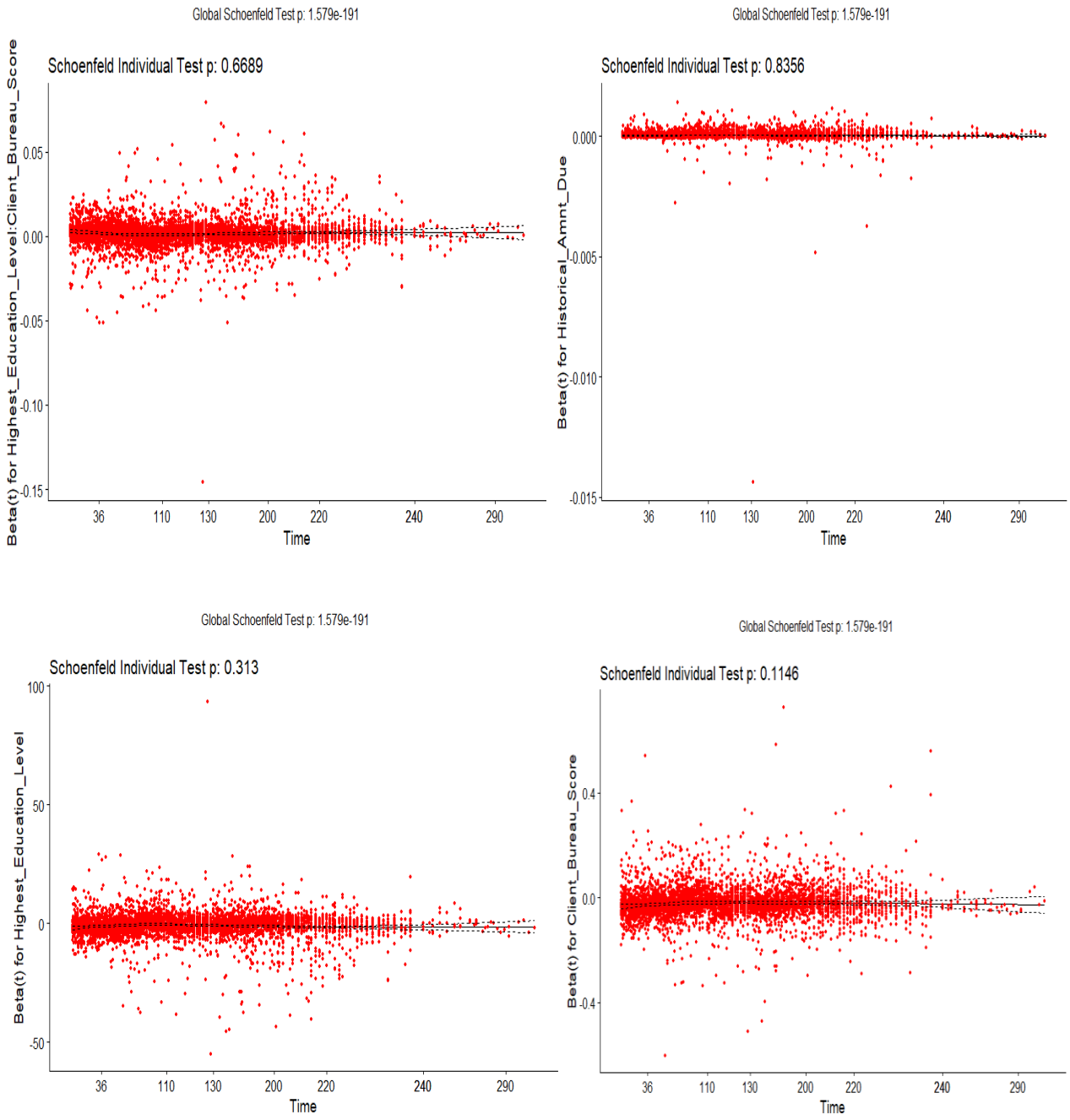


Figure 5. 19.: Graphs of the scaled Schoenfeld residuals and their Loess smooth curves for the covariates: (a) highest Education level and Client Bureau score interaction, (b) Historical Amount Due, (c) Education level, and (d) Client Bureau Score.

2. Identification of Influential diagnostics

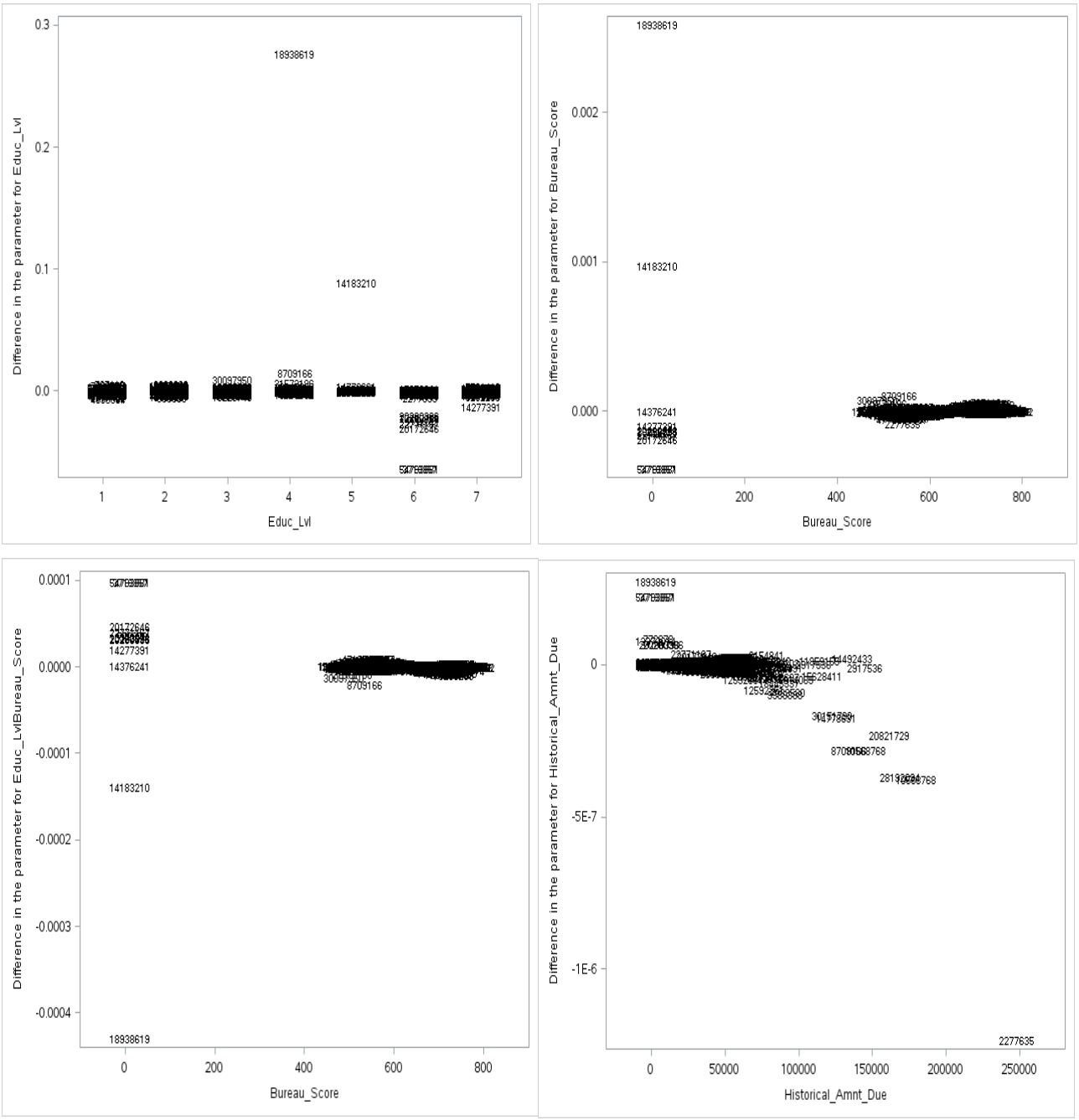


Figure 5. 20.: Plots of the score residuals for Credit Risk Score, Education level, Education level by Credit Risk Score interaction, and Past Due Amount.

The dfbetas for historical amount due and bureau risk score look small compared to regression coefficients themselves ($\hat{\beta}_{historicalamountdue} = 1.2e^{-05}$ and $\hat{\beta}_{bureauriskscore} = -0.01654$) for many parts: nevertheless, client id 18938619 has a large, negative dfbeta for historical amount due. 14183210 and 18938619 are identified as important on the linear education level coefficient ($\hat{\beta}_{educationlevel} = -1.04246$), and their huge positive dfbetas recommend they are drawing up the coefficient for education level when they are involved.

These two outcomes, client id = 18938619 and client id = 14183210, have little highest education level, 4 (Bachelor's and Graduate) and 5 (Diploma). Exclusion of these two outliers causes no change in coefficient ($\hat{\beta}_{educationlevel} = -1.04246$) even though their possessions are solid; thus, the

data of outliers are not at fault, and the consequences are that the whole effects are not affected if we do not include them, so they will be added in the model. The following tables illustrate two outliers we identified:

Obs	Client_ID	Highest_Highest_Education_Level	Client_Bureau_Score	Historical_Amount_Due
8345	14183210	5	524	38891.49
8346	14183210	5	0	32333.85
10964	18938619	4	0	0.00

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
Highest_Highest_Education_Level	1	-1.04246	0.10728	94.4148	<.0001	.	
Client_Bureau_Sscore	1	-0.01654	0.0009340	313.4575	<.0001	.	Client_Bureau_Sscore
Historical_Amount_Due	1	0.0000120	7.51124E-7	256.5860	<.0001	1.000	Last_Amount_Due
Highest_Highest_Education_Level * Client_Bureau_Sscore	1	0.00219	0.0001680	170.5549	<.0001	.	Highest_Highest_Education_Level * Client_Bureau_Sscore

Table 5. 16**.:** parameter estimates of the variables included in the final model

The intention is not only how important observations disturb coefficients, it's how they disturb the rest of the model. The probability Displacement Score measures how much the probability of the model, which is disturbed by the whole coefficients, changes when the observation is left out. The following graph shows the probability displacement scores versus remaining terms utilising **"proc sgplot"** in SAS.

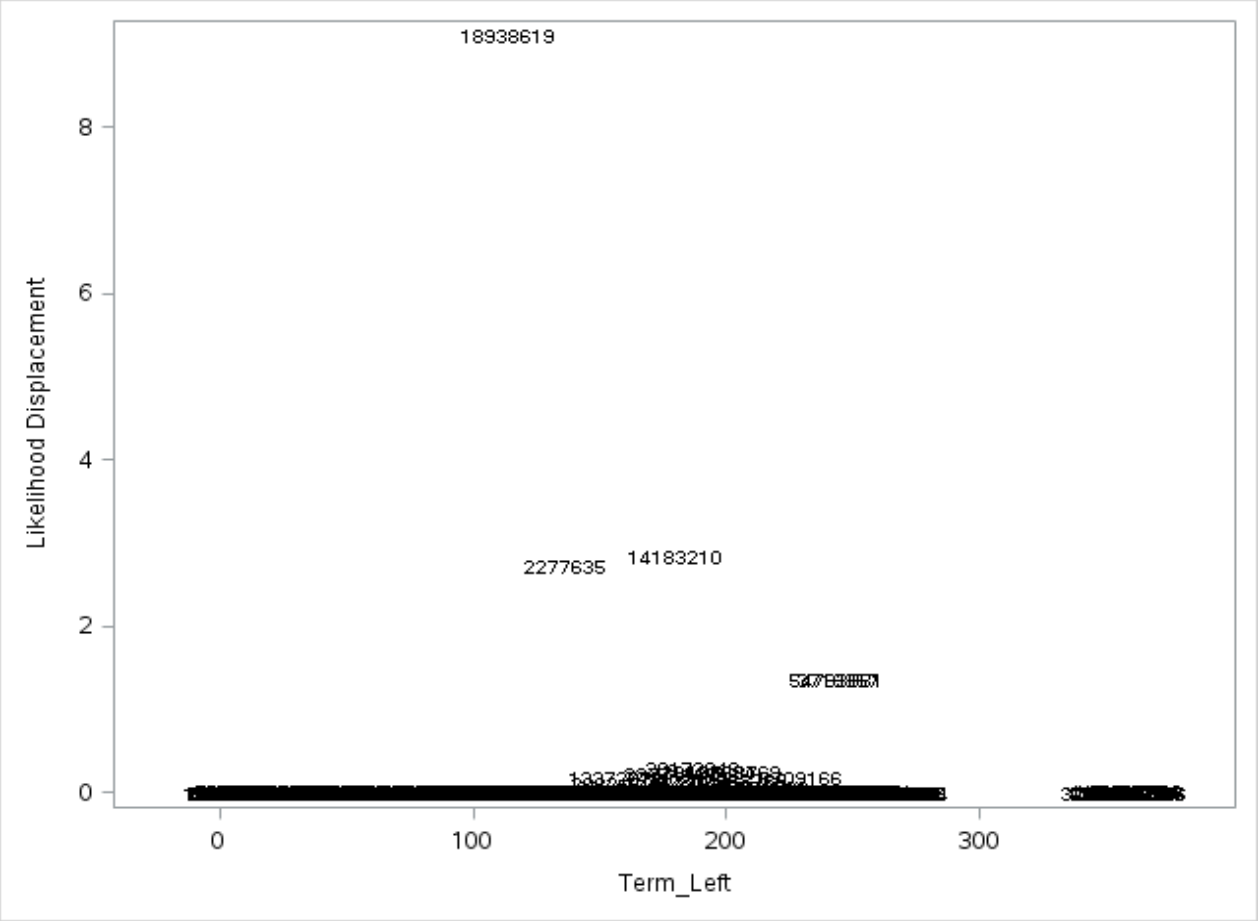


Figure 5. 21.**:** Likelihood displacement scores

Client id 14183210 and 18938619 are the outliers we identified earlier, as showing the biggest contribution on the final model, perhaps through their possessions on the education level coefficient, but, accepted that their covariate scores are realistic to keep all in the final model. In general, the plots in Figure 5.21 have shown that there is no strikingly large score residuals.

3. Poorly fit diagnostics

The cumulative hazard graph of the Cox-Snell residuals is shown in Figure 5.22. A residual is the variance between an actual data interval and projected points. A Cox-Snell residual studies the circulation and projected parameters from the lifetime regression model. The Cox-Snell residuals are identical to the negative of the natural log of the survival probability for each observation. By analysing Figure 5.22 below, we can see that the hazard function is a reasonably straight line that has a component slope and nil interception. Generally, in conclusion, the final model fits our data very well; therefore, the model, with estimates as given in Table 5.15, is the final model, highlighted in yellow.

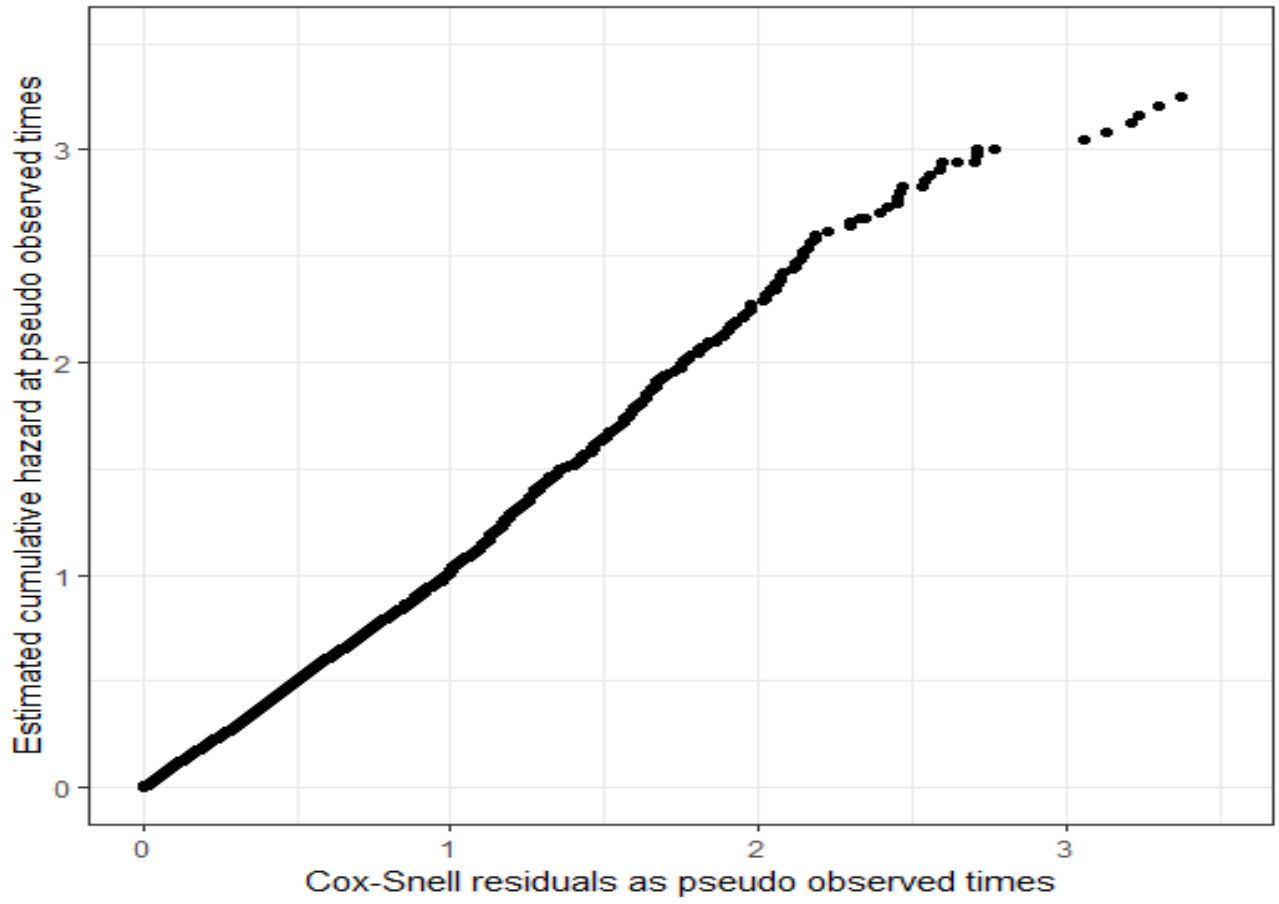


Figure 5. 22.**:** Cumulative hazard graph of the Cox Snell residuals of the proportional hazards Cox regression model in Table 5.15

The hazard function tails the forty-five-degree stripe identically, besides, for very large values of time. It is usual for models with censored information to make them wiggle at huge numbers of period, and it is not something which should cause much concern. By and large, infer that the last model fits the information exceptionally well. The following section will show how our final model is performing.

5.3. Model Performance

By hand-picking very good model from the pool of experimental statistical models, we compared the ROCs, AUROC, KS, Gini etc.

5.3.1. Logistic Regression

5.3.1.1. AUC

In Figure 5.24, AUROC curve is shown. The black stripe is the real model, and the blue stripe is the chance model. Thus, our AUROC is equal to 80.55% and greater than 80%, and is considered to be a good model.

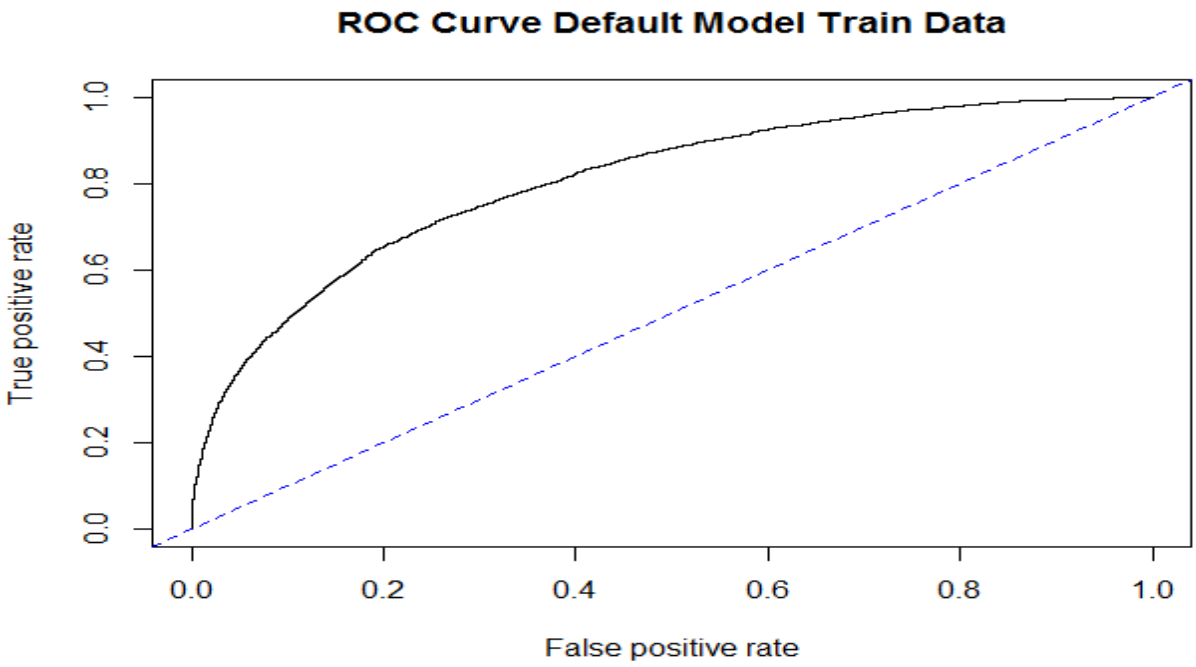


Figure 5. 23.: ROC Curve for Logistic Regression Model

5.3.1.2. Confusion Matrix

The following table illustrate the Confusion matrix in terms of defaulted mortgage loans. We tested the model's classification results against the actual observed classification. In this table it is the sensitivity "True Positive Rate (TP)" that agrees with the portion of Goods that are properly off the record (for instance in **table 5.17** below $\frac{7953}{(7953+2317)} = 0.7744$ and the specificity, "True Negative Rate" that matches to the portion of defaults that are properly off the record (can be calculated as $\frac{2538}{2538+1087} = 0.7001$).

```
Confusion Matrix and Statistics


          Observed = 0    Observed =1
Predicted = 0      7953         2317
Predicted = 1      1087         2538

                Accuracy:  0.755
                  95% CI: (0.7478, 0.7622)
     No Information Rate: 0.6506
     P-Value [Acc > NIR]: < 2.2e-16

              Kappa: 0.4276
Mcnemar's Test P-Value: < 2.2e-16

                Sensitivity: 0.8798
                Specificity: 0.5228
             Pos Pred Value: 0.7744
             Neg Pred Value: 0.7001
                 Prevalence: 0.6506
             Detection Rate: 0.5724
       Detection Prevalence: 0.7391
          Balanced Accuracy: 0.7013

           'Positive' Class: 0
```

Table 5. 17.: Confusion Matrix - Default Logistic Regression

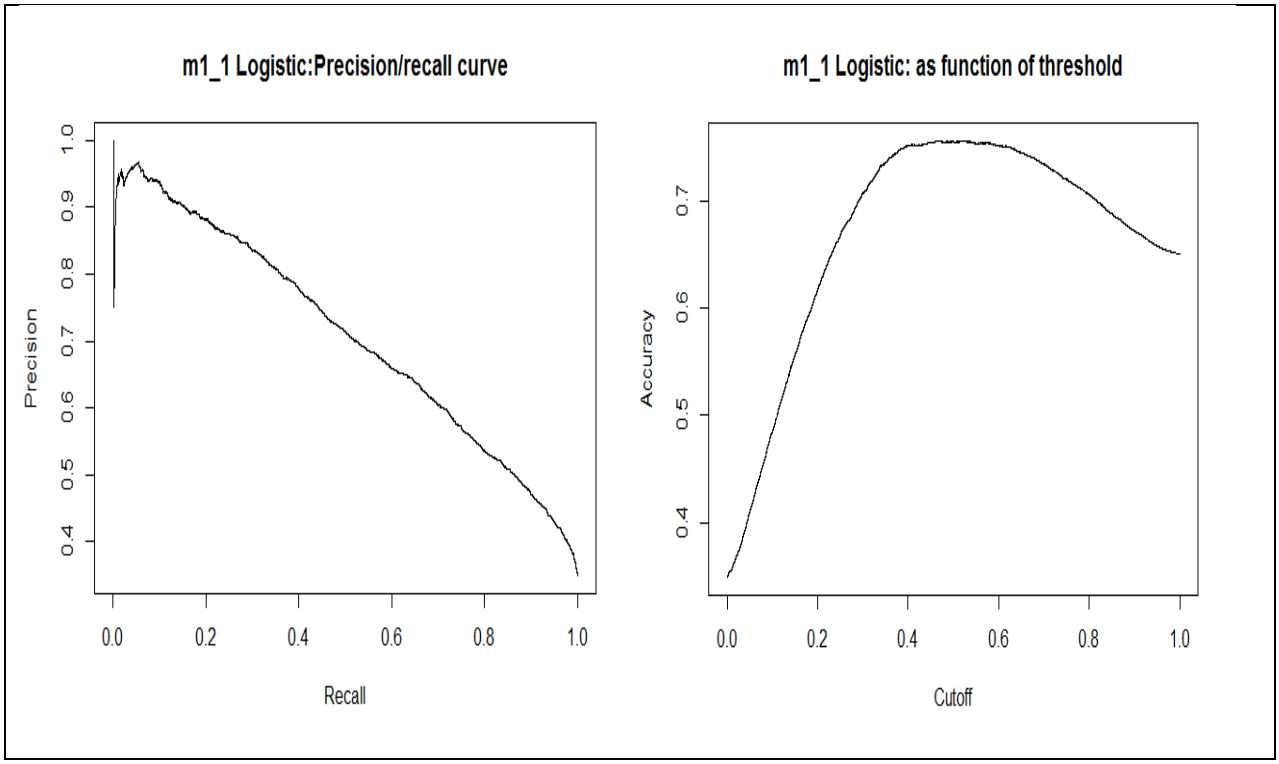


Figure 5. 24.: Left – Logistic Regression: Precision/recall curve and Right – Logistic regression: Accuracy as function of threshold

Accuracy	cutoff
0.7558834	0.4652670

The cut-off, which capitalises on the "true-positive rate" and "true-negative rate" is 0.4653. Generally, correctness of the model is 75.59%, and together true-positive and true-negative are high also; therefore, that model is correspondingly perfect at classifying credits that will go bad and credits that will not default.

5.3.1.3. Gains Table and Lift Chart

It defines as an amount of the success for classifying models intended as the fraction between the outcomes attained with and with no model. Gain and lift graphs are graphic benefits for assessing performance of classification models.

The following table shows a Gains table:

Depth of File	N	Cume N	Mean Resp	Cume Mean Resp	Cume % of Total Resp	Lift index	Cume Lift	Mean Model Score
10	1389	1389	0.86	0.86	24.60%	246	246	0.85
20	1390	2779	0.65	0.76	43.40%	187	217	0.66
30	1389	4168	0.51	0.67	57.80%	145	193	0.51
40	1390	5558	0.41	0.61	69.60%	117	174	0.4
50	1389	6947	0.31	0.55	78.30%	88	157	0.32
60	1390	8337	0.27	0.5	86.10%	78	144	0.26
70	1389	9726	0.2	0.46	91.70%	56	131	0.21
80	1390	11116	0.16	0.42	96.20%	44	120	0.15
90	1389	12505	0.09	0.38	98.90%	27	110	0.1
100	1390	13895	0.04	0.35	100.00%	11	100	0.04

Table 5. 18.: Gains table

Figure 5.25 illustrates a Lorenz curve computed from the training dataset. Each point of graph shows roughly number of specified score. Since assuming this number as cut-off point, someone can see the percentage of disallowed bad and good customers. We can see that by rejection of 25% of good customers, ~70% of rejection of bad customers at the same moment. The Gini-index is 0.39.

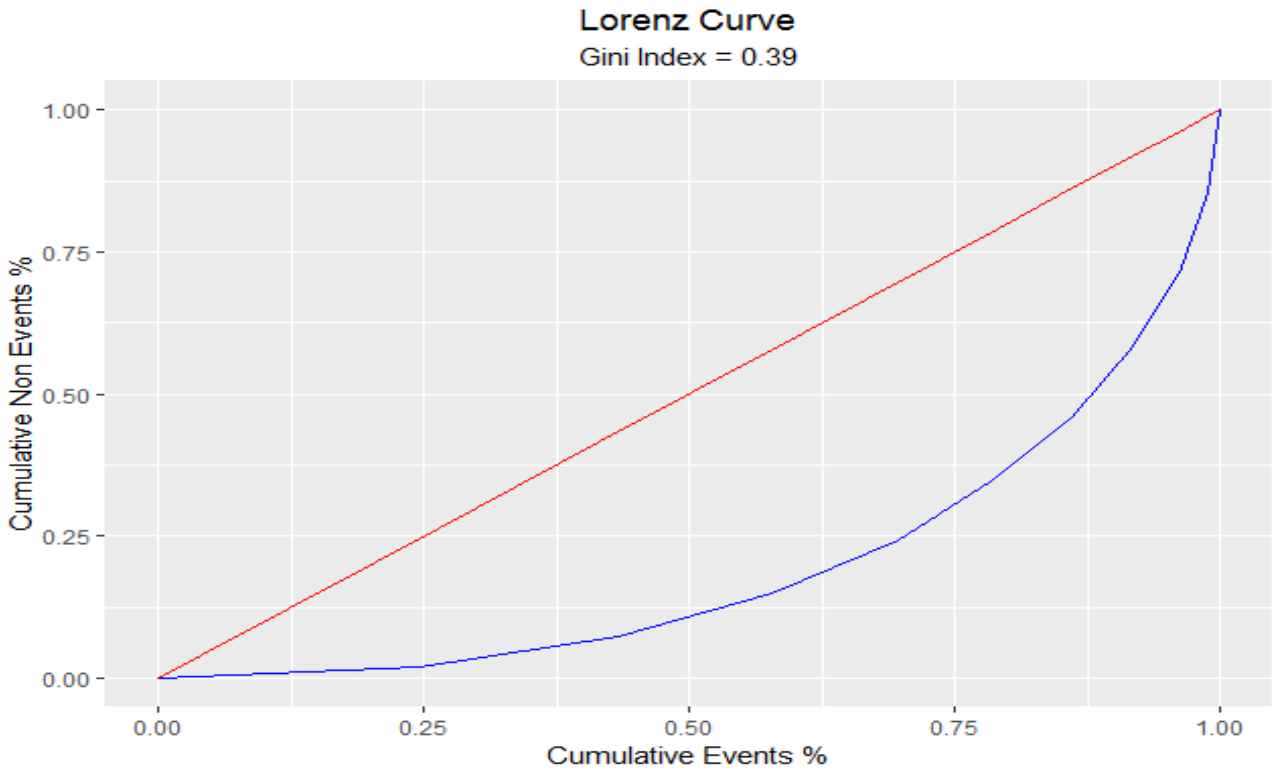


Figure 5. 25.: Lorenz Curve (ROC)

Figure 5.26 shows the Lift chart which shows quality assessment. In this instance, we have the proportion of the customers (%Population) on the horizontal axis and the proportion of default customers (%Cumulative 1s) on the vertical axis. The perfect model is signified by polyline from [0,0] through [pB, 1] to [1,1]. The benefit of this chart is that one can simply study the percentage of

rejected defaults vs. percentage of all disallowed. For example, in the event of Figure 5.26, we can realise that if we want to discard 50% of bad, we must discard about ~25% of all borrowers. That is, using the model, we can get 50% of the loan default for the top 25% of the population by only targeting the top 25% of scored customers, and this is what lift really refers to.

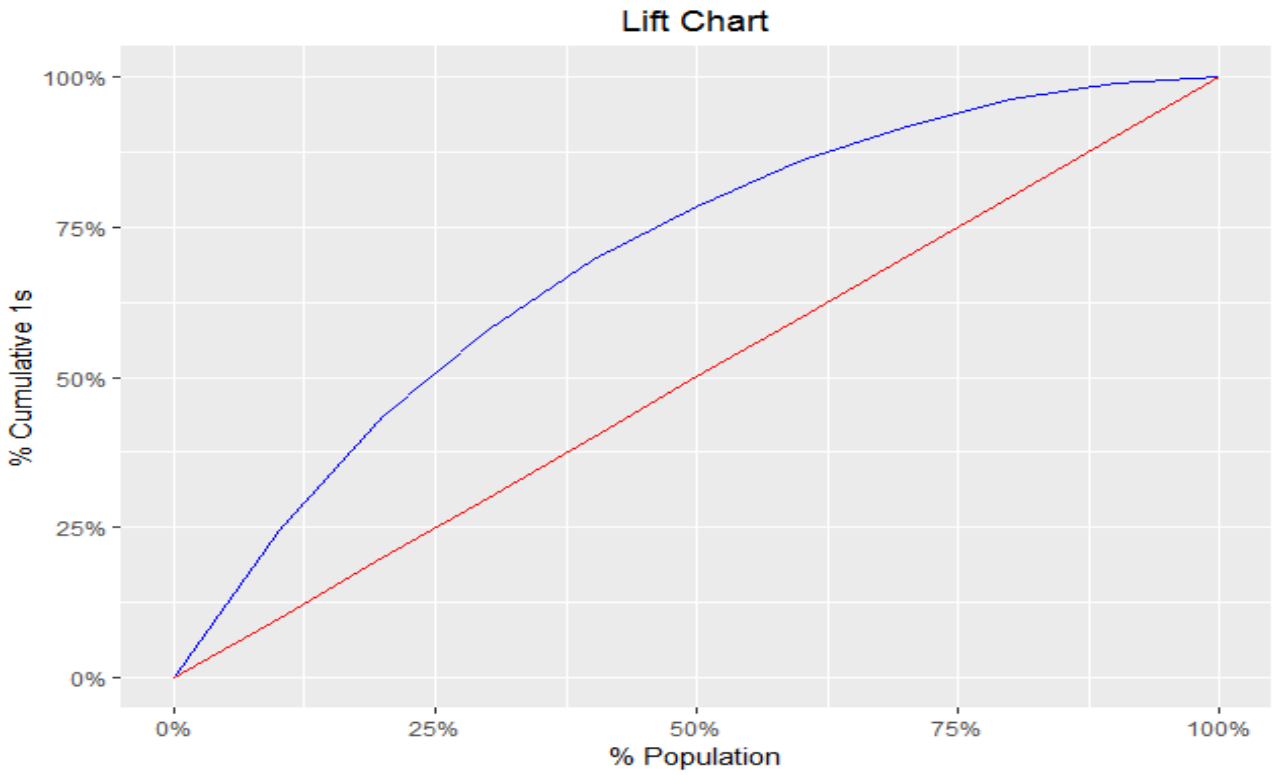


Figure 5. 26.: Lift Chart

5.3.1.4. Kolmogorov-Smirnov (KS) Goodness-of-fit Test

The calculation of the KS test is given below. Kolmogorov-Smirnov is extreme at fourth decile and K-S score is 45.4.

Decile	Total	`1`	`0`	Ks	Tp	tn	Fp	Fn	sensitivity	specificity	accuracy
1	1390	1197	193	22.5	1197	8847	193	3658	24.7	97.9	72.3
2	1390	908	482	35.9	2105	8365	675	2750	43.4	92.5	75.4
3	1390	703	687	42.8	2808	7678	1362	2047	57.8	84.9	75.5
4	1390	569	821	45.4	3377	6857	2183	1478	69.6	75.9	73.7
5	1390	426	964	43.5	3803	5893	3147	1052	78.3	65.2	69.8
6	1390	379	1011	40.1	4182	4882	4158	673	86.1	54	65.2
7	1390	275	1115	33.5	4457	3767	5273	398	91.8	41.7	59.2
8	1390	214	1176	24.9	4671	2591	6449	184	96.2	28.7	52.3
9	1390	130	1260	13.6	4801	1331	7709	54	98.9	14.7	44.1
10	1385	54	1331	0	4855	0	9040	0	100	0	34.9

Table 5. 19.: Logistic regression KS test

The following figure below shows KS-Test Comparison Cumulative fraction:

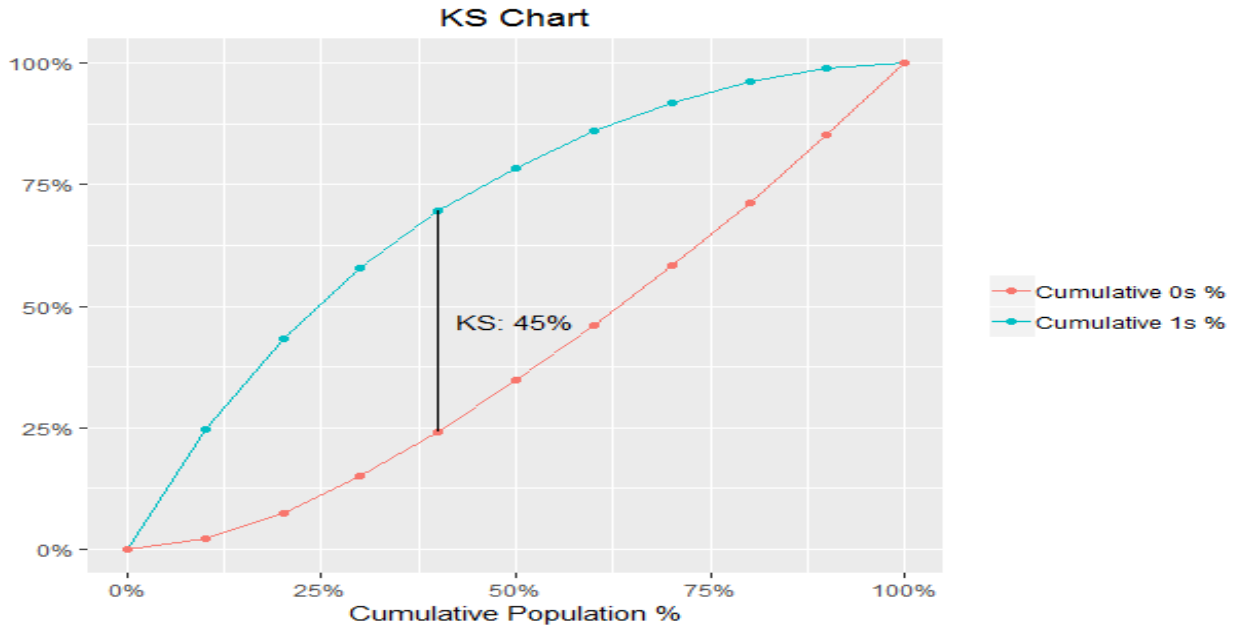


Figure 5. 27.: Empirical Distribution for KS test

The D statistic (highlighted in the above image) is the metrics that is used to report KS score. The distance statistics is the maximum change amid the cumulative distributions between Default (Y=1) and non-default (Y=0). In this logistic regression, **D=0.4583492**. Advanced the value of D; the good model distinguishes between defaults and non-defaults.

5.3.1.5. Gini Statistics

AUROC is equal to 80.55, therefore Gini = 2*0.8055 – 1 = 0.611 and therefore is greater than 0.4. The following table illustrates Model performance testing for logistic regression.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	80.6	Somers' D	0.611
Percent Discordant	19.4	Gamma	0.611
Percent Tied	0.0	Tau-a	0.278
Pairs	43889200	C	0.806

Table 5. 20.: Model performance testing for logistic regression

So, we have

- Percentage correct classification for train data - 76%
- Area under curve for train data - 0.8055
- Kolmogorov-Smirnov test value for train data - 0.458
- Hosmer-Lemeshow test for train data - p-value = 0.51 which is > 0.05
- Percentage of Concordance for train data – 80.55%

All the tests except Kolmorgorov-Smirnow test validate the model that it is good.

5.3.1.6. Measures of Accuracy

As discussed in Chapter 4.6 (Stratified Random Sampling), a random sample of 70% of observations (13,895) is selected for validation or training data, and the remaining 30% (6,152) of observations are used for testing. AUC is 0.806 for training dataset comparing with AUC of 0.802 in the testing dataset.

Model	AUC	KS	Gini
m1_1: Logistic regression Train Dataset	80.55	45.83	61.1
m1_2: Logistic regression Test Dataset	80.17	46.26	60.34

Table 5. 21.: Model Performance testing in both training and testing data for LR

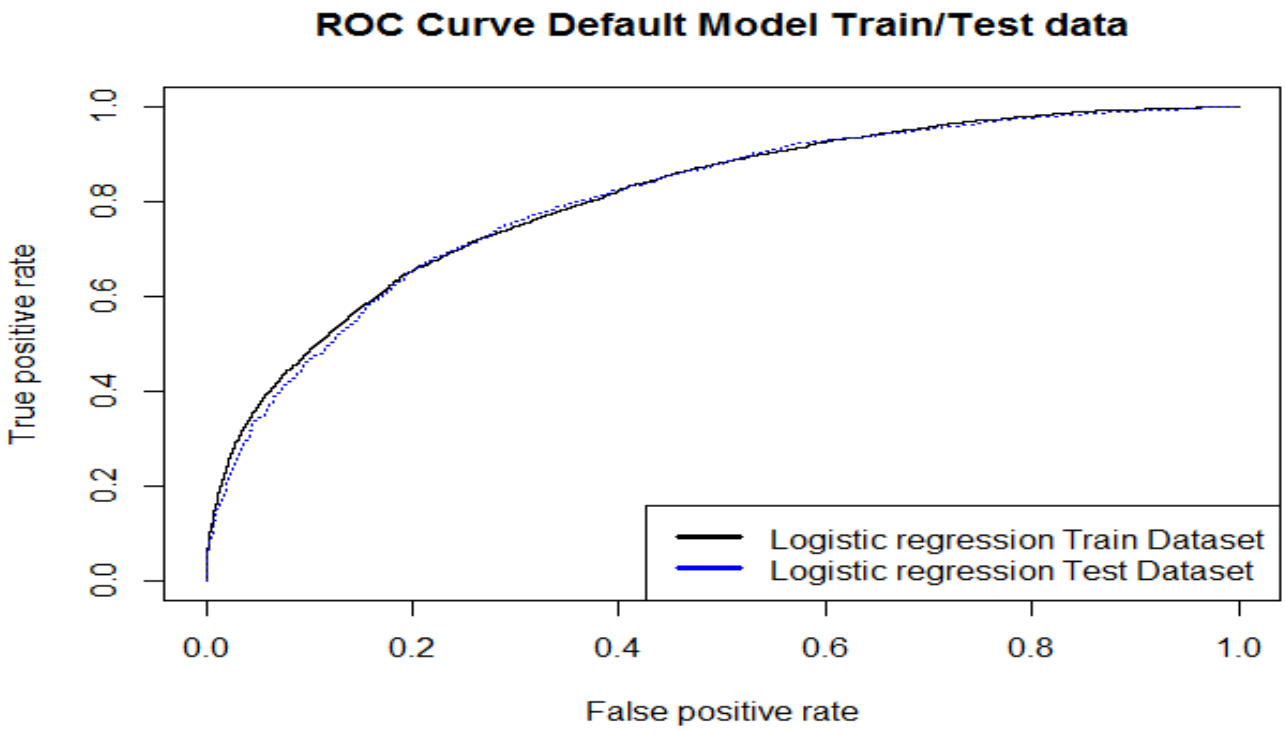


Figure 5. 28.: ROCs Model Performance Comparison for logistic regression

The coordinates of the points of the ROC curves Sensitivity and 1-Specificity are aimed at each of the Train dataset and Test dataset. The diagonal (45-degrees) line is a receiver operating characteristics (ROC) curve of random classification, and serves as a baseline. ROC curve displays the overall capability of using the score to categorise the condition. Train dataset is a better model than Test dataset, as can be seen from the above graph.

Now we can see the performance of the model on the training dataset as it is shown in the following score table:

Decile	Good	CountOfDecile	Bad	CuGood	CuBad	CuGoodPercent	CuBadPercent	CuBadAvoided
1	486	1390	904	486	904	0.1	0.1	0.9
2	505	1389	884	991	1788	0.2	0.2	0.8
3	460	1390	930	1451	2718	0.3	0.3	0.7
4	493	1389	896	1944	3614	0.4	0.4	0.6
5	507	1389	882	2451	4496	0.5	0.5	0.5
6	474	1390	916	2925	5412	0.6	0.6	0.4
7	500	1389	889	3425	6301	0.71	0.7	0.3
8	475	1390	915	3900	7216	0.8	0.8	0.2
9	479	1389	910	4379	8126	0.9	0.9	0.1
10	476	1390	914	4855	9040	1	1	0

Table 5. 22.: Score table

As we can see, the model identifies 486 good observations in decile 1 out of 1390, in decile 2, 505 good out of 1389, and so on. The difference is the bad observations. I have built variables such as

Cumulative Good, Cumulative Bad, Cumulative Good Percentage, Cumulative Bad Percentage, Cumulative Bad avoided and Profit, to obtain the conclusion for credit score. Here, Profit variable indicates that if we gain 100 for every good loan, and lose out 500 for every bad loan, we get a cut-off point to decide up to what percentage we can take the risk, up to what percentage we can give loans, and up to what point we get maximum profits. However, in this, we can take only the first decile, get a good loan of about 10% and avoid a bad loan of about 90%.

5.3.1.7. Cross Validation

Table 5.23 shows the cross-validation for the logistic regression, as well as the average coefficient estimate. Standard deviations for every factor in the model are also calculated in this table.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.536e+01	4.348e-01	35.326	< 2e-16 ***
Highest_Education_Level	3.098e-01	2.019e-02	15.345	< 2e-16 ***
Client_Bureau_Score	-2.702e-02	5.715e-04	-47.274	< 2e-16 ***
Previous_Amnt_Paid	-5.712e-06	9.499e-07	-6.013	1.82e-09 ***
Mortgage_Interest_Rate	7.111e-02	1.589e-02	4.475	7.66e-06 ***
PURCHASE_PRICE	-4.041e-07	5.143e-08	-7.857	3.95e-15 ***
Mnth_Repayment_Amnt	-6.606e-05	6.050e-06	-10.919	< 2e-16 ***

Table 5. 23.: The cross-validation for the logistic regression

The following figure is the model with logistic regression between model 1 logistic regression with k-fold cross-validation using caret, and model 2 – which is old logistic regression with training dataset.

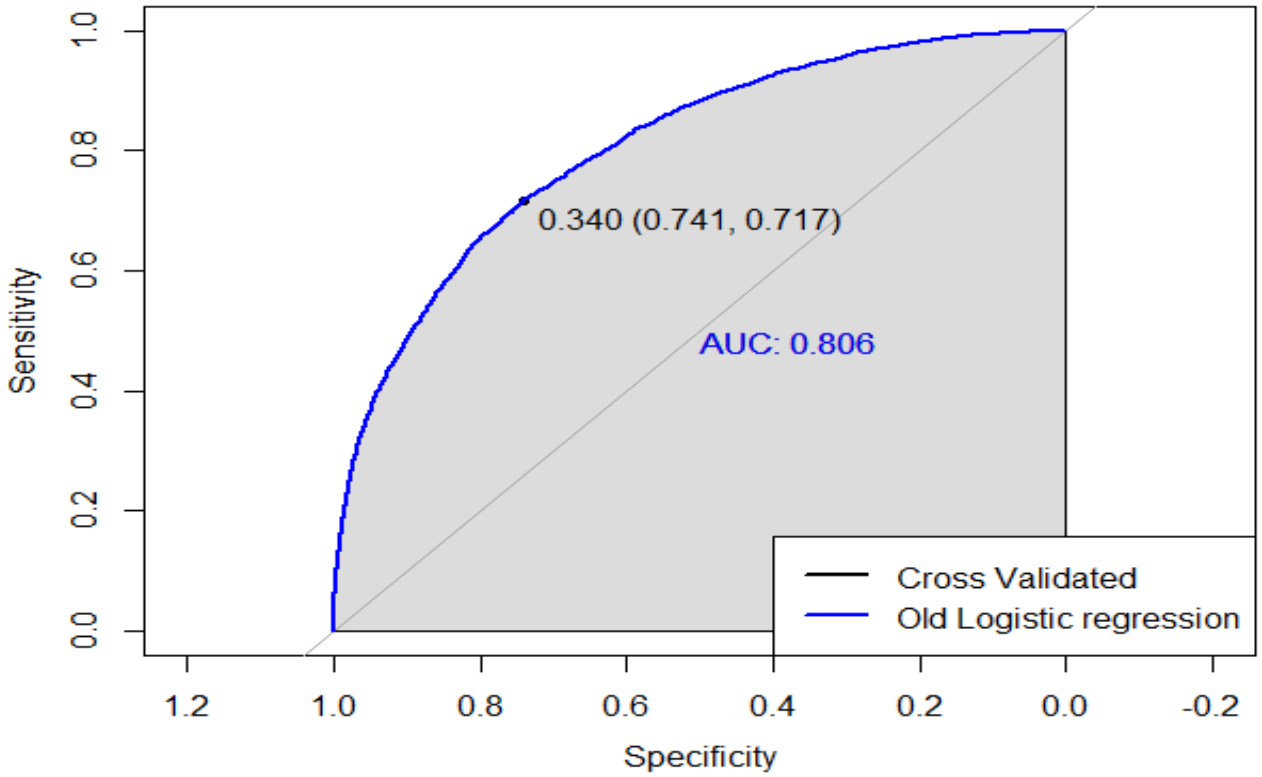


Figure 5. 29.: Comparison of logistic regression models

The models are to be identical – i.e. cross validation did not make a difference for the training data. The models have an accuracy of 76%, which I think is good for such a chaotic event. A probability threshold of 0.34 leads to the best result, according to the ROC plot.

For model stability we make prediction for test data as follows:

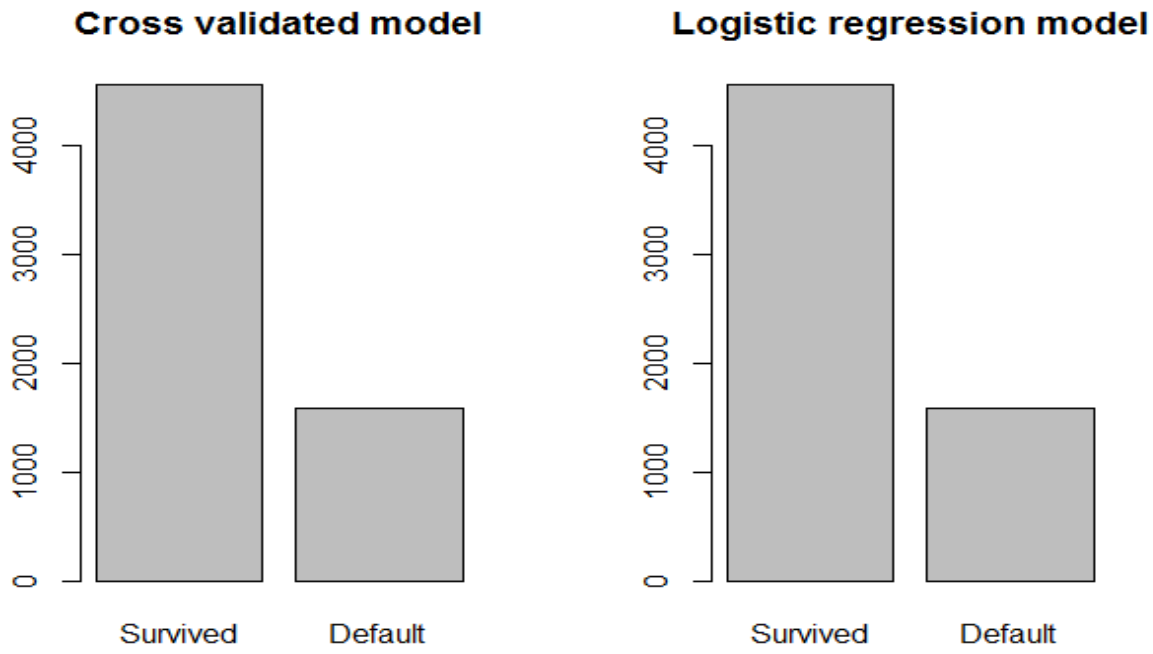


Figure 5. 30.: Prediction of test data

Exactly as was the case with the training data, cross-validation of the model did not make a difference for predicting survivors with the test data.

5.3.2. Survival Analysis

SA is defined as a division of data dealing with study of "time until an event occurs''. Several trials might be observed during the study, but only one event might be of interest. When more than one event of interest is to be considered, then the problem becomes a recurrent event or competing risk problem. Survival analysis will model for that.

5.3.2.1. AUC

The following graph shows the area under curve for Cox Regression with selected variables, using stepwise selection for training dataset. The AUROC is equal to 76.65% and greater than 70%, and is considered as being an acceptable model.

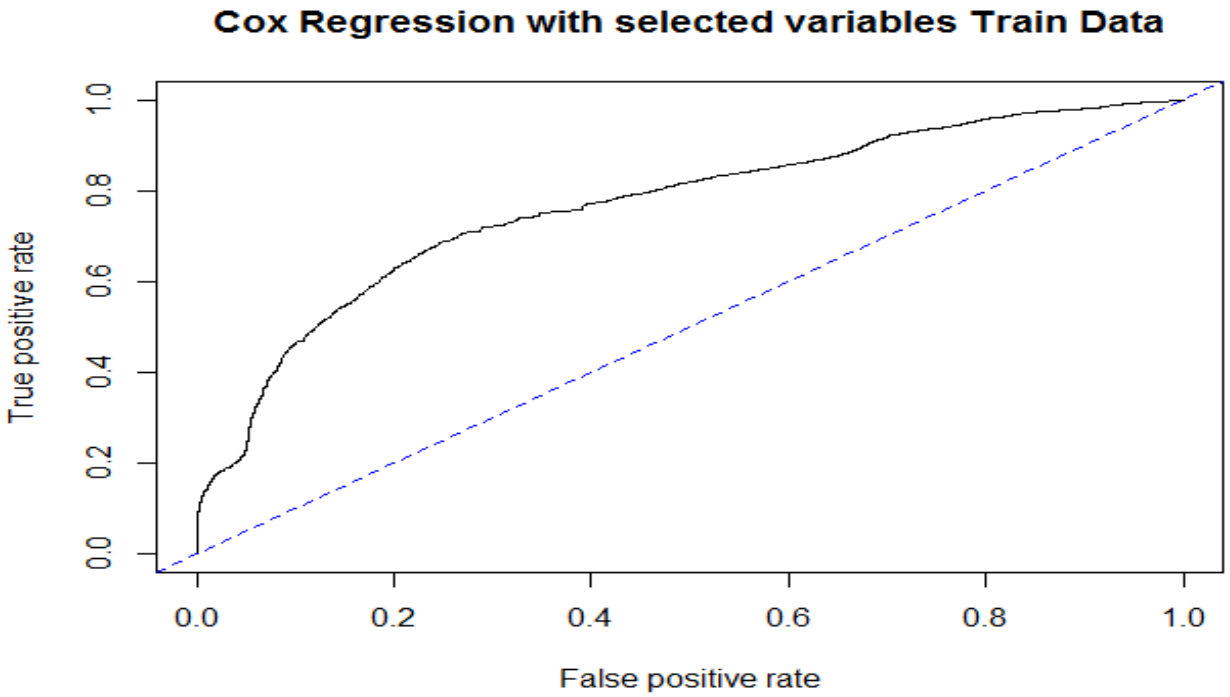


Figure 5. 31.**:** ROC Curve for Cox Regression Model

5.3.2.2. Confusion Matrix

We follow the same methodology applied in logistic regression to construct a Confusion matrix.

```
      Confusion Matrix and Statistics
train_1. pred2    0    1
              0 8559 3536
              1 481 1319

                Accuracy: 0.7109
                  95% CI: (0.7033, 0.7184)
     No Information Rate: 0.6506
     P-Value [Acc > NIR]: < 2.2e-16

                   Kappa: 0.2557
  Mcnemar's Test P-Value: < 2.2e-16
             Sensitivity: 0.9468
             Specificity: 0.2717
          Pos Pred Value: 0.7076
          Neg Pred Value: 0.7328
              Prevalence: 0.6506
          Detection Rate : 0.6160
    Detection Prevalence: 0.8705
       Balanced Accuracy: 0.6092
        'Positive' Class: 0
```

Table 5. 24.: Confusion Matrix - Default Cox Regression

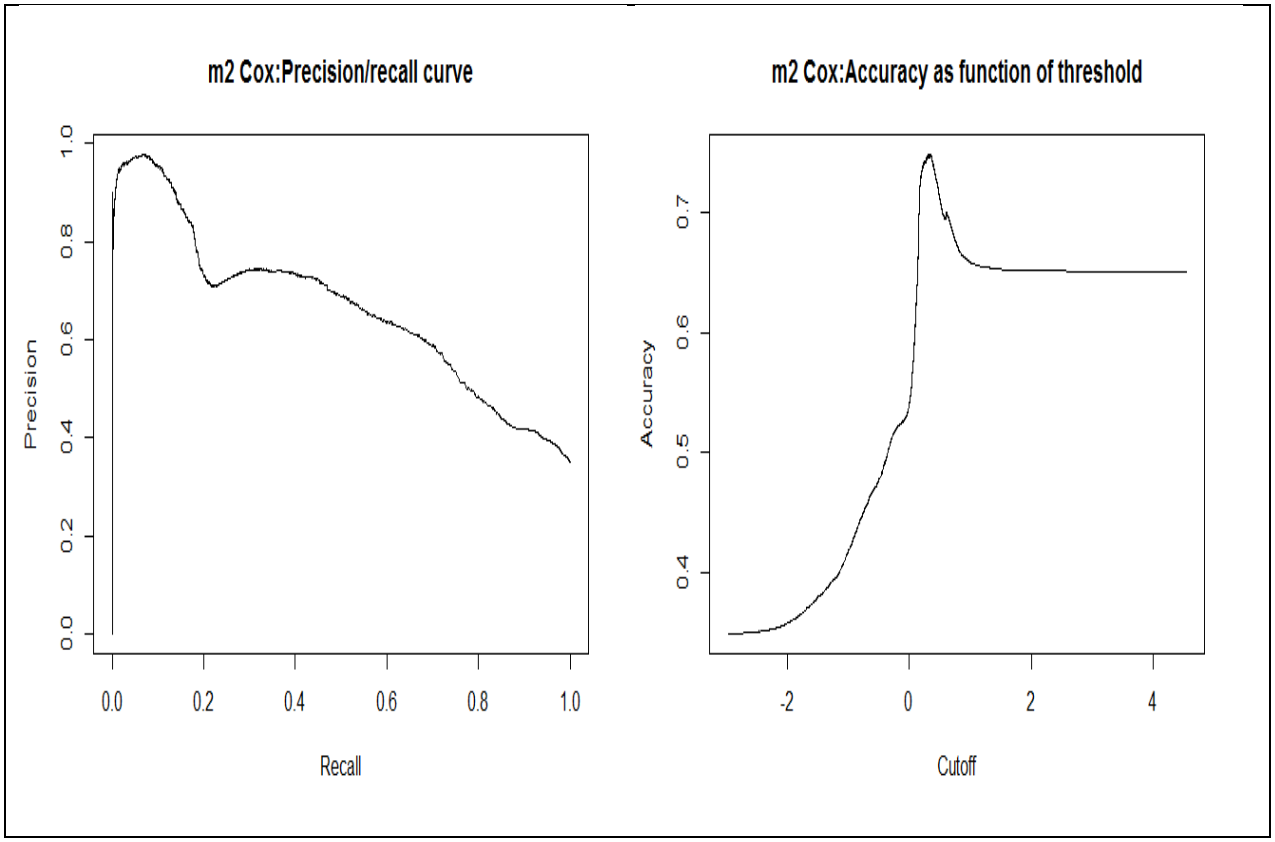


Figure 5. 32.**:** Left – Cox Regression: Precision/recall curve and Right – Cox regression: Accuracy as function of threshold

Accuracy	cutoff
0.7479669	0.3533661

The cut-off, which takes full advantage of the sensitivity (true-positive rate) and specificity (true-negative rate), is 0.3534. General correctness of the model is 74.80%, and together, sensitivity and specificity are as high, which proves that the model is the same as good at detecting loans that will go bad and loans that will not default.

5.3.2.3. Gini, KS Statistics

Survival Analysis GS is calculated as follows, **AUROCC** (Gini = 2*AUROCC - 1).

AUROC: 76.65
KS: 43.89
Gini: 53.3

So, we have

- Percentage of correct classification for train data: 71%
- Area under curve for train data: 0.7665
- KS statistic: 43.89%

All the tests thus validate the model that it is acceptable.

5.3.2.4. Measures of Accuracy

AUC is 0.78 for testing dataset comparing with AUC of 0.77 in the training dataset. Test dataset is a better model than train dataset, as can be seen from the graph and table, below.

Model	AUC	KS	Gini
m2: Cox regression Train Dataset	76.65	43.89	53.3
m5: Cox regression Test Dataset	77.87	42.52	55.74

Table 5. 25.: Model Performance testing in both training and testing data fox Cox regression

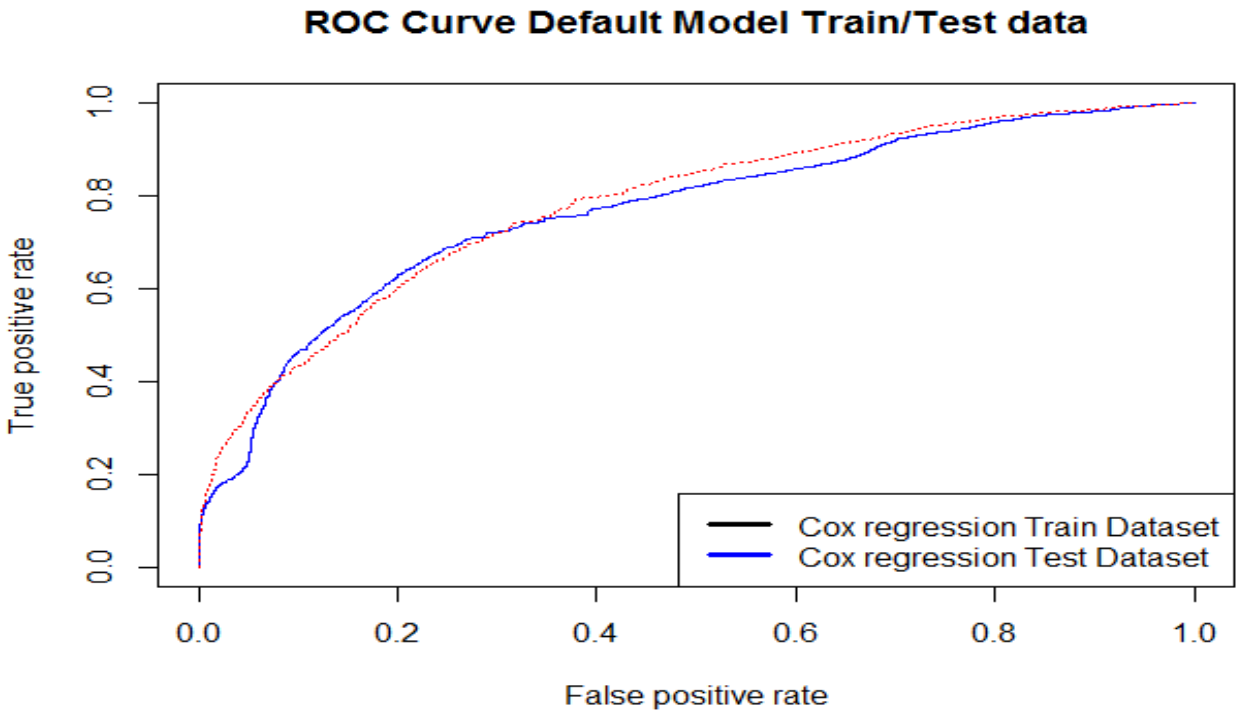


Figure 5. 33.: ROCs Model Performance Comparison for Cox regression

Now we can see performance of the model on the training dataset as shown in the following score table:

Decile	Good	CountOfDecile	Bad	CuGood	CuBad	CuGoodPercent	CuBadPercent	CuBadAvoided
1	512	1390	878	512	878	0.11	0.1	0.9
2	494	1389	895	1006	1773	0.21	0.2	0.8
3	466	1387	921	1472	2694	0.3	0.3	0.7
4	479	1391	912	1951	3606	0.4	0.4	0.6
5	489	1228	739	2440	4345	0.5	0.48	0.52
6	454	1535	1081	2894	5426	0.6	0.6	0.4
7	480	1388	908	3374	6334	0.69	0.7	0.3
8	521	1400	879	3895	7213	0.8	0.8	0.2
9	497	1385	888	4392	8101	0.9	0.9	0.1
10	463	1402	939	4855	9040	1	1	0

Table 5. 26.: Score Table for Cox regression

As we can see, the model identifies 512 good observations in decile 1 out of 1390, in decile 2, 494 good out of 1389 and so on. The difference is the bad observations. We build variables such as Cumulative Good, Cumulative Bad, Cumulative Good Percentage, Cumulative Bad Percentage, Cumulative Bad avoided and Profit to obtain the conclusion for credit score. Here, Profit variable indicates that if we gain 100 for every good loan, and lose out 500 for every bad loan, we get a cut-off point to decide up to what percentage we can take the risk, up to what percentage we can give loan, and up to what point we get maximum profits. However, in this, we can take only the first decile, get a good loan of about 11% and avoid a bad loan of about 89%.

5.4. Model Performance Comparison

The ROC plots for the Logistic regression (LR) and Cox proportional hazard regression models are shown in Figure 5.34.

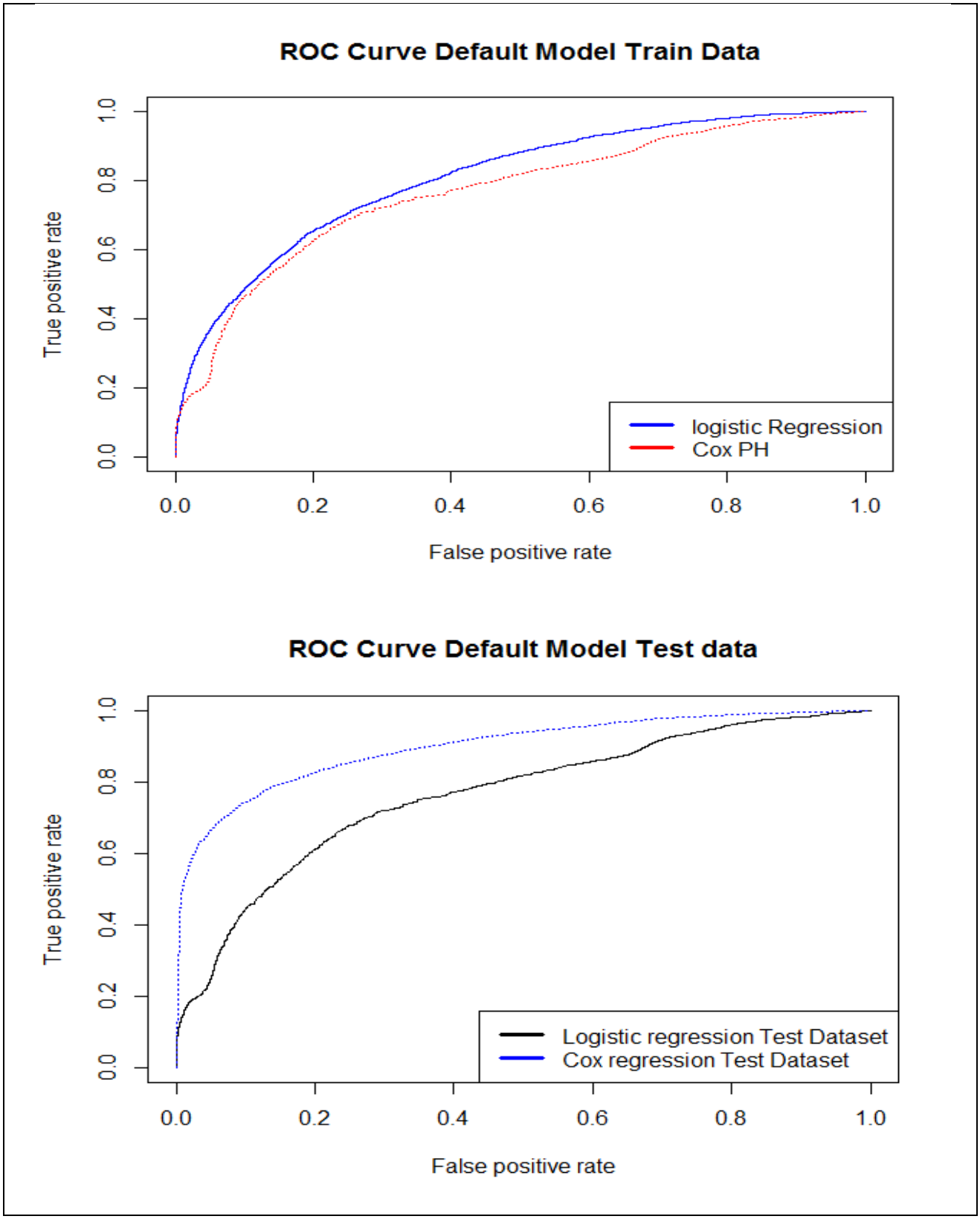


Figure 5. 34.: Receiver Operating Characteristics for both models plots

The intervals of the values of the Receiver Operating Characteristics Plots Sensitivity (True-Positive Rate) and 1-Specificity (False-Positive Rate) for individual of the single training dataset and testing dataset. The sloping (45 degrees) stripe is a receiver operating characteristics plot of chance classification, and aids as a reference point. Receiver Operating Characteristics plot expresses the general capability of utilising the score to order the condition.

Model	AUC	KS	Gini
m1_1: Logistic regression	80.51	45.8	61.02
m1: Cox regression	76.65	43.89	53.3

Table 5. 27.: Model Performance testing in training data for Logistic and Cox regression

Model	AUC	KS	Gini
m1_2: Logistic regression	89.86	64.9	79.72
m5: Cox regression	76.43	43.09	52.86

Table 5. 28.**:** Model Performance testing in testing data for Logistic and Cox regression

The results of the final modelling from both methods, namely Logistic Regression and Survival Analysis model, have shown the very same fit in terms of the AUROC, with the Logistic Regression showing better performance than the Cox regression model in the training and testing dataset. In terms of prediction of defaulted and non-defaulted home loan portfolios, the logistic regression model still outperforms survival analysis (Cox Regression) in the training and testing dataset, as can be seen in the tables above and in Figure 5.34.

5.5. Conclusion

The model building ethics were acceptably addressed in each model. All models have high Gini, implying that they can distinguish risk. An assessment of the actual and predicted observations was conducted. Model Gini, KS statistic, lift chart, gain tables and AUROC curves were calculated to determine the methodology striving to perform better for a consumer credit cohort data in the occurrence of competing risks and longstanding survivors. It was concluded that logistic regression performs far better than the Cox proportional hazard regression model.

CHAPTER 6: DISCUSSION AND RECOMMENDATIONS

6.1. Introduction

We conclude that our model-building ideals were encountered and addressed in each model. Logistic regression and survival model with time dependent were developed for modelling default behaviour of mortgage books. Survival analysis has been introduced by Narain (1992) as an alternative to logistic regression in the credit context (Dirick et al., 2017). The benefits of utilising survival analysis in a credit risk setting were that point to non-payment was modelled, and not just whether a borrower will pay or not. It gave a clear method of evaluating the expected success of a borrower, and survival models match the loan non-payment process, and subsequently integrate circumstances when a case has paid/not defaulted in the observation time (Dirick et al., 2017). To give the likelihood of non-payment in relation to the conditional distribution function of the period to non-payment, the non-parametric method was utilised (Đurović, 2017). As an alternative, the Logistic Regression model was utilised to develop likelihood of non-payment on mortgage loans.

6.2. Summary

Survival Analysis (SA)

The study of SA was an effort to discover the factors which disturb the survival of the customers with mortgage loans with a bank. I found that, from the Kaplan-Meier survival estimates, there was no statistically significant evidence that the survival distributions are not the same in gender groups. However, there were difference in survival curves of income band of customers. The log-rank test also disclosed that there was substantial transformation in survival experience among income band group (P-value less than .05). However, the log-rank test disclosed that the survival experience of customers in gender (log-rank test is p-value = .5436 aimed at the log-rank test and p-value = .1846 aimed at the Wilcoxon-test) does not differ significantly. The multivariable Cox proportional hazards regression results analysis indicated that the six covariates namely highest education level, client bureau risk score, historical amount due, mortgage interest rate, term paid and monthly repayment value were significantly associated with loan defaults among bank customers.

I have found that out of 13,895 bank mortgage loan customers, 9,040 customers were censored (65.60%) and 4,855 customers defaulted (34.94%) during the study; the middle follow up time is 222 months, 50% of the population is expected to have defaulted on a mortgage loan, 25% of the population defaulted had 126 months of the follow-up (upper quartile). This displays that the greatest of the events or loan non-payments happened in the earlier months of bank loan repayments.

Highest Education level has been acknowledged as a significant risk factor for loan defaults in bank mortgage customers (the hazard-ratio = 1.19, p value $\leq .001$). Model also identified debit interest

rate as risk factor for loan defaults in mortgage customers (HR = 1.11, p-value $\leq .001$) in the multivariable Cox PH regression model.

In this study, the covariates client bureau risk score, historical amount due, term paid and monthly repayment value were not found to be factors that affect the survival of the customers with mortgage loans in the multivariable Cox proportional regression model during mortgage loan repayment (Bureau Score HR = 0.99, Historical amount due HR = 1.00, term paid HR = 1.02 and monthly repayment value HR = 1.00 with all p-value ≤ 0.001). However, client bureau score with less than 500 score was a risk factor for default on mortgage loans, and was associated with survival of customers who begin repayment on mortgage loans.

Table 5.26 shows Score Table for the Cox regression model. In this table, variables such as Cumulative Good, Cumulative Bad, Cumulative Good Percentage, Cumulative Bad Percentage, Cumulative Bad avoided and Profit were developed to obtain the conclusion for credit score. In the first decile, we discovered about 11% of good loans were given and model avoided bad loans of about 89%.

Logistic Regression

Looking at the **Multivariable Logistic Regression model**, all variables are significant (p value < .05); my model found that coefficient estimate of client bureau risk score is negative, which means that the default rate is lower when the client bureau risk score is higher or increasing. The monthly repayment value has a negative sign, which clearly means that the default rate is lower when the repayment amount is higher or customers are making payments monthly as usual. The previous amount paid is found to have a negative coefficient, which means that the default rate is lower when the previous amount paid is higher, and Purchase price is also negative, which means that the default rate is lower when the mortgage purchase price is higher.

The two significant procedures to assess models are accuracy and error. The accuracy for the Logistic Regression model with cut-off is 0.4653, that capitalises on the true positive rate and true negative rate is 75.59% and which shows that the model is decent on detecting an applicant that will not pay and an applicant that will continue to pay. Thus, if we look at the true positive rate, it is 77.44%. This only explains that our model for loans which are defaulting, predicted correctly only in 77.44% cases. From the confusion matrix in Table 5.17, we can see that the model classified 7,953 loans as non-defaulters in training dataset, while in the test dataset there were, in total, 3,520 not defaulted off loans. Therefore, accuracy is good (77.5%), with perfect prediction of loans that will be paid (77.44%). This model would be very useful to use by banks or institutions. The main reason why banks should use this model is to avoid giving credit in loans that would not be fully paid. For this reason, the ROC curve gives adequate information about the quality of the model. Figure 5.23

shows the area under the curve of 0.855. Therefore, it is clear from this figure 5.23 that this model is good, as it can correctly classify loans that are defaulted or paid off.

In Table 5.18, the score for the logistic regression model is shown. This table explains to us that the model identifies 486 good observations in decile 1 out 1390, in decile 2, 505 good out of 1389, and so on. The difference in these deciles is the bad observations. We identified, in the first decile, that we have 10% of good loans and we must avoid about 90% of bad loans.

Looking at the Predictive performance for both models, the Logistic regression model outperformed survival analysis on both training and testing datasets. The main reason is AUROC on training dataset of logistic regression is equal to 80.55% and greater than 80%, and is considered as a good model, whereas AUROC of Cox proportional hazard model is equal to 76.55% and greater than 70%, and is being an acceptable model in this study as an alternative model to logistic regression.

6.3. Conclusions and Recommendations

In this paper, the models were using logistic regression and SA to forecast whether a borrower will refund the loan, or default on historical data provided by a financial institution, and helping a bank when deciding to which customers to approve a loan. After model development, testing and data analysis, we can make the following decisions:

1. Logistic regression and survival analysis models are perfect, and they displayed have good performance. The selection of cut-off value is very important if financial institutions are going to use these models to decide which applicants will get a loan or not. It is good, though, that financial institutions can decide the percentage of bad rate that they are willing to accept in their portfolio, and based on that they can decide very easily the percentage of the new loans that they want to finance.
2. In these models, it is better to use the ROC curve which shows true positive rates against false positive rates.
3. Loan risk models showed perfect performance in predicting true positive and true negative.
4. All limitations of Logistic Regression have been addressed on the Survival Analysis model, such as impact of changes in macro-economic variables, predicted probability of default that remains constant across outcome period and prediction of time to default, and handling of censored and truncated data.

Future research: a model proposed is the machine learning on survival data. After the improvement of technology related to Big Data, availability of data and power computing, most lenders or borrowing financial institutions are reintroducing their commercial models. Credit risk forecasts, monitoring, trustworthiness model and operative loan dispensation are important to decision-making

and transparency. It will be interesting to develop binary classifiers because of learning machine and profound knowledge models on factual data in forecasting loan non-payment likelihood. The to 10 important features from the models can be carefully chosen, and at that moment be utilised in the process of building a model to assess the steadiness of binary classifiers by relating performance on separate data. Our future model will observe that tree-based models as to whether they can be more constant than models based on multilayer artificial neural networks.

BIBLIOGRAPHY

Aalen, O. 1978. Non-parametric estimation of partial transition probabilities in multiple decrement models. *Annals of statistics,* 6:534-545.

Abdou, H.A. & Pointon, J. 2011. Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent systems in accounting, finance & management,* 18:59-88.

Banasik, J., Crook, J. & Thomas, L. 1999. Not if but when will borrowers default. *Journal of the operational research society,* 50:185-1190.

Basel, 2000. *Principles for the managment of credit risk.* s.l.: Basel Committee.

Bastos, J. 2008. *Credit scoring with boosted decision trees.* s.l.: School of Economics and Management (ISEG), Technical University of Lisbon.

Bellotti, T. & Crook, J. 2009. Credit scoring with macroeconomic variables using survival analysis. *Journal of the operational Research Society,* 6(12):1699-1707.

Brambor, T., Clark, W.R. & Golder, M. 2006. Understanding interaction models: improving empirical analyses. *Winter,* 14(1):63-82.

Brown, I. & Mues, C. 2012. An experimental comparison of classification algorithms for imbalanced credit. *Elsevier,* 39:3446-3453.

Capon, N. 1982. Credit scoring system: a critical analysis. *Journal of marketing,* 46:82-91, Spring.

Ptak-Chmielewska, A. 2016. Statistical models for corporate credit risk assessment – rating models. *Folia oeconomica,* 3(322):87-111.

Collett, D. 1994. *Modeling survival data in medical research.* London: Chapman & Hall.

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

Cox, D.R. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society, Series B (Methodological),* 20:215-242.

Crook, J. & Bellotti, T. 2009. Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society,* 60:1699-1707.

Dirick, L., Claeskens, G. & Baesens, B. 2017. Time to default in credit scoring using survival analysis. *Journal of the Operational Research Society,* 68:652-665.

Duncan, D.B. & Walker, S.H. 1967. Estimation of the probability of an event as a function of several independent variables. *Biometrika,* 54:167-179.

Đurović, A. 2017. Estimating probability of default on peer to peer market – survival analysis approach. *Journal of central banking theory and practice,* 2:149-167.

Estacion, A., DeLuca, S. & Plank, S. 2008. High school dropout and the role of career and technical education: a survival analysis of surviving high school. *Sage journals,* 81(4):345-370, October.

Featherstone, A.M., Roessler, L.M. & Barry, P.J. 2006. Determining the probability of default and risk-rating class for loans in the Seventh Farm Credit District Portfolio. *Applied economic perspectives and policy,* 28(1):4-23.

Ferreira, P.H., Louzada, F. & Diniz, C. 2015. Credit scoring modeling with state-dependent sample selection: a comparison study with the usual logistic modeling. *Pesquisa operacional*, 35(1):39-56, Jan/Apr.

Finkelstein, M. & Planck, M. 2008. Springer Series in Reliability Engineering. In: H. Pham (ed.) 2008. *Failure rate modelling for reliability and risk.* Bloemfontein: Springer Series in Reliability Engineering, pp. 19-27.

Gogtay, J.N. & Thatte, U.M. 2017. Survival analysis. *Journal of The Association of Physicians of India,* 65:80-84, May.

Gupta, V. 2017a. A survival approach to prediction of default drivers for Indian listed companies, Volume 7(2):116-138.

Gupta, V. 2017b. Identifying key predictors of default for Indian companies using Cox regression. *International journal of engineering technology science and research,* 4(2):97-111, February.

Györffy, B. et al. 2010. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast cancer research and treatment,* 123(3):725-731.

Hlongwane, T., Mdlongwa, P., Nare, H. & Moyo, I.L. 2014. Censored regression techniques for credit scoring: a case study for the Commercial Bank of Zimbabwe (Bulawayo). *International journal of economics and finance,* 6:1-16.

Hosmer, D.W. & Lemeshow, S. 1999. *Applied survival analysis: regression modeling of time to event data.* New York: Wiley.

Hosmer, D.W., Lemeshow, S. & Sturdivant, R.X. 2013. *Applied logistic regression.* 3rd ed. New York: Wiley.

Jaber, J.J. 2017. Credit risk assessment using survival analysis for progressive right-censored data: a case study in Jordan. *Journal of internet banking and commerce,* 22:2-18.

Jilek, O. 2008. Mathematical applications in credit risk modelling. *Journal of applied mathematics,* 1:432-438.

Kalbfleisch, J.D. & Prentice, R.L. 2002. *The statistical analysis of failure time data.* 2nd ed. s.l.:Wiley.

Kaplan, E.L. & Meier, P. 1958. Nonparametric estimation from Incomplete observations. *Journal of the American Statistical Association,* 53:457-481.

Kartsonaki, C. 2016. Diagnostic histopathology. *Mini-symposium: medical statistics,* 22(7):263-270.

Lechuga, G.P. & Sánchez, J.F.M. 2016. Assessment of a credit scoring system for popular bank savings and credit. *Contaduría y administración,* 61:391-417.

Lee, T.E. & Oscar, T.O. 1997. Survival analysis in public health. *Annual review of public health,* 18:105-134.

Mageto, D.K., Mwalili, S.M. & Waititu, A.G. 2015. Modelling of credit risk: Random Forests versus Cox Proportional Hazard Regression. *American journal of theoretical and applied statistics,* 4:247-253.

Memić, D. 2015. Assessing credit default using logistic regression and multiple discriminant analysis: empirical evidence from Bosnia and Herzegovina. *Interdisciplinary description of complex systems,* 13:128-153.

Nakartsonaki, C. 2016. Diagnosti chistopathology. *Elsevier,* 22:263-270.

Narain, B. 1992. Survival analysis and the credit granting decision. In: Thomas, L.C. & Edelman, D.B (eds.) 1992. *Credit scoring and credit control*. Oxford, Oxford University press, pp. 109-121.

Roy, D.G., Bindya, K. & Swati, K. 2013. Basel I to Basel II to Basel III: a risk management journey of Indian banks. *AIMA journal of management & research,* 7(0974):497.

Satagopan, J.M. et al. 2004. A note on competing risks in survival data analysis. *British journal of cancer,* 91:1229-1235.

Siddiqi, N. 2006. Credit risk scorecards. In: *Developing and implementing intelligent credit scoring.* s.l.:s.n., pp. 79-83.

Stepanova, M. & Thomas, L. 2002. Survival analysis methods for personal loan data. *Operations research,* 50(2):277-289, April.

Thomas, L.C., Edelman, D.B. & Crook, J.N. 2002. *Credit scoring and Its applications*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Volk, M. 2012. Estimating probability of default and comparing it to credit rating classification by banks. *Economic and business review,* 14(4):299-320.

Wekesa, O.A., Mwalili,S. & Mwita, P. 2012. Modelling credit risk for personal loans using Product-Limit Estimator. *International journal of financial research,* 3(1):22-32, January.

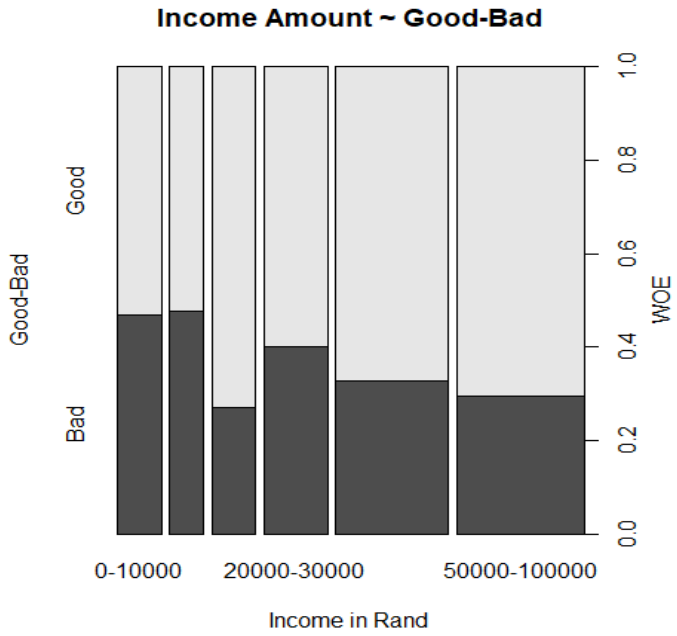
Yusuff, H., Mohamad, N., Ngah, U.K. & Yahaya, A.S. 2012. Breast cancer analysis using logistic regression. *IJRRAS,* 10(1):14-22, January.

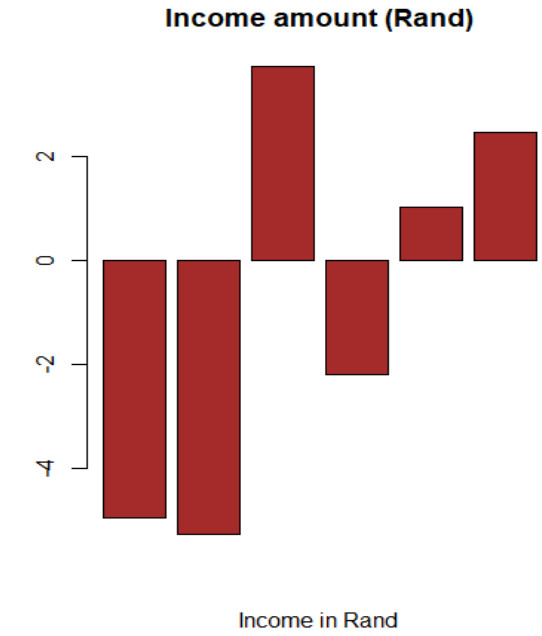
Zaghdoudi, T. 2013. Bank failure prediction with logistic regression. *International journal of economics and financial issues,* 3(2146-4138):537-543.

APPENDIX A

A.1 Univariate Analysis – Default Model

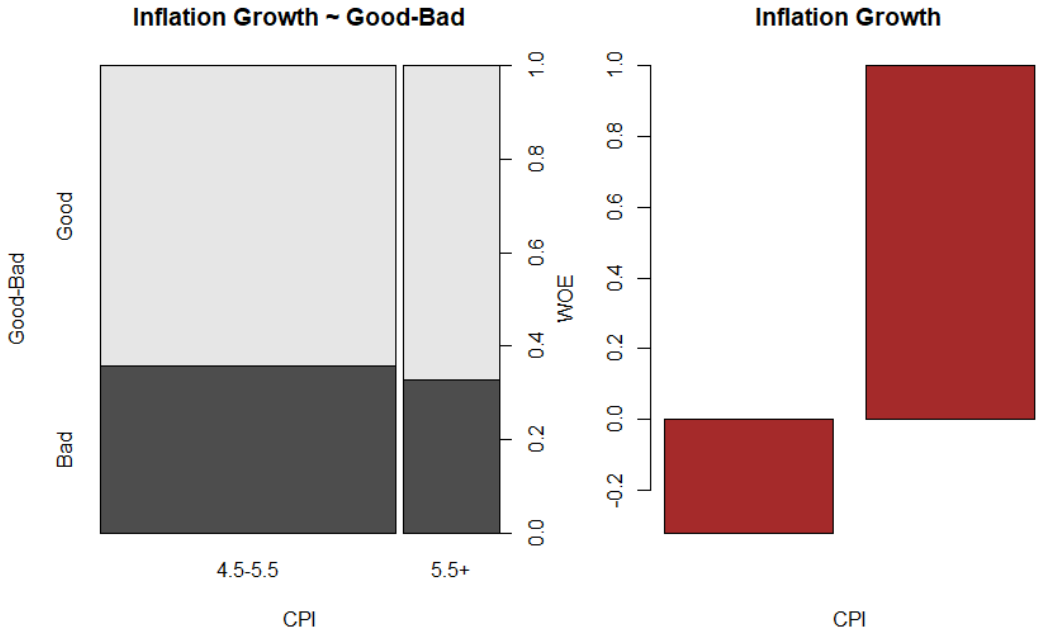
Variables which are weak predictors with their Event Rate, Score, Information Value and WoE Assessment





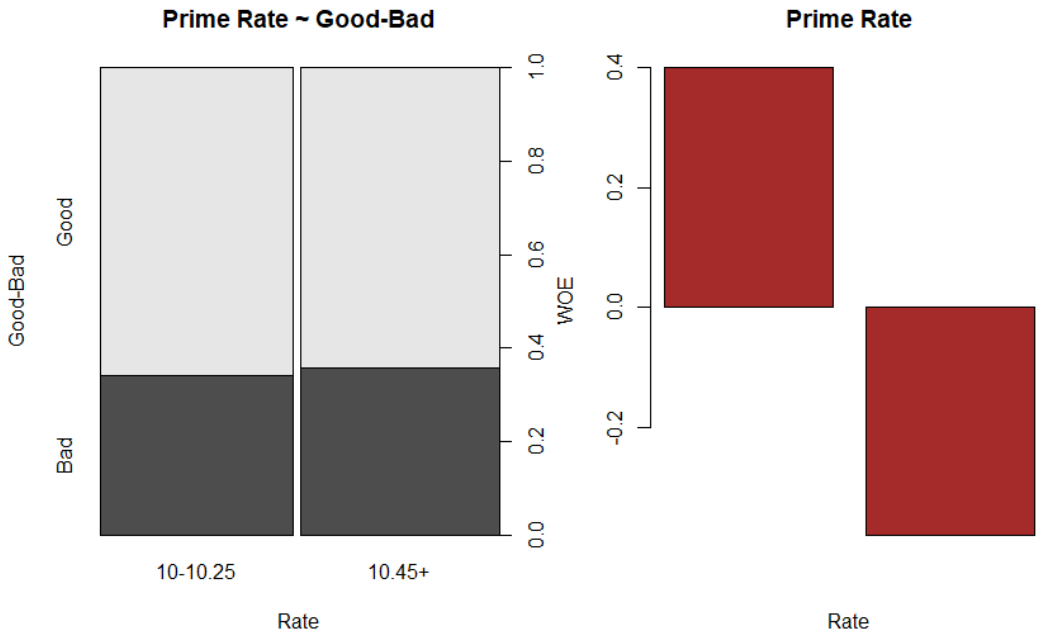
Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
0-10000	1092	962	8.37	13.75	2054	10.25	46.84	3.78	-4.96	2.66848	2.69
10000-20000	844	767	6.47	10.96	1611	8.04	47.61	3.71	-5.27	2.36623	2.245
100000+	1484	548	11.37	7.83	2032	10.14	26.97	5.92	3.73	1.32042	1.77
20000-30000	1809	1209	13.86	17.28	3018	15.05	40.06	4.45	-2.21	0.75582	1.71
30000-50000	3584	1734	27.46	24.79	5318	26.53	32.61	5.26	1.02	0.27234	1.335
50000-100000	4238	1776	32.47	25.39	6014	30	29.53	5.61	2.46	1.74168	3.54

Information Value is 9,13 and Efficiency is 13,29.



Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
4.5-5.5	9750	5395	74.71	77.12	15145	75.55	35.62	4.92	-0.32	0.07712	1.205
5.5+	3301	1601	25.29	22.88	4902	24.45	32.66	5.25	1	0.241	1.205

Information Value is 0,31812 and Efficiency is 2,41.



Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
10-10.25	6505	3350	49.84	47.88	9855	49.16	33.99	5.1	0.4	0.0784	0.98
10.45+	6546	3646	50.16	52.12	10192	50.84	35.77	4.9	-0.38	0.07448	0.98

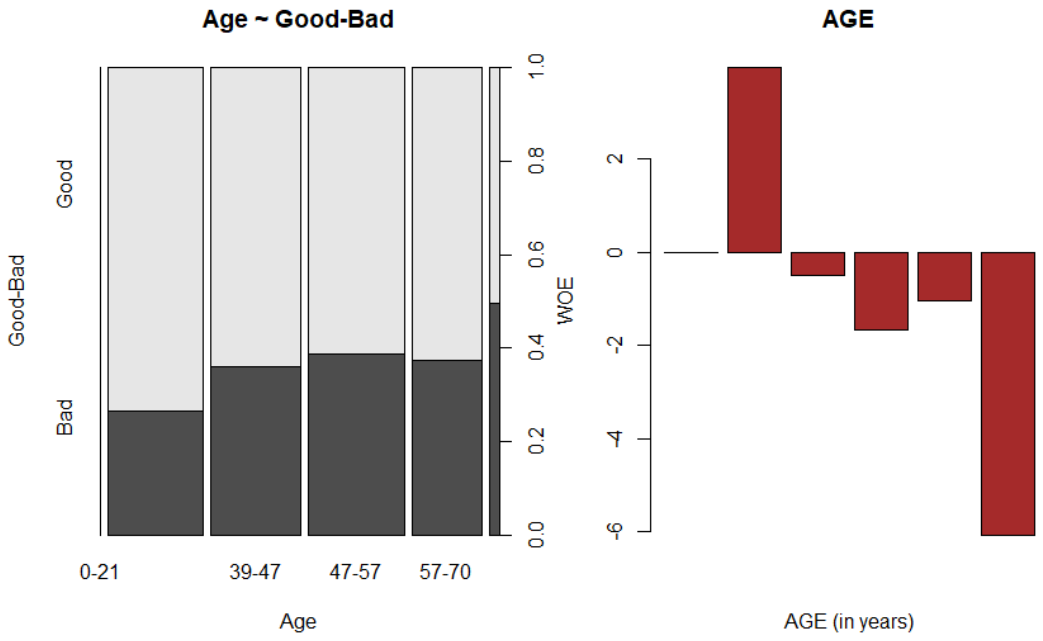
Information Value is 0,153 and Efficiency is 1,96.

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
03-Feb	2163	1006	16.57	14.38	3169	15.81	31.75	5.35	1.42	0.31098	1.095
04-Mar	5411	3229	41.46	46.15	8640	43.1	37.37	4.73	-1.07	0.50183	2.345
05-Apr	4375	2144	33.52	30.65	6519	32.52	32.89	5.22	0.9	0.2583	1.435
06-May	1102	617	8.44	8.82	1719	8.57	35.89	4.89	-0.44	0.01672	0.19

Information Value is 1,09 and Efficiency is 5,07.

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
0-0.5	5531	2976	42.38	42.54	8507	42.44	34.98	4.99	-0.04	0.00064	0.08
0.5-1	7520	4020	57.62	57.46	11540	57.56	34.84	5.01	0.03	0.00048	0.08

Information Value is 0.001 and Efficiency is 0,16.



Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
0-21	1	1	0.01	0.01	2	0.01	50	5	0	0	0
21-39	3886	1404	29.78	20.07	5290	26.39	26.54	5.97	3.95	3.83545	4.855
39-47	3195	1801	24.48	25.74	4996	24.92	36.05	4.87	-0.5	0.063	0.63
47-57	3259	2065	24.97	29.52	5324	26.56	38.79	4.58	-1.67	0.75985	2.275
57-70	2422	1442	18.56	20.61	3864	19.27	37.32	4.74	-1.05	0.21525	1.025
70+	288	283	2.21	4.05	571	2.85	49.56	3.53	-6.06	1.11504	0.92

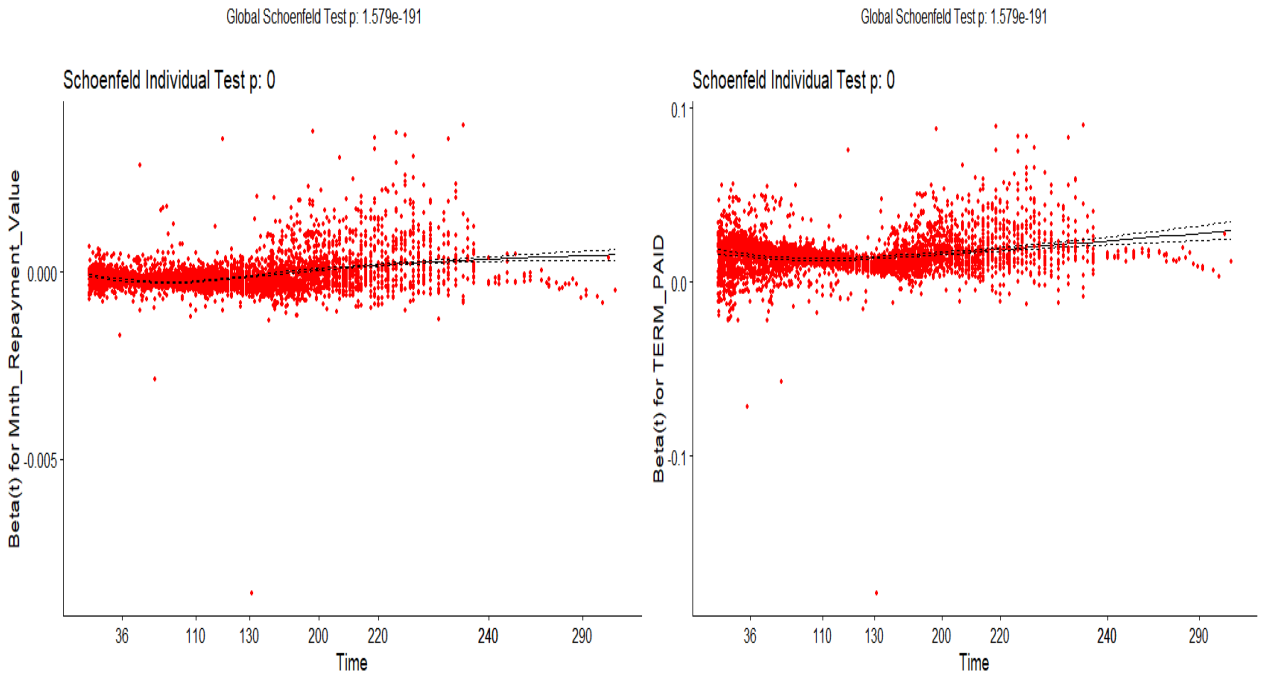
Information Value is 5,99 and Efficiency is 9,71.

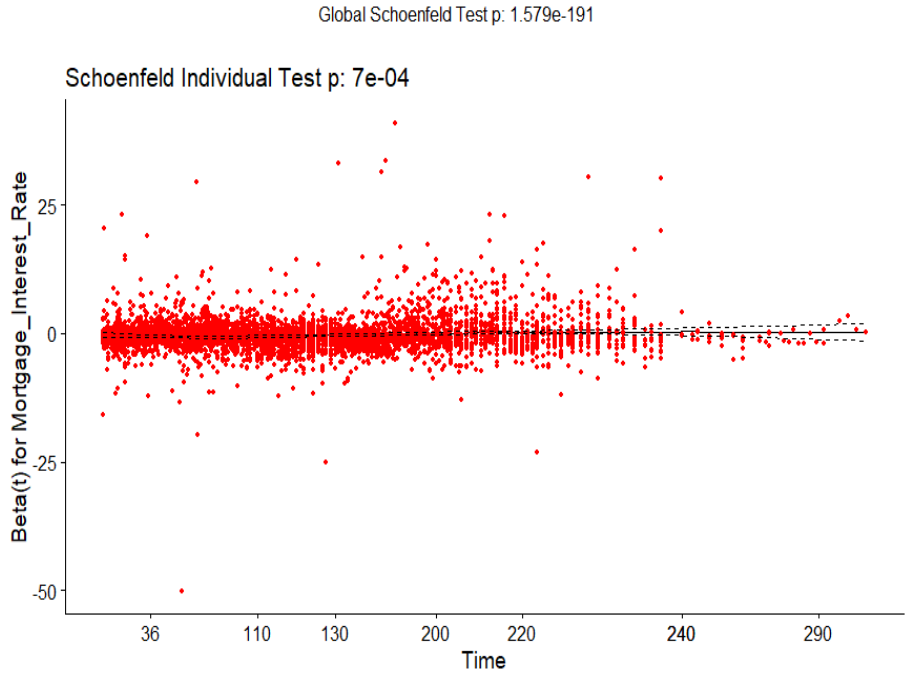
APPENDIX B

None Proportional Hazard Assumption

B.1 Assessment of the Proportional Hazard

The graphs below are of the Scaled Schoenfeld Residuals and lowness smooth curves that are not supporting the assumption of proportional hazards for individually covariates.





APPENDIX C THE R AND SAS CODE

C.1 Data Preparation

```
library(DT)          # For Data Tables
library(lattice)     # The lattice add-on of Trellis graphics for R
library(knitr)       # For Dynamic Report Generation in R
library(gplots)      # Various R Programming Tools for Plotting Data
library(ggplot2)     # An Implementation of the Grammar of Graphics
# install.packages("devtools")
devtools::install_github("r-lib/rlang", build_vignettes = TRUE)
library(rlang)
library(ClustOfVar)  # Clustering of variables
library(ape)         # Analyses of Phylogenetics and Evolution (as.phylo)
library(Information) # Data Exploration with Information Theory (Weight-of-Evidence and Information Value)
library(ROCR)        # Model Performance and ROC curve
library(caret)       # Classification and Regression Training -  for any machine learning algorithms
library(rpart)       # Recursive partitioning for classification, regression and survival trees
library(rpart.utils) # Tools for parsing and manipulating rpart objects, including generating machine readable
rules
library(rpart.plot)  # Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'
library(randomForest)# Leo Breiman and Cutler's Random Forests for Classification and Regression
library(party)       # A computational toolbox for recursive partitioning - Conditional inference Trees
library(bnlearn)     # Bayesian Network Structure Learning, Parameter Learning and Inference
library(DAAG)        # Data Analysis and Graphics Data and Functions
library(vcd)         # Visualizing Categorical Data
library(kernlab)     # Support Vector Machine
library(haven)
library(plotly)
# Install blorr from CRAN
install.packages("blorr")
library(blorr)
library(magrittr)
library(survival)
library("splines")
library("lattice")
library("JM")
library(shiny)
# Install blorr from CRAN
install.packages("blorr")

# Or the development version from GitHub
# install.packages("devtools")
```

```
devtools::install_github("rsquaredacademy/blorr")

# Or the development version from GitHub
# install.packages("devtools")
devtools::install_github("rsquaredacademy/blorr")
# Following libraries we have load for model 8 and model 9
#library(neuralnet)  # Neural Network
#library(lars)   # For Least Angle Regression, Lasso and Forward Stagewise
#library(glmnet) # Lasso and Elastic-Net Regularized Generalized Linear Models
pkg <- c("tidyverse", "survival", "ggfortify", "survminer", "plotly", "gridExtra",
        "Epi", "KMsurv", "gnm", "cmprsk", "mstate", "flexsurv", "splines",
        "epitools", "eha", "shiny", "ctqr", "scales")
new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
if (length(new.pkg)) install.packages(new.pkg, dependencies = TRUE)
sapply(pkg, require, character.only = TRUE)
train_1 <- read_sas("~/PERSONAL/modelling/train.sas7bdat", NULL)
train<- read_sas("~/PERSONAL/modelling/train.sas7bdat",  NULL)
test_1 <- read_sas("~/PERSONAL/modelling/test.sas7bdat",  NULL)
all <- read_sas("PERSONAL/modelling/all.sas7bdat", NULL)
train <- read_sas("~/PERSONAL/modelling/train.sas7bdat", NULL)
```

C.2 Data Analysis and Variable Creation

```
##Data analysis and variable creation

#2.0 Modify Variable types
all$CUST_NO  <- as.numeric(all$CUST_NO)
all$ACCT_NO   <-  as.numeric(all$ACCT_NO )
all$CUST_AGE <-  as.numeric( all$CUST_AGE)
all$INCOME_AMOUNT <-  as.numeric(all$INCOME_AMOUNT)
all$CUST_NO_CHILD        <-  as.numeric(all$CUST_NO_CHILD)
all$PURCHASE_PRICE    <-  as.numeric(all$PURCHASE_PRICE) #done
all$MORTGAGE_INTEREST_RATE<-  as.numeric(all$MORTGAGE_INTEREST_RATE) #done
all$BOND_AMT <-  as.numeric(all$BOND_AMT) #done
all$OUT_LN_BAL_AMT <-  as.numeric(all$OUT_LN_BAL_AMT)
all$PAST_DUE_AMT <-  as.numeric(all$PAST_DUE_AMT) #done
all$CURR_LOAN_TERM <-  as.numeric(all$CURR_LOAN_TERM)
all$MNTH_REPAYMENT_AMNT <-  as.numeric(all$MNTH_REPAYMENT_AMNT) #done
all$RMNG_TERM <-  as.numeric(all$RMNG_TERM)
all$PREVIOUS_AMNT_PAID <-  as.numeric(all$PREVIOUS_AMNT_PAID) #done
```

```
all$VOB_ACT_BAL <-  as.numeric(all$VOB_ACT_BAL)
all$CLIENT_BUREAU_SCORE <-  as.numeric(all$CLIENT_BUREAU_SCORE) #done
all$income_estimate_m <-  as.numeric(all$income_estimate_m) #done
all$Inflation_Growth_YoY <-  as.numeric(all$Inflation_Growth_YoY)
all$Prime_rate_level_eop <-  as.numeric(all$Prime_rate_level_eop)
all$HPI_Growth_Yoy_perc <-  as.numeric(all$HPI_Growth_Yoy_perc)
all$LN_TPE <- as.numeric(all$LN_TPE)
all$CUST_OCPTN_CDE <- as.numeric(all$CUST_OCPTN_CDE)
all$Marital_ind <- as.numeric(all$Marital_ind)
all$Gender <- as.numeric(all$Gender) #done
all$TERM_PAID <- as.numeric(all$TERM_PAID) #done
all$Highest_Education_Level <- as.numeric(all$Highest_Education_Level) #done
all$Default <- as.numeric(all$Default) #done
#Renaming variables
library(plyr)
rename(all, c("CUST_NO"="Client_ID", "ACCT_NO"="ACCOUNT_ID", "CUST_AGE"= "Client_Age"
            , "INCOME_AMOUNT"= "Client_Income"
            , "CUST_NO_CHILD"= "Client_NR_Child"
            , "MORTGAGE_INTEREST_RATE"= "Mortgage_Interest_Rate"
            , "PAST_DUE_AMT"= "Historical_Amount_Due"
            , "MNTH_REPAYMENT_AMNT"= "Repayment_Value"
            , "RMNG_TERM"= "Outstanding_term"
            , "CLIENT_BUREAU_SCORE"= "Client_Bureau_Score"
          ))
#2.1 Good-Bad and Univariate Analysis
#2.1.1 Analyse good_bad(1-good, 2-bad)
all$Default<-as.factor(ifelse(all$Default == 0, "Good", "Bad"))
pct(all$Default)
op<-par(mfrow=c(1,2), new=TRUE)
plot(as.numeric(all$Default), ylab="Good-Bad", xlab="n", main="Good ~ Bad")
hist(as.numeric(all$Default), breaks=2,
    xlab="Good(1) and Bad(2)", col="blue")
par(op)

#Univariate and bivariate Analysis
#Weight of Evidence(WOE), Information Value(IV) and Efficiency
#1. Term Paid
summary(all$TERM_PAID)
op2<-par(mfrow=c(1,2))
```

```
boxplot(all$TERM_PAID, ylab="Loan Duration(Month)", main="Boxplot:Loan Duration")
plot(all$TERM_PAID, all$Default,
    ylab="Good-Bad", xlab="Loan Duration(Month)",
    main="Loan Duration ~ Good-Bad ")
plot(as.factor(all$TERM_PAID), all$Default,
    ylab="Good-Bad", xlab="Loan Duration(Month)",
    main="Loan Duration(Before Groupping)")
# Create some groups from contious variables
all$TERM_PAID <-as.factor(ifelse(all$TERM_PAID<=60,'00-60',
                                 ifelse(all$TERM_PAID<=120,'60-120',
                                        ifelse(all$TERM_PAID<=180,'120-180',
                                               ifelse(all$TERM_PAID<=240,'180-240',
                                                      ifelse(all$TERM_PAID<=300,'240-300','300+'))))))

plot(all$TERM_PAID, all$Default,
    main="Loan Duration(after grouping) ",
    xlab="Loan Duration (Month)",
    ylab="Good-Bad")
par(op2)
A2<-gbpct(all$TERM_PAID)
barplot(A2$WOE, col="brown", names.arg=c(A2$Levels),
       main="Loan Duration",
       xlab="Duration(Months)",
       ylab="WOE"
)
kable(A2, caption = 'Loan Duration ~ Good-Bad')
#2.CLIENT_BUREAU_SCORE
# Attribute 3:  (numerical)
# Credit risk score
all$CLIENT_BUREAU_SCORE <- as.double(all$CLIENT_BUREAU_SCORE)
summary(all$CLIENT_BUREAU_SCORE)
boxplot(all$CLIENT_BUREAU_SCORE)

# Create groups based on their distribution
all$CLIENT_BUREAU_SCORE<-as.factor(ifelse(all$CLIENT_BUREAU_SCORE<=400,'0-400',
                                 ifelse(all$CLIENT_BUREAU_SCORE<=550,'400-550',
                                        ifelse(all$CLIENT_BUREAU_SCORE<=600,'550-600',
                                               ifelse(all$CLIENT_BUREAU_SCORE<=650,'600-650',
```

```
                                        ifelse(all$CLIENT_BUREAU_SCORE<=700,'650-
700','700+'))))))
A3<-gbpct(all$CLIENT_BUREAU_SCORE)
plot(all$CLIENT_BUREAU_SCORE, all$Default,
    main="Client Bureau Score (After Grouping) ~ Good-Bad",
    xlab="Scores",
    ylab="Good-Bad")
barplot(A3$WOE, col="brown", names.arg=c(A3$Levels),
       main="Client Bureau Score",
       xlab="Scores",
       ylab="WOE")
kable(A3, caption = "Client Bureau Score ~ Good-Bad")
#*************************************************************************************************
#3.PAST_DUE_AMT
all$PAST_DUE_AMT <- as.double(all$PAST_DUE_AMT)
summary(all$PAST_DUE_AMT)
boxplot(all$PAST_DUE_AMT)
# Create groups based on their distribution
all$PAST_DUE_AMT <-as.factor(ifelse(all$PAST_DUE_AMT<=10000,'0-10000',
                            ifelse(all$PAST_DUE_AMT<=20000,'10000-20000',
                                ifelse(all$PAST_DUE_AMT<=40000,'20000-40000',
                                    ifelse(all$PAST_DUE_AMT<=60000,'40000-60000',
                                        ifelse(all$PAST_DUE_AMT<=100000,'60000-
80000','80000+'))))))
A4<-gbpct(all$PAST_DUE_AMT)
plot(all$PAST_DUE_AMT, all$Default,
    main="Historical amount due ~ Good-Bad",
    xlab="Amount Due",
    ylab="Good-Bad")
barplot(A4$WOE, col="brown", names.arg=c(A4$Levels),
       main="Historical amount due (Rand)",
       xlab="Amount Due",
       ylab="WOE")
kable(A4, caption = "Historical amount due ~ Good-Bad")
#*************************************************************************************************
#4. PREVIOUS_AMNT_PAID
all$PREVIOUS_AMNT_PAID <- as.double(all$PREVIOUS_AMNT_PAID)
summary(all$PREVIOUS_AMNT_PAID)
boxplot(all$PREVIOUS_AMNT_PAID)
```

```
# Create groups based on their distribution
all$PREVIOUS_AMNT_PAID <-as.factor(ifelse(all$PREVIOUS_AMNT_PAID<=2000,'0-2000',
                                  ifelse(all$PREVIOUS_AMNT_PAID<=4000,'2000-4000',
                                      ifelse(all$PREVIOUS_AMNT_PAID<=6000,'4000-6000',
                                          ifelse(all$PREVIOUS_AMNT_PAID<=8000,'6000-8000',
                                              ifelse(all$PREVIOUS_AMNT_PAID<=10000,'8000-
10000','10000+'))))))
A5<-gbpct(all$PREVIOUS_AMNT_PAID)
plot(all$PREVIOUS_AMNT_PAID, all$Default,
    main="Previous Amount Paid(After Grouping)~Good-Bad",
    xlab="Amount",
    ylab="Good-Bad")
barplot(A5$WOE, col="brown", names.arg=c(A5$Levels),
       main="Previous Amount Paid (Rand)",
       xlab="Last Amount",
       ylab="WOE")
kable(A5, caption = "Previous Amount Paid ~ Good-Bad")
#*************************************************************************************************
#5. BOND_AMT
all$BOND_AMT <- as.double(all$BOND_AMT)
summary(all$BOND_AMT)
boxplot(all$BOND_AMT)
# Create groups based on their distribution
all$BOND_AMT <-as.factor(ifelse(all$BOND_AMT<=1000000,'0-1000000',
                            ifelse(all$BOND_AMT<=2000000,'1000000-2000000',
                                ifelse(all$BOND_AMT<=3000000,'2000000-3000000',
                                    ifelse(all$BOND_AMT<=4000000,'3000000-4000000',
                                        ifelse(all$BOND_AMT<=5000000,'4000000-
5000000','5000000+'))))))
A6<-gbpct(all$BOND_AMT)
plot(all$BOND_AMT, all$Default,
    main="Bond Amount ~ Good-Bad",
    xlab="Amount",
    ylab="Good-Bad")
barplot(A6$WOE, col="brown", names.arg=c(A6$Levels),
       main="Bond Amount (Rand)",
       xlab="Amount",
       ylab="WOE")
kable(A6, caption = "Bond Amount ~ Good-Bad")
```

```
#*************************************************************************************************
#6. Highest_Education_Level
summary(all$Highest_Education_Level)
op7 <- par(mfrow = c(1,2))
boxplot(all$Highest_Education_Level)
plot(as.factor(all$Highest_Education_Level),  all$Default,
    main = "Highest Education Level",
    xlab = "Education Level",
    ylab = "Good-Bad")
A7<-gbpct(all$Highest_Education_Level)
op7_1<-par(mfrow=c(1,2))
plot(as.factor(all$Highest_Education_Level),  all$Default,
    main="Highest Education Level ~Good-Bad",
    xlab="Education Level",
    ylab="Good-Bad")
barplot(A7$WOE, col="brown", names.arg=c(A7$Levels),
       main="Highest Education Level",
       xlab="Category",
       ylab="WOE")
par(op7)
kable(A7, caption = "Highest Education Level ~ Good-Bad")
#*************************************************************************************************
#7.PURCHASE_PRICE
all$PURCHASE_PRICE <- as.double(all$PURCHASE_PRICE)
summary(all$PURCHASE_PRICE)
boxplot(all$PURCHASE_PRICE)
# Create groups based on their distribution
all$PURCHASE_PRICE<-as.factor(ifelse(all$PURCHASE_PRICE<=500000,'0-500k',
                             ifelse(all$PURCHASE_PRICE<=1000000,'500k-1000k',
                                 ifelse(all$PURCHASE_PRICE<=2000000,'1000k-2000k',
                                     ifelse(all$PURCHASE_PRICE<=3000000,'2000k-3000k',
                                         ifelse(all$PURCHASE_PRICE<=4000000,'3000k-
4000k','4000k+'))))))
A8<-gbpct(all$PURCHASE_PRICE)
plot(all$PURCHASE_PRICE, all$Default,
    main="Purchase Price ~ Good-Bad",
    xlab="Price",
    ylab="Good-Bad")
barplot(A8$WOE, col="brown", names.arg=c(A8$Levels),
```

```
      main="Purchase Price",
      xlab="Price",
      ylab="WOE")
kable(A8, caption = "Purchase Price ~ Good-Bad")
#*************************************************************************************************************
#8. Debit interest rate
summary(all$MORTGAGE_INTEREST_RATE)
op9<-par(mfrow=c(1,2))
boxplot(all$MORTGAGE_INTEREST_RATE)
histogram(all$MORTGAGE_INTEREST_RATE,
        main = "Mortgage interest rate",
        xlab = "interest percent",
        ylab = "Percent Population")
par(op9)
# Create groups based on their distribution
all$MORTGAGE_INTEREST_RATE<-as.factor(ifelse(all$MORTGAGE_INTEREST_RATE<=9,'0-
9',
                              ifelse(all$MORTGAGE_INTEREST_RATE<=10,'9-10',
                                  ifelse(all$MORTGAGE_INTEREST_RATE<=11,'10-11',
                                      ifelse(all$MORTGAGE_INTEREST_RATE<=12,'11-12',
                                          ifelse(all$MORTGAGE_INTEREST_RATE<=13,'12-
13','13+'))))))
A9<-gbpct(all$MORTGAGE_INTEREST_RATE)
op9_1<-par(mfrow=c(1,2))
plot(as.factor(all$MORTGAGE_INTEREST_RATE), all$Default,
    main="Mortgage Interest Rate ~ Good-Bad",
    xlab="Percent",
    ylab="Good-Bad")
barplot(A8$WOE, col="brown", names.arg=c(A8$Levels),
      main="Mortgage interest rate",
      xlab="Percent",
      ylab="WOE")

kable(A9, caption = "Mortgage interest rate ~ Good-Bad")
#*************************************************************************************************************
#9. MNTH_REPAYMENT_AMNT
all$MNTH_REPAYMENT_AMNT <- as.double(all$MNTH_REPAYMENT_AMNT)
summary(all$MNTH_REPAYMENT_AMNT)
boxplot(all$MNTH_REPAYMENT_AMNT)
```

```
# Create groups based on their distribution
all$MNTH_REPAYMENT_AMNT<-as.factor(ifelse(all$MNTH_REPAYMENT_AMNT<=6000,'0-
6000',
                          ifelse(all$MNTH_REPAYMENT_AMNT<=8000,'6000-8000',
                              ifelse(all$MNTH_REPAYMENT_AMNT<=10000,'8000-10000',
                                  ifelse(all$MNTH_REPAYMENT_AMNT<=12000,'10000-12000',
                                      ifelse(all$MNTH_REPAYMENT_AMNT<=15000,'12000-
15000','15000+'))))))
A10<-gbpct(all$MNTH_REPAYMENT_AMNT)
plot(all$MNTH_REPAYMENT_AMNT, all$Default,
    main="Month Repay Amt ~ Good-Bad",
    xlab="Amount",
    ylab="Good-Bad")
barplot(A10$WOE, col="brown", names.arg=c(A10$Levels),
       main="Month Repay Amount",
       xlab="Amount",
       ylab="WOE")
kable(A10, caption = "Month Repay Amount ~ Good-Bad")
#*************************************************************************************************
#10.Loan To value
summary(all$LTV)
op11<-par(mfrow=c(1,2))
boxplot(all$LTV)
histogram(all$LTV,
         main = "instalment rate in percentage of disposable income",
         xlab = "instalment percent",
         ylab = "Percent Population")
par(op11)
op11_1<-par(mfrow=c(1,2))
plot(as.factor(all$LTV), all$Default,
    main="Loan To Value ~ Good-Bad",
    xlab="Ratio",
    ylab="Good-Bad")
# Create groups based on their distribution
all$LTV<-as.factor(ifelse(all$LTV<=0.5,'0-0.5',
                          ifelse(all$LTV <=0.75,'0.5-0.75',
                              ifelse(all$LTV<=1,'0.75-1',
                                  ifelse(all$LTV<=1.25,'1-1.25',
                                      ifelse(all$LTV<=1.5,'1.25-1.5','1.5+'))))))
```

```
A11<-gbpct(all$LTV)
plot(all$LTV, all$Default,
    main="LTV Ratio (After Grouping) ~ Good-Bad",
    xlab="LTV Ratio",
    ylab="Good-Bad")
barplot(A11$WOE, col="brown", names.arg=c(A11$Levels),
       main="LTV Ratio Ammount",
       xlab="Ratio",
       ylab="WOE")
kable(A11, caption = "LTV Ratio ~ Good-Bad")
#*****************************************************************************************************
#11. INCOME ESTIMATE
all$income_estimate_m <- as.double(all$income_estimate_m)
summary(all$income_estimate_m)
boxplot(all$income_estimate_m)
par(op15)
# Create groups based on their distribution
all$income_estimate_m <-as.factor(ifelse(all$income_estimate_m<=10000,'0-10000',
                                 ifelse(all$income_estimate_m<=20000,'10000-20000',
                                      ifelse(all$income_estimate_m<=30000,'20000-30000',
                                           ifelse(all$income_estimate_m<=50000,'30000-50000',
                                                ifelse(all$income_estimate_m<=100000,'50000-
100000','100000+'))))))
A11<-gbpct(all$income_estimate_m)
plot(all$income_estimate_m, all$Default,
    main="Income Amount ~ Good-Bad",
    xlab="Income in Rand",
    ylab="Good-Bad")
barplot(A11$WOE, col="brown", names.arg=c(A11$Levels),
       main="Income amount (Rand)",
       xlab="Income in Rand",
       ylab="WOE")
kable(A11, caption = "Income in Rand ~ Good-Bad")
#*****************************************************************************************************
#12. CPI
all$Inflation_Growth_YoY <- as.double(all$Inflation_Growth_YoY)
summary(all$Inflation_Growth_YoY)
boxplot(all$Inflation_Growth_YoY)
all$Inflation_Growth_YoY<-as.factor(ifelse(all$Inflation_Growth_YoY<=1.5,'0-1.4',
```

```
                    ifelse(all$Inflation_Growth_YoY <=2.5,'1.5-2.5',
                        ifelse(all$Inflation_Growth_YoY<=3.5,'2.5-3.5',
                            ifelse(all$Inflation_Growth_YoY<=4.5,'3.5-4.5',
                                ifelse(all$Inflation_Growth_YoY<=5.5,'4.5-5.5','5.5+'))))))
A12<-gbpct(all$Inflation_Growth_YoY)
plot(all$Inflation_Growth_YoY, all$Default,
    main="Inflation Growth ~ Good-Bad",
    xlab="CPI",
    ylab="Good-Bad")
barplot(A12$WOE, col="brown", names.arg=c(A12$Levels),
       main="Inflation Growth",
       xlab="CPI",
       ylab="WOE")
kable(A12, caption = "Inflation Growth ~ Good-Bad")
#*****************************************************************************************************************
**
#13. Prime Rate
all$Prime_rate_level_eop <- as.double(all$Prime_rate_level_eop)
summary(all$Prime_rate_level_eop)
boxplot(all$Prime_rate_level_eop)
all$Prime_rate_level_eop<-as.factor(ifelse(all$Prime_rate_level_eop<=10,'0-10',
                                ifelse(all$Prime_rate_level_eop <=10.25,'10-10.25',
                                    ifelse(all$Prime_rate_level_eop<=10.35,'10.25-10.35',
                                        ifelse(all$Prime_rate_level_eop<=10.40,'10.35-10.40',
                                            ifelse(all$Prime_rate_level_eop<=10.45,'10.40-
10.45','10.45+'))))))
A15<-gbpct(all$Prime_rate_level_eop)
plot(all$Prime_rate_level_eop, all$Default,
    main="Prime Rate ~ Good-Bad",
    xlab="Rate",
    ylab="Good-Bad")
barplot(A15$WOE, col="brown", names.arg=c(A15$Levels),
       main="Prime Rate",
       xlab="Rate",
       ylab="WOE")
kable(A15, caption = "Prime Rate ~ Good-Bad")
#*****************************************************************************
#14. HPI
all$HPI_Growth_Yoy_perc <- as.double(all$HPI_Growth_Yoy_perc)
```

```
summary(all$HPI_Growth_Yoy_perc)
boxplot(all$HPI_Growth_Yoy_perc)
all$HPI_Growth_Yoy_perc<-as.factor(ifelse(all$HPI_Growth_Yoy_perc<=2,'0-2',
                                  ifelse(all$HPI_Growth_Yoy_perc <=3,'2-3',
                                      ifelse(all$HPI_Growth_Yoy_perc<=4,'3-4',
                                          ifelse(all$HPI_Growth_Yoy_perc<=5,'4-5',
                                              ifelse(all$HPI_Growth_Yoy_perc<=6,'5-6','6+'))))))
A16<-gbpct(all$HPI_Growth_Yoy_perc)
plot(all$HPI_Growth_Yoy_perc, all$Default,
    main="HPI Growth ~ Good-Bad",
    xlab="HPI",
    ylab="Good-Bad")
barplot(A16$WOE, col="brown", names.arg=c(A16$Levels),
       main="HPI Growth",
       xlab="HPI",
       ylab="WOE")
kable(A16, caption = "HPI Growth ~ Good-Bad")
#*************************************************************************
#15. Sex
all$SEX <- as.double(all$SEX)
summary(all$SEX)
boxplot(all$SEX)
all$SEX<-as.factor(ifelse(all$SEX<=0.5,'0-0.5',
                          ifelse(all$SEX <=1,'0.5-1',
                              ifelse(all$SEX<=1.5,'1-1.5',
                                  ifelse(all$SEX<=2,'1.5-2',
                                      ifelse(all$SEX<=2.5,'2-2.5','2.5+'))))))
A18<-gbpct(all$SEX)
plot(all$SEX, all$Default,
    main="Gender ~ Good-Bad",
    xlab="Sex",
    ylab="Good-Bad")
barplot(A18$WOE, col="brown", names.arg=c(A18$Levels),
       main="Gender",
       xlab="Sex",
       ylab="WOE")
kable(A18, caption = "Gender ~ Good-Bad")
#*********************************************************************************************
#16. AGe
```

```
all$AGE <- as.double(all$AGE)
summary(all$AGE)
boxplot(all$AGE)
all$AGE<-as.factor(ifelse(all$AGE<=21,'0-21',
                  ifelse(all$AGE <=39,'21-39',
                      ifelse(all$AGE<=47,'39-47',
                          ifelse(all$AGE<=57,'47-57',
                              ifelse(all$AGE<=70,'57-70','70+'))))))
A19<-gbpct(all$AGE)
plot(all$AGE, all$Default,
    main="Age ~ Good-Bad",
    xlab="Age",
    ylab="Good-Bad")
barplot(A19$WOE, col="brown", names.arg=c(A19$Levels),
       main="AGE",
       xlab="AGE (in years)",
       ylab="WOE")
kable(A19, caption = "Age in years ~ Good-Bad")
************************************************************
#Remaining terms
summary(all$RMNG_TERM)
op13 <- par(mfrow = c(1,2))
boxplot(all$RMNG_TERM)
plot(as.factor(all$RMNG_TERM),  all$Default,
    main = "Remaining Term",
    xlab = "Age in Years",
    ylab = "Good-Bad")


par(op13)
# Group AGE - Coarse Classing (after some iterations in fine classing stage)
all$RMNG_TERM <- as.factor(ifelse(all$RMNG_TERM<=25, '0-25',
                          ifelse(all$RMNG_TERM<=50, '25-50',
                              ifelse(all$RMNG_TERM<=75, '50-75',
                                  ifelse(all$RMNG_TERM<=100, '75-100',
                                      ifelse(all$RMNG_TERM<=150, '100-150',
                                          ifelse(all$RMNG_TERM<=175, '150-175',
                                              ifelse(all$RMNG_TERM<=200, '175-200',
                                                  '200+')))))))
```

```
A13<-gbpct(all$RMNG_TERM)
op13_1<-par(mfrow=c(1,2))
plot(as.factor(all$RMNG_TERM),  all$Default,
    main="Remaining terms (After Grouping)",
    xlab="Terms",
    ylab="Good-Bad")
barplot(A13$WOE, col="brown", names.arg=c(A13$Levels),
       main="Remaining terms ",
       xlab="Category",
       ylab="WOE")
par(op13_1)
kable(A13,  caption = "Remaining Term (After Grouping) ~ Good-Bad")
##############################################################################
#######

summary(all$INCOME_ESTIMATE)

op13 <- par(mfrow = c(1,2))
boxplot(all$INCOME_ESTIMATE)
plot(as.factor(all$RMNG_TERM),  all$INCOME_ESTIMATE,
    main = "Income Estimate",
    xlab = "Income in Rand",
    ylab = "Good-Bad")
par(op13)
# Group AGE - Coarse Classing (after some iterations in fine classing stage)
all$RMNG_TERM <- as.factor(ifelse(all$RMNG_TERM<=25, '0-25',
                          ifelse(all$RMNG_TERM<=50, '25-50',
                               ifelse(all$RMNG_TERM<=75, '50-75',
                                    ifelse(all$RMNG_TERM<=100, '75-100',
                                         ifelse(all$RMNG_TERM<=150, '100-150',
                                              ifelse(all$RMNG_TERM<=175, '150-175',
                                                   ifelse(all$RMNG_TERM<=200, '175-200',
                                                        '200+')))))))
A13<-gbpct(all$RMNG_TERM)
op13_1<-par(mfrow=c(1,2))
plot(as.factor(all$RMNG_TERM),  all$Default,
    main="Remaining terms (After Grouping)",
    xlab="Terms",
    ylab="Good-Bad")
```

```
barplot(A13$WOE, col="brown", names.arg=c(A13$Levels),
        main="Remaining terms ",
        xlab="Category",
        ylab="WOE")
par(op13_1)
kable(A13,  caption = "Remaining Term (After Grouping) ~ Good-Bad")
```

C.3 Estimation of Survival Functions

```
#NON-PARAMETRIC ESTIMATORS

fit_km <- survfit(Surv(train$RMNG_TERM, train$Default) ~ 0, data = train)
print(fit_km, print.rmean = TRUE)

dat_km <- fortify(fit_km)
head(dat_km)
summary(fit_km, times = c(0, 5, 10,60, 110, 160, 210, 260, 359))

ggsurvplot(fit_km, risk.table = TRUE, xlab = "Time (months)", censor = T)

#Lifetable or actuarial estimator

cuts <- seq(0, 300, 25)
lifetab_dat <- train_1 %>%
  mutate(time_cat = cut(RMNG_TERM, cuts)) %>%
  group_by(time_cat) %>%
  summarise(nlost = sum(Default == 1),
            nevent = sum(Default == 0))

dat_lt <- with(lifetab_dat, lifetab(tis = cuts, ninit = nrow(train_1),
                                    nlost = nlost, nevent = nevent))
round(dat_lt, 4)
#Nelson-Aalen estimator
fit_fh <- survfit(su_obj ~ 0, data = train, type = "fleming-harrington", conf.type = "log-log")
dat_fh <- fortify(fit_fh)
head(dat_fh)
summary(fit_fh, times = c(0, 5, 10,60, 110, 160, 210, 260, 359))
summary(fit_fh, times = c(0,1,2,3,4, 5, 10,60, 110, 160, 210, 260, 359))
#Graphical comparison
```

```
ggplotly(
  ggplot() +
    geom_step(data = dat_km, aes(x = time, y = surv, colour = "K-M")) +
    geom_step(data = dat_fh, aes(x = time, y = surv, colour = "N-A")) +
    geom_step(data = dat_lt, aes(x = cuts[-length(cuts)], y = surv, colour = "LT")) +
    labs(x = "Time (months)", y = "Survival", colour = "Estimator") +
    theme_classic()
)
#Measures of central tendency
require("survival")
(mc <- data.frame(q = c(.25, .5, .75),
                  km = quantile(fit_km),
                  fh = quantile(fit_fh)))
ggsurvplot(fit_km, xlab = "Time (months)", censor = F)$plot +
  geom_segment(data = mc, aes(x = km.quantile, y = 1-q, xend = km.quantile, yend = 0), lty = 2) +
  geom_segment(data = mc, aes(x = 0, y = 1-q, xend = km.quantile, yend = 1-q), lty = 2)
```

C.4 Comparison of Survival Curves

```
#ci.exp(glm(all ~ 0 + stage, data = orca, family = "poisson", offset = log(time)))
group_by(train_1, train_1$INCOME_BAND) %>%
  summarise(
    D = sum(Default),
    Y = sum(RMNG_TERM)
  ) %>%
  cbind(
    pois.approx(x = .$D, pt = .$Y)
  )


su_stg  <- survfit(su_obj ~INCOME_BAND, data = train_1)
su_stg

ggsurvplot(su_stg, fun = "event", censor = F, xlab = "Time (months)")

ggsurvplot(su_stg, conf.int=TRUE, pval=TRUE, risk.table=TRUE,
           legend.labs=c("Male", "Female"), legend.title="Sex",
           palette=c("dodgerblue2", "orchid2"),
           title="Kaplan-Meier Curve for Mortgage loans Survival",
```

```
        risk.table.height=.15)


ggsurvplot(su_stg, conf.int = TRUE, risk.table.col = "strata",
        ggtheme = theme_bw(),
        palette = c("#E7B800", "#2E9FDF", "#0000FF"),fun = "cumhaz")

summary(su_stg)
lifetab_stg <- fortify(su_stg)
lifetab_stg %>%
  group_by(strata) %>%
  do(head(., n = 3))



glist <- list(
  ggsurvplot(su_stg, fun = "cumhaz"),
  ggsurvplot(su_stg, fun = "cloglog")
)
# plot(su_stg, fun = "cloglog")
arrange_ggsurvplots(glist, print = TRUE, ncol = 2, nrow = 1)

#Mantel-Haenszel logrank test
#

survdiff(su_obj ~ INCOME_BAND, data = train_1)

#Peto & Peto modification of the Gehan-Wilcoxon test

survdiff(su_obj ~ INCOME_BAND, data = train, rho = 1)

#-------------------------------------------------------------------------------------------

#ci.exp(glm(all ~ 0 + stage, data = orca, family = "poisson", offset = log(time)))
group_by(train_1, train_1$Gender) %>%
  summarise(
    D = sum(Default),
    Y = sum(RMNG_TERM)
  ) %>%
  cbind(
```

```
    pois.approx(x = .$D, pt = .$Y)
  )

su_stg  <- survfit(su_obj ~Gender, data = train)
su_stg

ggsurvplot(su_stg, fun = "event", censor = F, xlab = "Time (months)")

ggsurvplot(su_stg, conf.int=TRUE, pval=TRUE, risk.table=TRUE,
          legend.labs=c("Male", "Female"), legend.title="Sex",
          palette=c("dodgerblue2", "orchid2"),
          title="Kaplan-Meier Curve for Mortgage loans Survival",
          risk.table.height=.15)



summary(su_stg)
lifetab_stg <- fortify(su_stg)
lifetab_stg %>%
  group_by(strata) %>%
  do(head(., n = 3))

ggsurvplot(su_stg, conf.int = TRUE, risk.table.col = "strata",
          ggtheme = theme_bw(),
          palette = c("#E7B800", "#2E9FDF"),fun = "cumhaz")

#Mantel-Haenszel logrank test
#

survdiff(su_obj ~ Gender, data = train_1)
```

C.5 Cox PH Regression

SAS Code

```
/* Data exploration with proc univariate and proc corr*/

/*1. Data exploration with proc univariate and proc corr**/

/*data training;*/
/*set hl.whas500;*/
/*run;*/
/**/
```

```
/*proc corr data = training plots(maxpoints=none)=matrix(histogram);*/
/*var lenfol gender age bmi hr;*/
/*run;*/
/**/
/*proc lifetest data=training atrisk plots=survival(cb)
outs=outwhas500;*/
/*time lenfol*fstat(0);*/
/*run;*/


ods output fitstatistics = fitness;
ods output modelbuildingsummary = build;
ods output classlevelinfo = levels;
ods output ParameterEstimates = estimates;
proc phreg data = Train;
CLASS Gender ;
model rmng_term*default(0) =
PAST_DUE_AMT
CLIENT_BUREAU_SCORE
PREVIOUS_AMNT_PAID
TERM_PAID
/*RMNG_TERM*/
BOND_AMT
Highest_Education_Level
PURCHASE_PRICE
MORTGAGE_INTEREST_RATE
LTV
MNTH_REPAYMENT_AMNT
HPI_Growth_Yoy_perc
Inflation_Growth_YoY
Prime_rate_level_eop
/selection=stepwise  slentry=0.30 slstay=0.05  details;
;;
run;


/*Correlation*/

ods graphics on;
proc corr data = Train plots(maxpoints=none)=matrix(histogram);
var PAST_DUE_AMT
CLIENT_BUREAU_SCORE
PREVIOUS_AMNT_PAID
TERM_PAID
RMNG_TERM
BOND_AMT
Highest_Education_Level
PURCHASE_PRICE
MORTGAGE_INTEREST_RATE
LTV
MNTH_REPAYMENT_AMNT
;run;
ods graphics off;

/* Prepare the correlations coeff matrix: Pearson's r method */
%macro prepCorrData(in=,out=);
  /* Run corr matrix for input data, all numeric vars */
  proc corr data=&in. noprint
```

```
    pearson
    outp=work._tmpCorr
    vardef=df
  ;
  run;

  /* prep data for heat map */
data &out.;
  keep x y r;
  set work._tmpCorr(where=(_TYPE_="CORR"));
  array v{*} _numeric_;
  x = _NAME_;
  do i = dim(v) to 1 by -0.5;
    y = vname(v(i));
    r = v(i);
    /* creates a lower triangular matrix */
    if (i<_n_) then
      r=.;
    output;
  end;
run;

proc datasets lib=work nolist nowarn;
  delete _tmpcorr;
quit;
%mend;

  /* Create a heat map implementation of a correlation matrix */
ods path work.mystore(update) sashelp.tmplmst(read);

proc template;
  define statgraph corrHeatmap;
   dynamic _Title;
    begingraph;
      entrytitle _Title;
      rangeattrmap name='map';
      /* select a series of colors that represent a "diverging"  */
      /* range of values: stronger on the ends, weaker in middle */
      /* Get ideas from http://colorbrewer.org                   */
      range -0.5 - 1 / rangecolormodel=(cxD8B365 cxF5F5F5 cx5AB4AC);
      endrangeattrmap;
      rangeattrvar var=r attrvar=r attrmap='map';
      layout overlay /
        xaxisopts=(display=(line ticks tickvalues))
        yaxisopts=(display=(line ticks tickvalues));
        heatmapparm x = x y = y colorresponse = r /
          xbinaxis=false ybinaxis=false
          name = "heatmap" display=all;
        continuouslegend "heatmap" /
          orient = vertical location = outside title="Pearson
Correlation";
      endlayout;
    endgraph;
  end;
run;

/* Build the graphs */
ods graphics /height=800 width=1000 imagemap;
```

```
%prepCorrData(in=Train,out=cars_r);
proc sgrender data=cars_r template=corrHeatmap;
   dynamic _title="Corr matrix for Train";
run;


proc options option=jreoptions; run;

 ods graphics on;
/* example of dropping categorical numerics */
%prepCorrData(
  in=Train(keep = PAST_DUE_AMT
CLIENT_BUREAU_SCORE
PREVIOUS_AMNT_PAID
TERM_PAID
RMNG_TERM
BOND_AMT
Highest_Education_Level
PURCHASE_PRICE
MORTGAGE_INTEREST_RATE
LTV
MNTH_REPAYMENT_AMNT


),
  out=Train2);
proc sgrender data=Train2 template=corrHeatmap;
  dynamic _title="Corr matrix for Mortgage loans";
run;
 ods graphics off;

 *-------------------------------------------------------------------------
-------------------------;



/*variance inflation factors (VIF).*/
ods graphics on;
proc reg data=Train;
      model default=CLIENT_BUREAU_SCORE
AGE
PAST_DUE_AMT
PURCHASE_PRICE
MORTGAGE_INTEREST_RATE
Highest_Education_Level
BOND_AMT
MNTH_REPAYMENT_AMNT
PREVIOUS_AMNT_PAID
LTV
Marital_ind
CUST_NO_CHILD
 / tol vif collin;
run;
ods graphics off;

*8. Influence Diagnostics;
```

```
*8.1. Inspecting dfβ to assess influence of observations on individual
regression coefficients;

proc phreg data = Train;
class gender(ref=Last) ;
model RMNG_TERM*default(0)=
Highest_Education_Level
CLIENT_BUREAU_SCORE
PAST_DUE_AMT
Highest_Education_Level|CLIENT_BUREAU_SCORE;
output out = dfbeta dfbeta= dfHighest_Education_Level
dfCLIENT_BUREAU_SCORE dfPAST_DUE_AMT
dfHighest_Education_LevelCLIENT_BUREAU_SCORE;
run;

proc print data = train(where=(cust_no=18938619 or cust_no=14183210));
var cust_no Highest_Education_Level
CLIENT_BUREAU_SCORE
PAST_DUE_AMT;
run;

proc sgplot data = dfbeta;
scatter x = Gender y=dfGender / markerchar=cust_no;
run;


proc sgplot data = dfbeta;
scatter x = PREVIOUS_AMNT_PAID  y=dfPREVIOUS_AMNT_PAID /
markerchar=cust_no;
run;

proc sgplot data = dfbeta;
scatter x = HPI_Growth_Yoy_perc y=dfHPI_Growth_Yoy_perc/
markerchar=cust_no;
run;


proc sgplot data = dfbeta;
scatter x = Inflation_Growth_YoY y=dfInflation_Growth_YoY /
markerchar=cust_no;
run;

proc sgplot data = dfbeta;
scatter x = Prime_rate_level_eop y=dfPrime_rate_level_eop /
markerchar=cust_no;
run;



proc print data = train(where=(cust_no=18938619 or cust_no=14183210));
var cust_no Highest_Education_Level
CLIENT_BUREAU_SCORE
PAST_DUE_AMT;
run;


proc phreg data = Train (where=(cust_no^=18938619 or cust_no^=14183210));
```

```
class sex;
model RMNG_TERM*default(0) =  Highest_Education_Level
CLIENT_BUREAU_SCORE
PAST_DUE_AMT
Highest_Education_Level|CLIENT_BUREAU_SCORE;
output out = dfbeta dfbeta= dfHighest_Education_Level
dfCLIENT_BUREAU_SCORE dfPAST_DUE_AMT
dfHighest_Education_LevelCLIENT_BUREAU_SCORE;
run;



*8.2. Plotting likelihood displacement scores to assess influence on the
overall model ;

ods graphics on;
proc phreg data = Train;
class gender(ref=Last) ;
model RMNG_TERM*default(0)=
Highest_Education_Level
CLIENT_BUREAU_SCORE
PAST_DUE_AMT
Highest_Education_Level|CLIENT_BUREAU_SCORE;
output out=ld ld=ld;
run;

proc sgplot data=ld;
scatter x=RMNG_TERM y=ld/ markerchar=cust_no;
run;

ods graphics off;
```

R Code

```
#Stepwise - selected variables
#Model
m2 <-coxph(su_obj ~ Highest_Education_Level +
          CLIENT_BUREAU_SCORE +
          PAST_DUE_AMT +
          MORTGAGE_INTEREST_RATE+
          TERM_PAID +
          MNTH_REPAYMENT_AMNT
        ,
          data=train_1)
step(m2)
#Model Summary
summary(m2)
library(survminer)
library(plyr)
```

```
rename(train_1, c("Highest_Education_Level"="Highest_Education_Level",
"CLIENT_BUREAU_SCORE"="Client_Bureau_Score",
        "PAST_DUE_AMT" = "Historical_Amount_Due", "MNTH_REPAYMENT_AMNT" =
"Repayment_Value"))
require("survival")
model <- coxph( Surv(train_1$RMNG_TERM, Default) ~ Highest_Education_Level +
          Client_Bureau_Score +
          Historical_Amnt_Due +
          Mortgage_Interest_Rate +
          TERM_PAID +
          Mnth_Repayment_Value,
         data = train_1 )
ggforest(model)



m2 <- coxph(su_obj ~
        Highest_Education_Level*Client_Bureau_Score +
        Historical_Amnt_Due*Mnth_Repayment_Value +
        Mortgage_Interest_Rate*Client_Bureau_Score+
        TERM_PAID*Mnth_Repayment_Value
       , data = train_1
       )
summary(m2)

#Assess model fit
cox.zph.m2 <- cox.zph(m2)
cox.zph.m2
ggcoxzph(cox.zph.m2, var = 1)
ggcoxzph(cox.zph.m2, var = 2)
ggcoxzph(cox.zph.m2, var = 3)
ggcoxzph(cox.zph.m2, var = 4)
ggcoxzph(cox.zph.m2, var = 5)
ggcoxzph(cox.zph.m2, var = 6)
ggcoxzph(cox.zph.m2, var = 7)

#Martingale residual
m3 <- coxph(su_obj ~
        Highest_Education_Level*Client_Bureau_Score +
```

```
            Highest_Education_Level+
            Client_Bureau_Score+
            Historical_Amnt_Due
          , data = train_1
)
summary(m3)
library(tidyverse)
library(survival)
data(train_1)
head(heart)

train_1$resid_mart <- residuals(m3, type = "martingale")

ggplot(data = train_1, mapping = aes(x = AGE, y = resid_mart)) +
  geom_point() +
  geom_smooth() +
  labs(title = "AGE") +
  theme_bw() + theme(legend.key = element_blank())

## Cox-Snell residuals
train_1$resid_coxsnell <- -(train_1$resid_mart - train_1$Default)
## Fit model on Cox-Snell residuals (Approximately Expo(1) distributed under correct model)
fit_coxsnell <- coxph(formula = Surv(resid_coxsnell, Default) ~ 0,
                      data    = train_1,
                      ties    = c("efron","breslow","exact")[1])
## Nelson-Aalen estimator for baseline hazard (all covariates zero)
df_base_haz <- basehaz(fit_coxsnell, centered = FALSE)
head(df_base_haz)
## Plot
ggplot(data = df_base_haz, mapping = aes(x = time, y = hazard)) +
  geom_point() +
  scale_x_continuous(limit = c(0,3.5)) +
  scale_y_continuous(limit = c(0,3.5)) +
  labs(x = "Cox-Snell residuals as pseudo observed times",
       y = "Estimated cumulative hazard at pseudo observed times") +
  theme_bw() + theme(legend.key = element_blank())
#Testing influential observations
ggcoxdiagnostics(m3, type = , linear.predictions = TRUE)
ggcoxdiagnostics(m3, type = "dfbeta",
```

```
            linear.predictions = FALSE, ggtheme = theme_bw())
ggcoxdiagnostics(m3, type = "deviance",
            linear.predictions = FALSE, ggtheme = theme_bw())

#Non-linearity
m3 <- coxph(Surv(RMNG_TERM, Default) ~  Highest_Education_Level +
            CLIENT_BUREAU_SCORE +
            PAST_DUE_AMT +
            MORTGAGE_INTEREST_RATE+
            TERM_PAID +
            MNTH_REPAYMENT_AMNT + I(AGE-65) + I((AGE-65)^2), data = train_1)
summary(m3)
AGE <- seq(20, 80, 1)
hrtab <- ci.exp(m3, ctr.mat = cbind(0, AGE, AGE^2, 0, 0, 0, 0, 0))
ggplot(data.frame(hrtab), aes(x = AGE, y = exp.Est.., ymin = X2.5., ymax = X97.5.)) +
  geom_line() + geom_ribbon(alpha = .1) +
  scale_y_continuous(trans = "log", breaks =  pretty_breaks()) +
  labs(x = "AGE (years)", y = "Hazard ratio") + theme_classic() +
  geom_vline(xintercept = 65, lty = 2) + geom_hline(yintercept = 1, lty = 2)

#MODEL PERFORMANCE
# TRAIN
m4 <- coxph(su_obj ~
            Highest_Education_Level*CLIENT_BUREAU_SCORE +
            Highest_Education_Level+
            CLIENT_BUREAU_SCORE+
            PAST_DUE_AMT
           , data = train_1
)
summary(m4)
# Classification Table for train dataset
train_1$m2_score <- predict(m4,type='lp',train_1)
train_1.pred2 <- ifelse(train_1$m2_score > .5, 1, 0)
table(train_1.pred2, train_1$Default)
mean(train_1.pred2 == train_1$Default)
confusionMatrix(table(train_1.pred2, train_1$Default))
#ROC Curve for Train dataset. For this we require ROCR pacakge.
require(ROCR)
train_1.roc2 <- prediction(train_1$m2_score, train_1$Default)
```

```
plot(performance(train_1.roc2, "tpr", "fpr"), col = "Black", main = "Cox Regression with selected variables Train Data")
abline(0, 1, lty = 8, col = "blue")
train_1.auc2 <- performance(train_1.roc2, "auc")
slot(train_1.auc2, "y.values")
m1_1_pred <- prediction(train_1$m2_score,train_1$Default)
m1_1_perf <- performance(m1_1_pred,"tpr","fpr")

# Model Scoring
train_1$m2_score <- predict(m4,type='lp',train_1)
m2_pred <- prediction(train_1$m2_score,train_1$Default)
m2_perf <- performance(m2_pred,"tpr","fpr")
train_1.pred <- ifelse(train_1$m2_score > .5, "1", "0")
table(train_1.pred, train_1$Default)
mean( train_1.pred== train_1$Default)
train_1.roc2 <- prediction(train_1$m2_score, train_1$Default)
plot(performance(train_1.roc2, "tpr", "fpr"), col = "red", main = "ROC Curve Default Model Train Data")
abline(0, 1, lty = 8, col = "blue")
train_1.auc <- performance(train_1.roc, "auc")
slot(train_1.auc, "y.values")

#Model Performance plot
plot(m2_perf, lwd=2, colorize=TRUE,main = " ROC m2: Cox Regression with selected variables")
lines(x=c(0, 1), y=c(0, 1), col="red", lwd=1, lty=3);
lines(x=c(1, 0), y=c(0, 1), col="green", lwd=1, lty=4)
# lines(x=c(0.5, 0), y=c(0, 0.5), col="orange", lwd=1, lty=4)
# lines(x=c(1, 0.5), y=c(1, 0.5), col="blue", lwd=1, lty=4)

# Plot precision/recall curve
m2_perf_precision <- performance(m2_pred, measure = "prec", x.measure = "rec")
plot(m2_perf_precision, main="m2 Cox:Precision/recall curve")
# Plot accuracy as function of threshold
m2_perf_acc <- performance(m2_pred, measure = "acc")
plot(m2_perf_acc, main="m2 Cox:Accuracy as function of threshold")

##Cutoff and Accuracy
ind = which.max( slot(m2_perf_acc, "y.values")[[1]] )
acc = slot(m2_perf_acc, "y.values")[[1]][ind]
```

```
cutoff = slot(m2_perf_acc, "x.values")[[1]][ind]
print(c(accuracy= acc, cutoff = cutoff))
library(blorr)
library(magrittr)
#KS & AUC m2
m2_AUROC <- round(performance(m2_pred, measure = "auc")@y.values[[1]]*100, 2)
m2_KS <- round(max(attr(m2_perf,'y.values')[[1]]-attr(m2_perf,'x.values')[[1]])*100, 2)
m2_Gini <- (2*m2_AUROC - 100)
cat("AUROC: ",m2_AUROC,"\tKS: ", m2_KS, "\tGini:", m2_Gini, "\n")



#KS Statistic

ks.train_1 <- performance(m2_pred, "tpr", "fpr")
train_1.ks2 <- max(attr(ks.train_1, "y.values")[[1]] - (attr(ks.train_1, "x.values")[[1]]))
train_1.ks2

##Hosmer-Lemeshow test on train dataset;
##
require(ResourceSelection)
hl2.train_1 <- hoslem.test(train_1$Default, fitted(m4), g = 10)
hl2.train_1

#Percentage of Concordance for Train dataset. For this we need to run this function.

OptimisedConc=function(model)
{
  Data = cbind(model$y, model$fitted.values)
  ones = Data[Data[,1] == 1,]
  zeros = Data[Data[,1] == 0,]
  conc=matrix(0, dim(zeros)[1], dim(ones)[1])
  disc=matrix(0, dim(zeros)[1], dim(ones)[1])
  ties=matrix(0, dim(zeros)[1], dim(ones)[1])
  for (j in 1:dim(zeros)[1])
  {
    for (i in 1:dim(ones)[1])
    {
      if (ones[i,2]>zeros[j,2])
      {conc[j,i]=1}
```

```
    else if (ones[i,2]<zeros[j,2])
    {disc[j,i]=1}
    else if (ones[i,2]==zeros[j,2])
    {ties[j,i]=1}
   }
  }
  Pairs=dim(zeros)[1]*dim(ones)[1]
  PercentConcordance=(sum(conc)/Pairs)*100
  PercentDiscordance=(sum(disc)/Pairs)*100
  PercentTied=(sum(ties)/Pairs)*100
  return(list("Percent                                    Concordance"=PercentConcordance,"Percent
Discordance"=PercentDiscordance,"Percent Tied"=PercentTied,"Pairs"=Pairs))
}

OptimisedConc(m4)

# performance of the model on the train dataset score table
#
#We will do the credit score for train dataset as well
require(dplyr)
require(ggplot2)

train_1.final2 <- data.frame(train_1, train_1$m2_score)
train_1.final2$m2_score <- round(train_1.final2$m2_score, 2)
head(train_1.final2)

#Divide the train dataset into ten decile

train_1.f2 <- arrange(train_1.final2, desc(train_1$m2_score))
train_1.f2$decile <- with(train_1.f2, cut_number(train_1$m2_score, 10, labels = 10:1))
head(train_1.f2)
tail(train_1.f2)
#Score table for train dataset
train_1.score2 <- train_1.f2 %>% group_by(decile)
train_1.score1 <- train_1.score %>%
  summarise_each(funs(sum), Default) %>%
  arrange(desc(decile))
train_1.score2 <- train_1.f2 %>%
  group_by(decile) %>%
```

```
  summarise(Default = n()) %>%
  arrange(desc(decile))
train_1.table <- left_join(train_1.score1, train_1.score2, by = "decile")
train_1.table <- rename(train_1.table, Good = Default.x, CountOfDecile = Default.y)
train_1.table <- mutate(train_1.table, Bad = CountOfDecile - Good)
train_1.table <- mutate(train_1.table, CuGood = cumsum(Good))
train_1.table <- mutate(train_1.table, CuBad = cumsum(Bad))
train_1.table <- mutate(train_1.table, CuGoodPercent = CuGood/4855)
train_1.table$CuGoodPercent <- round(train_1.table$CuGoodPercent, 2)
train_1.table <- mutate(train_1.table, CuBadPercent = CuBad/9040)
train_1.table$CuBadPercent <- round(train_1.table$CuBadPercent, 2)
train_1.table <- mutate(train_1.table, CuBadAvoided = 1 - CuBadPercent)
train_1.table <- mutate(train_1.table, Profit = 100*CuGood - 500*CuBad)
train_1.table

#TEST
m5 <- coxph(Surv(test_1$RMNG_TERM, test_1$Default) ~
Highest_Education_Level*CLIENT_BUREAU_SCORE +
          Highest_Education_Level+
          CLIENT_BUREAU_SCORE+
          PAST_DUE_AMT

         , data = test_1
)
summary(m5)
test_1$m5_score <- predict(m5,type='lp',test_1)
m5_pred <- prediction(test_1$m5_score,test_1$Default)
m5_perf <- performance(m5_pred,"tpr","fpr")
test_1.pred2 <- ifelse(test_1$m5_score > .5, "1", "0")
table(test_1.pred2, test_1$Default)
test_1.roc3 <- prediction(test_1$m5_score, test_1$Default)
plot(performance(test_1.roc3, "tpr", "fpr"), col = "Yellow", main = "ROC Curve Default Model Test
Data")
abline(0, 1, lty = 8, col = "blue")
train_1.auc <- performance(test_1.roc3, "auc")
slot(test_1.auc, "y.values")

# Model Scoring
test_1$m5_score <- predict(m5,type='lp',test_1)
```

```
m5_pred <- prediction(test_1$m5_score,test_1$Default)
m5_perf <- performance(m5_pred,"tpr","fpr")
#Model Performance plot
plot(m5_perf, lwd=2, colorize=TRUE,main = " ROC m5: Cox Regression with selected variables")
lines(x=c(0, 1), y=c(0, 1), col="red", lwd=1, lty=3);
lines(x=c(1, 0), y=c(0, 1), col="green", lwd=1, lty=4)
# lines(x=c(0.5, 0), y=c(0, 0.5), col="orange", lwd=1, lty=4)
# lines(x=c(1, 0.5), y=c(1, 0.5), col="blue", lwd=1, lty=4)
# Plot precision/recall curve
m5_perf_precision <- performance(m5_pred, measure = "prec", x.measure = "rec")
plot(m5_perf_precision, main="m2 Cox:Precision/recall curve")
# Plot accuracy as function of threshold
m5_perf_acc <- performance(m5_pred, measure = "acc")
plot(m5_perf_acc, main="m5 Cox:Accuracy as function of threshold")
#KS & AUC m5
m5_AUROC <- round(performance(m5_pred, measure = "auc")@y.values[[1]]*100, 2)
m5_KS <- round(max(attr(m5_perf,'y.values')[[1]]-attr(m5_perf,'x.values')[[1]])*100, 2)
m5_Gini <- (2*m5_AUROC - 100)
cat("AUROC: ",m5_AUROC,"\tKS: ", m5_KS, "\tGini:", m5_Gini, "\n")
##MODEL COMPARISON
#Compare ROC Performance of Models
plot(m2_perf, col='Green', lty=1, main='ROCs: Model Performance Comparision') # Cox regression
plot(m5_perf, col='Yellow',lty=2, add=TRUE);
legend(0.6,0.5,
       c('m4:Cox regression: Train','m5:Cox regression: Test'),
       col=c('Green','Yellow'),
       lwd=3);
lines(c(0,1),c(0,1),col = "gray", lty = 4 ) # random line
# Performance Table
models <- c('m4:Cox regression: Train', 'm5:Cox regression: Test'
)
# AUCs
models_AUC <- c(m2_AUROC, m5_AUROC)
# KS
models_KS <- c(m2_KS, m5_KS)
# Gini
models_Gini <- c(m2_Gini, m5_Gini)
# Combine AUC and KS
```

model_performance_metric <- as.data.frame(cbind(models, models_AUC, models_KS, models_Gini))

Colnames

colnames(model_performance_metric) <- c("Model", "AUC", "KS", "Gini")

Display Performance Reports

kable(model_performance_metric, caption ="Comparision of Model Performances")

C.6 Logistic Regression

SAS Code

Sampling

```
LIBNAME HL "/grid/nfsshare/department19fs/DD_DAS/Lucky/MLS";
data all;
set hl.all;
run;

proc sort data=all;
by cust_no;
run;

data train test;
      set all;
      by cust_no;

      if first.cust_no then
            do;
                  if ranuni(12345) < 0.7 then
                        destination = 'train';
                  else destination = 'test';
                  retain destination;
            end;

      if destination = 'train' then
            output train;
      else output test;
      drop destination;
run;

data hl.train;
set train;
run;

data hl.test;
set test;
run;


* load data;

data train;
set train; /* insert correct path to file here */
run;
```

```
 title 'Stepwise Regression on Home loans data';
   proc logistic data=train outest=betas covout;
      model default =PREVIOUS_AMNT_PAID
Highest_Education_Level
LTV
CLIENT_BUREAU_SCORE
PAST_DUE_AMT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE
RMNG_TERM
TERM_PAID
MNTH_REPAYMENT_AMNT
BOND_AMT

                    / selection=stepwise
                     slentry=0.30 slstay=0.05
                      details
                      lackfit;
      output out=pred p=phat lower=lcl upper=ucl
             predprob=(individual crossvalidate);
   run;



* initial model;
proc logistic data=Train descending;
/*class Debit_Interest_Rate gender;*/
model default =  CLIENT_BUREAU_SCORE
PREVIOUS_AMNT_PAID
Highest_Education_Level
MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE

 / expb;
run;


*change to reference coding;
proc logistic data=train descending;
class  sex / param=ref;
model default =CLIENT_BUREAU_SCORE
PREVIOUS_AMNT_PAID
Highest_Education_Level
MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE

/ expb;
run;


*******ASSESSING MODEL FIT***************;
* deviance and pearson X2;
proc logistic data=train descending;
/*class sex ;*/
```

```
model default =CLIENT_BUREAU_SCORE
PREVIOUS_AMNT_PAID
Highest_Education_Level
MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE

/ expb aggregate scale=none;
run;


* H&L more appropriate if using continuous covariates;
proc logistic data=train descending;
class Gender / param=ref;
model default =  CLIENT_BUREAU_SCORE
PREVIOUS_AMNT_PAID
Highest_Education_Level
MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE
 / expb lackfit;
run;




* deltad;
proc logistic data=train descending;
/*class sex/ param=ref;*/
model default =CLIENT_BUREAU_SCORE
PREVIOUS_AMNT_PAID
Highest_Education_Level
MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE
/ expb lackfit;
output out=diag difdev=deltad predicted=predp;
run;


* plot of deltad vs predicted probabilities;
ods graphics on;
proc sgplot data=diag;
scatter x=predp y=deltad;
run;
ods graphics off;

* finding the 20 obs with highest residuals;
proc sort data=diag;
by descending deltad;
run;

proc print data=diag(obs=20);
run;
```

```
* redoing deltad vs probability plot split by gender and colored by
class;
proc sgpanel data=diag;
panelby gender;
scatter x=predp y=deltad / group=CLIENT_BUREAU_SCORE;
run;

* exploring nonlinear age;
* create format to split age into intervals;
proc format;
value agecat 0-<15 = '[0,15)'
             15-<30 = '[15,30)'
                  30-<45 = '[30,45)'
                  45-<60 = '[45,60)'
                  60-high = '60+';
run;

* estimate coefficients for age categories (by female) and then plot;
ods graphics on;
proc logistic data=train descending;
class gender(ref=first) age / param=glm;
model default = AGE GENDER CLIENT_BUREAU_SCORE
PREVIOUS_AMNT_PAID
Highest_Education_Level
MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE / expb;
lsmeans age*Gender /  plots=meanplot(join sliceby=Gender);
format age agecat.;
run;
ods graphics off;


******MODELING INTERACTIONS AND NONLINEAR EFFECTS*******;

* adding interactions;
proc logistic data=train descending;
class gender  / param=ref;
model default = Client_Bureau_Score|GENDER  MNTH_REPAYMENT_AMNT|GENDER
CLIENT_BUREAU_SCORE|AGE GENDER|Highest_Education_Level AGE|PURCHASE_PRICE
PREVIOUS_AMNT_PAID
MORTGAGE_INTEREST_RATE
 / expb;
run;


* plotting ORs for pclass across Gender;
* odds ratio statement;
ods graphics on;
proc logistic data=train descending;
class gender  Debit_Interest_Rate / param=ref;
model default = Client_Bureau_Score|GENDER  MNTH_REPAYMENT_AMNT|GENDER
CLIENT_BUREAU_SCORE|AGE GENDER|Highest_Education_Level AGE|PURCHASE_PRICE
PREVIOUS_AMNT_PAID
MORTGAGE_INTEREST_RATE/ expb;
oddsratio Debit_Interest_Rate;
run;
ods graphics off;
```

```
* putting variables on odds ratio graphs;
ods graphics on;
proc logistic data=train descending;
class gender  Debit_Interest_Rate / param=ref;
model default = Debit_Interest_Rate age|Gender CLIENT_BUREAU_SCORE
PURCHASE_PRICE
TERM_PAID
Highest_Education_Level
PREVIOUS_AMNT_PAID
MNTH_REPAYMENT_AMNT  / expb;
oddsratio CLIENT_BUREAU_SCORE;
oddsratio PURCHASE_PRICE;
oddsratio TERM_PAID;
oddsratio Highest_Education_Level;
oddsratio PREVIOUS_AMNT_PAID;
oddsratio MNTH_REPAYMENT_AMNT;
run;
ods graphics off;

* options for odds ratio graphs;
ods graphics on;
proc logistic data=train descending plots(only)=(oddsratio(logbase=e
type=horizontalstat)) ;
class gender  Debit_Interest_Rate / param=ref;
model default = Debit_Interest_Rate age|Gender CLIENT_BUREAU_SCORE
PURCHASE_PRICE
TERM_PAID
Highest_Education_Level
PREVIOUS_AMNT_PAID
MNTH_REPAYMENT_AMNT / expb;
oddsratio gender / at(Debit_Interest_Rate='[11 - 14.2]' age=1 5 18 30 50
70);
run;
ods graphics off;

* cubic splines for age by gender;

proc logistic data=train descending;
effect agesp = spline(age / basis=tpf(noint) naturalcubic
knotmethod=percentiles(5) details);
class gender  Debit_Interest_Rate / param=ref;
model default = Debit_Interest_Rate age|Gender CLIENT_BUREAU_SCORE
PURCHASE_PRICE
TERM_PAID
Highest_Education_Level
PREVIOUS_AMNT_PAID
MNTH_REPAYMENT_AMNT / expb;
run;

* plot for splines;
ods graphics on;
proc logistic data=train descending;
effect agesp = spline(age / basis=tpf(noint) naturalcubic
knotmethod=percentiles(5) details);
class gender  Debit_Interest_Rate / param=ref;
model default = Debit_Interest_Rate age|Gender CLIENT_BUREAU_SCORE
PURCHASE_PRICE
```

```
TERM_PAID
Highest_Education_Level
PREVIOUS_AMNT_PAID
MNTH_REPAYMENT_AMNT  / expb;
effectplot slicefit (x=age sliceby=Gender);
run;
ods graphics off;

* reassessing model fit;
proc logistic data=train descending;
class gender Debit_Interest_Rate / param=ref;
model default = Client_Bureau_Score|GENDER  MNTH_REPAYMENT_AMNT|GENDER
CLIENT_BUREAU_SCORE|AGE GENDER|Highest_Education_Level AGE|PURCHASE_PRICE
PREVIOUS_AMNT_PAID
MORTGAGE_INTEREST_RATE / expb lackfit;
run;


* influence plots;
ods graphics on;
proc logistic data=train descending plots(label)=(influence(unpack)
dfbetas(unpack)) PLOTS(MAXPOINTS=NONE);
class gender Debit_Interest_Rate / param=ref;
model default = CLIENT_BUREAU_SCORE
PREVIOUS_AMNT_PAID
Highest_Education_Level
MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE/ expb lackfit;
run;
ods graphics off;


* identifying influential observations;
data influence;
set train;
if  _N_ in (1860, 6768, 588, 3862, 2363, 8494, 570, 7845, 10964, 9439,
8522, 6769 ) then output;
run;

proc print data=influence (keep= CUST_NO CUST_NO_CHILD
CLIENT_BUREAU_SCORE PREVIOUS_AMNT_PAID HIGH_EDU_LVL MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATEPURCHASE_PRICE default)
 ;
run;

*removing them;
data noninfluence;
set train;
if  ^(_N_ in (570, 7845, 10964, 3439, 8522, 1461, 8864, 2363, 8494)) then
output;
run;

*rerunning - nothing too different NOT RUN IN SEMINAR;
proc logistic data=noninfluence descending;
/*class gender Debit_Interest_Rate / param=ref;*/
model default = CLIENT_BUREAU_SCORE
PREVIOUS_AMNT_PAID
```

```
Highest_Education_Level
MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE/ expb lackfit;
run;

*****************PREDICTIVE POWER************;

* rerunning model to get association table;
proc logistic data=train descending;
class gender Debit_Interest_Rate / param=ref;
model default = CLIENT_BUREAU_SCORE
PREVIOUS_AMNT_PAID
Highest_Education_Level
MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE
 / expb lackfit;
run;

* ROC curve for model;
ods graphics on;
proc logistic data=train descending plots(only)=roc;
class gender Debit_Interest_Rate / param=ref;
model default = CLIENT_BUREAU_SCORE
PREVIOUS_AMNT_PAID
Highest_Education_Level
MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE
/ expb lackfit;
run;
ods graphics off;


/*test*/
ods graphics on;
proc logistic data=test descending plots(only)=roc;
class gender Debit_Interest_Rate / param=ref;
model default = CLIENT_BUREAU_SCORE
PREVIOUS_AMNT_PAID
Highest_Education_Level
MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE
/ expb lackfit;
run;
ods graphics off;

* comparing ROC curve from model with no interactions;
ods graphics on;
proc logistic data=train descending;
class gender Debit_Interest_Rate / param=ref;
model default = age|Gender CLIENT_BUREAU_SCORE
PURCHASE_PRICE
TERM_PAID
Highest_Education_Level
PREVIOUS_AMNT_PAID
```

```
MNTH_REPAYMENT_AMNT / expb lackfit;
roc 'no interactions' CLIENT_BUREAU_SCORE
PURCHASE_PRICE
TERM_PAID
Highest_Education_Level
PREVIOUS_AMNT_PAID
MNTH_REPAYMENT_AMNT;
roccontrast 'no interactions';
run;
ods graphics off;

**************ESTIMATING AND REPORTING PREDICTED
PROBABILITIES*********************************;

* lsmeans for predicted probabilities;
proc logistic data=train descending;
class gender Debit_Interest_Rate / param=ref;
model default = CLIENT_BUREAU_SCORE gender|Debit_Interest_Rate
PREVIOUS_AMNT_PAID
Highest_Education_Level
MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE / expb lackfit;
lsmeans Gender * Debit_Interest_Rate /  ilink;
run;



* lsmeans for predicted probabilities, fixing covariates values;
proc logistic data=train descending;
class gender Debit_Interest_Rate / param=glm;
model default = CLIENT_BUREAU_SCORE gender|Debit_Interest_Rate
PREVIOUS_AMNT_PAID
Highest_Education_Level
MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE / expb lackfit;
lsmeans gender*Debit_Interest_Rate/ at(age CLIENT_BUREAU_SCORE
PURCHASE_PRICE )=(30 500 900000) ilink;
run;

* plots of covariate effects against predicted probabilities;
ods graphics on;
proc logistic data=train descending;
class Gender Debit_Interest_Rate/ param=glm;
model default = CLIENT_BUREAU_SCORE gender|Debit_Interest_Rate
PREVIOUS_AMNT_PAID
Highest_Education_Level
MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE/ expb lackfit;
effectplot interaction (x=Debit_Interest_Rate sliceby=Gender) / clm
connect noobs;
effectplot slicefit (x=Highest_Education_Level sliceby=Gender) / clm;
effectplot slicefit (x=MNTH_REPAYMENT_AMNT sliceby=Gender) / clm;
effectplot slicefit (x=PURCHASE_PRICE sliceby=Gender) / clm;
effectplot slicefit (x=PREVIOUS_AMNT_PAID sliceby=Gender) / clm;
effectplot slicefit (x=CLIENT_BUREAU_SCORE sliceby=Gender) / clm;
```

```
run;
ods graphics off;




**********CODE END**********;


*Cross Validation;

%macro k_fold_cv_rep(r=1,k=10);
ods select none;
%do r=1 %to &r;
proc surveyselect data=all group=&k out=have;
run;

%do i=1 %to &k ;
data training;
 set have(where=(cust_no ne &i)) ;
run;
data testing;
 set have(where=(cust_no eq &i));
run;

ods output
Association=native(keep=label2 nvalue2 rename=(nvalue2=native)
where=(label2='c'))
ScoreFitStat=true(keep=dataset freq auc rename=(auc=true));
proc logistic data=training
 outest=est(keep=_status_ _name_) ;
 class sex;
 model default(event='0')=CLIENT_BUREAU_SCORE
PREVIOUS_AMNT_PAID
Highest_Education_Level
MNTH_REPAYMENT_AMNT
MORTGAGE_INTEREST_RATE
PURCHASE_PRICE;
 score data=test fitstat;
run;

data score_r&r._&i;
 merge true native est;
 retain rep &r cust_no &i;
 optimism=native-true;
run;
%end;
%end;
data k_fold_cv_rep;
 set score_r:;
run;

ods select all;
%mend;

%k_fold_cv_rep(r=20,k=10);
```

```
/*******************/
data all;
 set k_fold_cv k_fold_cv_rep indsname=indsn;
 length indsname $ 32;
 indsname=indsn;
run;
proc summary data=all nway;
 class indsname;
 var optimism;
 output out=want mean=mean lclm=lclm uclm=uclm;
run;
```

R Code

```
#Stepwise - selected variables
#Model
m1_1 <- glm(Default~Highest_Education_Level+
          CLIENT_BUREAU_SCORE+
          PREVIOUS_AMNT_PAID+
          MORTGAGE_INTEREST_RATE+
          PURCHASE_PRICE +
          MNTH_REPAYMENT_AMNT,
        data=train_1,family=binomial())
step(m1_1)



#Model Summary
summary(m1_1)
library(DHARMa)
blr_regress(m1_1)
#dat.sim <- simulateResiduals(m1_1)
#dat.sim
#plotSimulatedResiduals(dat.sim)
##TRAIN DATA
# Model Scoring
# Classification Table for train dataset
train_1$m1_1_score <- predict(m1_1,type='response',train_1)
train_1.pred <- ifelse(train_1$m1_1_score > .5, 1, 0)
table(train_1.pred, train_1$Default)
mean(train_1.pred == train_1$Default)
confusionMatrix(table(train_1.pred, train_1$Default))
#ROC Curve for Train dataset. For this we require ROCR pacakge.
```

```
require(ROCR)
train_1.roc <- prediction(train_1$m1_1_score, train_1$Default)
plot(performance(train_1.roc, "tpr", "fpr"), col = "Black", main = "ROC Curve Default Model Train Data")
abline(0, 1, lty = 8, col = "blue")
train_1.auc <- performance(train_1.roc, "auc")
slot(train_1.auc, "y.values")
m1_1_pred <- prediction(train_1$m1_1_score,train_1$Default)
m1_1_perf <- performance(m1_1_pred,"tpr","fpr")
#Model Performance plot
plot(m1_1_perf, lwd=2, colorize=TRUE,main = " ROC m1_1: Logistic Regression with selected variables")
lines(x=c(0, 1), y=c(0, 1), col="red", lwd=1, lty=3);
lines(x=c(1, 0), y=c(0, 1), col="green", lwd=1, lty=4)
# lines(x=c(0.5, 0), y=c(0, 0.5), col="orange", lwd=1, lty=4)
# lines(x=c(1, 0.5), y=c(1, 0.5), col="blue", lwd=1, lty=4)
# Plot precision/recall curve
m1_1_perf_precision <- performance(m1_1_pred, measure = "prec", x.measure = "rec")
plot(m1_1_perf_precision, main="m1_1 Logistic:Precision/recall curve")
# Plot  as function of threshold
m1_1_perf_acc <- performance(m1_1_pred, measure = "acc")
plot(m1_1_perf_acc, main="m1_1 Logistic: as function of threshold")
ind = which.max( slot(m1_1_perf_acc, "y.values")[[1]] )
acc = slot(m1_1_perf_acc, "y.values")[[1]][ind]
cutoff = slot(m1_1_perf_acc, "x.values")[[1]][ind]
print(c(accuracy= acc, cutoff = cutoff))
#KS & AUC m1_1
m1_1_AUROC <- round(performance(m1_1_pred, measure = "auc")@y.values[[1]]*100, 2)
m1_1_KS <- round(max(attr(m1_1_perf,'y.values')[[1]]-attr(m1_1_perf,'x.values')[[1]])*100, 2)
m1_1_Gini <- (2*m1_1_AUROC - 100)
cat("AUROC: ",m1_1_AUROC,"\tKS: ", m1_1_KS, "\tGini:", m1_1_Gini, "\n")
auc = performance(m1_1_pred, "auc")
auc = unlist(auc@y.values)
auc
# Cross Validatio
#load Data Analysis And Graphics Package for R (DAAG)
library(DAAG)
install.packages("blorr")
library(remotes)
```

```
install_github("cran/blorr")
library(remotes)
install_version("blorr", "3.3")
library(blorr)
library(magrittr)
#calculate accuracy over 100 random folds of data for simple logit
m1_h <- CVbinary(obj=m1_1, rand=NULL, nfolds=100, print.details=TRUE)
#Gains Table
blr_gains_table(m1_1)
#KS Statistic
ks.train_1 <- performance(train_1.roc, "tpr", "fpr")
train_1.ks <- max(attr(ks.train_1, "y.values")[[1]] - (attr(ks.train_1, "x.values")[[1]]))
train_1.ks
#Lift Chart
m1_1 %>%
  blr_gains_table() %>%
  plot()
# Creating performance object
perf.obj <- prediction(predictions=train_1$m1_1_score,
                       labels=train_1$Default)
# Get data for ROC curve
lift.obj <- performance(perf.obj, measure="lift", x.measure="rpp")
plot(lift.obj,
     main="Cross-Sell - Lift Chart",
     xlab="% Populations",
     ylab="Lift",
     col="blue")
abline(1,0,col="grey")

install.packages("gains")
library(gains)
# gains table
actual <- ifelse(train$Default==1,1,0)
gains.cross <- gains(actual=actual ,
                     predicted=train_1$m1_1_score,
                     groups=10)
print(gains.cross)
#KS Chart
m1_1 %>%
```

```
  blr_gains_table() %>%
  blr_ks_chart()
#Lorenz Curve
blr_lorenz_curve(m1_1)
##Hosmer-Lemeshow test on train dataset;
require(ResourceSelection)
hl.train_1 <- hoslem.test(train_1$Default, fitted(m1_1), g = 10)
hl.train_1
blr_test_hosmer_lemeshow(m1_1)
#Gains table
blr_gains_table(m1_1)
# Creating performance object
perf.obj <- prediction(predictions=train_1$m1_1_score,
                       labels=train_1$Default)
#Lift Chart
# Get data for ROC curve
lift.obj <- performance(perf.obj, measure="lift", x.measure="rpp")
plot(lift.obj,
     main="Default - Lift Chart",
     xlab="% Populations",
     ylab="Lift",
     col="blue")
abline(1,0,col="grey")
#Cumulative Lift Chart using R;
install.packages("gains")
library(gains)
# gains table
actual <- ifelse(train_1$Default==1,1,0)
gains.cross <- gains(actual=actual ,
                     predicted=train_1$m1_1_score,
                     groups=10)
print(gains.cross)
#Percentage of Concordance for Train dataset. For this we need to run this function.
OptimisedConc=function(model)
{
  Data = cbind(model$y, model$fitted.values)
  ones = Data[Data[,1] == 1,]
  zeros = Data[Data[,1] == 0,]
  conc=matrix(0, dim(zeros)[1], dim(ones)[1])
```

```
 disc=matrix(0, dim(zeros)[1], dim(ones)[1])
 ties=matrix(0, dim(zeros)[1], dim(ones)[1])
 for (j in 1:dim(zeros)[1])
 {
  for (i in 1:dim(ones)[1])
  {
    if (ones[i,2]>zeros[j,2])
    {conc[j,i]=1}
    else if (ones[i,2]<zeros[j,2])
    {disc[j,i]=1}
    else if (ones[i,2]==zeros[j,2])
    {ties[j,i]=1}
  }
 }
 Pairs=dim(zeros)[1]*dim(ones)[1]
 PercentConcordance=(sum(conc)/Pairs)*100
 PercentDiscordance=(sum(disc)/Pairs)*100
 PercentTied=(sum(ties)/Pairs)*100
 return(list("Percent                    Concordance"=PercentConcordance,"Percent
Discordance"=PercentDiscordance,"Percent Tied"=PercentTied,"Pairs"=Pairs))
}
OptimisedConc(m2)
# performance of the model on the train dataset score table
#
#We will do the credit score for train dataset as well
require(dplyr)
require(ggplot2)
train_1.final <- data.frame(train_1, train_1$m1_1_score)
train_1.final$m1_1_score <- round(train_1.final$m1_1_score, 2)
head(train_1.final)
#Divide the train dataset into ten decile
train_1.f <- arrange(train_1.final, desc(train_1$m1_1_score))
train_1.f$decile <- with(train_1.f, cut_number(train_1$m1_1_score, 10, labels = 10:1))
head(train_1.f)
tail(train_1.f)
#Score table for train dataset
train_1.score <- train_1.f %>% group_by(decile)
train_1.score1 <- train_1.score %>%
 summarise_each(funs(sum), Default) %>%
```

```
  arrange(desc(decile))
train_1.score2 <- train_1.f %>%
  group_by(decile) %>%
  summarise(Default = n()) %>%
  arrange(desc(decile))
train_1.table <- left_join(train_1.score1, train_1.score2, by = "decile")
train_1.table <- rename(train_1.table, Good = Default.x, CountOfDecile = Default.y)
train_1.table <- mutate(train_1.table, Bad = CountOfDecile - Good)
train_1.table <- mutate(train_1.table, CuGood = cumsum(Good))
train_1.table <- mutate(train_1.table, CuBad = cumsum(Bad))
train_1.table <- mutate(train_1.table, CuGoodPercent = CuGood/4855)
train_1.table$CuGoodPercent <- round(train_1.table$CuGoodPercent, 2)
train_1.table <- mutate(train_1.table, CuBadPercent = CuBad/9040)
train_1.table$CuBadPercent <- round(train_1.table$CuBadPercent, 2)
train_1.table <- mutate(train_1.table, CuBadAvoided = 1 - CuBadPercent)
train_1.table <- mutate(train_1.table, Profit = 100*CuGood - 500*CuBad)
train_1.table
require(boot)
?cv.glm
glm.fit=glm(Default~Highest_Education_Level+
            CLIENT_BUREAU_SCORE+
            PREVIOUS_AMNT_PAID+
            MORTGAGE_INTEREST_RATE+
            PURCHASE_PRICE +
            MNTH_REPAYMENT_AMNT, data=test_1)

#LOOCV
cv.glm(test_1,glm.fit)$delta

loocv=function(fit){
  h=lm.influence(fit)$h
  mean((residuals(fit)/(1-h))^2)
}

# A vector for collecting the errors.
cv.error=vector(mode="numeric",length=5)
# The polynomial degree
degree=1:5
# A fit for each degree
```

```
for(d in degree){
  glm.fit=glm(Default~Highest_Education_Level+
                CLIENT_BUREAU_SCORE+
                PREVIOUS_AMNT_PAID+
                MORTGAGE_INTEREST_RATE+
                PURCHASE_PRICE +
                MNTH_REPAYMENT_AMNT, data=test_1)
  cv.error[d]=loocv(glm.fit)
}
# The plot of the errors
plot(degree,cv.error,type="b")
# load the library
library(caret)
# load the iris dataset
data(all)
# define training control
train_control <- trainControl(method="repeatedcv", number=10, repeats=3)
# train the model
model <- train(Default~., data=all, trControl=train_control, method="nb")
# summarize results
print(model)
```

```
#get results of terms in regression
g<-predict(m1_1,type='terms',test_1)
#function to pick top 3 reasons
#works by sorting coefficient terms in equation
# and selecting top 3 in sort for each loan scored
ftopk<- function(x,top=3){
```

```
  res=names(x)[order(x, decreasing = TRUE)][1:top]
  paste(res,collapse=";",sep="")
}
# Application of the function using the top 3 rows
topk=apply(g,1,ftopk,top=3)
#add reason list to scored tests sample
test_1<-cbind(test_1, topk)
#cross validation
auc = performance(m1_1_pred, "auc")
auc = unlist(auc@y.values)
#auc
#########################################################################################
########################################
#########################################################################################
########################################

##TESTING DATA
#Stepwise - selected variables
#Model
m1_3 <- glm(Default~Highest_Education_Level+
              CLIENT_BUREAU_SCORE+
              PREVIOUS_AMNT_PAID+
              MORTGAGE_INTEREST_RATE+
              PURCHASE_PRICE +
              MNTH_REPAYMENT_AMNT,
            data=test_1,family=binomial())
step(m1_3)
#Model Summary
summary(m1_3)
library(DHARMa)
blr_regress(m1_3)
#dat.sim <- simulateResiduals(m1_1)
#dat.sim
#plotSimulatedResiduals(dat.sim)
##TRAIN DATA
# Model Scoring
# Classification Table for train dataset
test_1$m1_3_score <- predict(m1_3,type='response',test_1)
test_1.pred <- ifelse(test_1$m1_3_score > .5, 1, 0)
```

```
table(test_1.pred, test_1$Default)
mean(test_1.pred == test_1$Default)
confusionMatrix(table(test_1.pred, test_1$Default))



#ROC Curve for Train dataset. For this we require ROCR pacakge.
require(ROCR)
train_1.roc <- prediction(train_1$m1_1_score, train_1$Default)
plot(performance(train_1.roc, "tpr", "fpr"), col = "Black", main = "ROC Curve Default Model Train Data")
abline(0, 1, lty = 8, col = "blue")
train_1.auc <- performance(train_1.roc, "auc")
slot(train_1.auc, "y.values")
m1_1_pred <- prediction(train_1$m1_1_score,train_1$Default)
m1_1_perf <- performance(m1_1_pred,"tpr","fpr")
#Model Performance plot
plot(m1_1_perf, lwd=2, colorize=TRUE,main = " ROC m1_1: Logistic Regression with selected variables")
lines(x=c(0, 1), y=c(0, 1), col="red", lwd=1, lty=3);
lines(x=c(1, 0), y=c(0, 1), col="green", lwd=1, lty=4)
# lines(x=c(0.5, 0), y=c(0, 0.5), col="orange", lwd=1, lty=4)
# lines(x=c(1, 0.5), y=c(1, 0.5), col="blue", lwd=1, lty=4)
# Plot precision/recall curve
m1_1_perf_precision <- performance(m1_1_pred, measure = "prec", x.measure = "rec")
plot(m1_1_perf_precision, main="m1_1 Logistic:Precision/recall curve")
# Plot  as function of threshold
m1_1_perf_acc <- performance(m1_1_pred, measure = "acc")
plot(m1_1_perf_acc, main="m1_1 Logistic: as function of threshold")
ind = which.max( slot(m1_1_perf_acc, "y.values")[[1]] )
acc = slot(m1_1_perf_acc, "y.values")[[1]][ind]
cutoff = slot(m1_1_perf_acc, "x.values")[[1]][ind]
print(c(accuracy= acc, cutoff = cutoff))
#KS & AUC m1_1
m1_1_AUROC <- round(performance(m1_1_pred, measure = "auc")@y.values[[1]]*100, 2)
m1_1_KS <- round(max(attr(m1_1_perf,'y.values')[[1]]-attr(m1_1_perf,'x.values')[[1]])*100, 2)
m1_1_Gini <- (2*m1_1_AUROC - 100)
cat("AUROC: ",m1_1_AUROC,"\tKS: ", m1_1_KS, "\tGini:", m1_1_Gini, "\n")
auc = performance(m1_1_pred, "auc")
auc = unlist(auc@y.values)
```

```
auc
# Cross Validatio
#load Data Analysis And Graphics Package for R (DAAG)
library(DAAG)
install.packages("blorr")
library(remotes)
install_github("cran/blorr")
library(remotes)
install_version("blorr", "3.3")
library(blorr)
library(magrittr)
#calculate accuracy over 100 random folds of data for simple logit
m1_h <- CVbinary(obj=m1_1, rand=NULL, nfolds=100, print.details=TRUE)
#Gains Table
blr_gains_table(m1_1)
#KS Statistic
ks.train_1 <- performance(train_1.roc, "tpr", "fpr")
train_1.ks <- max(attr(ks.train_1, "y.values")[[1]] - (attr(ks.train_1, "x.values")[[1]]))
train_1.ks

#Lift Chart
m1_1 %>%
  blr_gains_table() %>%
  plot()
# Creating performance object
perf.obj <- prediction(predictions=train_1$m1_1_score,
                       labels=train_1$Default)
# Get data for ROC curve
lift.obj <- performance(perf.obj, measure="lift", x.measure="rpp")
plot(lift.obj,
     main="Cross-Sell - Lift Chart",
     xlab="% Populations",
     ylab="Lift",
     col="blue")
abline(1,0,col="grey")
install.packages("gains")
library(gains)
# gains table
actual <- ifelse(train$Default==1,1,0)
```

```
gains.cross <- gains(actual=actual ,
                     predicted=train_1$m1_1_score,
                     groups=10)
print(gains.cross)
#KS Chart
m1_1 %>%
  blr_gains_table() %>%
  blr_ks_chart()
#Lorenz Curve
blr_lorenz_curve(m1_1)
##Hosmer-Lemeshow test on train dataset;
##
require(ResourceSelection)
hl.train_1 <- hoslem.test(train_1$Default, fitted(m1_1), g = 10)
hl.train_1
blr_test_hosmer_lemeshow(m1_1)
#Gains table
blr_gains_table(m1_1)
# Creating performance object
perf.obj <- prediction(predictions=train_1$m1_1_score,
                       labels=train_1$Default)
#Lift Chart
# Get data for ROC curve
lift.obj <- performance(perf.obj, measure="lift", x.measure="rpp")
plot(lift.obj,
     main="Default - Lift Chart",
     xlab="% Populations",
     ylab="Lift",
     col="blue")
abline(1,0,col="grey")
#Cumulative Lift Chart using R;
install.packages("gains")
library(gains)
# gains table
actual <- ifelse(train_1$Default==1,1,0)
gains.cross <- gains(actual=actual ,
                     predicted=train_1$m1_1_score,
                     groups=10)
print(gains.cross)
```

```
#Percentage of Concordance for Train dataset. For this we need to run this function.
OptimisedConc=function(model)
{
  Data = cbind(model$y, model$fitted.values)
  ones = Data[Data[,1] == 1,]
  zeros = Data[Data[,1] == 0,]
  conc=matrix(0, dim(zeros)[1], dim(ones)[1])
  disc=matrix(0, dim(zeros)[1], dim(ones)[1])
  ties=matrix(0, dim(zeros)[1], dim(ones)[1])
  for (j in 1:dim(zeros)[1])
  {
    for (i in 1:dim(ones)[1])
    {
      if (ones[i,2]>zeros[j,2])
      {conc[j,i]=1}
      else if (ones[i,2]<zeros[j,2])
      {disc[j,i]=1}
      else if (ones[i,2]==zeros[j,2])
      {ties[j,i]=1}
    }
  }
  Pairs=dim(zeros)[1]*dim(ones)[1]
  PercentConcordance=(sum(conc)/Pairs)*100
  PercentDiscordance=(sum(disc)/Pairs)*100
  PercentTied=(sum(ties)/Pairs)*100
  return(list("Percent                                    Concordance"=PercentConcordance,"Percent
Discordance"=PercentDiscordance,"Percent Tied"=PercentTied,"Pairs"=Pairs))
}
OptimisedConc(m2)
# performance of the model on the train dataset score table
#We will do the credit score for train dataset as well
require(dplyr)
require(ggplot2)
train_1.final <- data.frame(train_1, train_1$m1_1_score)
train_1.final$m1_1_score <- round(train_1.final$m1_1_score, 2)
head(train_1.final)
#Divide the train dataset into ten decile
train_1.f <- arrange(train_1.final, desc(train_1$m1_1_score))
train_1.f$decile <- with(train_1.f, cut_number(train_1$m1_1_score, 10, labels = 10:1))
```

```
head(train_1.f)
tail(train_1.f)
#Score table for train dataset
train_1.score <- train_1.f %>% group_by(decile)
train_1.score1 <- train_1.score %>%
  summarise_each(funs(sum), Default) %>%
  arrange(desc(decile))
train_1.score2 <- train_1.f %>%
  group_by(decile) %>%
  summarise(Default = n()) %>%
  arrange(desc(decile))
train_1.table <- left_join(train_1.score1, train_1.score2, by = "decile")
train_1.table <- rename(train_1.table, Good = Default.x, CountOfDecile = Default.y)
train_1.table <- mutate(train_1.table, Bad = CountOfDecile - Good)
train_1.table <- mutate(train_1.table, CuGood = cumsum(Good))
train_1.table <- mutate(train_1.table, CuBad = cumsum(Bad))
train_1.table <- mutate(train_1.table, CuGoodPercent = CuGood/4855)
train_1.table$CuGoodPercent <- round(train_1.table$CuGoodPercent, 2)
train_1.table <- mutate(train_1.table, CuBadPercent = CuBad/9040)
train_1.table$CuBadPercent <- round(train_1.table$CuBadPercent, 2)
train_1.table <- mutate(train_1.table, CuBadAvoided = 1 - CuBadPercent)
train_1.table <- mutate(train_1.table, Profit = 100*CuGood - 500*CuBad)
train_1.table
require(boot)
?cv.glm
glm.fit=glm(Default~Highest_Education_Level+
            CLIENT_BUREAU_SCORE+
            PREVIOUS_AMNT_PAID+
            MORTGAGE_INTEREST_RATE+
            PURCHASE_PRICE +
            MNTH_REPAYMENT_AMNT, data=test_1)

#LOOCV
cv.glm(test_1,glm.fit)$delta
loocv=function(fit){
  h=lm.influence(fit)$h
  mean((residuals(fit)/(1-h))^2)
}
# A vector for collecting the errors.
```

```
cv.error=vector(mode="numeric",length=5)
# The polynomial degree
degree=1:5
# A fit for each degree
for(d in degree){
  glm.fit=glm(Default~Highest_Education_Level+
                CLIENT_BUREAU_SCORE+
                PREVIOUS_AMNT_PAID+
                MORTGAGE_INTEREST_RATE+
                PURCHASE_PRICE +
                MNTH_REPAYMENT_AMNT, data=test_1)
  cv.error[d]=loocv(glm.fit)
}
# The plot of the errors
plot(degree,cv.error,type="b")



# load the library
library(caret)
# load the iris dataset
data(all)
# define training control
train_control <- trainControl(method="repeatedcv", number=10, repeats=3)
# train the model
model <- train(Default~., data=all, trControl=train_control, method="nb")
# summarize results
print(model)
```

C.7 Model Assessment and Comparisons

```
#Compare ROC Performance of Models Logistic regression training and test data
plot(m1_1_perf, col='Black', lty=1, main='ROC Curve Default Model Train/Test data') # logistic regression
plot(m1_2_perf, col='Blue',lty=3, add=TRUE); # Test Data
legend(x = "bottomright",
       c('Logistic regression Train Dataset','Logistic regression Test Dataset'),
       col=c('Black', 'Blue'),
       lwd=3);
```

```
#lines(c(0,1),c(0,1),col = "gray", lty = 4 ) # random line
# Performance Table
models <- c('m1_1:Logistic regression Train Dataset', 'm1_2: Logistic regression Test Dataset')
# AUCs
models_AUC <- c(m1_1_AUROC, m1_2_AUROC)
# KS
models_KS <- c(m1_1_KS, m1_2_KS)
# Gini
models_Gini <- c(m1_1_Gini, m1_2_Gini)
# Combine AUC and KS
model_performance_metric <- as.data.frame(cbind(models, models_AUC, models_KS, models_Gini))
# Colnames
colnames(model_performance_metric) <- c("Model", "AUC", "KS", "Gini")
# Display Performance Reports
kable(model_performance_metric, caption ="Comparision of Model Performances")
#Cox PH
#Compare ROC Performance of Models Cox regression training and test data
plot(m2_perf, col='Blue', lty=1, main='ROC Curve Default Model Train/Test data') # logistic regression
plot(m5_perf, col='Red',lty=3, add=TRUE); # Test Data
legend(x = "bottomright",
      c('Cox regression Train Dataset','Cox regression Test Dataset'),
      col=c('Black', 'Blue'),
      lwd=3);
#lines(c(0,1),c(0,1),col = "gray", lty = 4 ) # random line
# Performance Table
models <- c('m2:Cox regression Train Dataset', 'm5: Cox regression Test Dataset')
# AUCs
models_AUC <- c(m2_AUROC, m5_AUROC)
# KS
models_KS <- c(m2_KS, m5_KS)
# Gini
models_Gini <- c(m2_Gini, m5_Gini)
# Combine AUC and KS
model_performance_metric <- as.data.frame(cbind(models, models_AUC, models_KS, models_Gini))
# Colnames
colnames(model_performance_metric) <- c("Model", "AUC", "KS", "Gini")
```

```
# Display Performance Reports
kable(model_performance_metric, caption ="Comparision of Model Performances")
```