RABOBANK INTERNATIONAL

Survival analysis in credit scoring

A framework for PD estimation

R. Man

5/9/2014

Document properties

Title	Survival analysis in credit scoring: A framework for PD estimation
Date	May 9, 2014
On behalf of	Rabobank International – Quantitative Risk Analytics & University of Twente
Author	Ramon Man
Rabobank Supervisor	V. Tchistiakov
First Supervisor	B. Roorda
Second Supervisor	R.A.M.G. Joosten
Contact adtress	Ramon.man@rabobank.com / ramon.man90@gmail.com

Management summary

The goal of this thesis is to develop a survival model framework for PD estimation. This framework should use the steps of the current framework. Finally should this model be benchmarked to the currently used logistic regression model. The deliverables are a framework for PD estimation and a survival model to benchmark the current model. Currently a framework for logistic regression is available, however recent papers show that survival analysis might improve PD estimations. Therefore Rabobank International wants to develop a framework for survival models and asses the performance of the model.

Credit scoring systems try to answer the question how likely an applicant for credit is to default within a certain period. The models use scores and ratios (called factors) of the clients that indicate the clients creditworthiness. There are many models available, currently the most commonly used is the logistic regression (LR) approach. This model tries to estimate the number of defaults within a fixed time interval (typically 1 year).
In the recent years survival analysis has been introduced into credit scoring. Survival analysis is the area of statistics that deals with the analysis of lifetime data. The variable of interest is the time to event. This model tries to estimate the survival probability over the entire dataset. The major advantage of survival analysis is the capability to incorporate censored data. For these observations no default is observed during the study period.

The model must be unbiased and therefore as few as possible observations should be removed from the dataset. In the current logistic regression approach censored observations with a short time to maturity (less than one year), are removed from the dataset. Since survival analysis is capable of incorporating these observations, it might provide better results.

The development of the survival model in the PD framework demands some new procedures. The new procedures are:

- Bucketing approach based on a survival function. Factors are divided into buckets depending on their survival function instead of default rate.
- A new transformation of variables based on the survival function is developed. This procedure is called the logrank transformation.
- Another new procedure for the assessment of the predictive power of an individual factor is developed, again based on the survival function.
- Selection procedures are changed to select the most predicting factors for the survival model.
- Procedures for the incorporation of different observations per facility due to new assessments. Facilities in a portfolio are assessed on a interval bases, methodologies for the incorporation of this data per facility are developed.

The survival model developed is a Cox proportional hazard model. The performance of the model is compared to the logistic regression techniques. The logrank transformation outperforms the logistic transformation and the statistical optimal approach, because it is more significant in predicting the survival probability based on the Wald test statistic. Furthermore, the results of the ROC/AUC, power statistic and KS statistic showed there is little difference in the performance of the survival models and the logistic regression.

The model shows no improvement in performance but has certain advantages compared to the current model. This model requires significant less data cleaning because of the model estimates the survival probability over the entire data set, in contrast to logistic regression that only estimates the survival probability for a fixed time interval.

Some remarks for further research are the incorporation of truncation into the survival functions. This is another type of missing data and is not developed because it was beyond the scope of this thesis. Furthermore, the logrank transformation has some problems in case of low number of variables. In this case the logistic transformation outperforms the logrank transformation and is recommended. This should be researched further.
Finally the incorporation of macroeconomic variables, as investigated by Bellotti & Crook [2008], could improve the model significantly and could be further researched.

Contents

1 Introduction

With the introduction of Basel II, banks are allowed to use internally developed models for calculating its regulatory capital. This is called the Internal Ratings-Based (IRB) approach. Otherwise, the bank should calculate its regulatory capital based on the standardized approach which results in higher capital requirements. The purpose of the development of rating models is to identify and combine those factors that differentiate between facilities the best in terms of riskiness. The terms used to describe risk of facilities are probability of default (PD), loss given default (LGD) and exposure at default (EAD). Within Rabobank International (RI) models have been developed by the Modelling and Research department.

In order to determine PD, credit scoring systems were built. They try to answer the question how likely an applicant for credit is to default within a certain period. Many models are available of which currently the most commonly used is the logistic regression (LR) approach, see e.g. Stepanova & Thomas, [2002]. In recent years, survival analysis has been introduced into credit scoring. This collection of statistical methods tries to model the time to default and has advantages compared to other credit scoring systems. The goal of this thesis is to develop a framework for the incorporation of survival models in PD estimation and scorecard development.

First there will be a short background on capital requirements and current PD models, next the research objectives will be described and a further outline of the thesis is given.

1.1 Background

All credit scoring models share the same basic ideas. The first step is to classify a sample of previous customers based on historical repayment performance, into either good or bad. Next step is to link the characteristics (factors) of these clients to the default status of the client. Many different techniques are available for building such systems, for example discriminant analysis, expert systems and logistic regression. Currently logistic regression is an industry standard, see e.g. Stepanova & Thomas [2002] and Tong et al. [2012], for a regression on an outcome variable which is binary.

In recent years survival analysis has been introduced into credit scoring. Survival analysis is the area of statistics that deals with the analysis of lifetime data. The variable of interest is the time to the occurrence of an event. It is commonly used in medical drug studies and reliability studies in engineering see e.g., Hosmer et al. [2008]. For example in medical studies the effect of a drug on the lifetime of a patient with a certain disease can be studied. The time to event in this case is defined as time till death.

In the case of credit risk the event of interest is default. The major advantage of survival analysis compared to other credit scoring models, is that the model is capable of including censored and truncated data in the development sample. In the current logistic regression approach these observations are removed from the dataset. Right censoring is the most common type of censoring and states that the event is not observed within the study period. In the case of credit risk: a customer who doesn't default. Because most of the customers do not default, a lot of the data is right censored.

	2013											2014
	Jan	Feb	Mar	May	June	July	Aug	Sept	Oct	Nov	Dec	Jan
Facility 1	G	G	G	G	G	G	G	G	G	G	G	G
Facility 2	G	G	G	G	G	G	G	G	G	B	B	B
Facility 3	B	B	B	D								
Facility 4	G	G	G	G								
Facility 5	G	G	G	G	G	G	G	G	G			

Figure 1: Example of portfolio with censored observations.

In Figure 1 an example of a portfolio is given. Facility 3, 4 and 5 leave the portfolio during the year (Facility 3 defaulted; Facility 4 and 5 are for example sold). Facility 3 is the event of interest in PD modelling and remains in the portfolio. On the other hand are Facility 4 and 5 removed from the dataset because they are not in the portfolio for 1 year. These are called censored observations.

Another type of missing data is truncated data, of which left-truncation is the most common type. In this case no information from the start of the loan until it is observed in the dataset is available (Yan, 2004). For example if a loan is initiated before the study period, there is no observation of the start of this loan.

The use of survival analysis instead of logistic regression demands some new procedures. For example if a facility is assessed on a yearly basis, the last observation has more information about the current creditworthiness of a facility and should be incorporated into the model. A procedure for these different observation periods has to be developed.

Currently RI is not using this kind of modelling and wants to develop a survival analysis model and compare it with the logistic regression models.

1.2 Research objective and approach

The objective of this thesis is to develop a framework for PD estimation and scorecard development using survival model approach. In order to guarantee easy adoption of the model, it should consist of the same steps of the current framework used within Rabobank International. Although survival models have been already widely used in credit scoring systems for a couple of years, Rabobank International hasn't yet developed such a model. However Rabobank International sees the potential of such a model and believes it can improve the PD estimations. This thesis focuses on developing a framework for PD estimations using survival analysis. Furthermore a survival model will be developed and its performance will be compared with the currently used logistic regression model. The research goal of this thesis is:

> Developing a survival model for PD estimation and scorecard development and setup a framework for Rabobank International. The new developed survival model will be benchmarked to the currently used logistic regression model.

From this research goal the research question and research sub questions are derived. The research question is:

> How can survival analysis be used for PD estimation and scorecard development within the framework of Rabobank International, and how does the model perform compared to the currently used logistic regression?

In order to reach this goal first the current logistic regression approach has to be understood. Next a literature study on survival models has to be completed. Taking these objectives into account the first sub question is:

What model is currently used and what are its shortcomings?

This sub question will be how the current models within RI are developed using logistic regression. The bottlenecks of this model with regard to for example censored data are detected and will form the basis of this research. In the next sub question the survival models will be investigated. The sub question is given by:

What survival models are currently available for credit risk estimation?

This sub question is a literature study on the currently available survival models used in credit risk estimation. First the general idea behind survival analysis and the implementation in credit risk will be explained followed by the different survival models. The incorporation of different observation periods data is investigated in the next sub question:

What is the effect of using different observation periods?

Most datasets have multiple observations of each loan. The time to event is for every observation different. This results in an ability to insert the data in different ways. For example, take only the first observation of each loan and it's the time to event. Another possibility is to take every observation and look one year ahead if the loan is in default. The effect of changing different observation periods is investigated in order to setup a framework for using survival models. Next, the model is benchmarked against the current model as given in the next sub question:

How to compare survival models with logistic regression models?

After both models, survival model and logistic regression model, have been implemented, the model performance should be compared. Literature is used in order to find good benchmarking techniques to compare the models. Next, the models are compared and the findings are documented. The last sub question is a conclusion on the new developed model. The sub question is given as:

What are the advantages and disadvantages of survival model compared to the current model?

In this sub question first the shortcomings of the current logistic regression are investigated as well as the advantages. Next, the results are compared to the survival model.

1.3 Outline

Chapter 2 of this thesis describes the current model approach. In this chapter the modelling steps are explained. It starts with the singlefactor analysis where the individual factors are selected based upon the predicting properties of the factor. The selected factors are then transferred to the multifactor analysis where the best combination of factors is selected. Finally in the calibration stage the transformation from score to rating is explained.

Chapter 3 starts with a general introduction on survival analysis. After the introduction different types of survival models are explained and compared. Finally a model is selected.

Chapter 4 is the development of the model. This chapter can be seen as a framework for survival models. It describes the different steps to take in order to come up with a model. It starts with the data preparation and how the time to event is defined. Next the singlefactor analysis, multifactor analysis and calibration of the model are explained.

Chapter 5 describes the results of model and compares them to the logistic model. The chapter starts with the explanation of the different datasets used for model development. Next the results of both models are compared after which a conclusion is drawn.

Chapter 6 sums up the conclusions about the PD framework and the performance of the survival model. In this chapter also some topics for further research are discussed.

Chapter 7 contains the bibliography and Chapter 8 contains the appendices.

2 Current model

2.1 Introduction

In this chapter we describe the currently used model building approach. The new framework for the use of survival analysis in PD estimation is based upon the currently used framework and therefore will contain the same steps as described in Section 2.2. The focus of this chapter is on the PD estimation and scorecard development for both retail and corporate exposures using the good-bad approach. Main difference between the two exposures is that corporate exposures are modelled on an individual client bases whereas with retail exposures are bucketed into risk categories.

In Section 2.2 the different steps in the current modelling process are described. In the Section 2.2.1 the first stage of the development process is explained in more detail. This is called the singlefactor analysis where the individual factors are tested on predicting properties. In Section 2.2.2 the second stage of the development process is explained: the multifactor analysis. Here, the best set of predicting variables is determined. Finally in Section 2.2.3 the score of the model is transformed to a rating. This stage is called the calibration of the model. In Section 2.3 the logistic regression and the shortcomings of the current modelling process are described.

2.2 Current modelling process

The general modelling framework for PD estimation consists of five steps and is presented in Figure 2. With the development of the survival analysis framework the focus is on the two phases: scorecard development and calibration. These are the stages that will differ from the current modelling process.

The initial set-up or preliminary study provides an overview of the development of the model. It names all people in the development team, objective of the model and regulatory requirements and finally the methodology.

In the scorecard preparation phase it is all about factors. Factors are explanatory variables known as covariates. This stage identifies, collects, links and cleans these factors in order to make them ready for regression.

Figure 2: General modelling framework for PD.[1]

The third stage, scorecard development, will be redeveloped and consists of singlefactor analysis (SFA) and multifactor analysis (MFA). The SFA examines the standalone discriminatory power and predictive power of the individual factors. In order to reduce the number of factors in the multifactor analysis only the best predictive factors are selected. Next, the factors are transformed into interpretable scores, i.e., between 0-10. The SFA is followed by the MFA, this stage combines the best standalone predictive factors into a model. Correlation between factors is taken into account in order to create a stable and robust final model. The output of the MFA is not a credit rating but a credit worthiness score. The higher the score, the lower the expected PD of the facilities.

The fourth stage is the calibration of the model. The scorecard generally only results in a model that ranks the clients in terms of creditworthiness, but does not assign rating grades. In this stage the scores are mapped to rating grades in order to make capital estimations.

[1] Retail PD modelling consists of an additional step after the calibration called the client acceptance threshold. This step determines when the facility is accepted into the portfolio and when it is rejected.

The fifth stage is the final stage and consists of the acceptance of the model by experts.

The model described in this chapter is built using the good-bad approach which uses actually observed defaults as input for model development. Since a default is the ultimate measure of creditworthiness it is the best target variable for a rating model. In the end, the model should predict/quantify the likelihood of a default occurring.

2.2.1 Singlefactor analysis

The first stage in the model development is the singlefactor analysis. In this stage the factors that have predictive power for defaults are selected and transformed. For example, a defaulted counterparty prior to the default is likely to have factor scores that are significantly lower or higher than other comparable counterparties that did not default. The main task is to find the factors for which either high or low values correspond to high PDs.

The goal of the SFA is twofold:

1. The selection of factors for further modelling, which is done by analysing the standalone discriminatory powers of the factors and expert opinion.
2. The transformation of the factor values into interpretable scores, i.e. between 0-10. This transformation is generally based on the relation between the factors and the goal variable.

Figure 3: Overview of steps in the singlefactor analysis.

The SFA consists of several steps as is shown in Figure 3. First step is the bucketing techniques which is not always required but has certain advantages. First of all, the relation between creditworthiness and factor values can be assessed more easily by experts and second some performance tests require buckets such as information value and weight of evidence. The last advantage is that it is helpful for the transformation that needs to be performed. The bucketing techniques are described in Appendix B.

Next step is the performance measurement of each individual factor. This is the measure of the predictive power of an individual factor. This measure is used in the selection and transformation of factors. Four measures of predictive power are given: Power Statistic, Weight of Evidence (WoE), Information value and trend analysis (Siddiqi, 2005). The most popular used within RI are the power statistic and trend figures because they require a minimum amount of assumptions, see Herel et al. [2010]. These measures are explained in Appendix B.

After the performance measurement the factors are transformed. This transformation of a factor brings the ratios into a standard interval from 0 to 10. The goal of a transformation is twofold:

1. In order to be able to compare (the coefficients of) different ratios in the multifactor analysis.
2. A transformation suppresses the impact of outliers in the development process and in daily use of the model.

The transformation of factors can be completed on the based on creditworthiness of the facility or distribution of the factor. Within RI the preferred transformation is based on the creditworthiness and this approach is used in retail modelling. For corporate modelling a transformation based on the distribution of the factor is used because a transformation on the basis of creditworthiness is not available.

The preferred approach for transformation of continuous factors based on the distribution of the factor is the logistic transformation approach (Herel et al., 2012). Transformation based on the credit

worthiness used within RI is called linear transformation. Both techniques are described in Appendix B.

2.2.2 Multifactor analysis

After the singlefactor analysis, the multifactor analysis is started. The multifactor analysis determines how the individual risk drivers, identified in the SFA, is incorporated into the final model. The goal of the MFA is to come up with a final model based on the best combined explanatory factors, taking into account redundancy/dependence between the factors. Herel et al., [2012] the multifactor analysis consists of the following steps:

Figure 4: Overview of steps in multifactor analysis.

The first step is to select a regression method for the multifactor analysis. The goal variable is either good (0) or bad (1). Because the goal variable is binary, logistic regression is used (Siddiqi, 2005). This regression method is discussed in more depth in Section 2.3.

The selection of factors consists of two steps: the stepwise multifactor analysis and sampling. Stepwise multifactor analysis finds the best set of predicting variables given a dataset. Next is sampling used to generate subsets of the total dataset and come up with a robust model. The procedures are given in Appendix B.

After the selection of factors is completed and the model performs well on the holdout samples, it should be tested whether it also performs well on other subsamples. This stage is called the scorecard performance. A more detailed description of this step is given in Appendix B.

Next step is feedback from experts. If experts strongly disagree with the statistical model, the weights can be changed accordingly to the experts and the power statistic is compared. If the difference between the two models is not significant, the model with the new expert weights is selected, otherwise the results should be discussed with the experts. The procedure is described in Appendix B.

2.2.3 Calibration

After the SFA and MFA have been finished a score for each loan in the portfolio is calculated. The higher the score, the lower the expected PD of the facilities. Next step is to bucket the scores based on homogeneous scores and assign PD values to each individual bucket. This bucketing uses the technique as explained in Appendix B.

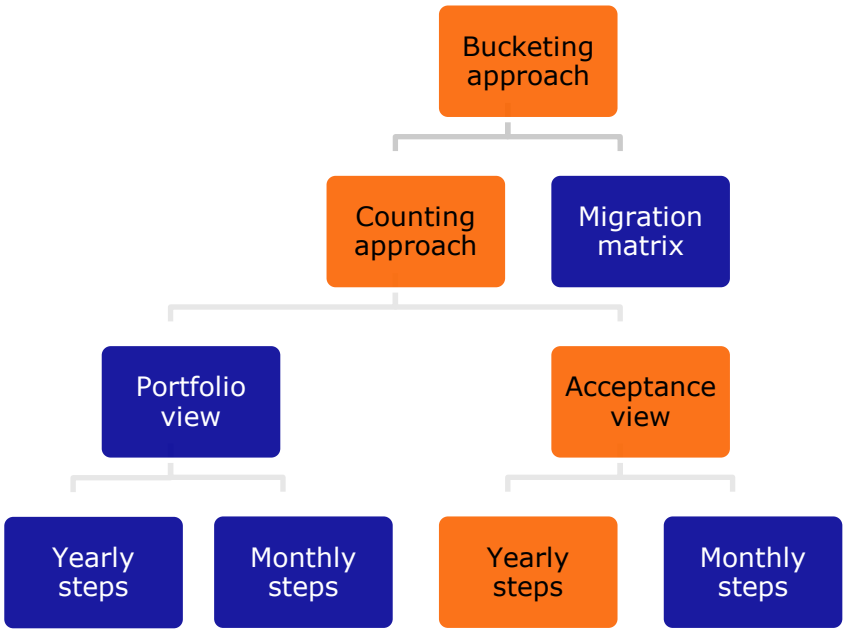

Figure 5: Overview of the different approaches for calibration (Orange tiles are preferred).

In Figure 5, different approaches for calibration are shown. The preferred approach is stated in orange. There are basically two techniques: migration matrix and counting approach. The most used approach is the counting approach because it is the most simple. The counting approach counts the number of facilities that went into default within the forecasting horizon and divides this number by the total number of facilities.

The acceptance view takes a snapshot when the facility enters the portfolio. From the moment the facility entered the portfolio a yearly snapshot is taken until the end of the facility. Next step is to divide the number of bads by the total number of facilities and calculate the PD. An example of the portfolio view and acceptance view is given in Appendix B.

2.3 Logistic regression

Current PD estimations are based upon logistic regression. This methodology consists of taking a sample of previous customers and classify them into good or bad (good-bad modelling). The classification is based on the repayment performance over a given period. The goal of this regression modelling is to estimate credit risk and to extract variables that are important in credit risk prediction.

The observations in the portfolio are marked as good or bad based on certain criteria. These criteria are mainly based on payment behaviour. For example, a typical criterion for a bad observation is a payment delay of more than 90 days. A good observation is then an observation for which payments have been received on time. Finally, observations for which it cannot be determined whether they are good or bad, are marked as undefined and left out of the modelling.

The logistic regression is different from regular regression because dependent variable is binary. A measure of the probability of the outcome is given by the odds of occurrence an event. If the probability of defaulting of a facility is given by P, then the probability of a facility not defaulting is given by 1-P. The odds of default are given by:

$$odds\ of\ default = \frac{P}{1-P} \tag{1}$$

The probability that a facility ends up in default is modelled by the logistic model and is given by:

$$P(Default|x_i) = \pi(x) = \frac{e^{\alpha+\sum \beta_i x_i}}{1+e^{\alpha+\sum \beta_i x_i}} = \frac{1}{1+e^{-(\alpha+\sum \beta_i x_i)}} \tag{2}$$

The shape of the logistic function from Equation 2 is given in Figure 6. The shape suggests that the probability for low predictor values x is low, and then there is some threshold value of the predictor at which the estimated probability of event begins to increase. For high predictor values the probability is high.

The popularity of the logistic model is mainly caused by this shape and the fact the function ranges between 0 and 1. The model is used to describe the probability of default of a facility, which is always between 0 and 1.

In order to estimate the PD for a fixed time period such as in the logistic regression, the data must be prepared for one year observations. This consists of taking an observation and observe if the facility ended up good or bad.

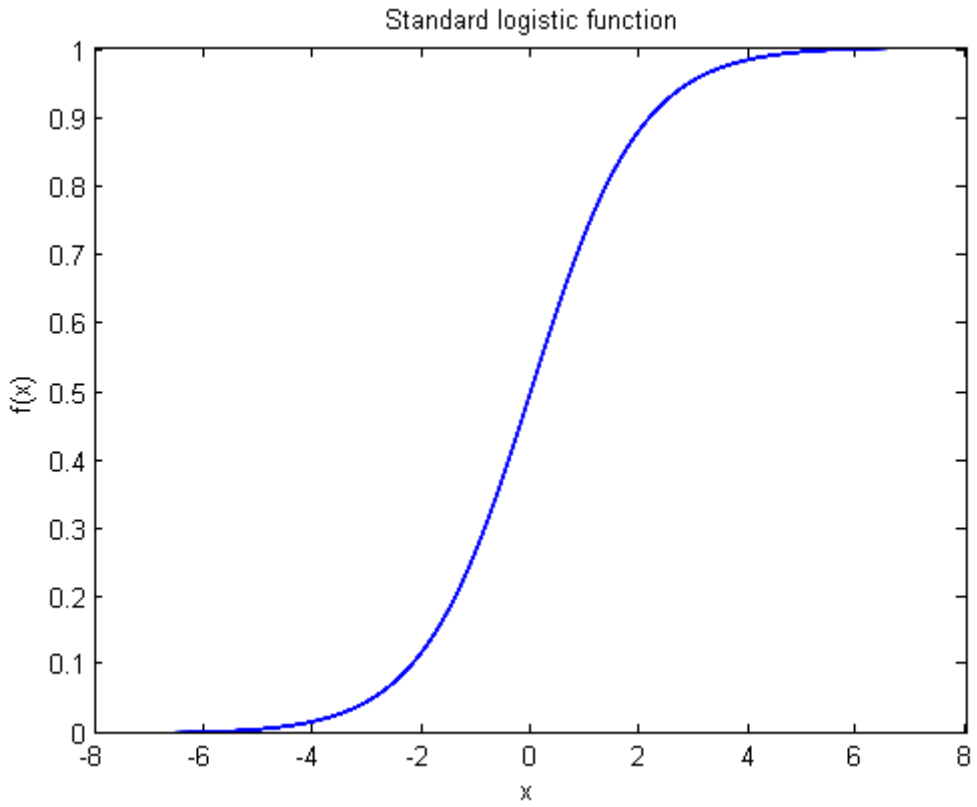

Figure 6: Standard logistic function.

Some facilities leave the portfolio before the one year period expires, for example because the loan is sold. In Figure 7, Facility 1 and 2 are still in the portfolio after one year and can therefore be measured. Facility 3 is in default after 4 months and is therefore marked as bad. Facility 4 and 5 are still marked as good when they leave (for example loan is sold) the portfolio after 4 and 10 months. If the time between the first observation and the facility leaving the portfolio is between 1-6 months the facility is classified as undefined. If the facility leaves between 7-12 months, the facility is classified as good. Because of this Facility 4 is left out of the modelling and facility 5 is marked as good.

	2013											2014
	Jan	Feb	Mar	May	June	July	Aug	Sept	Oct	Nov	Dec	Jan
Facility 1	G	G	G	G	G	G	G	G	G	G	G	G
Facility 2	G	G	G	G	G	G	G	G	G	B	B	B
Facility 3	B	B	B	D								
Facility 4	G	G	G	G								
Facility 5	G	G	G	G	G	G	G	G	G			

Figure 7: Example portfolio with different times.

This poses a problem with the logistic regression since the goal is to generate an unbiased model, as few facilities as possible should be removed from the dataset. Observations like facility 4 are removed from the dataset while they do contain information. They are not in the dataset for one year but with the values of Facility 4, the facility survived at least 4 months. Facility 5 is only in the portfolio for 10 months but in the regression this observation is assumed to be for the full year present.

Furthermore, logistic regression estimates the PD over a fixed time horizon (usually 1 year). For which the data must be prepared, a time consuming activity.

A possible solution to these shortcomings might be the use of survival analysis. Survival analysis can incorporate incomplete observations as censored data. These are observations that for which the event is not observed during the study period, as will be further explained in Chapter 3. Survival analysis offers more advantages since it focuses on estimating the survival distribution, it can estimate the default risk over any future time horizon.

3 Survival analysis

3.1 Introduction

Survival analysis is a statistical method for data analysis where the outcome variable of interest is the time to the occurrence of an event, often referred to as a failure time, survival time, or event time, see e.g., Kleinbaum [1998]. Survival analysis is used in different fields for example: medicine and engineering. In drug studies the time till death is modelled and in engineering time till failure in mechanical systems is studied, see e.g., Lawless [1982].

Survival analysis in credit scoring was introduced by Narain [1992], see e.g., Sthepanova & Thomas [2002]. It was further developed by Thomas et al. [1999]. The event of interest is default. Narain [1992] applied the accelerated life exponential model to personal loan data. He found that this model estimated the number of failures well at each time interval. Next, he showed that credit granting decisions could be improved by using the methods of survival analysis as compared to multiple regression. Finally, the author argued that survival analysis can be used in all credit operations in which there are predictor variables and the time to event is of interest.
Thomas et al. [1999] made a comparison of the performance of exponential, Weibull, and Cox models with logistic regression and found that survival-analysis methods are competitive with, and sometimes superior to, the traditional logistic-regression approach.
This indicates that survival analysis may be useful for accurate PD estimation for a fixed 12 months horizon for various types of loans, which is useful for PD estimation within the Basel II Accord (Tong et al., 2012).

In Section 3.2 the basics of survival analysis are given such as notation and common problems. Next the most common survival models are explained in Section 3.3. And finally comparison of the models is made and a model is selected for the implementation of a survival model within Rabobank International.

3.2 Survival analysis

Suppose that T is the length of time before a facility defaults. The randomness of T can be described in three standard ways, see e.g. Kalbfleisch & Prentice [1980] and Leemis [1995]:

Distribution function $F(t)$ describes the probability the time to event (T) is smaller or equal compared to a fixed time (t) and is given as:

$$F(t) = P(T \leq t) \tag{3}$$

From this, the survival function $S(t)$, the probability the time to event (T) is larger compared to a fixed time (t), can be derived as:

$$S(t) = 1 - F(t) \tag{4}$$

The second way is the density function $f(t)$. This is the probability that the failure time occurs at exactly time t (out of all possible times) and is given as:

$$f(t) = \lim_{\Delta t \downarrow 0} \frac{Prob(t \leq T \leq t + \Delta t)}{\Delta t} \tag{5}$$

And the last description is given by the hazard function $h(t)$. This is the probability that if a facility survives up till time t, it will experience the event in the next instant and is given by:

$$h(t) = \lambda(t) = \lim_{\Delta t \downarrow 0} \frac{\Pr\,(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} \qquad (6)$$

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} \qquad (7)$$

The interpretation of the survival function $S(t)$ is the most obvious. It is a plot with on the left axis the proportion of the population still alive and on the x axis the time. The hazard function $h(t)$, also called incidence rate, instantaneous risk or force of mortality, is the event rate at t among those at risk at time t. The interpretation is straight forward, for example:

$$h(t) = \ 1\% \text{ at } t = 6 \ \ months$$

This states that after six months, facilities are defaulting at a rate of 1% per month. Or in other words: at six months, if the facility is still in the portfolio, the chance of defaulting in the following month is 1%. This is often plotted as the cumulative hazard rate on the y-axis and the time on the x-axis.

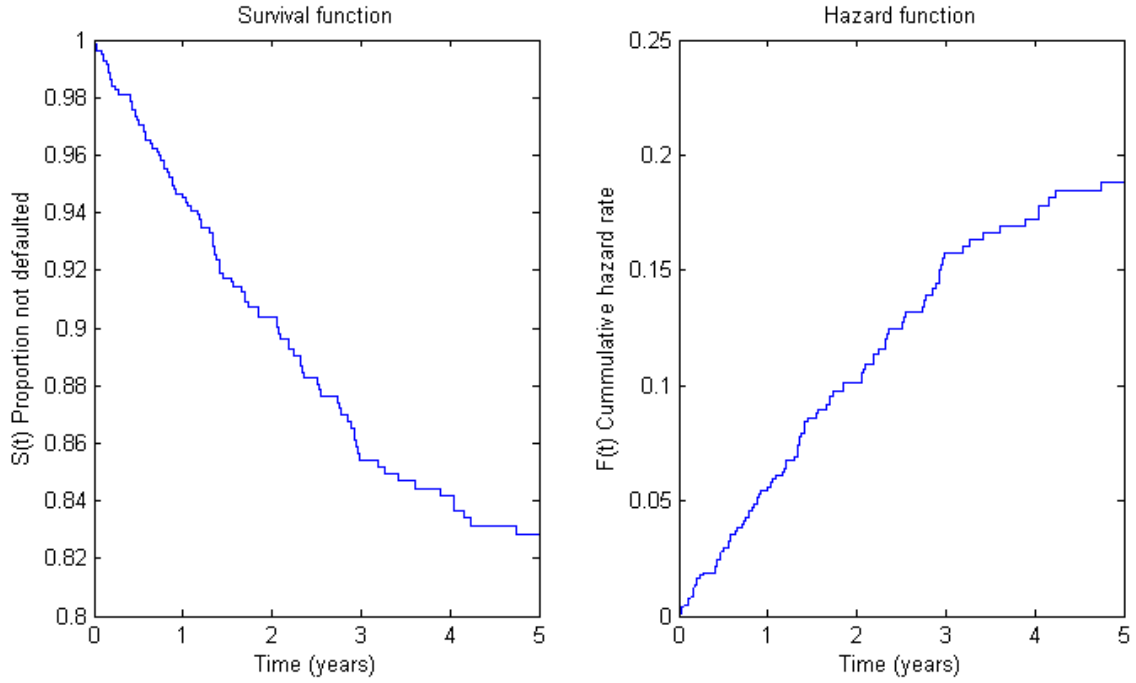

Figure 8: Survival and hazard function.

3.2.1 Censored data

A common problem in survival analysis is censored data or also called the missing data problem. The ideal dataset contains the begin and end dates of all the facilities in the portfolio of which the lifetime is determined. If the end date is not available in the dataset it is denoted as right censored data. In survival analysis right censored data are a common problem in estimating survivor and hazard function, see e.g. Thomas et al. [1999]. This occurs when data are collected over a finite period of time and consequently the event may not be observed for all facilities. Within a portfolio are typically two reasons for right censoring:

- The event of a facility is not observed within the study period. Facility is still in the portfolio at the end of the study.
- Facility left the portfolio during the study period; for example facility is sold.

In Figure 9 the dataset of a portfolio is shown where the study period (indicated by vertical, striped black line) is from t = 0 until t = 3, furthermore three facilities are shown:

1. Facility 1 is a censored observation: the facility left the portfolio during the study period.
2. Facility 2 is a censored observation: the default date is not observed during the study period.
3. Facility 3 is an uncensored observation: both the starting data and the default date are within the study period.

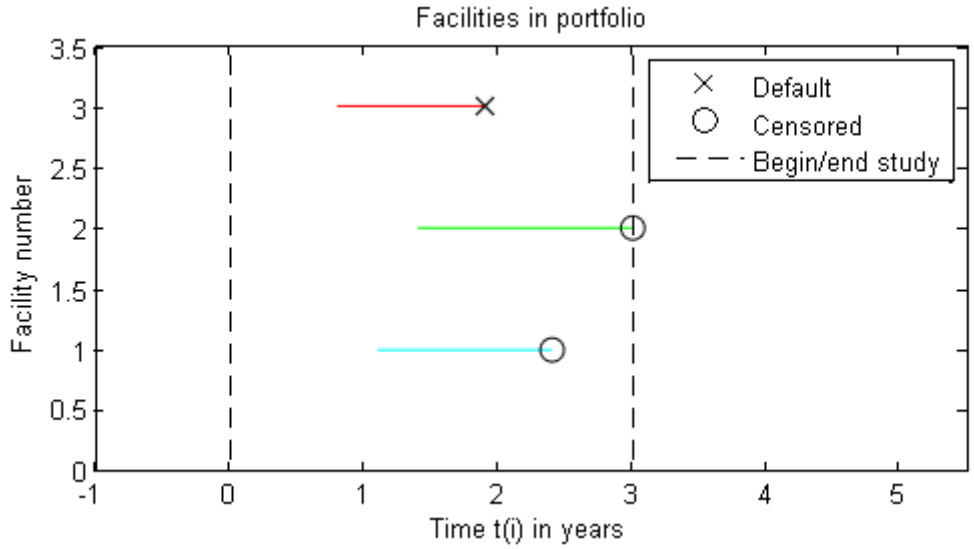

Figure 9: Example portfolio with censored data.

For 1 year PD estimations using logistic regression, facilities should survive at least 1 year. Survival analysis estimates the survival function over the entire data period. If the observations that leave the portfolio are left out, the number of facilities in the dataset is significantly reduced. Therefore it is of importance that the model includes all observations, also the censored observations which are not in the portfolio the entire data period.

An important assumption using censored data is that the censoring is non-informative censoring. This assumption states that the censored facilities are at the same risk of subsequent failure as those who are alive and uncensored. Or in other words facilities that drop out of the portfolio should do so due to reasons unrelated to the study. In this thesis, censoring is assumed to be non-informative. The facilities in the portfolio at any point in time should be representative of all facilities at the same time.

3.2.2 Truncated data

Truncation is another part of missing data of which left truncation is the most common. It occurs when the loans have been at risk before entering the study. Truncation is a condition other than the event of interest that is, for example, used to screen respondents or patients, see e.g. Klein & Moeschberger [1997]. This is very common in datasets, for example facilities enter the portfolio at a certain point in time because loans are bought. In that case loans are at risk before entering the portfolio and data is available.

An example from the medical world is the death times of elderly residents of a retirement community. The time from entry in the retirement community until the moment of death is studied. Only the elderly people of a certain age can be admitted into the community. People died before this age cannot be observed.

The difference between censoring and truncated data is often confused. Strictly speaking censoring is the case when facilities are known to default within a certain time, but the exact time to default is not known. Truncation is the case when facilities aren't in the dataset because they are not observed.

The incorporation of truncated data is beyond the scope of this thesis.

3.3 Types of survival models

For modelling survival data, different models can be applied. In the following sections the most common models of non-parametric, semi-parametric and fully parametric models are explained. In Section 3.3.1 the non-parametric Kaplan-Meier estimate is explained and in Section 3.3.2 the parametric models are explained followed by the Accelerate Failure Time models in Section 3.3.3. In Section 3.3.4 the proportional hazards models are explained and the Cox proportional hazards model is described in Section 3.3.5.

3.3.1 Kaplan-Meier estimator

In case of censored data, raw empirical estimators will not produce good results. In order to determine distribution function of these data, two basic techniques can be applied: the Kaplan-Meier (KM) or product limit estimator (Kaplan & Meier, 1958) and the Nelson-Aalen (NA) estimator (Aalen, 1978). The KM estimator estimates the median survival distribution function whereas the NA estimator estimates the cumulative hazard rate function. The advantage of these estimators is that these methods take censored data into account. It is the limit of the life-table estimator when intervals are taken so small that at most one distinct observation occurs within an interval. In this thesis only the KM estimator is used.

Suppose r individuals experience events in a group of individuals. Let the observed event times be given by $0 \leq t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(r)} \leq \infty$. Let n_i be the number of individuals at risk (uncensored and alive) just before t_i. And let d_i be the number of observed deaths at t_i. The KM estimator of the survival function $S(t)$ is defined by

$$KM(t) = \hat{S}(t) = \prod_{x<t} \left(1 - \frac{d(x)}{n(x)}\right) \tag{8}$$

Next, estimate $\Lambda(t)$ by $\widehat{\Lambda_{KM}}(t) = -log[KM(t)]$. Their variance is estimated by:

$$\widehat{Var}\{\widehat{\Lambda_{km}}(t)\} = \sum_{x<t} \left[\frac{dN(x)}{\left[Y(x) - \frac{w(x)}{2}\right]\left[Y(x) - dN(x) - \frac{w(x)}{2}\right]} \right] \tag{9}$$

$$se\{\widehat{\Lambda_{km}(t)}\} = \sqrt{\widehat{Var}\{\widehat{\Lambda_{km}}(t)\}} \tag{10}$$

The confidence interval of the KM estimator is given by:

$$CI[S(t)] = KM(t) * e^{\pm z\alpha/_2 * se[\widehat{\Lambda_{\mathrm{km}}(t)}]} \tag{11}$$

Where

- $d(x)$ is the number of events at time x, generally either zero or one, but in case of tied survival times $d(x) \geq 1$,
- $n(x)$ is the number of items at risk at time x,

- $\widehat{\Lambda_{KM}}(t)$ is the unbiased estimator of the cumulative hazard rate at t,
- $dN(x)$ number of observed events occurring in [x; x+Δx],
- $Y(x)$ is the number of items at risk at time x,
- $w(x)$ is the number of censored at time x,
- $z_{\alpha/2}$ is the $(1 - \alpha/2)$th quintile of a standard normal distribution.

Note that the KM estimator is a step function that does not change between events, nor at time censorings occur, it only changes at the time of each event. In Figure 10 a KM survival estimator is illustrated. The 'x' on the time axes determines an event and the 'o' determines a censored data.

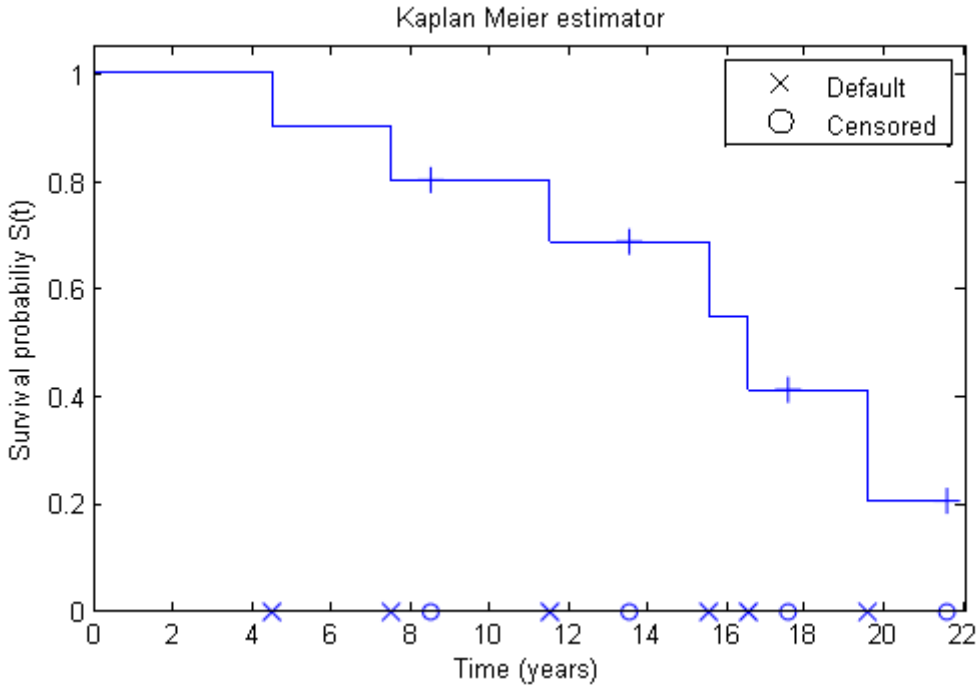

Figure 10: Kaplan Meier estimator.

The KM estimator can be used to describe the event of default. Then n_j is equal to the number of facilities in the portfolio and d_j is the number of observed defaults at time j. When dealing with problems that use this estimator, the goal is to understand how n_j changes with respect to censoring or truncation. If for example a facility leaves the portfolio, it will have an effect on the number at risk, but not on the observed death. This is the same for new facilities that enter at time t, they will be part of the risk set at time $t + 1$.

The KM estimator is capable of using stratification of variables. This concept consists of splitting the population into two or more groups based on some criteria. For example purpose of the loan: mortgage or otherwise. Plotting the KM estimator of both groups can give insight in whether one group has a higher survival probability than the other.

The Nelson-Aalen estimator is used to estimate the cumulative hazard rate function $\widehat{H}(t)$ and is given by:

$$\widehat{H}(t) = \sum_{i:t_i<t} \frac{d_i}{n_i} \tag{12}$$

Using Greenwoods formula (Greenwood, 1926), the confidence intervals of the survival time can be calculated. The variance is estimated by:

$$\widehat{Var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{i:t_i<t} \frac{d_i}{n_i(n_i - d_i)} \quad (13)$$

The KM estimator has certain advantages starting with the simplicity: it can be computed by hand. Furthermore the possibility of incorporating stratification factors in order to asses different survival probabilities of different groups. The limitations of the KM estimator are that it is a mainly descriptive estimator and doesn't control for covariates. Furthermore no sensible interpretation of competing risks can be made.

3.3.2 Parametric models

Although KM estimator is a very useful tool for estimating the survival function, sometimes we want to model the data in more detail. One solution is fitting a parametric model to the data. Popular distributions for estimating the survival curves are for example exponential, Weibull and Log-logistic distribution. In order to estimate S (t), maximum likelihood estimation is used.
In Figure 11 three models are fitted to the survival data: KM estimator, exponential and Weibull.

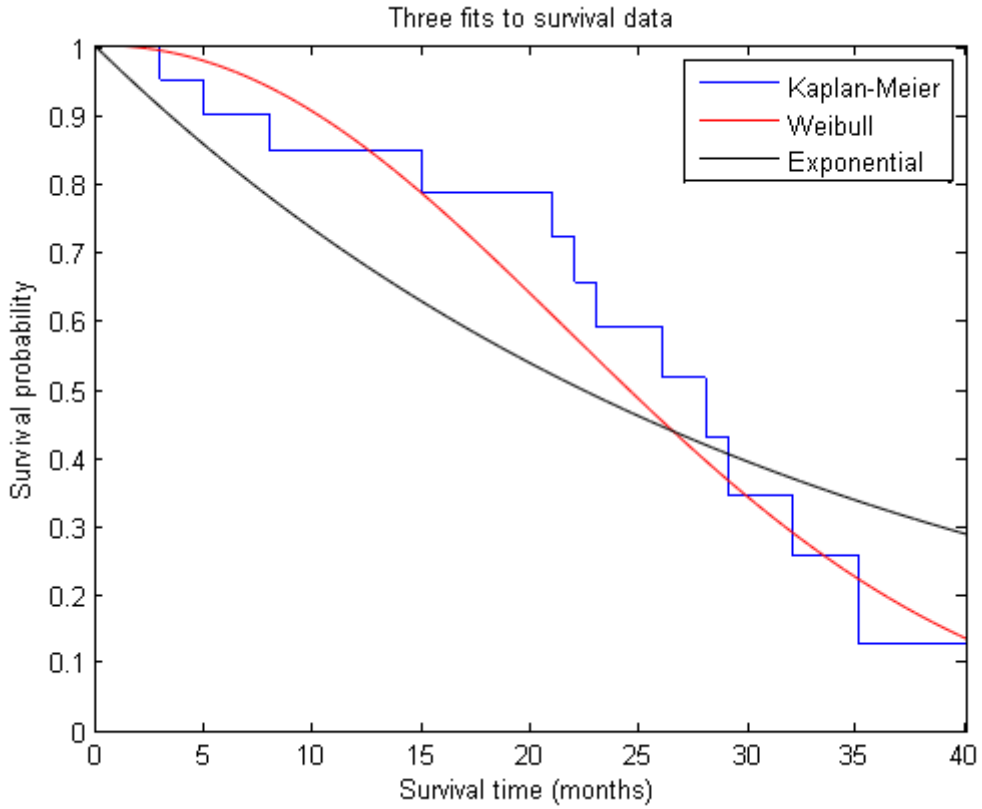

Figure 11: Three fits on the survival data: Kaplan Meier, Weibull and Exponential.

The exponential distribution with parameter λ is given by:

$$F(t) = 1 - e^{-\lambda t} \quad (14)$$

$$f(t) = \lambda e^{-\lambda t} \quad (15)$$

$$h(t) = \lambda \quad (16)$$

And the Weibull distribution with parameters scale λ and shape k is given by:

$$F(t) = 1 - e^{-(\lambda t)^k} \tag{17}$$

$$f(t) = k\lambda^k t^{k-1} e^{-(\lambda t)} \tag{18}$$

$$h(t) = \ k\lambda^k t^{k-1} \tag{19}$$

An overview of the most common parametric models is given in Table 1.

Table 1: Overview of the most common parametric models.

	Exponential $\lambda > 0$	**Weibull** $\lambda > 0, k > 0$	**Log-logistic**
$\boldsymbol{f(t)}$	$\lambda e^{-\lambda t}$	$k\lambda^k t^{k-1} e^{-(\lambda t)}$	$\frac{\lambda\kappa t^{\kappa-1}}{(1+\lambda t^{\kappa})^2}$
$\boldsymbol{F(t)}$	$1 - e^{-\lambda t}$	$1 - e^{-(\lambda t)^k}$	$1 - \frac{1}{1+\lambda t^{\kappa}}$
$\boldsymbol{S(t)}$	$e^{-\lambda t}$	$e^{-(\lambda t)^k}$	$\frac{1}{1+\lambda t^{\kappa}}$
$\boldsymbol{h(t)}$	λ	$k\lambda^k t^{k-1}$	$\frac{\lambda\kappa t^{\kappa-1}}{1+\lambda t^{\kappa}}$

Fitting a parametric distribution to the data has certain advantages. First of all the survival $S(t)$ and density functions $h(t)$ are fully specified. Using these estimates it is easier to compute quintiles of the different distributions and tests for differences between parameters are more powerful.

3.3.3 Accelerated failure time

An accelerated failure time (AFT) model assumes that the effect of a covariate is to accelerate or decelerate the time to event of a facility by some constant, see e.g. Kalbfleisch & Prentice [1980]. The AFT model states that the relationship between two survival functions $S_1(t)$ and $S_2(t)$ is given by:

$$S_1(t) = S_2(ct) \quad for\ all\ t \geq 0 \quad where\ constant\ c\ > 0$$

This model implies that the aging rate of population 1 is c times as much as that of population 2. When explanatory variables for the time to event are available, one could also use an AFT model in order to model the time to event. The AFT model states that the predicted event time can be multiplied by some constant, in order for a covariate to take effect. With this model the direct effect of the explanatory variables on the survival time is measured. This results in an easy interpretation of estimated parameters, because the parameters measure the effect of the corresponding covariate on the mean survival time.

The simplest accelerated failure time models assume time-independent covariates. This class of AFT models assumes that the survival distribution is given by:

$$S(t) = S_0(\psi(z)t) = S_0(\exp(\beta x)t) \tag{20}$$

$$h(t) = \psi(z)h_0(t\psi(z)) \tag{21}$$

Where $\psi(z)$ is proportionality constant and a function of the covariates z by which the lifetime is decreased. Often $\psi(z)$ is assumed to be log linear $\psi(z) = \exp(\beta x)$ and S_0 and h_0 are the baseline survivor function and hazard rate function.

The interpretation of $\psi(z)$ is quite simple. If for example a facility has a proportionality constant of $\psi(z) = 2$, the aging of the loan is estimated to be twice as fast as the baseline. As a result of this the estimated time to event for the facility is estimated at half of the baseline time. However the hazard function $h(t)$ is not twice as high over the total life span. This is a property of Proportional Hazard models as explained in the next section.

3.3.4 Fully parametric proportional hazards model

Next to the AFT model, proportional hazards models (PH model) are used in survival analysis. Cox [1972] described the proportional hazards model in JRSSB in 1972 and is now one of the most quoted statistical papers in history. In contrast with the AFT model, the PH model estimates the hazard rate. Proportional hazards models are built from two parts: the baseline hazard function and the effect parameters $\exp(\beta' x)$. The baseline hazard function describes how the hazard changes over time at the baseline. The baseline is where all the covariates are zero. The effect parameters describe how the hazard changes in response to explanatory covariates. The proportional hazards model is given by:

$$h(t|x) = h_0(t)\exp\left(\beta_1 x_1 + \beta_2 x_3 + \cdots + \beta_p x_p\right) = h_0(t)\exp(\beta' x) \qquad (22)$$

$$\log h(t|x) = \log h_0(t) + \beta_1 x_1 + \cdots + \beta_k x_k \qquad (23)$$

Where $h_0(t)$ is the underlying hazard function which can take on any form and $exp(\beta' x)$ is referred to as a relative hazard function or relative risk.

Different PH models are available such as the Exponential model and Weibull. The simplest distribution is the exponential model and is appropriate if the hazard function is constant, $h(t) = \lambda$. This means that the age of the facility has no influence on the survival probability.

Another frequently used model is the Weibull model. In contrast with the exponential model the Weibull model has no constant hazard function. Both models are given in Section 3.3.2.

The parametric models can be compared using the likelihood ratio test or the Akaike information criterion (AIC) (Akaike, 1974). If the models are nested, the likelihood ratio test can be used, otherwise the AIC should be used to select the best fitted model.

3.3.5 Cox proportional hazards model

In general if proportional hazards models are used, the Cox proportional hazards or Cox model is used. The Cox PH model is the same as the fully parametric model given in Equation 23. In "partial likelihood", Cox [1975] found that the parameters β can be estimated without knowing the baseline hazard function h_0 and is therefore called a semi-parametric model. This model uses the rank of the event and censored times instead of the actual times.

The difference between a Cox PH model and a normal PH model is that the Cox model does not make any assumptions about the baseline hazard function $h_0(t)$. This is the non-parametric part of the model. The model does assume parametric form for the effect of the predictor variables on the hazard. This is the parametric part of the model. Because of this, the model is referred to as a semi-parametric model and the model estimates relative risk instead of absolute risk. Because no assumptions are made about the baseline this model is very flexible.

In order for the Cox model to be valid the assumption of proportional hazards has to hold. This assumption states that the risk of default of different groups is constant over time. In essence this means that the effect of a covariate is the same at any moment. For example if at the start facility 1 has a risk of default twice as high as facility 2, then the risk of default for facility 1 should be twice as high everywhere in time.

In case the proportional hazards assumption does not hold the model is capable of incorporating time depending covariates. An explanatory variable is time-dependent if the values change over time. Another solution to this problem might be the use of the stratified Cox model. This consists of defining the factor into buckets and use dummy variables for each strata.

In order to estimate the coefficients without making assumptions about the baseline, Cox proposed partial likelihood. The estimation of the Cox model is explained in Appendix C.

3.4 Comparison

The most common survival models have been discussed in previous sections and all have the advantage of being able to handle censored data. Although between the different models there are some similarities, the models are very different.

The advantages of the KM estimator are that it is easy to compute and to interpret. Furthermore it doesn't require any assumptions about a baseline. One of the main drawbacks of this estimator is that it doesn't account for variables that are related to the survival time. It is a descriptive estimator and only describes the estimation of the survival function of the population. Therefore it is only an applicable to homogeneous samples. This model can be used in order to get a quick impression of the survival function of a population.

The AFT model assumes that a covariate is able to accelerate or decelerate the time to a certain event by some constant. These models have two main advantages: they are very easy interpreted (Kay & Kinnersley, 2002) and are more robust to omitted covariates and the less affected by the choice of probability distribution compared to proportional hazards model (Keiding & Andersen, 1997).

The basic idea behind proportional hazard models is that the effect of the covariate is to multiply the baseline hazard by some constant. In order to use these models the proportional hazards assumption should hold. This assumption states that the risk of default of different groups is constant over time. For example if at the start facility 1 has a risk of default twice as high as Facility 2, then the risk of default for Facility 1 should be twice as high everywhere in time. There are two types of PH models: parametric PH models and Cox PH models. The difference is that the parametric PH models assume the baseline hazard function follows a specific distribution whereas the Cox model does not make assumptions about the baseline. The Cox model makes estimations on the basis of the rank of the survival times.

For the popularity of the Cox PH model are several reasons. First the model does not require any assumptions about the baseline, the model is robust, flexible and a safe choice in many cases. Furthermore, the model is capable of handling discrete and continuous measures of event times and is it possible to incorporate time-dependent covariates, in order to correct for changes in value of covariates over the course of the observation periods.

The Cox PH model is chosen for the development of the model.

4 Implementation

4.1 Introduction

In this chapter the PD scorecard development framework using survival analysis is described. The framework of the PD modelling uses the same steps as in the current approach used within RI. This will ensure easy adoption and implementation within RI. The general modelling approach used within the RI consists of four stages as given in Figure 12.

Figure 12: Overview of steps in the survival model development.

In order for the model to fit in the procedure depicted in Figure 12, some new procedures are to be developed. For the transformation of variables, as explained in Section 4.3, a new bucketing technique as well as a transformation procedure is to be developed. Furthermore a new procedure for the selection of variables is implemented even as methods for comparison between the performances of different models.

We start by describing which steps are to be taken to prepare the data for the development of a PD scorecard. Next section describes the SFA where the individual performance of different factors on the survival time is assessed and the factors are transformed. In Section 4.4, the MFA, how the best set of predicting variables is selected, is described. In the final stage of the model the calibration, the mapping of the individual scores to PD, is described.

4.2 Data

The model development starts with a few checks on the data. The number of missing values and other incorrect values in the factors are first checked. If only a small part of the values is missing or incorrect, the values are replaced with the median value of the factor.

Figure 13: Overview of steps in the survival model development (Data highlighted).

Next step is to incorporate the weights of the facilities. Weights are used to compensate for inconsistencies in the dataset compared to the portfolio. This discrepancy is a result of cleaning of the dataset and changes in the portfolio as a result of strategy changes. For example, in the portfolio 15% of the facilities is a grocery store, but in the dataset only 5% is a grocery store. The dataset is made representative for the portfolio by giving this 5% in the dataset a weight of 3. On the other hand, some facilities in the dataset do not have quantitative data and should be left out. These observations have weight zero and are not incorporated into the regression.

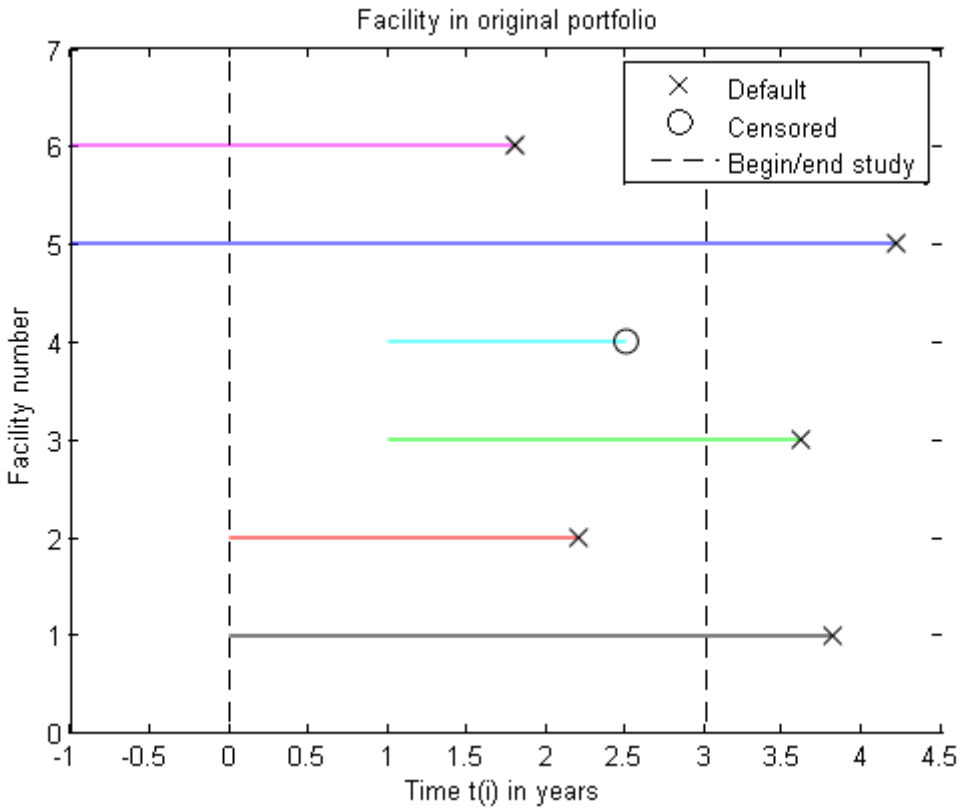

Figure 14: Facilities in original portfolio.

The facilities in the portfolio have different begin and end dates as is illustrated in Figure 14. In the figure six different facilities are illustrated. The study period is illustrated by the striped, vertical black line. In the portfolio Facilities 1 and 2 determine the start the study period. Facility 2 went into default just after 2 years, but facility 1 is still in the portfolio at the end of the study and is therefore censored after 3 years. Facilities 3 and 4 enter the portfolio after 1 year. Facility 4 left the portfolio around 2.5 years and has therefore survived for about 1.5 years and ended as censored observation. Facility 3 is still in the portfolio at the end of the study and is therefore censored after 2 years. Facility 5 is in the portfolio before the start of the dataset, but because no data is available the time before the study period is ignored. The facility is still in the portfolio at the end of the dataset and therefore censored after 3 years. Facility 6 is straightforward and defaults after around 2 years.

Truncation is not available in the functions developed in this thesis. Therefore it assumed that the first observation of each facility was at the start of the study. In essence the first observation of each facility is shifted to the start of the study period. The result of this shifting on the example portfolio in Figure 14, is show in Figure 15.

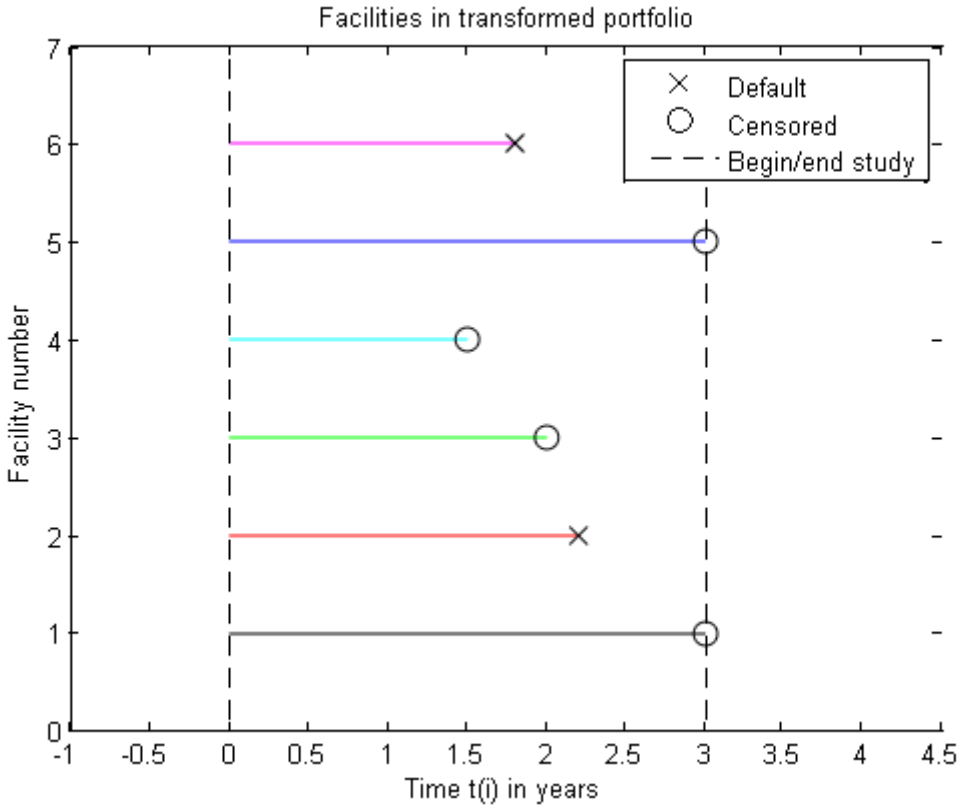

Figure 15: Facilities in shifted portfolio.

In datasets it is common that there is more than one observation per facility. For example corporate clients are assessed on a yearly interval which is illustrated in Figure 16.

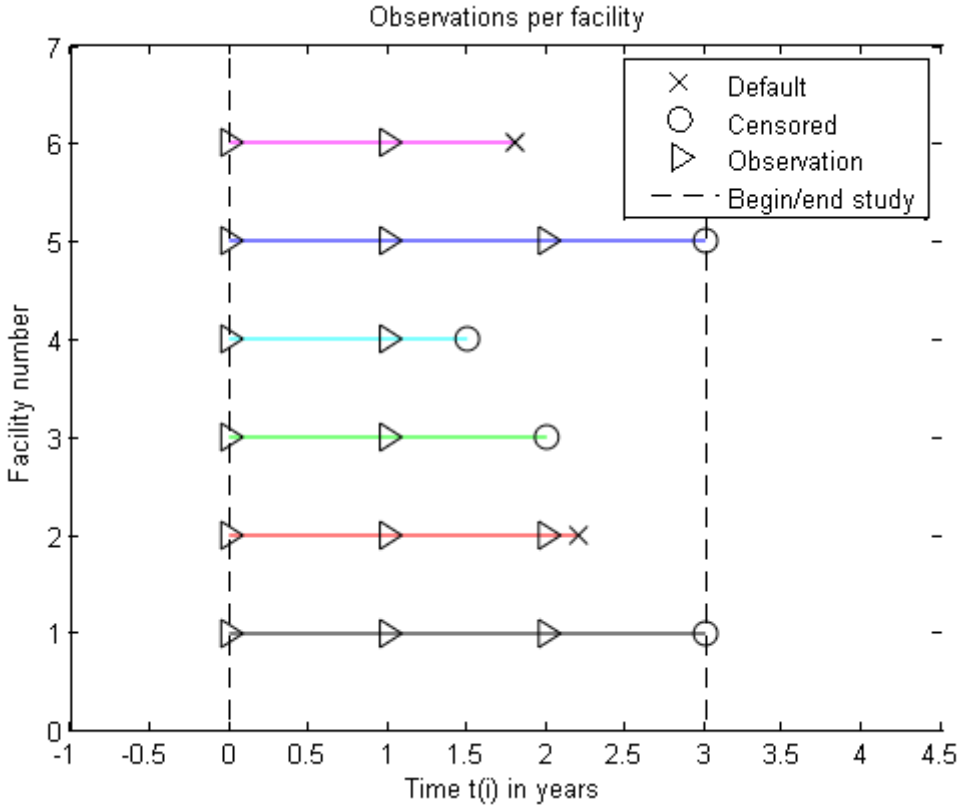

Figure 16: Observations per facility.

There are three different methodologies of incorporating these different observations per facility. This is explained using facility 2 from Figure 16 as example. The first option is to only use the first observation available in the portfolio and calculate the time to event from this observation. This option removes observations and is therefore not the preferred approach. This is illustrated in Figure 17.

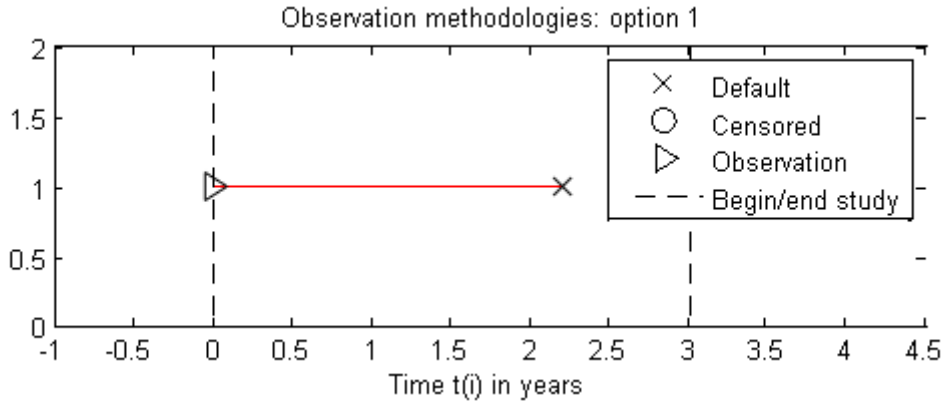

Figure 17: Observation methodologies (Option 1).

The second option is to use all the observations and calculate the time to the next occurrence in the portfolio. In case of an observation, the time to this observation is calculated and observation is marked as censored. The result is a portfolio as illustrated in Figure 18.

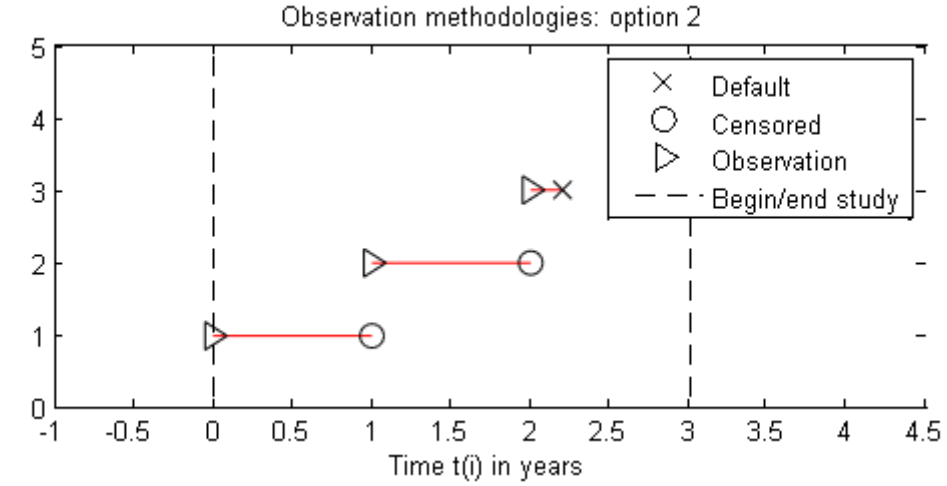

Figure 18: Observation methodologies (Option 2).

This second type results in problems by using the datasets within RI without truncation. The retail datasets contain an update every 3 months on every facility. Furthermore, the datasets are prepared for good-bad analysis, and if a facility goes into default, the last observation between 3 and 15 months before default, is marked as bad. Without using truncation, this dataset will contain no defaults up till 3 months and the 3 to 15 months will contain all defaults. The incorporation of this type of data is beyond the scope of this thesis and therefore not implemented.

The third option is to use all the observations and calculate the time to event for each observation as the time to the last observation. The different observations are weighted by assigning each observation an equal weight so that the sum of the weights is equal to 1. The portfolio is illustrated in Figure 19.

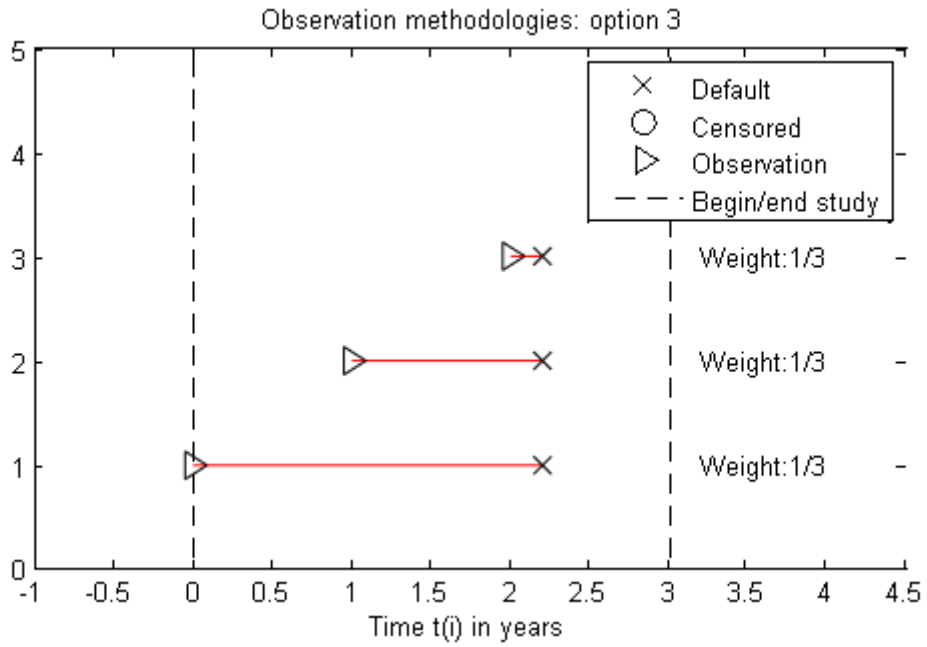

Figure 19: Observation methodologies (Option 3).

The basic idea behind this methodology is that with the first factor values the facility survived around two years. After one year there is a new assessment, the factor values are worsened and the facility survives only for about 1 year. The last assessment is just before the default date and the factor values should be worse.

4.3 Singlefactor analysis

The goal of the SFA is to find financial ratios and qualitative information that have predictive power for defaults. For example, the financial ratios prior to a default of a counter party might be different compared to other counter parties that did not default. In essence the correlation between the factors and default is examined. These factors can be classified into different groups: descriptive, predictive and random. In credit risk modelling the goal is to find the predictive variables since these are used to predict defaults on a given time horizon.

Figure 20: Overview of steps in the survival model development (Singlefactor analysis highlighted).

The singlefactor analysis consists of two steps:

1. The selection of factors for further modelling, which is done by analysing the standalone discriminatory powers of the factors and expert opinion.
2. The transformation of the factor values into interpretable scores between 0 and 10.

The singlefactor analysis for survival analysis demands new procedures for bucketing, performance measurement and transformation of variables. These procedures are developed during this thesis and explained in this chapter. First, the bucketing techniques of continuous variables will be explained, next the performance measurement and last section is the transformation of the variables.

4.3.1 Bucketing techniques

In order to make credit-scoring systems robust, it is an industry standard to split continuous variables into bins. This process is also called bucketing. The goal of this process is threefold:

1. The relation between the creditworthiness indicator and factor value can be assessed more easily by experts.
2. Bucketing can be useful in transformation.
3. Some performance tests require buckets.

In traditional approaches the split of the factors was based upon good-bad ratio (default rate) or similar measures. The definition of bad in these measures is defined as a default before the time horizon, while the other observations are considered good. Using survival analysis, not the default rate within a fixed time horizon is of interest, but the time to the actual event. Therefore, the survival function is used for different buckets.

In literature techniques for bucketing of continuous variables are mostly based on manual operations. (Stepanova & Thomas, 2002) suggested the following procedure:
1. Split the characteristic into 10 to 20 equal bands.
2. Create a binary variable for each band.
3. Fit Cox's proportional hazard model to these binary variables.
4. Chart parameter estimates for all bands.
5. Choose the splits based on similarity of parameter estimates.

The procedure for discrete variables is similar: a binary variable for every discrete answer is created. Next the procedure above is used.

.

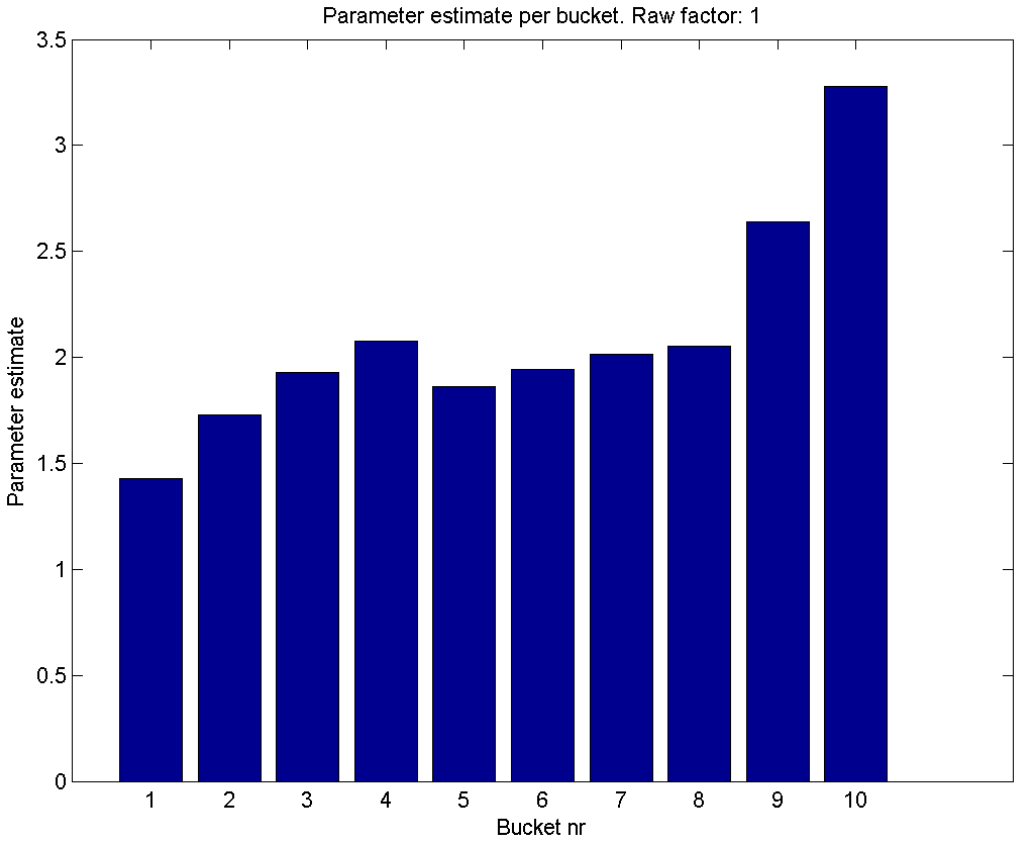

Figure 21: Parameter estimates per bucket.

In Figure 21 the results of this approach to the first factor of the simulation data is shown. The buckets are ranked on the factors scores from low (Bucket 1) to high (Bucket 10). As can be observed the trend is that higher factor scores get higher parameter estimates, which result in a worse survival probability.

The goal is to generate an automated process for bucketing variables. Because this approach consists of manual steps, we developed a new bucketing procedure in this thesis. This procedure is developed

based on the difference in survival function. The procedure starts with splitting the factor into equal buckets and plot the KM estimators of each bucket. Next compare the survival functions of neighbouring buckets using the Logrank test, see Appendix E. This test compares the survival functions and tests whether they are significantly different from each other or not. If they are not, the two buckets are merged.

The bucketing process consists of the following steps:

1. Rank factor scores.
2. Split data in equal buckets.
3. (optional) For small buckets it might be necessary to check on trend and merge bucket that do not comply with trend. The following steps are taken to correct for this:
    a. Determine the trend of the factor by ranking the scores and splitting the sample into 5 buckets. Next plot the KM estimator of the first and the last bucket and fit an exponential model to it.
    b. Next plot the KM estimator and fit an exponential model to the individual buckets generated in Step 2.
    c. Merge buckets that do not comply with the trend, determined in step a.
4. Compute the KM estimator and confidence interval for every bucket.
5. Compare survival functions of the buckets using the Logrank test (see Appendix E)
6. Merge the buckets with the highest p-value and above significance level (typically 0.05).
7. Repeat Steps 4 and 6 until all buckets are significantly different.

The optional Step 3, ordering the buckets to create a trend in the survival functions, is only required if the buckets are very small. In the case of very small buckets, one bucket may be a very biased bucket. As a result of this, the neighbouring buckets are never merged on the basis of the Logrank test.

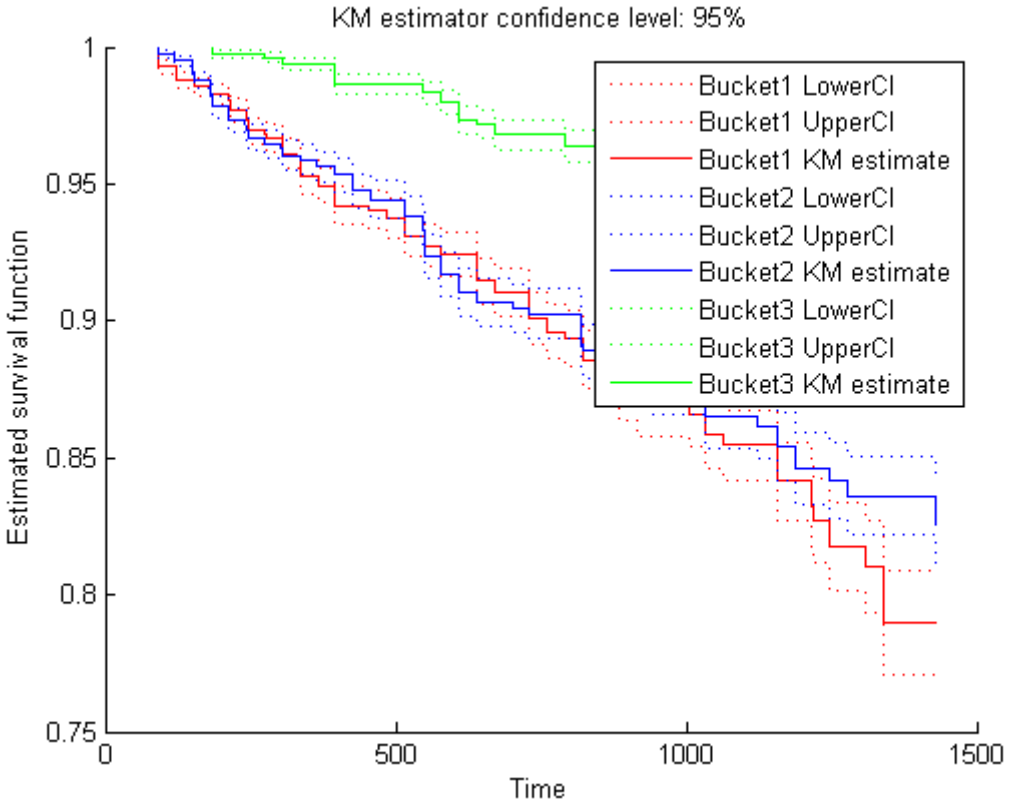

Figure 22: KM estimator for 3 buckets.

For example, the factor is split into three buckets and the KM estimator is plotted, as show in Figure 22. As can be observed bucket 1 and 2 have quite similar survival distributions and will be tested using the logrank test. If the survival functions are not significantly different, the buckets are merged

with a result as given in Figure 23. As observed from the figure, the survival functions of both buckets are significantly different.

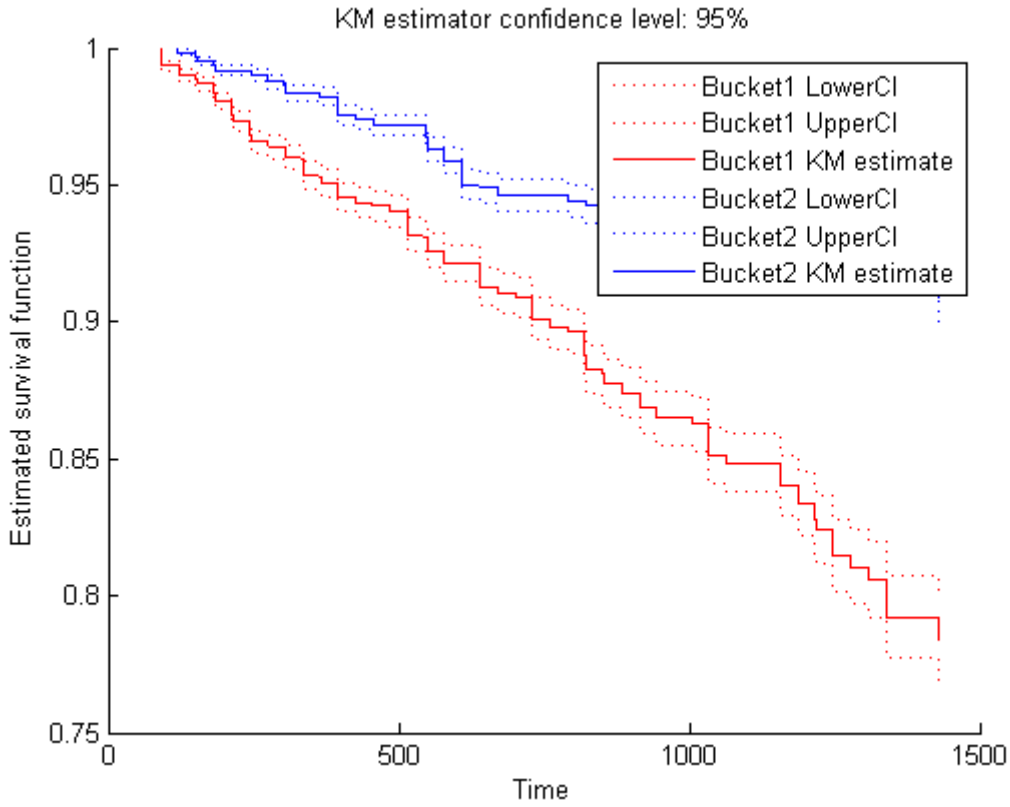

Figure 23: KM estimator for 2 buckets.

4.3.2 Performance measurement

In this section the predictive power of a single factor is determined. This is used in the selection of factors for the multifactor analysis and transformation of the factors.

The basic idea behind this measurement is that the more a factor is discriminating, the more difference exists between the survival functions. Continuous factors are ranked based on the scores and split into two groups. Next step is to plot a KM-estimator for every group and compare them. The factor with the most difference between the two KM-estimators discriminates the best. In order to quantify the predictive power of a single factor, an exponential model is fitted on the KM-estimator. Furthermore is the intuitiveness of the factor assessed. The intuitiveness of a factor determines if high factors scores should result in good survival probabilities or if the low scores should result in good survival probabilities.

The performance test based on the exponential model in steps:

1. Rank factor scores.
2. Split data in two equal buckets.
3. Compute the KM estimator.
4. Fit an exponential model to the KM estimator.
5. Subtract the ln(λ-estimate) of the bucket with the lowest 50% of the scores, from the ln(λ-estimate) of the bucket with the highest 50% of the scores.
6. Performance: The more the absolute difference between both ln(λ-estimates) is, the more discriminatory the factor is.

7. Intuitiveness: If this difference between both ln(λ-estimates) is negative, the higher scores have a better survival probability and the other way around.

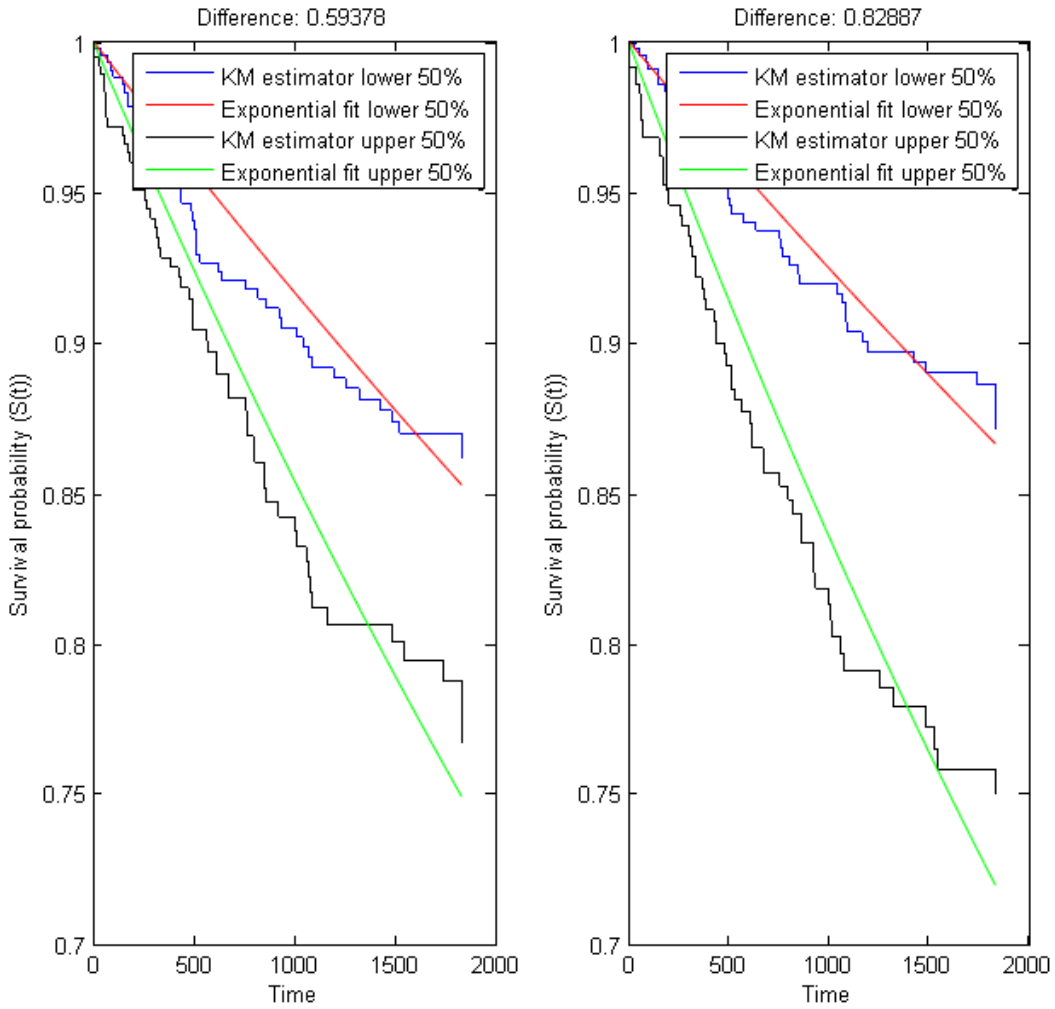

Figure 24: Performance measurement example.

In Figure 24 an example of the performance test on two factors is shown. As can be observed, the difference between the KM estimators is larger in the second factor (right image) compared to the first factor (left image). As a result the second factor is more discriminatory than the first one. The difference on top of the graph is the difference between the ln(λ-estimate). In both factors, higher factor scores result in a lower survival probability. This can be seen from the case that the bucket with the highest 50% of the scores (black line) is below the bucket with the lowest 50% of the scores (blue line). If the difference on top of the factor is positive, lower scores have better survival probabilities and vice versa. The absolute value is a measure of the predictive power of a single factor: the higher the value, the more predictive the factor is.

4.3.3 Transformation of data

A transformation of a factor brings the ratio into a standard interval, such as [0, 10], so all transformed ratios will correspond to scores between 0 and 10.

The goal of a transformation is twofold:

- In order to be able to compare (the coefficients of) different ratios in the multifactor analysis,
- A transformation suppresses the impact of outliers in the development process and in daily use of the model.

Transformation can be based on two types, one based on the creditworthiness of the facilities and another based on the distribution of the factor (e.g., the first 10% in the Bucket 1, the second 10% Bucket 2 and so on). The transformation based on the creditworthiness is preferred and mainly used for retail modelling. Transformation based on the distribution of the de factor scores is used in the corporate modelling.

For the survival analysis the standard logistic transformation (see Appendix B) is used and a new transformation approach is developed, based upon the creditworthiness of the factor as explained in Section 4.3.3.1.

4.3.3.1 Logrank transformation

Another new developed procedure in this thesis in order to transform scores is named the logrank transformation. This transformation is based on the creditworthiness of the facilities and uses the bucketing approach as explained in Section 4.3.1. The transformation consists of the following steps:

1. Bucket the factor scores using the bucketing approach explained in Section 4.3.1.
2. Fit an exponential model to the KM estimators.
3. Create a list of the λ-estimates and use this to transform the variables.
4. The lowest ln(λ-estimate) has the best survival function and should be given value 10 and vice versa.

For example: there are three buckets with different survival functions. The exponential model fitted to the KM estimators resulted in three λ-estimates as given in Table 2. The transformation is based upon the LN of the λ-estimates, and is a logrank transformation.

Table 2: Logrank transformation using exponential model.

Bucket	λ-estimate	LN(λ-estimate)	Transformation value
1	1	0	10
2	2	0.693	3.6907
3	3	1.099	0

In Figure 25 the exponential estimates from Table 2 are plotted. Using only the λ-estimates, the model assigns the values 10, 5 and 0 to the buckets. This is not representative because Buckets 2 and 3 are more related than for example Bucket 1 and 2. Therefore, as can be observed from the graph, the ln(λ-estimates) are used for the actual transformation.

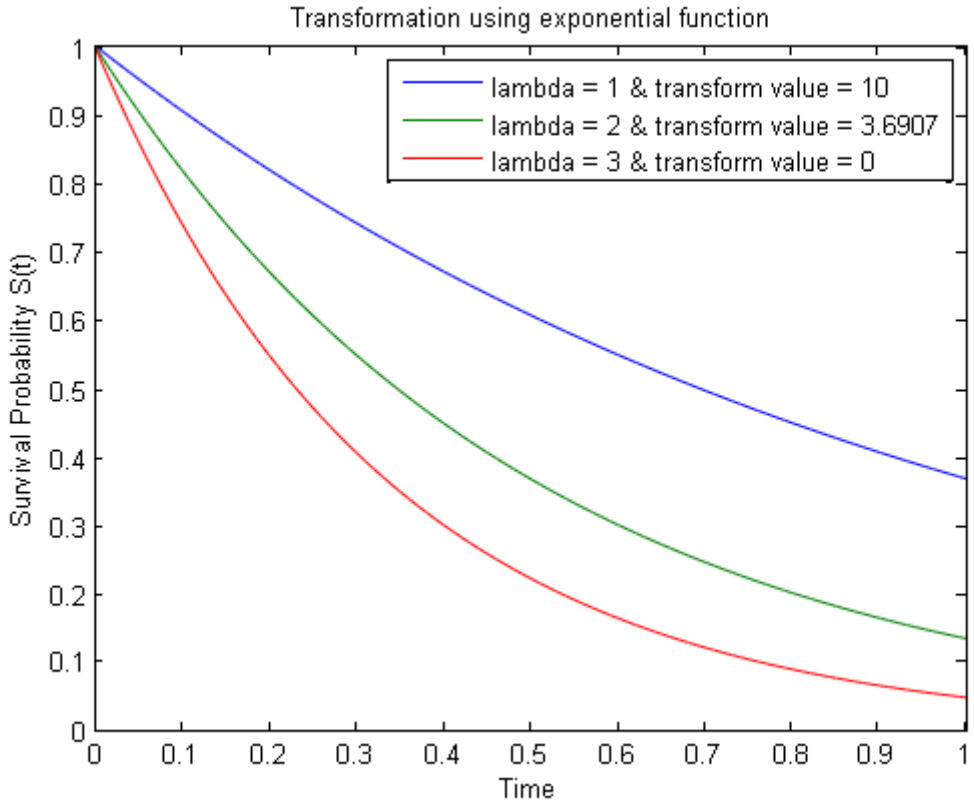

Figure 25: Logrank transformation using exponential model.

An output of this logrank transformation is given in Figure 26 and Figure 27. In Figure 26 the survival functions and the ln(λ-estimates) are plotted. In Figure 27 the number of observations per bucket is shown. The size of the buckets should not vary too much.

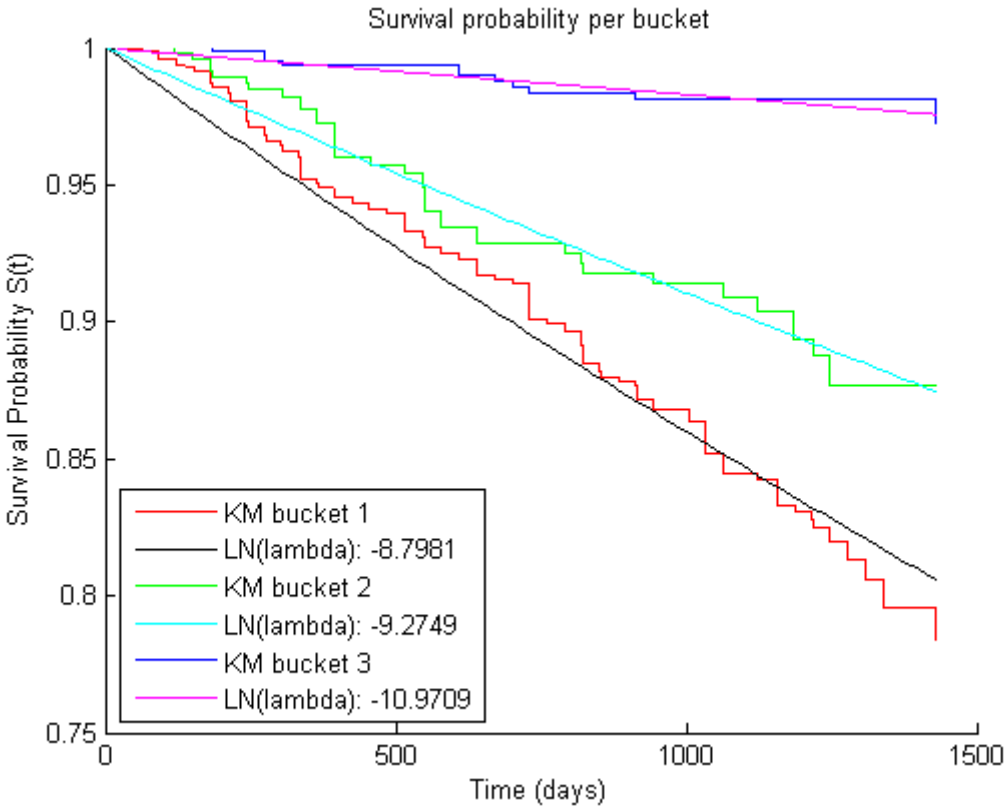

Figure 26: Logrank transformation, survival function plot per bucket.

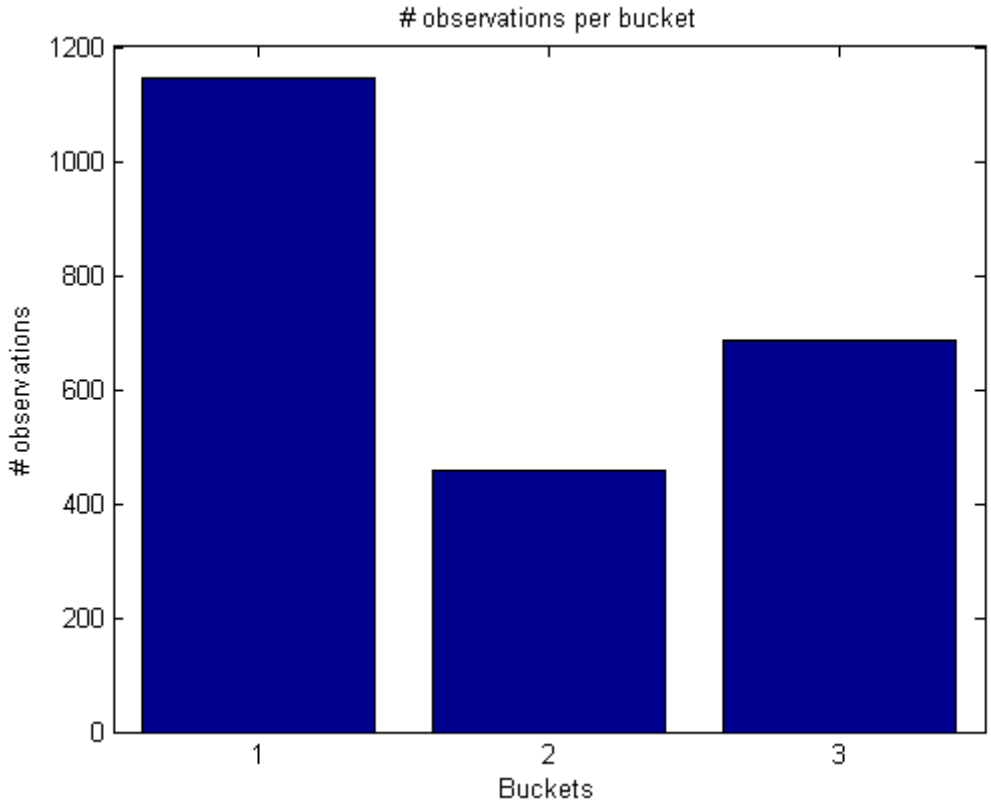

Figure 27: Number of observations per bucket.

4.3.3.2 *Monotonous*

It occurs that the factor scores are non-monotonous and form a risk for the transformation. An example of non-monotonous factor scores if both: very good and very bad scores (U-shape) are considered as potential risk factors. In such case, first it should be determined by experts what shape the relation between score and default rate is and where approximately the optimum is located. Once this shape is determined one solution is to split the factor into buckets and use dummy variables which each have their own coefficient in order to model the data. This is a standard procedure in credit scoring e.g., Tomas et al. [1999].

4.4 *Multifactor analysis*

The multifactor analysis starts with explaining the different types of stepwise regression used to find the best set of factors from a dataset. Next the selection of the final set of factors is explained and finally the scorecard performance is assessed.

Figure 28: Overview of steps in the survival model development (Multifactor analysis highlighted).

4.4.1 *Stepwise regression*

Stepwise regression is used to find the best set of predicting variables. Different regression processes can be developed for survival analysis. The most common are backwards elimination, forward selection and selection-elimination regression. For the selection of factors the Wald test statistic is

used (see Appendix D) and for the selection of the best model, the Akaike Information Criterion (AIC) is used (see Appendix D). These different selection methods are described in the following sections and compared in Chapter 5.

Correlation
After the selection the factors should be checked upon correlation. Too highly correlated factors can lead to multicollinearity. In this case, two or more predicting variables can be linearly predicted from each other. As a result of this, the coefficients estimates can be unstable. This is that value of the coefficients can change substantial by removing only a few observations. In general, the final model should contain only risk factors, whose correlation coefficients are not too high, as indication RMVM uses 75%. In case of higher correlations, this should be discussed with experts and preferably one of the highly correlated factors should be chosen.

4.4.1.1 Backward elimination

Backward elimination starts with all factors and deleting them one at the time. First the least predicting factor is removed on the basis of the p-value. The p-value is the calculated as the Wald test statistic and is used to test the true value of the parameter based on the sample estimate. The factor with the highest p-value that is above the significance level (typically 0.05) is removed and the model is fitted again until only significant factors remain. This process is called backward elimination.

The process in phases:

1. First a Cox PH model is fitted with all factors on the goal variable and create a list with all the p-values.
2. Next compare the p-values from the Wald test statistic. The highest p-value that is above the significance level is removed from the list with selected variables. The significance level is usually set at 0.05.
3. If one of the factors becomes counter intuitive the factor is removed. Intuitive means that if a positive influence on the score is expected also a positive coefficient should come out of the regression.
4. Repeat Step 1-3 with the remainder of the variables until all variables have a p-value lower than the significance level.

The main disadvantage of this model is the efficiency of the model. This factor selection procedure demands a lot of time if many factors are present.

4.4.1.2 Forward Selection

The forward selection starts with the selecting the most significant factor for predicting the outcome based on the Wald statistic. In each following step one factor is added that has the lowest p-value and a p-value lower compared to the significance level. The procedure is given by:

1. First a standalone regression of all the individual factors on the goal variable is preformed and create a list with all the p-values.
2. Next compare the p-values and the lowest p-value that is below the significance level is added to the list with selected variables. The significance level is usually set at 0.05.
3. If one of the selected factors has a p-value below the threshold or becomes counter intuitive it is deleted.
4. Fit a Cox model with the selected variables and each individual remaining factors.
5. Repeat Step 2 - 4 until no further factors meet the requirements from Step 2.

This process has one a main advantage: the performance. This process is efficient compared to, for example backwards elimination which results in a faster factor selection process. A disadvantage of this process is that it often selects too many factors which are often correlated.

4.4.1.3 Selection-Elimination Regression

The selection elimination regression, which is currently used for the logistic model, is also developed for the survival model. This procedure is changed to handle survival data using the Wald test statistic and AIC. The selection elimination regression multifactor analysis selects step-by-step the most predictive factors:

1. First a standalone regression of all the individual factors on the goal variable is done.
2. A list is created for all the factors that have an intuitive coefficient and a p-value below the significance level set by the modeller. Usually a significance level of 0.05 is chosen.
3. Select the factor with the lowest AIC value from the list created in Step 2.
4. Do a combined regression of the previous selected factors, combined with the other individual factors.
5. Repeat Step 2 until no further factors meet the requirements from Step 2.
6. If one of the previously selected factors has a p-value below the threshold or becomes counter intuitive it is deleted.
7. The AIC value of the previous regression round is compared with the new AIC values. If the AIC value of the previous round is better than the new round, the factors of previous round are chosen and the regression stops.
8. In case the new AIC is better, the factor with the lowest combined AIC value is selected from the list created in Step 2.
9. Repeat Step 4 to 8 until no further factors meet the requirements from Step 2.

4.4.2 Selection of factors

The dataset is split into two groups: 80% for the development of the model and 20% holdout sample. In the case of large datasets a 50%/50% distribution of the data can be considered. The model is fitted on the development sample and tested on the holdout sample. It might accidentally happen that a very biased holdout sample was chosen, because the holdout sample is generally very small. In order to avoid these problems and gain more insight in the data set, the creation of a development and holdout sample is repeated multiple times (e.g., 100 times).

This way the sensitivity of the model is tested regarding:

- Selection of factors; factors that are in for example 80% of the models selected, are the most important factors. Factors that are selected in 30%-80% might also contain significant information and can be discussed with experts.
- The coefficients; coefficients will vary for each development sample, since the estimation depends on the data. Asses the stability of coefficients. If they vary too much it can be considered to remove the factors from the sample.
- Performance; The models that are estimated on the development samples can immediately be tested on the corresponding holdout samples. This way there is not one performance number (determined on only one holdout sample), but there are many more. This provides better insight in the average performance as well as the stability of the performance

1. The selection of factors consists of 2 steps. First the steps for selection of factors:
   a. Produce a development sample and a holdout sample, generally 80%/20%, of randomly selected records of the complete dataset.
   b. Apply stepwise regression on the development sample.
   c. Apply Step a - b 100 times and identify which factors have been incorporated in 80% of the 100 model estimates.

d. The final factor selection is based on a combination of this analysis and expert feedback.

2. After the selection of the factors, the performance and stability of the model is examined. The steps are given as:
   a. Produce a development sample and a holdout sample, generally 80/20%, of randomly selected records of the complete dataset.
   b. Fit Cox model on the development sample and store the coefficients.
   c. Next calculate the log likelihood using the 20% holdout sample and the coefficients of the development sample.
   d. Next fit a Cox model on the holdout sample and calculate the log likelihood.
   e. Calculate and store the difference between the log likelihood from Step c and d. This step assesses the performance of the model estimated in Step c. Basic view: Step d is the best fit, given the holdout data, step c is the best fit using the development data. If both data sets were completely identical the log likelihood values would be exactly the same and so the difference would be 0. If the holdout sample set is by accident very biased, it will show in the log likelihood comparison. The lower this difference, the better the model fits.
   f. Apply Step a-e 100 times and store the coefficients for each factor and the performance (log likelihood difference)
   g. Assess the coefficients. The following are per factor assessed:
      a. standard deviations
      b. minimum
      c. maximum
      d. mean
      e. median
      f. weighted average based on difference in performance (log likelihood difference) between model based on development sample and fitted on holdout sample.

3. The final coefficients are assessed by fitting a Cox model to the complete data set.

4.4.3 Scorecard performance

After the factors have been selected and their coefficients estimated, the performance of the scorecard is evaluated.

The performance of the scorecard will be looked upon for the following two points:

- Performance on subgroups of the portfolio.
- Expert based adjustments.

In order to visualize the performance of the model, the observed default rate and the estimated default probability per time interval (e.g., quarterly) is plotted. In order to determine the defaults estimated by the model, average values for the factors are taken (close to the 50^{th} percentile).

In Figure 29 the empirical pdf (observed default rate) and the Cox model are plotted for the complete dataset. The time interval is set to 90 days and the first 90 days does not contain any defaults because of the data cleaning as explained in Section 4.2. The model performs well if the median and mean estimated default rates are similar to the realized default rate. As can be observed, the Cox model predicts the number of defaults quite well in every time interval.

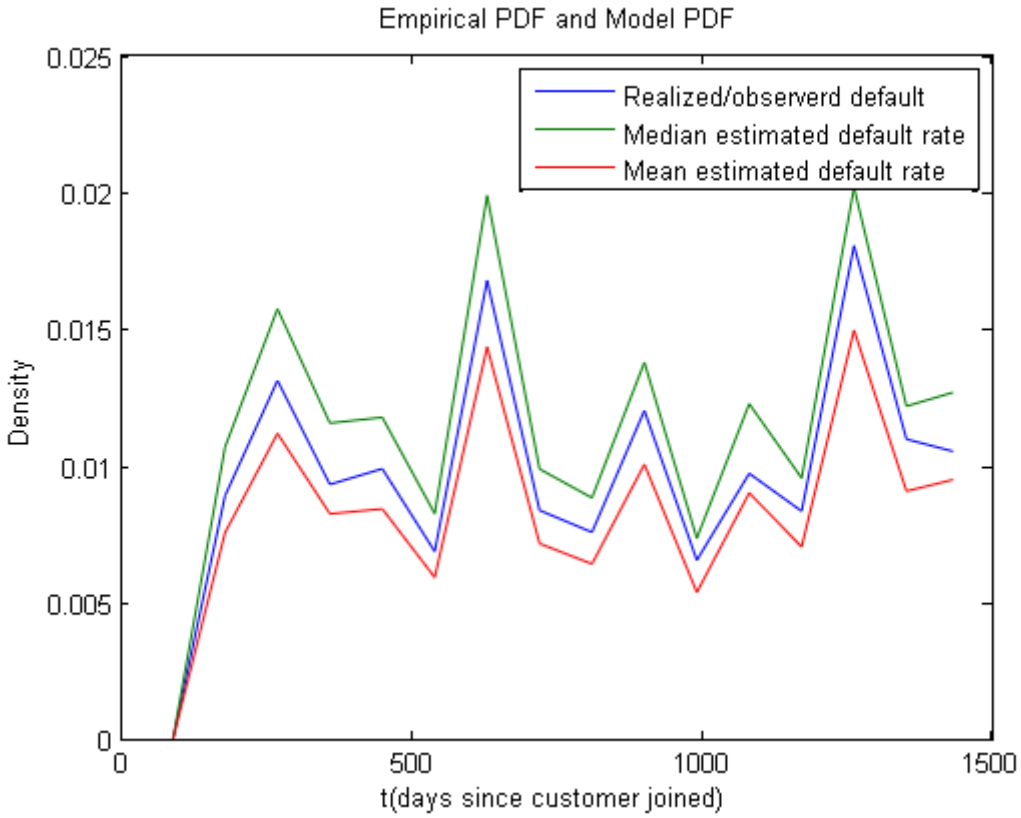

Figure 29: Empirical scorecard performance (Full dataset).

This performance test can also be used to test subsamples as shown in Figure 30. The estimated default rate is quite similar to the observed default rate. This indicates that the model fits the dataset well.

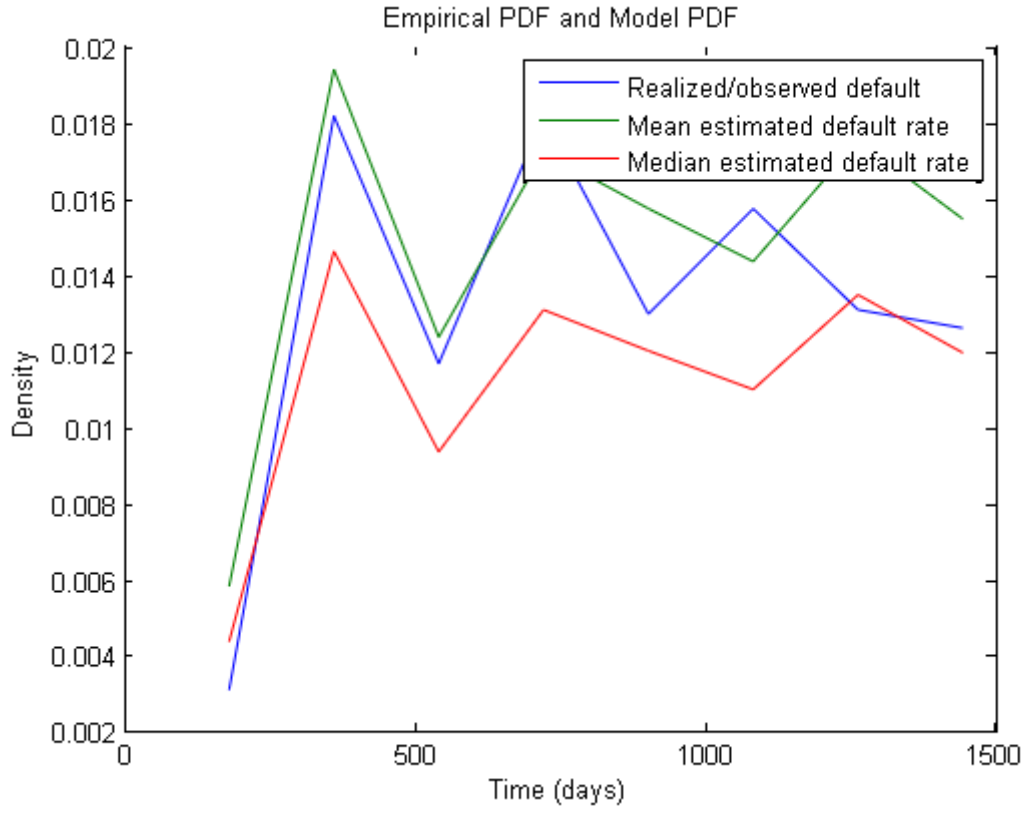

Figure 30: Empirical scorecard performance (Subsample).

Another technique to assess the performance of the model on subsamples is to stratify a variable and plot the KM estimator for the subsample together with the Cox survival function of the sample e.g., Tong et al. [2012].

Tong et al. [2012] stratified on the homeownership of clients, and by looking at both KM-plots the conclusion was that home owners had higher survival rates across all time points and therefore home owners have less risk. Because all the discrete factors are transformed to continuous variables, the stratification in this case is based upon a threshold of a factor score. In Figure 31 the result of this stratification is shown. The results are similar to the results in the paper of Tong, Mues & Tomas. The red and black line represent the KM estimator, and the blue and green line the corresponding Cox estimates. As can be observed, the Cox model distinguishes between good and bad factor values because by comparing the Cox fits, the green line is closer to the black line compared to the blue line.

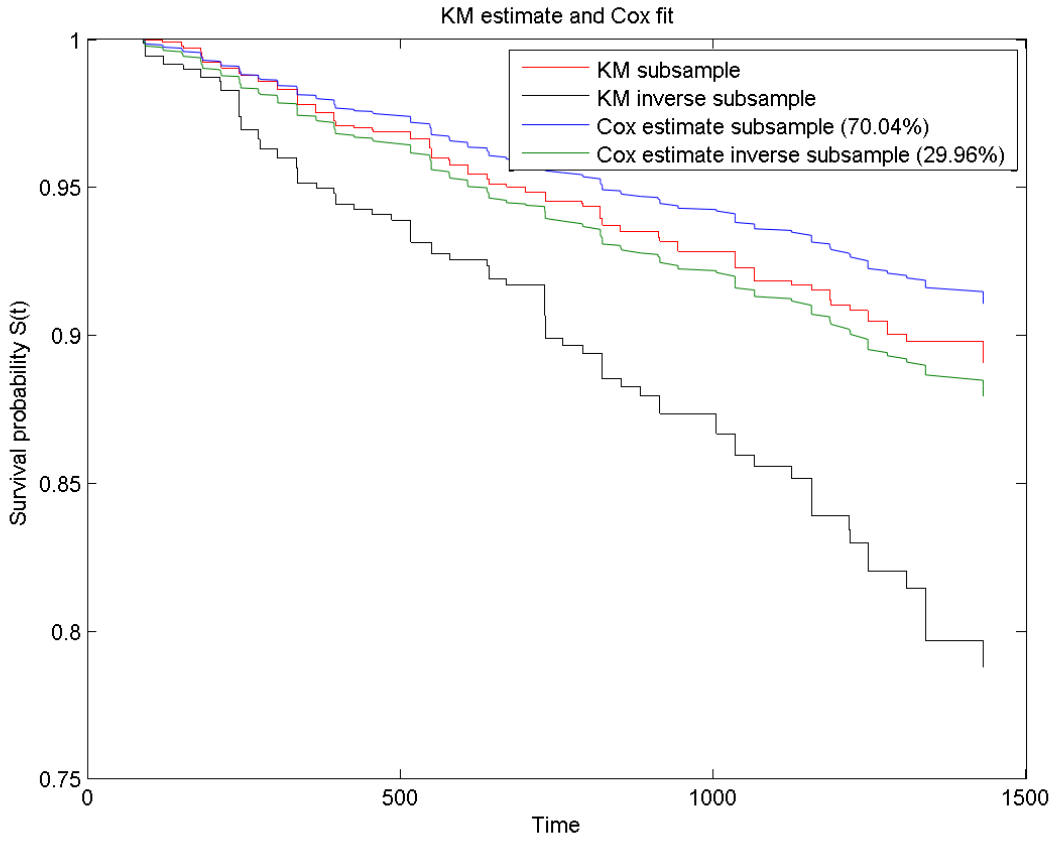

Figure 31: Marginal survival function stratified by setting a threshold of a continuous factor.

4.5 Calibration

After the multifactor analysis, the next step is calibration. Calibration consists of mapping the scores in the standard range to either a rating or a probability of default in order to calculate capital requirements.

Figure 32: Overview of steps in the survival model development (Calibration highlighted).

The score of the Cox proportional hazards model is easy to extract from the formula. The model is given by:

$$h(t|x) = h_0(t)\exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right) = h_0(t)\exp\left(\beta' x\right) \qquad (24)$$

Where

- $h_0(t)$ is the baseline hazard rate at time t.
- x is the observation.
- β is the estimated coefficient for observation x.

The score can be determined as only the term $\left(\beta_1 x_1 + \beta_2 x_3 + \cdots + \beta_p x_p\right)$ because the baseline hazard $h_0(t)$ is the same for every observation at time t. The use of this term of the formula as the score has some pros and cons. First of all it realizes easy comparison between different loans because they can be compared without baseline estimation. The comparison of different portfolios on the other hand, becomes impossible since the scores are relative to the baseline function. One can determine which facility in the portfolio is most at risk but in order to compare different portfolios, the baseline estimate is also required.

The calibration process starts with bucketing the scores in order to create groups with homogeneous exposures, as prescribed by the DNB regulations.[2] The final scores are bucketed using the approach given in Section 4.3.1. Next the model is used to estimate the survival probability within a specific time horizon. The time horizon should be set to 24 or 36 months because a 12 months cut off point might be too vigorous for the estimation. This estimation might give biased results since the survival analysis tries to estimate the total survival time over the entire study period and the data of the first year might not represent the next few years. The estimation is done by calculating the mean of the scores for all facilities in the bucket. Next calculate the survival probability over the time horizon. Assume the default probability is linear and calculate the default probability of 1 year.

So for example if after 36 months the survival probability is 94%, the change of defaulting within 36 months is 6%. Assuming the default probability is equal over every year, the PD for 1 year (x) is calculated as:

$$x + x(1-x) + x(1-x)^2 = 6\% \qquad \rightarrow \quad x = 2.71\%$$

4.6 Performance assessment

The performance of the survival model is compared to the logistic model. There are several techniques available to compare models such as Receiver Operating Characteristic curve (ROC), Kolmogorov-Smirnov statistic (K-S or KS) and Power Statistic. In order to compare the logistic model with the survival model the time to event and censoring are transformed to good or bad within a specific time horizon (see Section 5.3).

4.6.1 ROC curve

The industry standard for comparing two or more scoring algorithms are the ROC curves (Thomas, Edelman, & Crook, 2004). The ROC curve is created by first ranking the scores for all the facilities from bad to good. Next on the y-axis the percentage of defaults out of the total number of defaults is plotted also called hit rate. On the x-axes the false alarm rate is plotted, this is the percentage of goods out of the total number of goods. The quality of the model can be quantified by the area under curve value (AUC). The higher the AUC value is (closer to 1) the better the model estimate is.

[2] The BIS-II requirements for retail development are summarized in the PD, LGD and EAD checklists from (Rabobank, Checklist for EAD models, 2008), (Rabobank, Checklist for LGD models, 2008) & (Rabobank, Checklist for PD models, 2008)

$$Hit\ Rate\ (c) = \frac{\sum_{i=\min(c)}^{c} Number\ of\ Defaults_i}{Total\ Number\ of\ defaults} \tag{25}$$

$$False\ Alarm\ Rate\ (c) = \frac{\sum_{i=\min(c)}^{c} Number\ of\ Non\ Defaults_i}{Total\ Number\ of\ Non\ Defaults} \tag{26}$$

In Figure 33 an example of a ROC curve is given. The blue line is the actual model, the red line is the perfect model and the green line is the random model. The AUC in this example is 0.77 which is a good result.

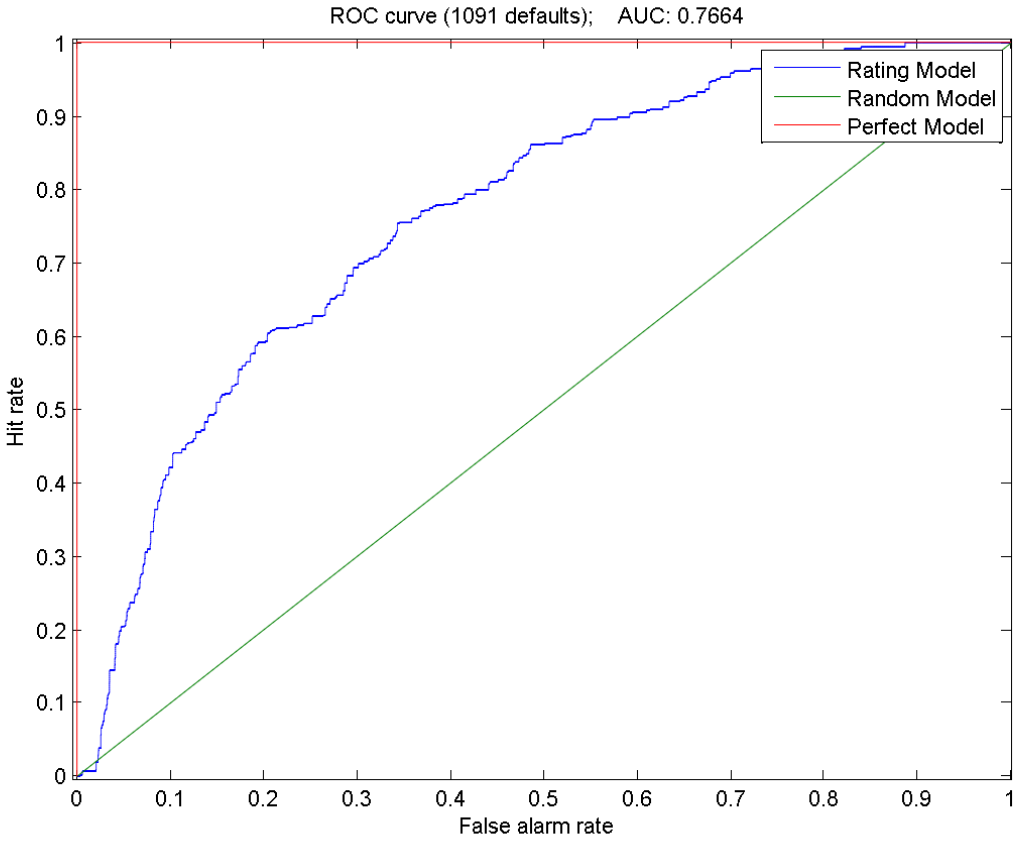

Figure 33: ROC curve with AUC of 0.77.

4.6.2 KS statistic

Another popular measure to assess the performance of the model is the KS statistic. This measure is in essence the difference between the empirical distribution of the good values and the bad values. The empirical distributions are given by:

$$D_K = \begin{cases} 1, & Client\ is\ good \\ 0, & otherwise \end{cases}$$

$$F_{m,BAD}(a) = \frac{1}{m}\sum_{i=1}^{m} I(s_i \le a \wedge D_K = 0)\,, a \in [L, H] \tag{27}$$

$$F_{m,,GOOD}(a) = \frac{1}{n}\sum_{i=1}^{n} I(s_i \leq a \wedge D_K = 1), a \in [L, H] \quad (28)$$

Where s_i is the score of the ith client, n is the number of good clients, m the number of bad clients and I is the indicator function if $I(true) = 1$ and $I(false) = 0$.

The KS statistic is given by:

$$KS = \max_{a \in [L,H]} \left| F_{m,BAD}(a) - F_{m,,GOOD}(a) \right| \quad (29)$$

Figure 34 gives an example of estimation of distribution functions of good and bad clients. Furthermore a KS statistic is calculated. The empirical distribution of the bad observations is the red line, while the empirical distribution of the good observations is plotted as the green line. The example is based upon a survival model as can be seen from the fact that higher scores result in a lower creditworthiness (red line is below the green line). The KS statistic is determined as 0.41 at around the score of -3.5. It can be seen that the score around -3.5 has a population of around 75% of the good clients and 30% of the bad clients.

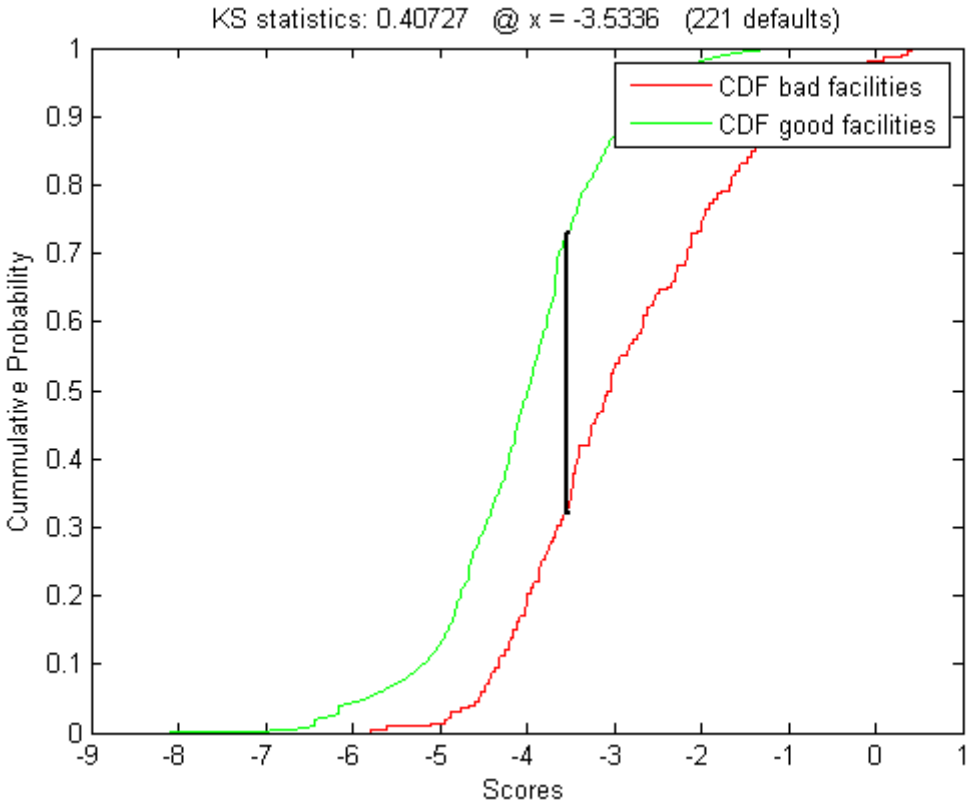

Figure 34: KS statistic with value of 0.41.

4.6.3 Power statistic

Power statistic is the last performance assessment (see Appendix B). In Figure 35 an example of a power stat is given. The red line, Crystal Ball, is the perfect model estimate and the blue line is the model estimate. Furthermore the green line is the random line. The closer the model estimate is to the Crystal ball the more discriminatory the model is and thus a better fit. In order to quantify the power statistic the area under the blue line to the green line is calculated and divided by the area under the Crystal Ball to the green line.

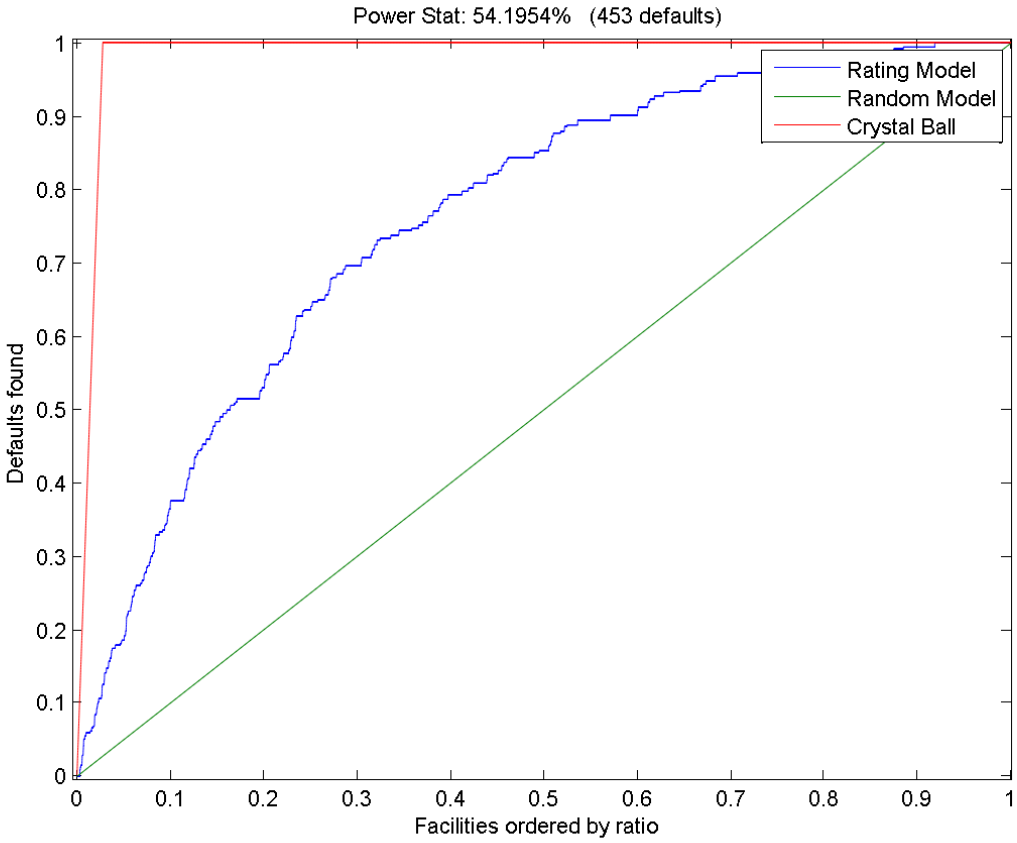

Figure 35: Power statistic with value 53.1%.

5 Results

5.1 Introduction

The survival analysis model developed is compared to the logistic regression method. In the first section the different functions per step in the framework are explained and how the different functions are compared. The second section explains which data are used and the characteristics of this data. In Section 5.4 the results of the different models, procedures and datasets are shown. And finally in Section 5.5 the results of the models are summed up.

5.2 Methodology

The procedure described in Chapter 4 is compared to the current model: the logistic regression. An overview of all the functions used to develop both models is given in Figure 36. Both, the logistic and survival model, have specific functions only available for that model. But some functions can be used in both models. These functions are presented in the middle: logistic transformation, selection elimination, ROC plot, power statistic and the KS statistic.

The model development starts with the data which are the same for both the logistic and survival model in order to be able to compare both models. The data consist of B2C, SME and simulation data (see Section 5.3).

The singlefactor analysis contains the performance measurement and transformation of the factors. The transformation for the logistic model can be done in two ways: statistical optimal bucketing and logistic transformation, but in this thesis only the logistic transformation is developed. The survival model transforms variables using the logistic transformation as well as the newly developed logrank transformation.

The logistic model requires the selection elimination method for the multifactor analysis. The survival model uses the selection elimination method, backwards elimination method and forward selection. The maximum number of factors is determined by experts. In this thesis no maximum number of factors is assigned.

The calibration buckets the scores into different rating classes. This step uses the bucketing procedure of the singlefactor analysis. The logistic model uses the statistically optimal approach were the survival model uses the logrank transformation (see Appendix B). The calibration step is not required in order to compare the performance of the survival model to the logistic model.

The last stage is the comparison between the logistic model and survival model. The comparison is based upon the ROC plot, Power statistic and the KS statistic as explained in Section 4.6. The performance of the model is measured over one, two and three years because survival analysis is built to predict the time to event over the total portfolio length rather than a fixed time interval.

The focus of these results is on the performance of the newly developed procedures such as the logrank transformation and incorporation of different observations per facility. Furthermore the factor selection procedure and the performance difference between the logistic and survival model will be highlighted.
The comparison between the logistic model and the survival model is commonly discussed in literature. For predicting the probability of default within a single period, the survival model has no or little advantage over the logistic model. For example, Stepanova & Thomas [2002] found that the differences between the models were nearly indistinguishable.

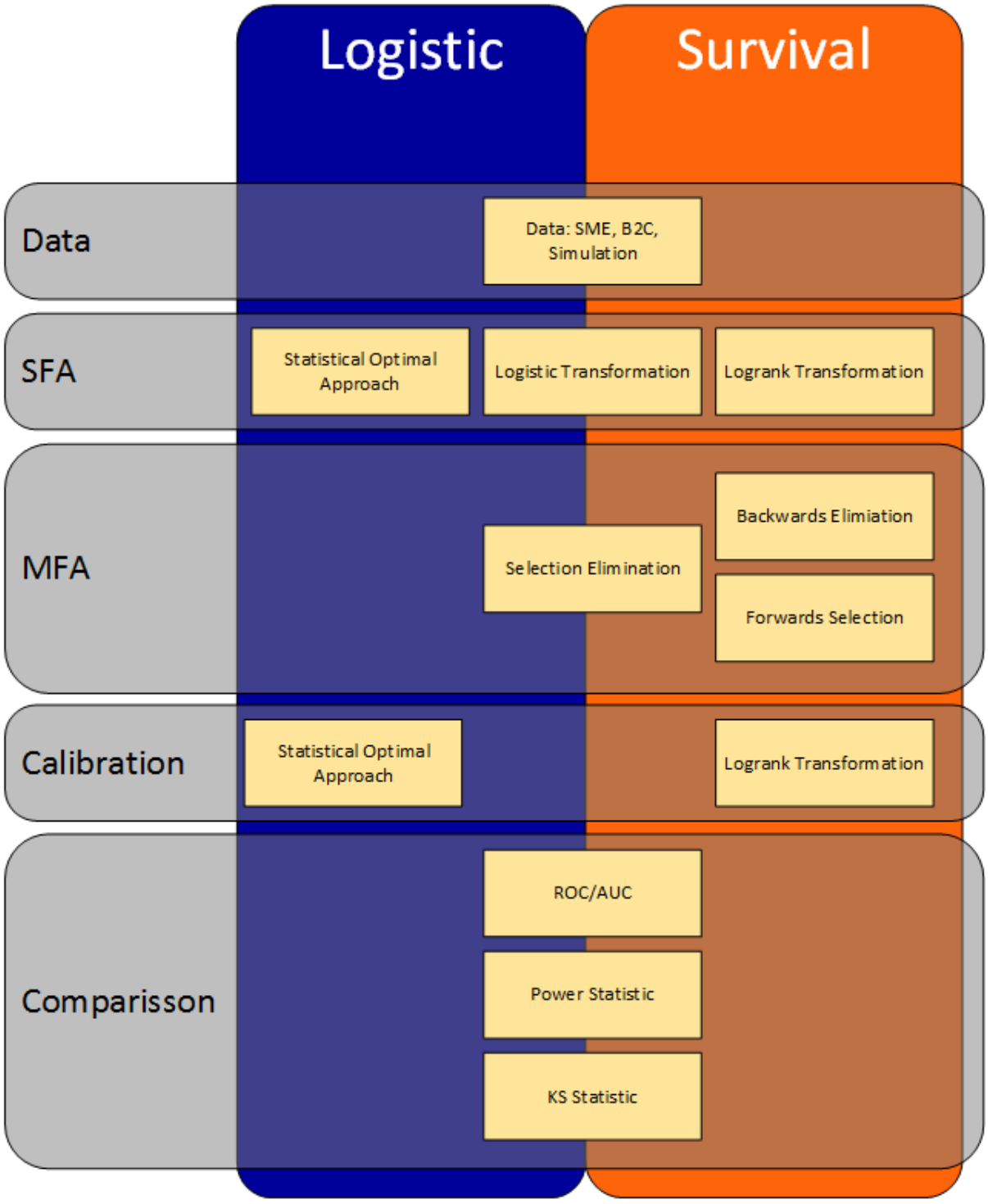

Figure 36: Overview of the functions of the logistic and survival model.

5.3 Data

For the development of the models three different sets of data are used of different types. There is a retail dataset (B2C), corporate dataset (SME) and a simulation dataset. A summary of the dataset is given in

Table 3. The B2C and SME are internal datasets on the historical loan performance. Simulation dataset is an artificial dataset created according to a simple survival model.

Table 3: Overview properties of the datasets

	B2C	SME	Simulation
Type	Retail	Corporate	-
# observations	16383	14953	1000
# facilities	2290	6789	1000
# defaults	195	347	106
% defaults	8.52%	5.11%	10,60%
# censored	2095	6442	894
# raw factors	1	0	2
# scored factors	9	78	0
# total factors	10	78	2
Start date	31-01-2008	21-04-2003	0
End date	31-07-2012	12-10-2012	1825 days
Time span	1430 days	3789 days	1825 days
Weights available	Yes	Yes	No

To be able to compare the models, both the B2C and SME datasets are cleaned for the good-bad analysis (logistic regression). As a result no defaults are available in the first three months. If a facility defaults, the last observation of the facility, marked as bad, should be between 3 and 15 months before the actual default date. Facilities with only one observation, defaulting within the first 3 months after these observations, are removed from the dataset. As a result the minimum time to event in this dataset is 3 months.

Every facility in the B2C dataset is reviewed on a 3 months interval and the SME dataset is reviewed on a 1 year interval. The simulation dataset contains only one observation per facility.

The B2C dataset has 9 scored variables and 1 raw variable. Scored variable 1 is the transformed score of the raw variable using the statistical optimal approach. In order to test the new developed logrank transformation the raw variable is transformed again using the logistic and logrank transformation, and added to the scored variables. This is scored variable number 10.

To compare the logistic model with the survival analysis the time to event and censored observation in the datasets are transformed to a good-bad variable. The following procedure is used:

$$Good/Bad = \begin{cases} 1\ (bad), tte < (365 * years)\ AND\ cens = 0 \\ 0\ (good), otherwise \end{cases}$$

Where years is the number of years the PD is estimated for, tte is the time to default and cens is a flag if the observation is censored (1) or defaulted (0).

Models calculated for:

- 0-1 year
- 0-2 year
- 0-3 year

In Thomas et al. [1999] the data are transformed for 0-1 year and 1-2 year. This has certain advantages if for example the estimation for the first year is not good, this is neglected in the results for the second part. In our research the model is not estimated for this part, because it is averaged over a couple of years. For example, the one year PD is estimated by calculating the 3 year PD and divided by 3.

5.4 Models

The performance of the models is calculated for every dataset. Both methodologies for incorporating different observations per facility are used in the B2C data and the SME data. Section 5.4.1 describes the results of the B2C data, Section 5.4.2 contains the results for the SME data and Section 5.4.3 contains the results of the simulation data. The survival model is build using the three types of selection types: forwards selection, selection elimination and backwards elimination.

5.4.1 B2C

The B2C dataset has nine scored factors and one raw factor. The nine scored factors are transformed using the statistical optimal approach. The raw factor was already transformed into scored factor one. In order to compare to newly developed logrank transformation to the logistic transformation, the raw factors is transformed using both techniques and added as scored factor number 10. So each dataset contains 10 scored factors of which numbers 1 and 10 contain the same underlying raw variable. The results are given in Table 4.

Table 4: Results B2C data (1 observation per facility).

Model type	Survival		Survival		Survival		Logistic
Selection type	Backward elimination		Forward selection		Selection elimination		Selection elimination
Transformation Type	Log rank	Logistic	Log rank	Logistic	Log rank	Logistic	Logistic
Selected factors	3, 6, 7, 8, 9, 10	1, 3, 6, 7, 8, 9, 10	3, 6, 9, 10	1, 3, 6, 7, 8, 9	3, 6, 9, 10	1, 3, 6, 7, 8, 9	1, 3, 6, 7, 8, 9
1 Year							
ROC/AUC	0.76	0.76	0.75	0.76	0.75	0.76	0.77
Power statistic	51.15%	52.49%	49.03%	52.85%	50.46%	52.85%	54.56%
KS Statistic	0.40	0.43	0.40	0.45	0.39	0.45	0.45
2 Year							
ROC/AUC	0.74	0.76	0.74	0.73	0.74	0.73	0.72
Power statistic	48.53%	45.18%	47.14%	45.24%	48.69%	45.24%	44.59%
KS Statistic	0.36	0.34	0.35	0.34	0.35	0.34	0.35
3 Year							
ROC/AUC	0.77	0.75	0.76	0.75	0.77	0.75	0.75
Power statistic	54.11%	49.65%	53.85%	49.94%	53.47%	49.95%	49.49%
KS Statistic	0.40	0.38	0.39	0.38	0.38	0.38	0.37

As can be observed in Table 4 and Table 5, when the logrank transformation is applied, the scored factor 10 is in all cases selected and factor 1 is rejected. On the other hand when scored factor 10 is transformed using the logistic transformation, the factor is only selected in 1 out of 4 methods. This concludes that the logrank transformation outperforms the logistic transformation and statistical optimal approach.

Different selection procedures on this data set show some differences. The backwards elimination more factors compared to the forward selection and selection elimination procedure. This is caused by the selection of highly correlated factors which can result in an unstable model. The performance of the backwards elimination is comparable to the other models.

Table 5: Results B2C data (all observations).

Model type	Survival		Survival		Survival		Logistic
Selection type	Backward elimination		Forward selection		Selection elimination		Selection elimination
Transformation Type	Log-rank	Log-istic	Log-rank	Log-istic	Log rank	Log-istic	Logistic
Selected factors	3, 4, 5, 6, 9, 10	1, 3, 4, 5, 6, 9, 10	3, 4, 5, 6, 9, 10	1, 3, 4, 5, 6, 9, 10	3, 5, 6, 9, 10	3, 5, 6, 9, 10	1 year: 2, 3, 5, 6, 7, 9, 10 2 year: 2, 3, 4, 5, 6, 9, 10 3 year: 1, 3, 5, 6, 7, 9, 10
1 Year							
ROC/AUC	0.77	0.75	0.77	0.75	0.76	0.75	0.76
Power statistic	53.49%	50.67%	53.49%	50.67%	52.37%	49.52%	52.19%
KS Statistic	0.43	0.40	0.43	0.40	0.41	0.40	0.43
2 Year							
ROC/AUC	0.77	0.76	0.77	0.76	0.77	0.75	0.75
Power statistic	54.58%	51.04%	54.58%	51.04%	53.92%	50.30%	50.64%
KS Statistic	0.45	0.40	0.45	0.40	0.43	0.40	0.39
3 Year							
ROC/AUC	0.78	0.77	0.78	0.77	0.78	0.77	0.76
Power statistic	57.10%	53.18%	57.10%	53.18%	56.96%	53.28%	51.52%
KS Statistic	0.48	0.43	0.48	0.43	0.46	0.42	0.40

Furthermore the different observations methodologies in Table 4 and Table 5 have no significant difference in the results. The difference between the two tables is the selection of different factors. Some factors selected in Table 4 were also selected in Table 5, and remain in the model, but some new factors entered the model.

Different selection procedures on this data set show some differences. First of all the selection elimination procedure didn't select the fourth factor from the dataset where the other procedures did. Although the performance of the final model is comparable with the other variables.

The performance over the different years shows no significant change of predicting quality of the different models. In Table 4 the performance of the estimation of the second year, is a little lower compared to the first and third year, but that trend is visible in all the model, possibly due to the dataset.

The results of the ROC/AUC, power statistic and KS statistic shows there is little difference in the performance of the survival models and the logistic regression

5.4.2 SME

The second dataset is the SME dataset which has a lot of different scored factors with different observations per facility. In Table 6 the results of the inclusion of only one observation per facility is given and in Table 7 the results of inclusion of all the observations are given.

Table 6: Results SME data (1 observation per facility).

Model type	Survival	Survival	Survival	Logistic
Selection type	Backward elimination	Forward selection	Selection elimination	Selection elimination
Selected factors	1, 5, 11, 13, 14, 21, 22, 26, 27, 30, 38, 39, 46, 48, 49, 50, 56, 63, 64, 65, 66, 67	1, 5, 13, 27, 39, 48, 76	1, 5, 11, 13, 27, 39, 44, 46, 48, 76	1 year: 1, 5, 13, 27, 76 2 year: 1, 5, 13, 27, 37, 39, 48, 76 3 year: 1, 5, 13, 27, 29, 34, 38, 39, 44, 46, 48, 67, 70, 76
1 Year				
ROC/AUC	0.77	0.77	0.75	0.73
Power statistic	54.95%	53.67%	50.68%	46.04%
KS Statistic	0.43	0.44	0.40	0.37
2 Year				
ROC/AUC	0.77	0.76	0.75	0.79
Power statistic	53.27%	52.84%	50.46%	57.24%
KS Statistic	0.40	0.43	0.40	0.45
3 Year				
ROC/AUC	0.75	0.74	0.73	0.78
Power statistic	49.56%	48.82%	46.38%	56.11%
KS Statistic	0.37	0.40	0.37	0.44

Table 7: Results SME data (all observations).

Model type	Survival	Survival	Survival	Logistic
Selection type	Backward elimination	Forward selection	Selection elimination	Selection elimination
Selected factors	1, 3, 4, 5, 7, 8, 10, 13, 15, 21, 29, 32, 37, 38, 43, 45, 47, 61, 67,78	1, 4, 5, 7, 13, 29, 38, 43, 47, 67	1, 4, 5, 7, 13, 15, 29, 38, 43, 47, 67	1 year: 1, 4, 5, 13, 27, 76 2 year: 1, 4, 5, 13, 27, 37, 39, 48, 76 3 year: 1, 4, 5, 13, 27, 29, 34, 38, 39, 44, 46, 48, 67, 70, 76
1 Year				
ROC/AUC	0.77	0.75	0.75	0.73
Power statistic	54.95%	49.62%	49.86%	47.23%
KS Statistic	0.44	0.41	0.40	0.36
2 Year				
ROC/AUC	0.77	0.74	0.74	0.74
Power statistic	53.67%	48.00%	48.19%	50.64%
KS Statistic	0.43	0.39	0.38	0.39
3 Year				
ROC/AUC	0.76	0.73	0.73	0.73
Power statistic	51.69%	46.64%	46.70%	46.74%
KS Statistic	0.41	0.38	0.37	0.38

The tables show similar results as the results of the B2C data. The logistic regression performs slightly better the second and third year compared to the first year. Furthermore the results of different observation methods resulted in no significant improvement of the model. The overall comparison between the survival function and the logistic regression is that both models perform similarly.

The different selection procedures resulted in different selected factors. The backwards elimination selects more factors, but performance similar compared to forward or selection elimination procedure.

5.4.3 Simulation

The simulation data contains two raw factors which are transformed using the logrank and logistic methodology. Only one observation per facility is available, therefore contains Table 8 all the results.

Table 8: Results Simulation data (1 observation per facility).

Model type	Survival		Survival		Survival		Logistic
Selection type	Backward elimination		Forward selection		Selection elimination		Selection elimination
Transformation Type	Log-rank	Log-istic	Log-rank	Log-istic	Log rank	Log-istic	Logistic
Selected factors	1, 2	1, 2	1, 2	1, 2	1,2	1, 2	1, 2
1 Year							
ROC/AUC	0.62	0.67	0.62	0.67	0.62	0.67	0.67
Power statistic	23.55%	33.57%	23.55%	33.57%	23.55%	33.57%	33.70%
KS Statistic	0.21	0.35	0.21	0.35	0.21	0.35	0.37
2 Year							
ROC/AUC	0.60	0.65	0.60	0.65	0.60	0.65	0.65
Power statistic	20.20%	30.96%	20.20%	30.96%	20.20%	30.96%	29.20%
KS Statistic	0.19	0.32	0.19	0.32	0.19	0.32	0.32
3 Year							
ROC/AUC	0.60	0.63	0.60	0.63	0.60	0.63	0.67
Power statistic	20.91%	26.32%	20.91%	26.32%	20.91%	26.32%	26.55%
KS Statistic	0.18	0.26	0.18	0.26	0.18	0.26	0.29

The results for the logrank transformation of the variables are rather biased. From the ROC/AUC, Power statistic and KS statistic in Table 8 can be observed that the performance of the models using the logrank transformation is lower compared to the models using the logistic transformation. This is due to the fact that there are only two variables and both are divided into two buckets. That results in $2^2 = 4$ possible scores as can be observed from Figure 37 where both CDF plots contain 4 jumps.

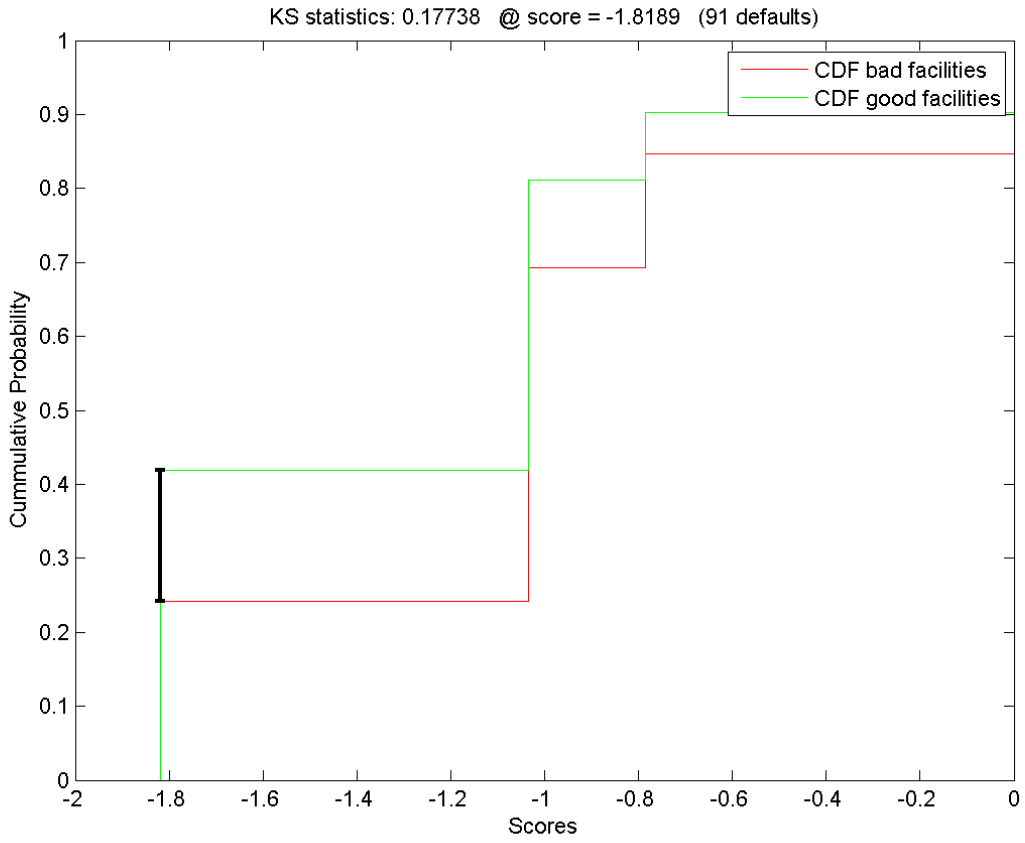

Figure 37: KS statistic based on logrank transformation.

The ROC and power statistic rank the scores and calculate the number of defaults. The ranking only distinguish between the 4 scores, but the ranking of defaults within these groups is random. One could assume that all defaults occur first within the group, which will improve the result, this is a rather positive view on reality. The cause of this problem is the transformation that is not able to distinguish several different groups. This can be partially solved by changing the parameters for the logrank test and the number of buckets the transformation starts with. This problem must be further researched. Recommended is the use of a logistic transformation in case of low number of variables to avoid results as given in Table 8.

5.5 Conclusion

The logrank transformation was tested using two datasets: B2C and the simulation dataset. In Section 5.4.1 was shown that the logrank transformation outperformed the logistic transformation. The raw variable was transformed using both methodologies and added to the dataset. The selection procedure selected the logrank transformed variable in all models. In Section 5.4.3 we found that the procedure has some shortcomings when a small number of factors is available. Both variables were split into 2 buckets which resulted in only unique 4 scores. This resulted in worse model scores.

The difference between the logistic model and the survival model was tested using all the datasets available. The comparison of the different ROC/AUC, power stat and KS statistic were not significant different between the survival models and the logistic regression. If the models were compared over time: 1 year, 2 year and 3 year, the survival model performs comparable to the logistic regression. This confirms the findings of (Stepanova & Thomas, 2002) that differences between survival and logistic models for a fixed time period are nearly indistinguishable.

The incorporation of different observation methodologies was applied to the B2C and SME dataset. The incorporation of more observations per facility could have resulted in better performance, since more data is incorporated, but these results are comparable with only one observation per facility.

The different selection procedures resulted in the selection of different factors. The selection elimination provided a good trade-off between the complexity of the model and the performance. This selection procedure selected the far less factors but the performance was similar.

The inclusion of different observations per facility resulted not in a better performing model. The performance of the models was quite similar compared to the model with only one observation.

6 Conclusion and further research

The main goal of this thesis was defined as: development of a survival model for PD estimations, using the steps of the current framework for easy implementation, and benchmark this model to the currently used logistic regression model. In order for the survival model to fit into the steps of the current framework, some new procedures are developed. These new developed procedures are tested against current procedures. The final survival model is compared to the current logistic model.

The first new procedure was the bucketing of factors, based on the survival function rather than on the default rate in a fixed period. The procedure starts with a fixed number of buckets and compares the survival function of the neighbouring buckets using the logrank test. If the survival functions are not significantly different, the buckets are merged.

The second new procedure was the transformation of raw variables by means of the so-called logrank transformation. This procedure uses the bucketing approach and fits an exponential model to survival function of each bucket. The estimates are used for the actual transformation. This transformation was compared to the logistic transformation and statistical optimal approach. A raw variable was transformed using the different transformations, next the selection method of the multifactor analysis selected in all cases the logrank transformation.

Another new procedure for the assessment of the predictive power of factor was developed based on the survival function. The factor scores are ranked and split into two buckets. The exponential model was fitted to the survival functions of both buckets and compared. The more difference between the estimates the more discriminatory between good and bad survival rates, the factor is.

The procedures for selecting the best set of predicting variables are changed for survival analysis. The Wald test statistic and the Akaike information criteria are used for this selection. Three selection procedures were described: backwards elimination, forward selection and stepwise selection elimination. The stepwise selection elimination created tradeoffs between the complexity and performance of the model. Although the number of selected factors was low, the performance of the model was similar to the performance of the other selection methods.

The incorporation of different observations per facility was not that successful. Three methods are given, but because the inclusion of truncated data in the model was beyond the scope of this thesis, only two methods were compared. The difference between the methods was not significant.

Finally the difference between the logistic model and the survival model was tested. The comparison of the different ROC/AUC, power stat and KS statistic were not significant different between the survival models and the logistic regression. If the models were compared over time: 1 year, 2 year and 3 year, the survival model performs comparable to the logistic regression. This confirms the findings of Stepanova & Thomas [2002] that differences between survival and logistic models for a fixed time period are nearly indistinguishable.

Although the survival model performs similar to the logistic regression, the survival model provides certain advantages compared to the logistic model. The main advantage is that the cleaning of the dataset is less extensive. For example: if a facility has not got a payment delay of more than 90 days overdue, but defaults at once, the last observation between 3 months (90 days) and 15 months is marked as bad. This short time to maturity can be incorporated into survival models without any change to the dataset, since only the time to the event is of interest.

Further research
Some suggestions for further research are given to improve the performance of the survival model.

The logrank transformation performs not well in case there are a low number of variables in the dataset. An example was given in Section 5.4.3 were the simulation dataset with two variables was transformed using the logistic transformation and the logrank transformation. The logrank transformation splits both variables into two buckets resulting in $2^2=4$ possible scores. A procedure should be developed to guarantee a minimum number of buckets in this procedure. This will prevent problems as experienced in this dataset.

Second the incorporation of truncated data might result in a better model. The incorporation was beyond the scope of this thesis and has therefore not been developed. The incorporation of this type of data allows the incorporation of different observations per facility using methodology 2 (see Section 4.2).

Finally, many extensions are available to the survival analysis. One of which consists of incorporating macro-economic variables, see e.g. (Bellotti & Crook, 2008). In this paper they suggest that the inclusion of macroeconomic variables gives a statistically significant improvement in predictive performance. The inclusion of these variables into the model might be further researched.

7 Bibliography

Aalen, O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models. *The Annals of Statistics* , 534-545.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* , 716-723.

Anderson, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation.* New York: Oxford University Press Inc.

Bellotti, T., & Crook, J. (2008). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society* , 1699-1707.

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics 30* , 89-99.

Cox, D. R. (1975). Partial likelihood. *Biometrika 62* , 269-276.

Cox, D. R. (1972). Regression models and life-tables. *J. Royal Statist. Society* , 187-220.

Cox, D., & Oakes, D. (1984). *Analysis of survival data.* CRC Press.

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183* , 1447-1465.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Amerian Statistical Association 72* , 557-565.

Gehan, E. A. (1965). A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. *Biometrika 52* , 203.

Greenwood, M. (1926). A Report on the Natural Duration of Cancer. *Reports on Public Health and Medical Subjects* , 1-26.

Herel, M. v., Hoek, B. v., & Vedder, R. (2010). *QRA Rating Model Development Guideline.* Rabobank internal document.

Herel, M. v., Hoek, B. v., & Vedder, R. (2012). *Retail Modelling Development Guidelines.* Rabobank internal document.

Hosmer, D., Lemeshow, S., & May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data* (second ed.). Wiley-Interscience.

International, Q. d. (2011). *PDLGDresultsv3.* Rabobank internal document.

Kalbfleisch, J., & Prentice, R. (1980). *The statistical analysis of failure time data.* New York: Wiley.

Kalbfleisch, J., & Prentice, R. (2002). *The Statistical Analysis of Failure time Data.* (2nd edition ed.). Hoboken, New Jersey: John Wiley Sons, Inc.

Kaplan, E., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of American Statistical Association 53* , 457-481.

Kay, R., & Kinnersley, N. (2002). On the use of the accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data: A case study in influenza. *Drug Information Journal* , 571-579.

Keiding, N., & Andersen, P. &. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in medicine* , 215-224.

Kleinbaum, D. (1998). Survival Analysis, a Self-Learning Text. *Biometrical Journal* , 107-108.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data Analysis.* New York: Wiley.

Leemis, L. (1995). *Reliability.* Englewood Cliffs: Prentice-Hall.

Narain, B. (1992). Survival analysis and the credit granting decision.

Rabobank, R. d. (2008). *Checklist for EAD models.* Rabobank internal document.

Rabobank, R. d. (2008). *Checklist for LGD models.* Rabobank internal document.

Rabobank, R. d. (2008). *Checklist for PD models.* Rabobank internal document.

Řezáč, M., & Řezáč, F. (2011). How to measure the quality of credit scoring. *Czech Journal of Economics and Finance* , 486–507.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tets for comparing survival distributions. *Biometrika 68* , 316-319.

Siddiqi, N. (2005). *Credit risk scorecards: developing and implementing intelligent credit scororing.* Wiley.

Stepanova, M., & Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research 50(2)* , 277–289.
Thomas, L. C., Crook, J. N., Edelman, D. B., & eds. (1992). *Credit Scoring and Credit Control.* Oxford University Press.
Thomas, L., Banasik, J., & Crook, J. (1999). Not if but when loans default. *J. Oper. Res. Soc. 50* , 1185–1190.
Thomas, L., Edelman, D., & Crook, J. (2004). *Reading in Credit Scoring: Recent Developments, Advances, and Aims.* Oxford, UK: Oxford University Press.
Tong, E. N., Mues, C., & Tomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research , 218*, 132-139.
Wedderburn, J. N. (1972). *Generalized Linear Models.*
Yan, J. (2004). Survival Analysis: Techniques for Censored and Truncated Data. *Journal of the American Statistical Association , 99*, 900-901.
Yitzhaki, S. (1979). Relative Deprivation and the Gini coefficient. *The Quarterly Journal of Economics* , 321-324.

8 Appendices

Appendix A: List of figures

Appendix B: Current model

B 1 Singlefactor analysis

The first stage in the model development is the singlefactor analysis (SFA). In this stage the factors that have predictive power for defaults are selected and transformed. For example, a defaulted counterparty prior to the default is likely to have factor scores that are significantly lower or higher than other comparable counterparties that did not default. The main task is to find the factors that for which either high or low values correspond to high PDs.

The goal of the SFA is twofold:

3. The selection of factors for further modelling, which is done by analysing the standalone discriminatory powers of the factors and expert opinion.
4. The transformation of the factor values into interpretable scores, i.e. between 0-10. This transformation is generally based on the relation between the factors and the goal variable.

Figure 38: Overview of steps in the singlefactor analysis

The SFA consists of several steps as is show in Figure 3. This chapter starts with explaining the different bucketing techniques used within RI and is followed by explaining the different performance measurement techniques. The chapter will end with the transformation of the factors.

B 1.1 Bucketing techniques

Before assessing the performance of the individual factors, it should be considered to bucket the factors. Bucketing of factors has certain advantages, firstly the relation between creditworthiness and factor values can be assessed more easily by experts. Secondly some performance tests require buckets such as information value and weight of evidence. Lastly is it helpful for the transformation that needs to be performed.

Figure 39: Overview of steps in the singlefactor analysis (Bucketing techniques highlighted)

Basically there are three types of bucketing approaches:

- Statistical significance or statistical optimal bucketing, this is the preferred approach within RI and discriminates on statistically significant difference in the creditworthiness between two buckets. This procedure is further explained below.
- Equally sized buckets, this creates buckets with similar size. For example a factor with a score from 0-100 the buckets can be made for scores from 0-20, 20-40 and so on. (see Figure 40)
- Equally distributed, this approach creates buckets with the same number of facilities in each bucket. (see Figure 41)

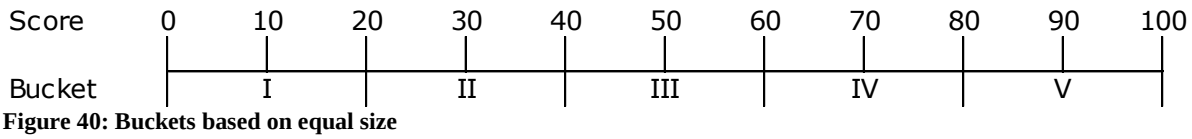

Figure 40: Buckets based on equal size

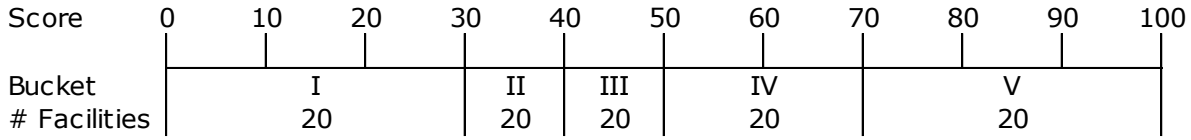

Figure 41: Buckets based on distribution

The preferred approach for bucketing is the statistical optimal bucketing approach. A bucket contains those observations that have factor values between the corresponding bucket boundaries. This approach starts with a large number of buckets and calculates the average default frequency in each bucket. Because buckets should discriminate neighbouring buckets on basis of default frequency, a t-test is used to determine if two samples have the same average default frequency. If they don't have significantly different average default frequency the buckets are merged. The tests will be applied with a specified significance level (95%). This procedure will continue until all tests are significant or when there is only one bucket left.

After the bucketing process a histogram is generated showing the number of observations per bucket. Next the average default frequency per bucket is calculated and plotted in the middle of the bucket. Next a trend line is fitted on the default frequency per bucket. This type of bucketing can detect positive and negative linear trends and U-shaped trends. In U-shaped trends both very high and very low factor values correspond to worse (or better) creditworthiness.

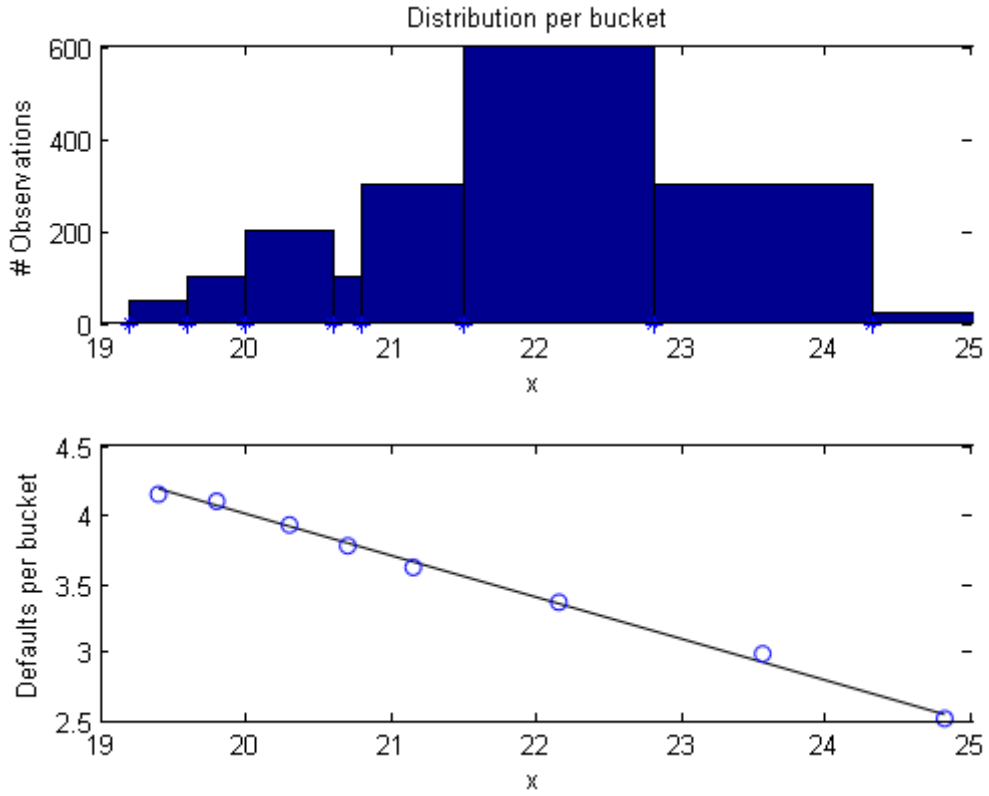

Figure 42: Bucketing continuous variable

B 1.2 Performance measurement

An important part of the SFA is the measure of the predictive power of an individual factor. This measure is used in the selection and transformation of factors. Four measures of predictive power will be given: Power Statistic, Weight of Evidence (WoE), Information value and trend (Siddiqi, 2005). The most popular used within RI are the power statistic and trend figures because they require a

minimum amount of assumptions. (Herel, Hoek, & Vedder, QRA Rating Model Development Guideline, 2010)

Figure 43: Overview of steps in the singlefactor analysis (Performance measurement highlighted)

B 1.2.1 Power statistic

The power statistic closely related to Gini coefficient, visualizes and quantifies the predictive power of individual factors. The idea is that the worst factor values should correspond to the worst observations (bads).

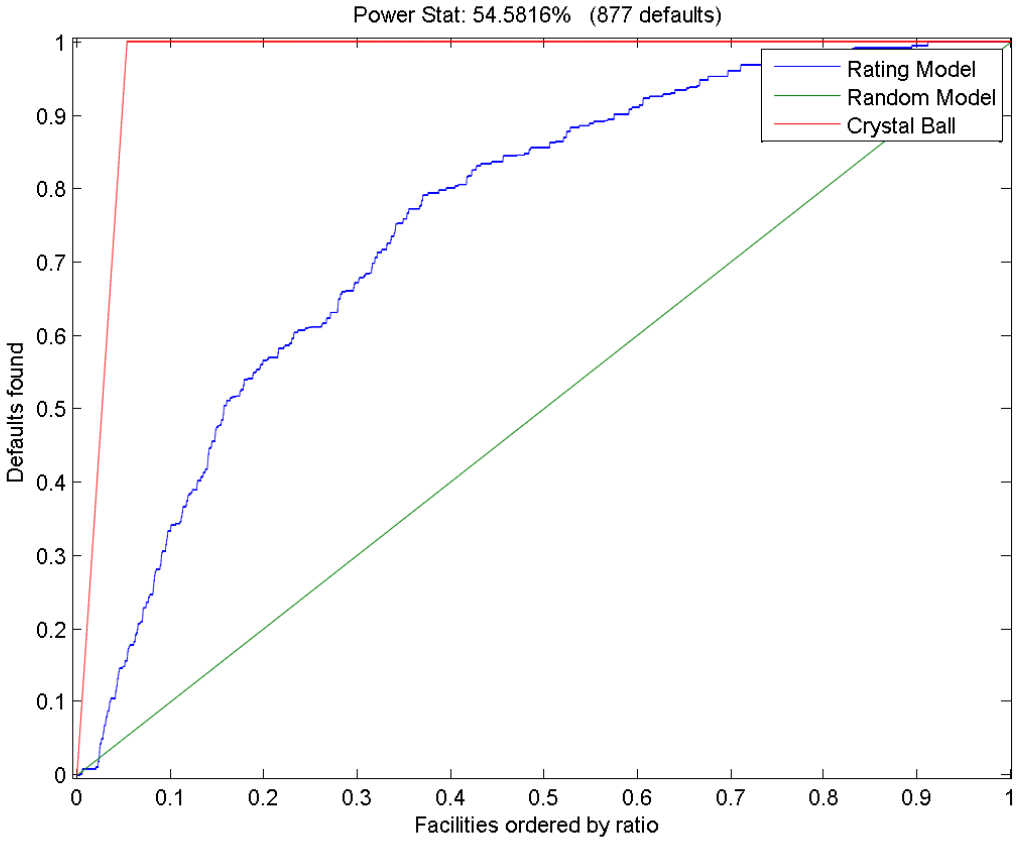

Figure 44: Power Statistic

As can be seen from Figure 44, the red line has perfect discriminatory power: all bads are with the worst factor values. The green line is the random line, this means there is no relation between the factor values and the percentage of bad. To construct this graph, rank all factor values from low to high and plot the them on the x-axis. Next compute the cumulative percentage of bads and plot them on the y-axis. The closer the rating model is to the Crystal ball line, the more discriminatory the factor is. The closer the factor is to the green line, the less discriminatory the factor is.

The power statistic can be computed as the surface between the blue line and the green line, divided by the area between red line and the green line. The larger the quotient, the higher discriminatory power. A negative quotients indicates an inverse relationship: the higher the factor values the lower the creditworthiness.

B 1.2.2 Trend

Another tool used by RI for visualizing and determining the predictive power of a singlefactor is the trend analysis. This is type of analysis is especially useful for factors that are already bucketed, such as the qualitative questions with a fixed number of answering possibilities. Continuous variables should be bucketed before they are analysed. This figure shows the average observed default rate and the distribution per bucket. The buckets should be ranked from bad to good and the trend of the observed default rate is expected to decline. With this technique it is easier to identify non-monotone relationships.

Figure 45: Trend for number of employees per company

The trend line shows that the observed default frequency is monotone increasing for smaller companies, which indicates that the 'number of employees' question is a factor that does distinguish companies in terms of creditworthiness

Finally the expectations and power statistic are compared with the observed trend and based upon this analysis it can be decided whether to include or exclude certain ratios from the short list.

B 1.2.3 Weight of Evidence

Another performance measure which requires bucketed factors is Weight of Evidence (WoE). (Herel, Hoek, & Vedder, Retail Modelling Development Guidelines, 2012) This measures the difference between the proportion of goods and bads within each factor category. It is a measure of how well each bucket of a particular risk factor separates good accounts from bad accounts. The formula is give by:

$$WoE_j = ln\left(\frac{\left(N_{Gj} + \frac{1}{2}\right)/(N_G + 1)}{\left(N_{Bj} + \frac{1}{2}\right)/(N_B + 1)}\right)$$

Factor cross table with k buckets							
Status / Bucket	1	2	…	j	…	k	Row totals
'Bads'	N_{B1}	N_{B2}	…	N_{Bj}	…	N_{Bk}	N_B
'Goods'	N_{G1}	N_{G2}	…	N_{Gj}	…	N_{Gk}	N_G
Column n totals	N_1	N_2	…	N_j	…	N_k	N

Table 9: WoE calculations

A negative WoE value indicates that the proportion bads in the bucket is larger than the proportion goods and vice versa. The WoE values should not become too large or too small as it indicates not properly defined buckets. In such a case a bucket has too little bads or goods in order to be discriminatory. As a rule of thumb one can say: values greater than 3 or smaller than -3 are considered questionable.

B 1.2.4 Information Value

After the WoE is calculated the Information Value (IV) (Herel, Hoek, & Vedder, Retail Modelling Development Guidelines, 2012) can be calculated. This performance measure compares the bad rate in the bucket compared to the bad rate in the total factor. The IV for each bucket is calculated using the following formula:

$$IV_j = \left(\frac{N_{Gj}}{N_G} - \frac{N_{Bj}}{N_B}\right) \cdot WoE_j$$

The information value is the sum of the individual information values of each bucket and given by:

$$IV = \sum_j IV_j$$

Information values are always positive. A high Information Value for a specific bucket (IV_j) indicates that there is a large difference between the bad rate in the bucket and the bad rate in the total factor.

The following scale is used in order to map the IV to discriminatory power:

- IVj < 0.02; these factors are regarded as non-discriminatory.
- 0.02 < IVj < 0.1; these factors are regarded as weakly discriminatory.
- 0.1 < IVj < 0.3; these factors are regarded as average discriminatory.
- IV_j > 0.3; these factors are regarded as strongly discriminatory.

The factor IV values should be between 0.1 and 0.5 (Anderson, 2007) (International, 2011). A higher IV indicates that there are errors in the data or the number of buckets is too high.

B 1.3 Transformation

After the performance measurement the factors are transformed. This transformation of a factor brings the ratios into a standard interval from 0 to 10. The goal of a transformation is twofold:

- In order to be able to compare (the coefficients of) different ratios in the multifactor analysis.
- A transformation suppresses the impact of outliers in the development process and in daily use of the model.

Figure 46: Overview of steps in the singlefactor analysis (Transformation highlighted)

The transformation of factors can be completed on the based on creditworthiness or distribution of the factor. Within RI the preferred transformation is based on the creditworthiness and this approach is used in retail modelling. For corporate modelling transformation based on the distribution of the factor is used because transformation on the basis of creditworthiness is not available.

B 1.3.1 Logistic Transformation

The preferred approach for transformation of continuous factors based on the distribution of the factor is the Logistic Transformation approach. (Herel, Hoek, & Vedder, Retail Modelling Development Guidelines, 2012) This transformation contains two steps: fitting a logistic function to the empirical function and next use this logistic function to transform the variables.

The logistic function is given by:

$$T(x) = \frac{1}{1 + e^{Slope \cdot (Midpoint - x)}} \tag{30}$$

Where:

- x is an observation from a particular ratio

Midpoint is determined by:

$$Midpoint = \frac{\left(x^{5th} + x^{95th}\right)}{2} \tag{31}$$

Where:

- x^{5th} is the value at the 5^{th} percentile
- x^{95th} is the value at the 95^{th} percentile

$$Slope = \frac{k}{x^{95th} - midpoint} \tag{32}$$

Where:

- k is determined by solving $T\left(x^{95th}\right) = 0.95$

The standard logistic given in Figure 47, is given by the function:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{33}$$

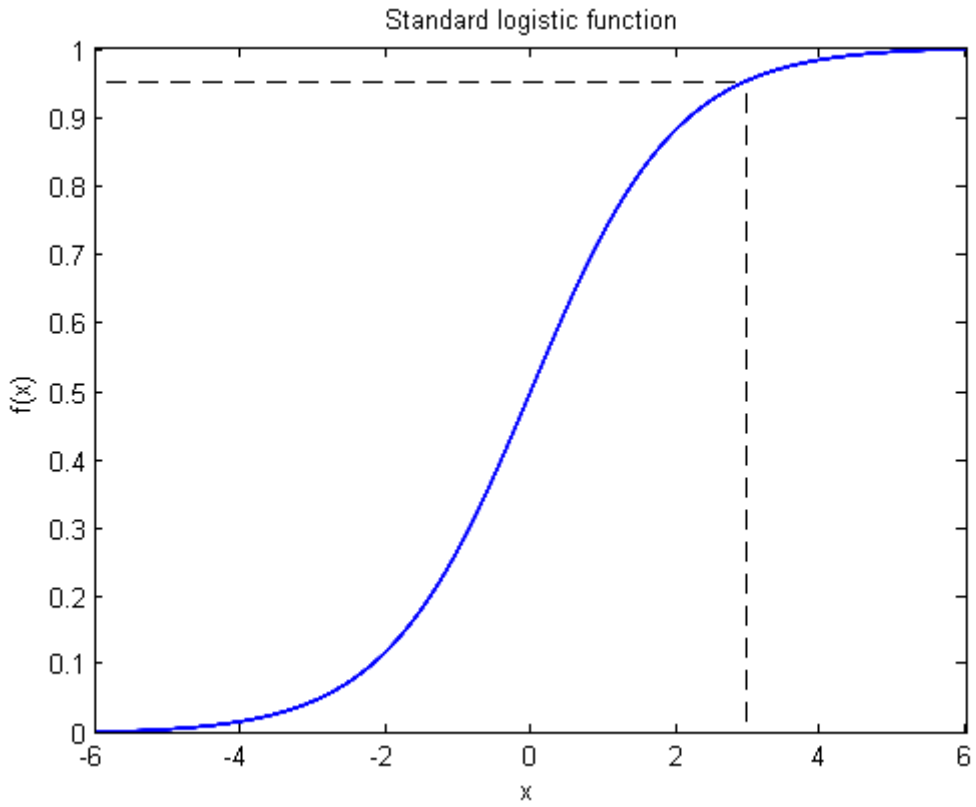

Figure 47: Standard logistic function with f(x)=0.95

The variable k is determined by solving the equation:

$$f(x) = \frac{1}{1 + e^{-x}} = 0.95 \qquad \rightarrow \qquad \mathrm{x} = -\ln\left(\frac{1}{0.95} - 1\right) = 2.94 = k$$

In Figure 48 an example of a transformation is shown. The factor scores range from 0 to 6 and must be transformed into the range of 0 to 10. First the 5th percentile and 95th percentile are calculated, where after the slope and midpoint are calculated. Next the factor values are transformed according to Equation 31.

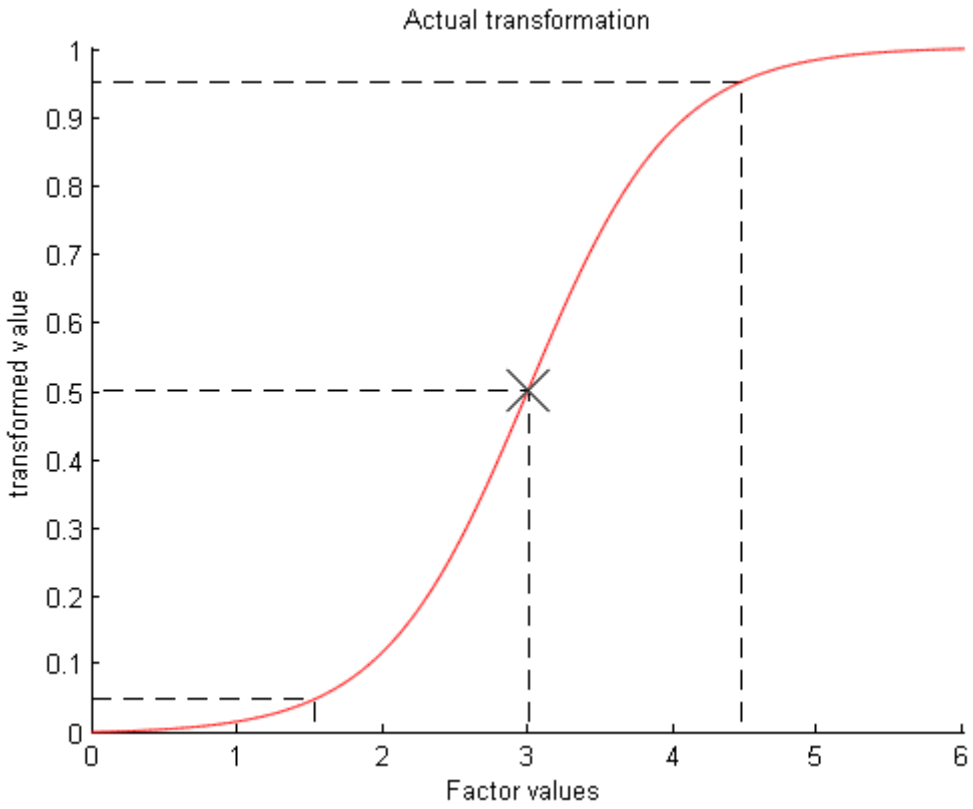

Figure 48: Logistic distribution example

B 1.3.2 Linear transformation

Another transformation which is based on the credit worthiness is called linear transformation. First the factor should be put into buckets and next calculate the WoE value for each bucket. Next a linear transformation on the WoE value is used to transform the factor scores into the range of [0,10]. The highest WoE value is defined as 10 and the lowest WoE value as a 0.

An example is given in Table 10 where the factor is split into three buckets with each 30 observations. Next the distribution goods and bads are calculated from where the WoE value is calculated. The highest value is 91.63 and the observations in this bucket get the score 10. The lowest WoE value is -69.31 and the observations in this bucket get the score 0. The other observations are linear interpolated. So the observations in bucket with WoE value of 0 get the score 4.3.

Category	Obs	Distr	Good	Bad	BadRate	Distr Good N_{Gi}/N_G	Distr Bad N_{Bi}/N_B	Distr Good /Distr Bad	WoE	IV	WoE Score
1	30	33.3%	25	5	16.7%	41.7%	16.7%	250.0%	91.63	0.23	10.0
2	30	33.3%	20	10	33.3%	33.3%	33.3%	100.0%	0.00	0.00	4.3
3	30	33.3%	15	15	50.0%	25.0%	50.0%	50.0%	-69.31	0.17	0.0
Total	90		60	30	33.3%					0.40	

Table 10: Linear transformation using WoE

B 2 Multifactor analysis

After the singlefactor analysis, the multifactor analysis is started. The multifactor analysis (MFA) determines how the individual risk drivers, identified in the SFA, are incorporated into the final model.

The goal of the MFA is to come up with a final model based on the best combined explanatory factors taking into account redundancy/dependence between the factors. (Herel, Hoek, & Vedder, Retail Modelling Development Guidelines, 2012) The multifactor analysis consists of the following steps:

Figure 49: Overview of steps in multifactor analysis

The first step is to select a regression method for the multifactor analysis. Next step is to find the best set of predicting factors that have high predictive power using the regression method. The third step test the stability and robustness of the scorecard performance. Resulting in the final model. Next step is feedback from experts. This is discussed and implemented to end up in the last step with the actual model.

B 2.1 Regression method

The goal variable is either good(0) or bad(1). Because the goal variable is binary, logistic regression is used. (Siddiqi, 2005)

Figure 50: Overview of steps in the multifactor analysis (Regression method highlighted).

B 2.2 Selection of factors

The selection of factors section consists of two steps: the stepwise multifactor analysis and sampling. Stepwise multifactor analysis finds the best set of predicting variables given a dataset. Next a sampling is used to generate subsets of the total dataset and generate a robust model.

Figure 51: Overview of steps in the multifactor analysis (Selection of factors highlighted).

B 2.2.1 Stepwise multifactor analysis

To come up with the best set of predicting factors, stepwise multifactor analysis is used. This is a heuristic approach. A sampling method is used in order to establish a robust model and furthermore the correlation between factors is checked. The final model should satisfy three requirements:

1. Good overall explanatory power
2. Limited number of factors
3. Each factor in the model has significant predictive power

The stepwise multifactor analysis consists of the following steps:

1. First a standalone logistic regression of all the factors on the goal variable is done.
2. Next a list is created of all the factors that have an intuitive weight and a p-value above the threshold, usually a p-value of 0.05 is chosen.
3. From the list select the factor with the highest F-statistic. This is a statistical test used to identify the model that best fits the population from which the data were sampled.

4. Make a new regression with the selected factors and each individual remaining factor.
5. Repeat step 2.
6. If a factor has a p-value below threshold or becomes counterintuitive the factor is deleted.
7. Select the set of factors with the highest F-statistic.
8. Repeat step 4 to 7 until no factors are left.

The result of these steps is a set of predicting variables with weights that predict the goal variable best.

B 2.2.2 Sampling

The dataset is divided into a randomly selected development sample and a holdout sample. Next the model is fitted on the development sample and tested on the independent holdout sample. Because the holdout sample is not used in the optimisation procedure this testing is very powerful. (Herel, Hoek, & Vedder, Retail Modelling Development Guidelines, 2012)

A typical ratio between the development and holdout sample is 80%/20%. Because the holdout sample is small it occasionally happens that the holdout sample is very biased. If the model, estimated on the development sample is tested on this holdout sample it would result in a incorrect result about the real performance of the model. In order to cover for this the process is 100 times repeated. With this technique the model is tested regarding:

1. Selection of factors; factors that are in for example 80% of the models selected, are the most important factors. Factors that are selected in 30-80% might also contain significant information and can be discussed with experts.
2. The weight; weight will vary for each development sample, since the estimation depends on the data. Asses the stability of weights. If they vary too much it can be considered to remove them from the sample.
3. Performance; The models that are estimated on the development samples can immediately be tested on the corresponding holdout samples. In this way there is not one performance number (determined on only one holdout sample), but there are many more. This provides better insight in the average performance as well as the stability of the performance

The final selection of factors is done in two steps: selection of factors and model performance and stability. In the first step the model is fitted 100 times and the most selected factors are selected. In the second step the model is fitted 100 times to the selected factors and the weight estimates are checked on stability.

1. Selection of factors:
    a) Produce a development sample and a holdout sample, generally 80/20%, of randomly selected records of the complete dataset.
    b) Apply stepwise regression on the development sample.
    c) Apply step 1-2 100 times and identify which factors have been incorporated in the 100 model estimates, plus the frequencies.
    d) The final factor selection is based on a combination of this analysis and expert feedback.

2. Model performance and stability
    a) Produce a development sample and a holdout sample, generally 80/20%, of randomly selected records of the complete dataset.
    b) Apply stepwise regression on the development sample.
    c) Apply step 1-2 100 times and store the weight and performance of the model.
    d) Asses the average weight and minimum, maximum and standard deviation.
    e) Assess the average performance and for example the standard deviations, minima and maxima of the performance..

Finally the factors are tested for correlation since too high correlation can lead to multicollinearity. This means that the one factor can be linearly predicted from the other, resulting in unstable weights. Another result of correlation between factors might be the selection of redundant factors. This stage eliminates these factors.

B 2.3 Scorecard performance

After the selection of factors is completed and the model performs well on the holdout samples it should be tested whether it also performs well on other subsamples. This stage is called the scorecard performance.

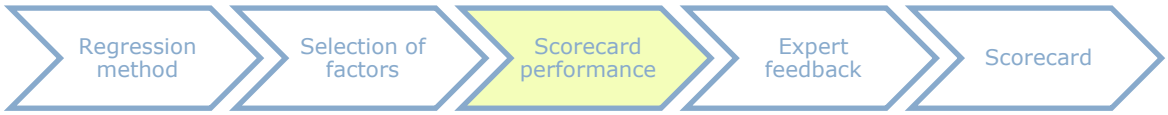

Figure 52: Overview of steps in the multifactor analysis (Scorecard performance highlighted).

After the model is completed and performs well on the holdout samples it should be tested whether it also performs well on other subsamples. The list of subsamples is created with experts. Some examples of subsamples are:

- New facilities
- Different product groups
- Application observations

.

Another way of testing the performance is to develop a model on a subgroup and test whether is model performs much better on this subgroup compared to the model developed on the whole portfolio. In case the model is significantly better, it confirms that one model for different product groups is acceptable.

B 2.4 Expert feedback

If experts strongly disagree with the statistical model, the weights can be changed accordingly to the experts and the power statistic is compared. If the difference between the two models is not significant the model with the new expert weights is selected, otherwise the results should be discussed with the experts.

Figure 53: Overview of steps in the multifactor analysis (Expert feedback highlighted).

The scorecard is completed when both, the modellers and the experts, are satisfied with the model.

B 3 Calibration

After the SFA and MFA are finished a score for each loan in the portfolio can be calculated. The higher the score, the lower the expected PD of the facilities. Next step is to bucket the scores based on homogeneous scores and assign PD values to each individual bucket. This bucketing uses the same technique as explained in Appendix B 1.1 Bucketing techniques.

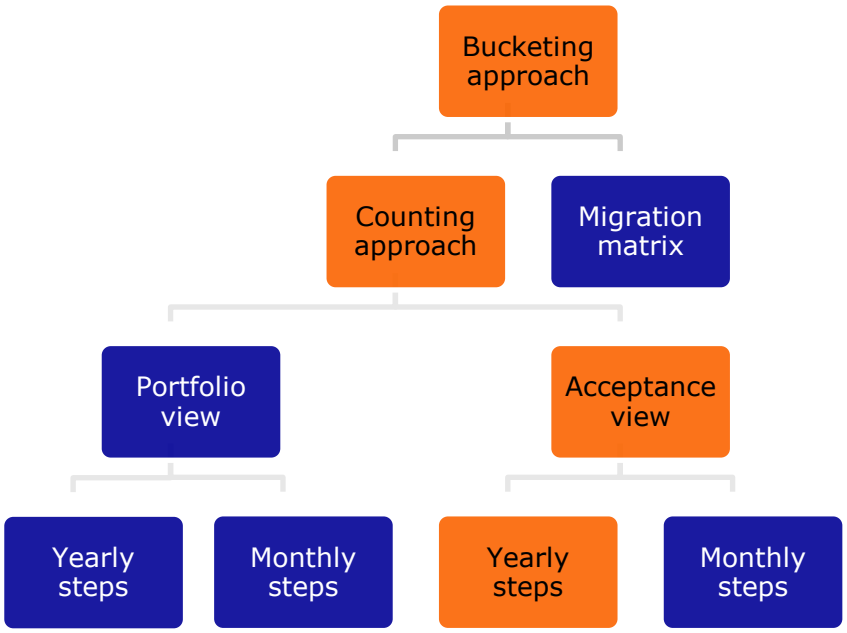

Figure 54: Overview of the different approaches for calibration (Orange tiles are preferred)

As shown in Figure 5 different approaches for calibration are available. The preferred approach is stated in orange. There are basically two techniques: migration matrix and counting approach. The most used is the counting approach because it is the most simple. The counting approach counts the number of facilities that went into default within the forecasting horizon and divides this number by the total number of facilities.

The acceptance view takes a snapshot when the facility enters the portfolio. From the moment the facility entered the portfolio a yearly snapshot is taken until the end of the facility. Next step is to divide the number of bads by the total number of facilities and calculate the PD.

B 3.1 Portfolio View

The portfolio view is explained with an example. In Figure 55 a stylized portfolio is given. The portfolio view takes a snapshot at a yearly interval, at January in this example, and counts the number of goods and bads at that time. The G in the cells means "good", the B means "bad" and the D means "default".

	2012												2013												2014		
	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec	jan	feb	mar
Facility 1	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G			
Facility 2			G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	
Facility 3	G	G	G	G	G	G	G	G	G	G	G	G															
Facility 4	G	G	G																								
Facility 5		G	G	B	B	B	B	B	B	B	B	B	B	B	B	D											
Facility 6					B	B	B	B	B	D																	

Figure 55: Stylized portfolio to explain the portfolio view calibration

In January 2012: 3 good facilities are counted, next year in January 2013: 2 good facilities and 1 bad facility and finally in January 2014: 1 good facility is counted. The PD estimate in this example is 1/(6+1) = 14%

As can be seen from Figure 55 two facilities went into default: facility 5 and 6, but only one is counted. This because facility 6 entered the portfolio in may 2012 and went in to default at October 2012. This short time to maturity and time to default are currently not taken into account.

B 3.2 Acceptance View

The acceptance view doesn't take a snapshot at a yearly interval but considers the facilities when they enter the portfolio. From the moment the facility entered the portfolio a yearly snapshot is taken until the end of the facility.

	2012												2013												2014		
	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec	jan	feb	mar
Facility 1	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G			
Facility 2			G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	
Facility 3	G	G	G	G	G	G	G	G	G	G	G	G															
Facility 4	G	G	G																								
Facility 5		G	G	B	B	B	B	B	B	B	B	B	B	B	B	D											
Facility 6					B	B	B	B	B	D																	

Figure 56: Stylized portfolio to explain the acceptance view

As can be observed from Figure 56good observations can be present for different times. For example facility 4 is only good for 3 months. In order compensate for this observations are weighted.

Each facility contributes the following number of "goods" and "bads" :

- Facility 1: 2 goods
- Facility 2: 2 goods
- Facility 3: 1 good
- Facility 4: 0.25 good
- Facility 5: 1 good and 1 bad
- Facility 6: 1 bad

Appendix C: Cox PH model estimation

C 1 Partial likelihood

The basic idea behind the partial likelihood proposed by Cox (Cox D. R., Regression models and life-tables, 1972), is that the full likelihood can be written as the product of a series of conditional likelihoods. The partial likelihood for estimating the coefficients β of the proportional hazard model of Equation 23 is given by:

$$PL_p(\beta) = \prod_{\{all\ grid\ pt\ u\}} P\{I(u) = i(u) | \mathcal{F}(u) = f(u); \lambda_0(\cdot), \beta\} = \prod_{i=1}^{m} L_i \tag{34}$$

The $I(u)$ term can be either 0 (no default) or 1(default). In case of $I(u) = 0$ the probability is 1 and therefore has no effect on the partial likelihood. In case of $I(u) = 1$ the probability is the answer to the question: what is the probability that the observed default happened to the ith facility rather than to other facilities alive at time u.

A problem with the partial likelihood is it is not able to deal with ties in the data. A solution for this problem is explained in the next section.

C 1.1 Partial likelihood example

The partial likelihood is explained using a example. In Table 11 survival times $t_1, t_2, \dots, t_6$ are given as:

Table 11: Example sample for partial likelihood estimation

Observation nr:	1	2	3	4	5	6
Time	1	3	4	10	12	18
Censored	0	0	0	1	0	0

The general partial likelihood formula is given by:

$$PL_p(\beta) = \prod_{\{all\ grid\ pt\ u\}} P\{I(u) = i(u) | \mathcal{F}(u) = f(u); \lambda_0(\cdot), \beta\} = \prod_{i=1}^{m} L_i$$

The partial likelihood for this example is given by:

$$L_p(\beta) = \prod_{i=1}^{m} L_i = \left(\frac{h_1(1)}{h_1(1) + h_2(1) + h_3(1) + h_4(1) + h_5(1) + h_6(1)}\right) \cdot \left(\frac{h_2(3)}{h_2(3) + h_3(3) + h_4(3) + h_5(3) + h_6(3)}\right) \cdot \left(\frac{h_3(4)}{h_3(4) + h_4(4) + h_5(4) + h_6(4)}\right) \cdot \left(\frac{h_5(12)}{h_5(12) + h_6(12)}\right) \cdot \left(\frac{h_6(18)}{h_6(18)}\right)$$

Where $h_j(t) = \lambda_0(t) e^{\beta x_j}$

$$I(u) = 1$$

$$L_p(\beta) = \prod_{i=1}^{m} L_i = \left(\frac{\lambda_0(1)e^{\beta x_1}}{\lambda_0(1)e^{\beta x_1} + \lambda_0(1)e^{\beta x_2} + \lambda_0(1)e^{\beta x_3} + \lambda_0(1)e^{\beta x_4} + \lambda_0(1)e^{\beta x_5} + \lambda_0(1)e^{\beta x_6}}\right)$$
$$\cdot\left(\frac{\lambda_0(3)e^{\beta x_2}}{\lambda_0(3)e^{\beta x_2} + \lambda_0(3)e^{\beta x_3} + \lambda_0(3)e^{\beta x_4} + \lambda_0(3)e^{\beta x_5}\lambda_0(3)e^{\beta x_6}}\right)$$
$$\cdot\left(\frac{\lambda_0(4)e^{\beta x_3}}{\lambda_0(4)e^{\beta x_3} + \lambda_0(4)e^{\beta x_4} + \lambda_0(4)e^{\beta x_5}\lambda_0(4)e^{\beta x_6}}\right)\cdot\left(\frac{\lambda_0(12)e^{\beta x_5}}{\lambda_0(12)e^{\beta x_5} + \lambda_0(12)e^{\beta x_6}}\right)$$
$$\cdot\left(\frac{\lambda_0(18)e^{\beta x_6}}{\lambda_0(18)e^{\beta x_6}}\right)$$

Next remove $\lambda_0(t)$ terms

$$L_p(\beta) = \prod_{i=1}^{m} L_i = \left(\frac{e^{\beta x_1}}{e^{\beta x_1} + e^{\beta x_2} + e^{\beta x_3} + e^{\beta x_4} + e^{\beta x_5} + e^{\beta x_6}}\right)$$
$$\cdot\left(\frac{e^{\beta x_2}}{e^{\beta x_2} + e^{\beta x_3} + e^{\beta x_4} + e^{\beta x_5} + e^{\beta x_6}}\right)\cdot\left(\frac{e^{\beta x_3}}{e^{\beta x_3} + e^{\beta x_4} + e^{\beta x_5} + e^{\beta x_6}}\right)$$
$$\cdot\left(\frac{e^{\beta x_5}}{e^{\beta x_5} + e^{\beta x_6}}\right)\cdot\left(\frac{e^{\beta x_6}}{e^{\beta x_6}}\right)$$

$$L_p(\beta) = \prod_{i=1}^{m}\left(\frac{e^{\beta x_j}}{\sum_{j\in R(t_i)} e^{\beta x_j}}\right)^{\delta_j}$$

Where δ_j is the censoring variable $(1 = event, 0 = censoring)$ and $R(t_i)$ is the risk set at time t_i.

$$\log L_p(\beta) = \sum_{i=1}^{m} \delta_j \left[\beta x_j - \log\left(\sum_{j\in R(t_i)} e^{\beta x_j}\right)\right]$$

In order to estimate the β the log of the partial likelihood is maximized which consists of the following steps:

1. Take the derivative of the function
2. Set the derivative to 0
3. Solve for the most likely values of β

C 2 Ties

Ties are the case if there are two or more observations with equal survival times. In real life datasets it is common that censoring and or defaults occur at the same time. For example in the loan data borrowers often default on repayment date, which is often at the end of the month. Ties are a problem in the partial likelihood estimation. There are several methods available for handling ties. (Kalbfleisch & Prentice, The Statistical Analysis of Failure time Data., 2002) suggests 4 methods: exact method, Breslow approximation (Breslow, 1974), Efron approximation (Efron, 1977) and discrete method (Cox & Oakes, 1984).

The exact method assumes that ties result from imprecise measurement of time but there is an ordering of events. Suppose there are 15 facilities (ID's = 13, 16, 28, 32, 52, 54, 69, 72, 78, 79, 82, 83, 93, 96, 100) with the same survival time. Assuming there is an ordering results in 15! = $1.3*10^{12}$.

$$L = \sum_{i=1}^{15!} P(O_i)$$

$$P(O_{15!}) = \left(\frac{e^{\beta x_{100}}}{e^{\beta x_1} + e^{\beta x_2} + \cdots + e^{\beta x_{100}}}\right) \cdot \left(\frac{e^{\beta x_{96}}}{e^{\beta x_1} + e^{\beta x_2} + \cdots + e^{\beta x_{99}}}\right) \cdot \left(\frac{e^{\beta x_{93}}}{e^{\beta x_1} + e^{\beta x_2} + \cdots + e^{\beta x_{95}} + e^{\beta x_{97}} + e^{\beta x_{98}} + e^{\beta x_{99}}}\right) \ldots$$

As can be observed this is a huge computation and therefore should be approximated. Two popular estimations are Breslow and Efron.

The Efron likelihood is given by:

$$L_E(\beta) = \prod_{i=1}^{k} \frac{\exp\left(s'_{D_i}\beta\right)}{\prod_{j=1}^{d_i}\left[\sum_{l \epsilon R(t_{(i)})} \exp\left(x'_l\beta\right) - \frac{j-1}{d_i}\sum_{l \epsilon D_i} \exp(x'_l\beta)\right]} \qquad (35)$$

The Breslow likelihood is given by:

$$L_B(\beta) = \prod_{i=1}^{k} \frac{\exp\left(s'_{D_i}\beta\right)}{\left[\sum_{l \epsilon R(t_{(i)})} \exp(x'_l\beta)\right]^{d_i}} \qquad (36)$$

Where

- d_i is the number of failures at t_i
- $R(t_{(i)}; d_i)$ is the set of all subset of d_i individuals taken from the risk set $R(t_{(i)})$
- $R \in R(t_{(i)}; d_i)$ is the set of d_i individuals who might have failed at $t_{(i)}$
- $S_R = \sum_{l \epsilon R} x_l$ is the sum of the covariate vectors x over the individual in set R.
- D_i the set of d_i individuals failing at t_i
- $S_{D_i} = \sum_{l \epsilon D} xl$ sum of covariate vectors of these individuals

C 3 Hazard ratio

After the estimations of the coefficients using partial likelihood the predictors can interpreted using the hazard ratio (HR). The hazard rate is defined as the ratio of the hazard rates corresponding to the conditions described by two levels explanatory variables. As an example, in case of loan study, the population who is divorced is twice as likely to default as the non-divorced people. In this case the hazard ratio would be two. Or in formula form for binary predictors:

$$HR_{default/divorce} = \frac{h_0(t)e^{\beta_{divorce}(1)+\beta_{age}(50)}}{h_0(t)e^{\beta_{divorce}(0)+\beta_{age}(50)}} = e^{\beta_{divorce}(1-0)} = e^{\beta_{divorce}}$$

The same is true for continuous predictors. Illustrated below is the hazard ratio for a 1-year increase in age, adjusted for divorce.

$$HR_{default/age} = \frac{h_0(t)e^{\beta_{divorce}(0)+\beta_{age}(60)}}{h_0(t)e^{\beta_{divorce}(0)+\beta_{age}(50)}} = e^{\beta_{age}(60-50)} = e^{\beta_{age}(10)}$$

Suppose that the age of an applicant for credit is an predictor for the hazard rate. Using the Cox regression the estimated coefficient is $\beta = 0.092$. The hazard rate is $HR = e^{0.092} = 1.096$. This states that for every year older in age, results in an 9.6% increase in default rate.

C 4 Baseline estimations

The Cox PH model is a semi-parametric model and therefore doesn't specify a model for the baseline. In order to make estimation about the absolute time to default and not just the ordering of the facilities, an estimation about the baseline can be calculated. For estimation of the baseline hazard function the Kaplan-Meier method is used. The estimate for the baseline hazard function at time $t_{(i)}$ of the ith event is given by:

$$\widehat{h_0}(t_{(i)}) = \frac{d_{(i)}}{\sum_{j \epsilon R(t_{(i)})} \exp(\hat{\beta}^T x_j)} \tag{37}$$

Where $d_{(i)}$ is the number of defaults at time $i$and $R(t_{(i)})$ is the set of facilities at risk at that time.

The estimation of the baseline survival function is straightforward since the survival and hazard function are linked: $S(t) = \exp\left(-\int_0^t h(t)dt\right)$. Because the baseline hazard function is a discrete function the integral changes to: $\exp\left(-\int_{t_{(i-1)}}^{t_i} \hat{h}_0(t_{(i)})dt\right)$. The final estimate of the baseline survival function given by:

$$\hat{S}_0(t_{(i)}) = exp\left[-\sum_{j \leq i} \hat{h}_0(t_{(j)})\right] \tag{38}$$

Appendix D: Wald test statistic and Akaike information criterion (AIC)

D 1 Wald test statistic

The Wald test statistic is used to test the significance of a coefficient. If the data can be described using a statistical model with parameter that are estimated of a sample, the Wald test is used to test the true value of the parameter based on the sample estimate. Basically it test if there is a relation between the factor and the dataset. So there is no relationship ($\beta = 0$) or there is a relationship ($\beta \neq 0$). This test statistic is the ratio between the estimated coefficient and its estimated standard error.

The hypothesis of this test statistic are given as:
H_0: $\beta = 0$
H_1: $\beta \neq 0$

Next the observed information is calculated as follows:

$$I(\beta) = \frac{\partial^2 L_p(\beta)}{\partial \beta^2} = -\sum_{i=1}^{m} \sum_{t_j \geq t_{(i)}} w_{ij}\left(x_j - \overline{x_{w_i}}\right)^2 \tag{39}$$

$$w_{ij}(\beta) = \frac{e^{x_j \beta}}{\sum_{t_i \geq t(i)} e^{x_j \beta}} \tag{40}$$

$$\overline{x_{w_i}} = \sum_{t_j \geq t(i)} w_{ij}(\beta) x_j \tag{41}$$

Where

- $x_1, \ldots, x_m$ are the covariate values with survival times $t_{(1)} \leq \cdots \leq t_{(m)}$
- β is the coefficient value
- $L_p(\beta)$ is the log partial likelihood

The estimated standard error is given by:

$$\widehat{SE}\left(\hat{\beta}\right) = \sqrt{\widehat{Var}\left(\hat{\beta}\right)} = \sqrt{I\left(\hat{\beta}\right)^{-1}} \tag{42}$$

The Wald test statistic is given by:

$$z = \frac{\hat{\beta}}{\widehat{SE}\left(\hat{\beta}\right)} = N(0,1) \tag{43}$$

D 2 The Akaike information criterion (AIC)

The AIC is a statistical test used to identify the model that best fits the population from which the data were sampled. This measure creates a trade-off between the goodness of fit and complexity of the model. The complexity of the model is expressed in number of factors. Formula is given by:

$$AIC = 2k - 2\ln(L) \tag{44}$$

Where:

- k is the number of factors
- L is the log likelihood

The goodness of fit is explained by the log likelihood and the number of factors is the measure of complexity. The lowest AIC value is the preferred model. The use of this measure for selection of the model, prevents overfitting of the model.

Appendix E: Logrank test statistic

The statistical test used to compare survival functions of two buckets is called the logrank test statistic (Schoenfeld, 1981). This test compares the survival functions and tests if they are significantly different from each other or not. The basic idea is if a group is divided into 2/3 in group 1 and 1/3 into group 2, then on average 2/3 of the events observed on that interval should occur in group 1 and the other 1/3 in group 2. Unless one of the groups has significantly different survival probabilities.

The log rank test the hypothesis:

H_0: $S_0(t) = S_1(t)$
H_1: $S_0(t) \neq S_1(t)$

The logrank test is given by:

$$T(w) = \frac{\sum_x \left[dN_1(x) - \frac{dN(x) \cdot Y_1(x)}{Y(x)}\right]^2}{\sum_x \left[\frac{Y_1(x)Y_0(x)dN(x)[Y(x) - dN(x)]}{Y^2(x)[Y(x) - 1]}\right]} = \chi^2_{logrank} \qquad (45)$$

An extension of the logrank test is the weighted logrank test. With this test it is possible to give specific weights to certain observations. The weighted log rank test is given by:

$$T(w) = \frac{\sum_x w(x)\left[dN_1(x) - \frac{dN(x) \cdot Y_1(x)}{Y(x)}\right]}{\left\{\sum_x w^2(x)\left[\frac{Y_1(x)Y_0(x)dN(x)[Y(x) - dN(x)]}{Y^2(x)[Y(x) - 1]}\right]\right\}^{1/2}} \approx N(0,1) \qquad (46)$$

Where:

- $dN(x)$ is the number of deaths observed at time x for bucket 0 and 1
- $dN_1(x)$ is the number of deaths observed at time x for bucket 1
- $w(x)$ is the weight function
- $Y(x)$ is the number at risk at time x for bucket 0 and 1
- $Y_0(x)$ is the number at risk at time x for bucket 0
- $Y_1(x)$ is the number at risk at time x for bucket 1

Table 12: Logrank statistic for different buckets

	Bucket		
	0	1	Total
# of events at x	$dN_0(x)$	$dN_1(x)$	$dN(x)$
# of no events at x	$Y_0(x) - dN_0(x)$	$Y_1(x) - dN_1(x)$	$Y(x) - dN(x)$
# at risk at x	$Y_0(x)$	$Y_1(x)$	$Y(x)$

The weight function $w(x)$ can be used to emphasize differences in the hazard rates over time according to their relative values. If the difference between buckets in early survival times is of importance, one could chose a weight function that emphasises on these early survival times.

Types of weighted logrank tests suggested by literature are:

1. $w(x) = 1$ for all x
2. $w(x) = Y(x)$; Gehan's generalization of Wilcoxon test (Gehan, 1965)
3. $w(x) = \prod_{u \leq x}\left[1 - \frac{dN(u)}{Y(U)}\right]$; Peto-Prentice's generalization of Wilcoxon test

The first weight function is the standard logrank test where every observation get an equal weight. The Gehan's and Peto-Prentice's weighted logrank test put more emphasis on early survival times. The Peto-Prentice's test is considered the most robust version, while the Gehan's test is the most commonly used. (Breslow, 1974; Tarone and Ware, 1977; Kalbfleisch and Prentice, 1980; Miller, 1981; Hosmer and Lemeshow 1999).