

UNIVERSITY OF TWENTE

FINANCIAL ENGINEERING AND MANAGEMENT

MASTER THESIS

Survival Analysis in LGL Modelling for Retail Mortgage Portfolios

Author:

A.M.M. ARENTS

Supervisors:

B. ROORDA

R.A.M.G. JOOSTEN

Examination date:

JULY 12, 2019

External supervisors:

P. MIRONCHYK

V.L. TCHISTIAKOV



PUBLIC VERSION

Abstract

Loss given default (LGD) is one of the key parameters banks need in order to estimate expected and unexpected losses. These losses are necessary for credit pricing and for the calculation of the regulatory requirements regarding Basel III. Loss given cure (LGC) and loss given liquidation (LGL) are the components used in the LGD model for retail mortgage portfolios of Rabobank. In particular, this study focuses on the modelling of the LGL component. LGL depends on the recovery cash flow data of defaulted loans. The main difficulty in modelling LGL is incorporating the incomplete recovery cash flow data.

This study makes use of the statistical technique of survival analysis (SA) in order to solve this main difficulty. It uses the modelling and validation choices from earlier studies that satisfy regulatory requirements. The two used SA methods, the Cox Proportional Hazards model and the Extended Cox model, examine the effects of risk drivers on the length of repayment of a monetary unit.

With the models, Rabobank can estimate LGL for newly defaulted loans. At the same time, Rabobank gets insight into what drives high LGL estimations and what drives low LGL estimations. The models scored high on discriminatory power and calibration. Therefore, our results show high potential of applying SA in the modelling of the LGL component, used in the LGD model for retail mortgage portfolios.

Keywords— LGD Modelling, Retail Mortgage Portfolios, Survival Analysis, Cox Proportional Hazards Model, Loss Given Default

Acknowledgements

This thesis has been written as part of the Master's degree 'Industrial Engineering and Management' with a specialization in 'Financial Engineering and Management' at the University of Twente. Most of the work has been done at the Risk Analytics department of the Rabobank in Utrecht.

I would like to thank my supervisors at the Rabobank, Pavel Mironchyk and Viktor Tchistiakov, for giving me this opportunity and helping me finishing my Master's degree. They always took the time to guide me through the process and helped me shape this study. I really enjoyed my time at their department and appreciate the openness and welcome of the team members. I also would like to thank my supervisors from the University of Twente, Berend Roorda & Reinoud Joosten for their supervision, feedback and new ideas on the subject. Lastly, I would like to thank my family and loved ones for their unconditional support.

Annelieke Arents

July 2019

Contents

1	Introduction	1
1.1	Background	1
1.1.1	Rabobank	1
1.1.2	Risk management framework	1
1.1.3	Credit risk model framework	1
1.1.4	LGD model	2
1.2	Research proposal	2
1.2.1	Problem statement	2
1.2.2	Research goal	3
1.2.3	Research questions	4
1.2.4	Methodology	4
2	Theory	5
2.1	Workout method	5
2.1.1	Loss rate	6
2.1.2	Costs	7
2.1.3	Discount factor	7
2.2	LGL Model	7
2.3	Unresolved cases	8
2.4	Survival analysis	8
3	Modelling methodology	9
3.1	Recovery Rates	9
3.2	Negative cash flows	10
3.3	Survival analysis	10
3.3.1	Cox Proportional Hazards (PH) model	11
3.3.2	Partial likelihood	12
3.4	Censoring	13
3.4.1	Censoring method - recovery rates	14
3.4.2	Censoring method - recovery rates with categorization	15
3.4.3	Censoring method - unresolved cases	15
3.5	Explanatory variables	15
3.6	Variable selection	16
4	Test methodology	18
4.1	Data splitting	18

4.2	Testing of performing models	18
4.3	Testing of non-performing models	18
4.3.1	Realized recoveries	19
5	Validation methodology	20
5.1	Discriminatory power	20
5.1.1	Loss Capture Ratio	20
5.1.2	Spearman's rank correlation coefficient	21
5.1.3	Kendall's rank correlation coefficient	22
5.1.4	Concordance index	22
5.2	Calibration	23
5.2.1	Loss Shortfall	23
6	Results	24
6.1	Lifetime data	24
6.2	Cox Proportional Hazards model	26
6.2.1	Stepwise variable selection	26
6.2.2	Selected variable estimations	27
6.3	Extended Cox model	28
6.3.1	Stepwise time-dependent variable selection	28
6.3.2	Selected time-dependent variable estimations	29
6.4	Model validation	29
7	Discussion	31
7.1	Limitations	32
7.2	Suggestions for further research	33
8	Conclusion	35
A	Survival Analysis	38
A.1	Basic concepts	38
A.1.1	Cumulative Distribution Function	38
A.1.2	Survival function	38
A.1.3	Hazard function	38
A.1.4	Cumulative hazard function	39
A.2	Parametric and non-parametric methods	39
A.3	Kaplan-Meier estimator	40
A.4	Cox Proportional Hazards (PH) model	40
A.4.1	Partial likelihood	41

A.4.2	Hazard ratio	41
A.4.3	Breslow or Efron approximations	42
A.5	Time-dependent variables	43
A.5.1	Hazard ratio for Extended Cox model	43
B	Model selection and parameter estimates	44
B.1	Cox Proportional Hazards model	44
B.2	Extended Cox model	47
B.3	Survival curves	48
C	Inclusion of realized recoveries	52

List of Figures

1	Structure of Loss Given Default Model.	2
2	Discounted cash flows.	6
3	LGL calculation.	8
4	Structure of recovery data.	9
5	Censored and uncensored data.	14
6	Test methodology.	19
7	The loss capture curve.	21
8	Lifetime table curves.	25
9	Estimated survival curves for different values of selected variables.	27
10	95% Confidence interval for estimated regression coefficients.	28
11	Loss rates.	30
B.1	Estimated survival curves for different values of selected variables - non-performing model segment 1.	48
B.2	Estimated survival curves for different values of selected variables - performing model segment 1.	49
B.3	Estimated survival curves for different values of selected variables - non-performing model segment 2.	50
B.4	Estimated survival curves for different values of selected variables - performing model segment 2.	51
C.1	Distribution of LR at time t_{random}	52

List of Tables

1	Lifetime table.	25
2	Stepwise variable selection.	26
3	Selected variable estimations.	27
4	Stepwise time-dependent variable selection.	29
5	Selected time-dependent variable estimations.	29
6	Performance measurements.	30
B.1	Stepwise variable selection - performing model segment 1.	44
B.2	Selected variable estimations - performing model segment 1.	44
B.3	Stepwise variable selection - non-performing model segment 2.	45
B.4	Selected variable estimations - non-performing model segment 2.	45
B.5	Stepwise variable selection - performing model segment 2.	46
B.6	Selected variable estimations - performing model segment 2.	46
B.7	Stepwise variable selection - non-performing time-dependent model segment 2.	47
B.8	Selected variable estimations - non-performing time-dependent model segment 2.	47
C.1	Performance measurements incorporating recoveries realized so far.	52

Executive summary

The goal of this study is to determine the potential of the application of survival time analysis as an alternative to the state of the art loss given default (LGD) model for retail mortgage portfolios of Rabobank. LGD is one of the key parameters banks need in order to estimate expected and unexpected losses. In general, there are two approaches to model LGD, a structural and a non-structural approach. Rabobank uses a structural approach to estimate LGD and for this purpose, it models two different scenarios for mortgages, a cure scenario, and a liquidation scenario. For both scenarios, it models the probability of the scenario being realized and the expected loss. This study focuses on the expected loss in the liquidation scenario, the loss given liquidation (LGL) component of the LGD model.

Rabobank uses the recovery processes of resolved defaulted cases in the model estimation and calibration for LGL. The regulator requires to incorporate incomplete recovery processes of unresolved defaulted cases as well. However, this is more difficult since only a part of the recovery process is known. Therefore, this study examines the application of survival time analysis in the model estimation and calibration for LGL.

Data preparation

Survival analysis is a branch of statistics that models the time until an event happens. This technique allows incorporating incomplete recovery processes as censored data. First, we need to transform the original retail mortgage dataset of Rabobank in one suitable for survival analysis methods. The original one contains information on the repayments of monetary units of defaulted loans, as well as information on possible risk drivers (explanatory variables). The main data preparations done in this study are the following.

- Transformation of cash flows
- Treatment of negative cash flows
- Censoring of events

Modelling choices

This study uses three different survival analysis methods. The first method examines the overall survival of the retail mortgage data whether the other two examine the effects of explanatory variables on the survival rate. The latter methods use a stepwise selection in order to choose the variables. The three survival methods used are the following.

- Lifetime methods
- Cox Proportional Hazards model

- Extended Cox model

Validation choices

This study uses two fundamental aspects to validate the models: discrimination and calibration. Discrimination refers to the rank order of different estimations whereas calibration refers to the accuracy of the estimations on average. The data are randomly split into an 80% training set and a 20% test set in order to train, test and validate the models. The performance metrics used are the following.

- Loss Capture Ratio (LCR) ([Li et al., 2009](#))
- Spearman's rank correlation coefficient ([Zwillinger & Kokoska, 2000](#))
- Kendall's rank correlation coefficient ([PSU, 2019](#))
- Concordance index (C-index) ([Harrel et al., 1982](#))
- Loss Shortfall (LS) ([Li et al., 2009](#))

Results and conclusion

Our results suggest that the Cox Proportional Hazards model scores better on discriminatory power than the Extended Cox model. However, based on calibration the Extended Cox model scores better than the Cox Proportional Hazards model. Overall, the models scored very high on discriminatory power and calibration in comparison with other credit portfolios of Rabobank.

Returning to the goal of this study, our results show high potential of applying survival analysis in the modelling of the LGL component, used in the LGD model for the retail mortgage portfolios. This results in an alternative to the state of the art LGD model, Rabobank uses, satisfying regulatory requirements. The outline of this study gives insight into the major data preparations and modelling choices used in order to apply survival analysis methods in the modelling of the LGL component.

1 Introduction

1.1 Background

Loss given default (LGD) is one of the key parameters banks need in order to estimate expected and unexpected credit losses. These losses are necessary for credit pricing and for the calculation of the regulatory requirements regarding Basel III. Financial authorities determine which models banks should use for these calculations. Currently, Rabobank uses an internal rating based (IRB) approach for LGD estimations. With this approach, banks can use their internal rating systems to determine credit risk ([BCBS, 2017](#)).

1.1.1 Rabobank

Rabobank Group is a cooperative international financial services provider. It offers services in different sectors with a focus on retail banking, wholesale banking, and food & agriculture internationally. The organization currently operates in 44 countries with currently 106 local banks in the Netherlands. Rabobank provides a full range of financial services to over 7.6 million Dutch individuals ([Rabobank, 2019](#)).

1.1.2 Risk management framework

Rabobank maintains a risk management framework to identify, assess, manage, monitor and report risks. Therefore, it develops models for various risk types. The models most widely used are the ones developed for credit, market and operational risk ([Rabobank, 2018](#)). The Group Credit Models (GCM) department is responsible for the design and maintenance of the credit risk models. This group consists of ten different teams. One of these is the Mortgages and Consumer Finance team which is responsible for the development of the LGD model constructed for the calculation of the required capital for the retail mortgage portfolio.

1.1.3 Credit risk model framework

Credit risk within Rabobank is defined as: “the risk of the bank facing an economic loss because the bank’s counterparties cannot fulfill their contractual obligations” ([Rabobank, 2018](#)). Credit risk models quantify the risk related to a credit contract. The credit risk model framework used within Rabobank consists of three different parts. First, the probability of default (PD) which estimates the probability that a client will be unable to meet its debt obligations in the next 12 months. Second, the exposure at default (EAD) which represents the expected exposure at the moment of default. Lastly, the loss given default (LGD) which estimates how much of the exposure at default the bank could expect to lose.

1.1.4 LGD model

In general, there are two approaches to model LGD, a structural and a non-structural approach. A non-structural approach estimates LGD directly by relating observed historical losses and recoveries with risk drivers of defaulted loans. A structural approach splits the model into several components and for each component, a separate model is developed. The final model combines the probabilities and outcomes of the components to estimate LGD.

Rabobank uses a structural approach to estimate LGD. It models two different scenarios for defaulted mortgages, a cure scenario, and a liquidation scenario, in case of other portfolios there could be more scenarios. Cure is the scenario where a defaulted facility (non-performing portfolio) returns to a performing portfolio through the repayment of arrears and the completion of a probation period. The liquidation scenario refers to the liquidation of the collateral, pledged savings and a collection of other cash-flows.

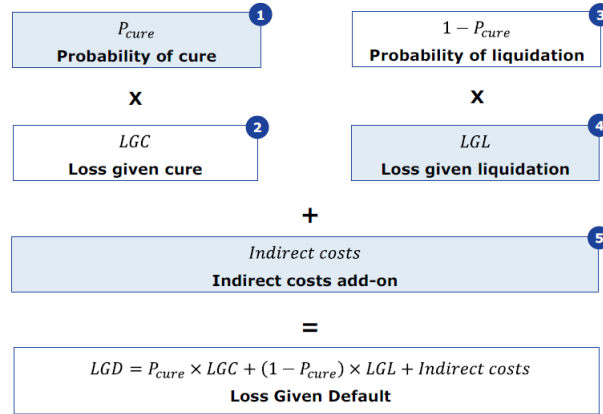


Figure 1: Structure of Loss Given Default Model.

Figure 1 shows the structure of the LGD model. The probability of cure (P_{cure}) refers to the probability of the cure scenario being realized. The loss given cure (LGC) refers to the expected loss in that scenario. The probability of liquidation ($1 - P_{cure}$) refers to the probability of the liquidation scenario being realized. The loss given liquidation (LGL) refers to the expected loss in that scenario. The indirect costs component is added on as an overall adjustment to the LGD and is not estimated separately for cure and liquidation scenarios.

1.2 Research proposal

1.2.1 Problem statement

Loss given cure (LGC) and loss given liquidation (LGL) are the components used in the LGD model for retail mortgage portfolios of Rabobank. The [EBA/GL/2017/08 \(2017\)](#) require that

banks compute these loss components from cash flow data. This data consist of all (potential) post-default cash flows. These cash flows should be discounted to the point of default because of the time lag between default and recovery.

One difficulty with LGL is estimating the potential post-default cash flows. This must be done for performing and non-performing portfolios. Performing portfolios might go into default some time in the future and therefore, estimations on potential post-default cash flows need to be made. Non-performing portfolios are currently in default but are not yet resolved (an unresolved case). This means that not all defaults in the model development dataset are currently solved in either a cure or a loss. Therefore, in order to estimate economic loss for unresolved cases, estimations on potential future cash flows need to be made as well.

The recovery process of resolved cases is used in the model estimation and calibration for LGL. Although the behavior and distribution of unresolved cases can be different from the set of resolved cases, e.g., due to business cycles, the regulator requires to incorporate these unresolved cases (incomplete recovery processes) as well ([EBA/GL/2017/16](#), [2017a](#)). This is because excluding these unresolved cases from the dataset can cause a significant bias and estimation error. Currently, these unresolved cases are included in the model with adjustments to the scale parameters.

For credit pricing and for the calculation of the regulatory requirements regarding Basel III it is important to have good and accurate estimations on potential post-default cash flows for both performing and non-performing portfolios. The main difficulty in modelling LGL is incorporating the incomplete recovery process of unresolved cases. Therefore, the main problem statement is defined as follows: *Incorporating unresolved cases in the model development and calibration of LGL can be difficult.*

Earlier studies showed the potential of the statistical technique of survival time analysis to incorporate unresolved cases in the modelling of LGD. An advantage of this technique is that it allows utilizing censored default data in order to make LGD estimations ([Witzany et al., 2010](#), [Privara et al., 2013](#), [Zhang & Thomas, 2012](#)).

1.2.2 Research goal

The goal of this study is to determine the potential of applying survival time analysis techniques in the LGD model estimation and calibration for retail mortgage portfolios. Survival analysis (SA) is a branch of statistics that models the time until an event happens. The event can be for example a recovery cash flow of a monetary unit, liquidation or cure. The time until such an event happens is called survival time. The subjects in SA are the persons exposed to the risk of the event. One strength of SA is that it can handle the censoring of observations.

Censored observations concern subjects that have survived until a point in time but no further information is available, which applies for unresolved cases. For the LGL component, the total amount of EAD can be seen as the subjects in the study where the repayment of each monetary unit of a defaulted loan can be seen as the event of interest and the survival time can then be seen as the time to the repayment of each monetary unit.

1.2.3 Research questions

In order to reach the goal of this study, we formulate the following main research question and sub-research questions.

Main research question: *What is the potential of applying survival analysis methods for estimating LGL and which metrics should be used to evaluate the performance of these methods?*

Sub research questions:

- *What is the current LGL model development procedure?*
- *What is survival analysis and its major advantage with respect to the modelling of LGL?*
- *Which modelling choices should be made when applying survival analysis in the LGL model?*
- *How to include time-dependent variables, including recoveries realized so far, into the modelling of LGL?*
- *Which metrics should be used to evaluate the performance of the LGL models?*

1.2.4 Methodology

In order to reach the goal of this study, first, we explain the structure of the currently used LGD model and workout method to obtain some understanding of the currently used methods and modelling choices. Afterwards we discuss which modelling choices should be made when applying survival analysis in the LGL model. Rabobank provides data on retail mortgage portfolios and we need to transform this dataset in one suitable for survival analysis methods. Next, we use survival regression in order to estimate LGL. Then, we use different tools and performance metrics to measure the performances of the models. Eventually, the results give some insights in the application of survival analysis as an alternative to the state of the art LGD model satisfying regulatory requirements.

2 Theory

As discussed earlier, loss given default (LGD) is the full economic loss in case of a default. Rabobank expresses this as a percentage of the exposure at default (EAD). This means that LGD is the difference between the EAD and the sum of all (potential) post-default cash flows. The cash flows should be discounted to the point of default because of the time lag between default and recovery. The bank estimates LGD at facility level, which is a credit obligation or a set of credit obligations.

LGD estimations need to be made for performing and non-performing portfolios. Performing portfolios might go into default some time in the future while non-performing ones are currently in default and may enter cure or liquidation phase. LGD estimations for performing portfolios need to be appropriate for an economic downturn while for non-performing ones it should reflect the sum of expected loss under current economic circumstances and possible unexpected loss that might occur within the recovery period.

Rabobank uses a structural approach to model LGD. A structural approach estimates LGD by calculating and combining estimates of all possible resolution scenarios (e.g., cure, restructuring, liquidation). Normally, a structural LGD model can have several components. Within these components, Rabobank estimates the probability of a certain scenario and the expected loss in that scenario. *Figure 1* showed the structure of the LGD model for the retail mortgage portfolio. It models two different scenarios, a cure scenario, and a liquidation scenario.

A facility is in default when at least one of the default triggers has occurred. This definition of default is in line with the CRR requirements and the recent EBA Guidelines on Definition of Default ([EBA/GL/2017/16](#), 2017b). The facility in default is cured when arrears go to zero, costs are fully repaid to the bank and customer passes the probation period. The facility in default is regarded liquidated when the house is sold, a facility is then resolved when the collateral has been sold and all the retail mortgage loans have been terminated. Up until the moment the house is sold it is not possible to differentiate between cure or liquidation process, there is no legal condition to state that a customer has entered irreversible liquidation phases.

In particular, this study focuses on the loss given liquidation (LGL) component and how to estimate this component using cash flow data. Therefore, in the remainder of this chapter, we explain the current LGL model development procedure.

2.1 Workout method

The loss rate (LR) is the observed loss expressed as a percentage of the EAD, it basically describes the LGL as a percentage of the EAD. This means that the LR is the difference

between the EAD and the sum of all (potential) post-default cash flows. For every default, cash flows can occur within a certain workout period. This period lasts from the moment a portfolio goes into default until the default is resolved. The cash flows obtained during the workout period need to be discounted back to the Default Event Date (DED). The complement of the LR is the recovery rate (RR) and it is used to define the observed recovery expressed as a percentage of the EAD.

2.1.1 Loss rate

To determine the LR, all cash flows within the workout period should be taken into account. An overview of the different types of cash flows will be given later. The cash flows should be discounted to the point of default, the Default Event Date (DED), to define the economic loss and obtain the LR. *Figure 2* illustrates this process, here $PV(CF_1)$ represents the present value at the DED of cash flow CF_1 , and $PV(CF_2)$ represents the present value at the DED of cash flow CF_2 . The loss rate for every facility is defined as follows

$$LR = 1 - \frac{1}{EAD} \sum_{t=1}^T \frac{CF_t}{(1+r)^{t-t_0}} \quad (1)$$

where CF_t is the cash flow at time t , r is the discount rate of the facility, $t - t_0$ is the time between default date t_0 and cash flow date t and T is the total number of cash flows. Incoming cash flows will decrease the eventual loss of a default, while outgoing cash flows increase the amount that needs to be repaid by the client.

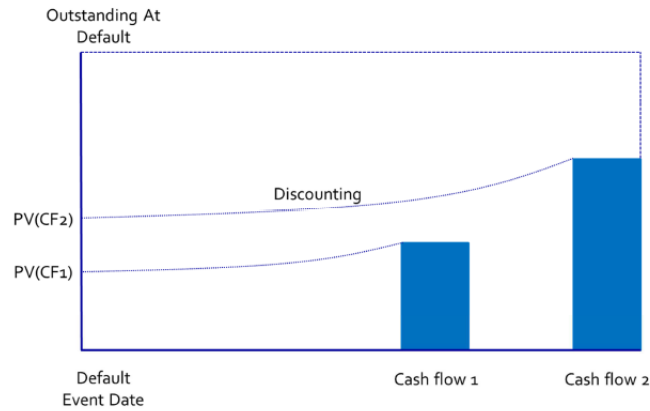


Figure 2: Discounted cash flows.

2.1.2 Costs

Rabobank should also include costs in the LR estimations. Costs can be split into direct costs and indirect costs. Direct costs are the fees and expenses paid to external parties while indirect costs are made within the bank during the workout period. The bank includes direct costs directly in the LR estimations and it includes indirect costs as an add on top of the LGD.

2.1.3 Discount factor

The regulator defines which discount factor should be used for calculating the present value of the cash flows. Currently, this is the twelve-months Euribor +500 bps ([EBA/GL/2017/16, 2017a](#)).

2.2 LGL Model

Rabobank uses all possible types of cash flows available for retail mortgage portfolios in the workout method to determine the LR and RR. Historical data are not available on a more granular level than the following categories:

- *Collateral (i.e., property) sale* - This is normally based on the most recent collateral valuation in combination with the expected market value changes.
- *Pledged savings* - The bank has a legal claim on the value of the savings in the account of the defaulted client.
- *NHG claim* - If a client has an NHG claim then the bank can make a claim on the NHG insurance.
- *Insurance* - These additional recoveries may come from other insurance policies.
- *Other recoveries* - Related to recoveries other than residential real estate objects and savings.
- *Regress recoveries* - These cash flows are associated with "unsecured" recoveries.
- *Direct costs* - These are the costs directly associated with the recovery process.
- *Drawdowns* - This refers to the revolving mortgage exposure increase during the closure process due to additional drawings.

As with the just described workout method, these cash flows should be discounted to the point of default to produce an economic LGL. *Figure 3* visualizes this process.

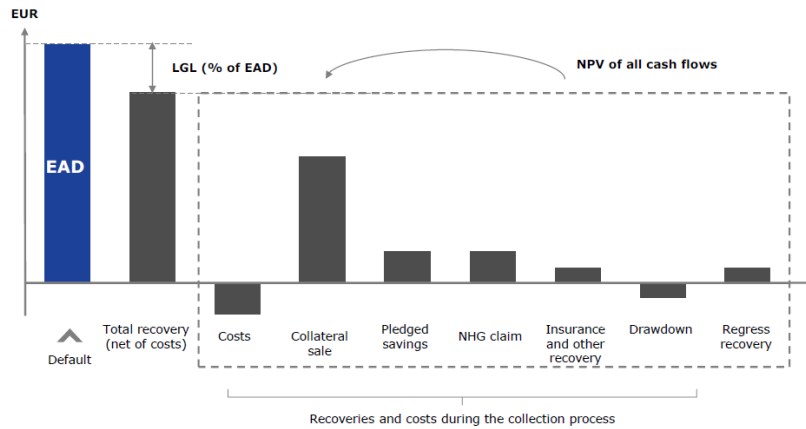


Figure 3: LGL calculation.

2.3 Unresolved cases

One problem in determining the LR, and RR, is that not all cases are resolved, but still these unresolved cases should be taken into account. Therefore, unresolved cases in the historical default dataset need specific treatment, otherwise they will bias the LR if they were left untreated. Currently, these unresolved cases are included in the dataset in the same way as the other model parameters but with some adjustments to the scale parameters. However, earlier studies showed the potential of the statistical technique of survival time analysis to incorporate these unresolved cases (incomplete recovery processes) in the modelling of LGD (Witzany et al., 2010, Privara et al., 2013, Zhang & Thomas, 2012).

2.4 Survival analysis

Survival analysis is a branch of statistics that models the time until an event occurs. In survival analysis, the time until an event occurs is called the survival time, because it gives the time that a subject has 'survived' over a certain time period. An event in survival analysis is called a failure, because the event of interest is usually the death of the subject or another negative event. Survival analysis can be used in different fields, as for example in reliability analysis in engineering, duration analysis in economics or event history analysis in sociology. Witzany et al. (2010), Privara et al. (2013), Zhang & Thomas (2012) proposed to model the recovery process of a defaulted loan as a survival process. The total amount of EAD can then be seen as the subjects in the study where the repayment of each monetary unit of a defaulted loan can be seen as the event of interest. The survival time can then be seen as the time to the repayment of each monetary unit. Since the event of interest is a repayment of a monetary unit the failure is, in this case, a positive event. In the next chapter, we explain survival analysis and the modelling choices made in more detail.

3 Modelling methodology

In this chapter, we introduce the concepts and statistical notations that are essential in order to apply survival analysis to the modelling of LGL. First, we provide an introduction in recovery rates and an explanation of the structure of the recovery data.

3.1 Recovery Rates

The bank makes a distinction between realized and expected recovery rates (RR) together with the complementary loss rate (LR). It calculates the realized RR from historical data with the workout method, as discussed in the previous chapter, while it estimates the expected RR for performing and unresolved non-performing cases. The workout method looks at the net recovery cash flows. These cash flows should be discounted to the moment of default with a discount rate r of 12-months Euribor + 500 bps (EBA/GL/2017/16, 2017a). As discussed earlier, LR and RR are expressed as a percentage of the exposure at default (EAD), where $LR = 1 - RR$. The recovery rate for every facility is defined as follows

$$RR = \frac{1}{EAD} \sum_{t=1}^T \frac{CF_t}{(1+r)^{t-t_0}} \quad (2)$$

where CF_t is the cash flow at time t , r is the discount rate of the facility, $t - t_0$ is the time between default date t_0 and cash flow date t and T is the total number of cash flows. We assume that the recovery rate can never be smaller than 0 or larger than 1 and therefore $LR = 1 - RR$ should lie in the interval $[0, 1]$.

Figure 4 illustrates the typical situation of recovery cash flow data (Privara et al., 2013). The vertical axis represents the specific month a loan went into default while the horizontal axis represents the months after default. Typically, there is a maximum number of months, t_{max} , for which we can observe recoveries. Here t_{start} and t_{end} denote the first observed default and last observed default, while t_a denotes an observed default in the middle of the study period.

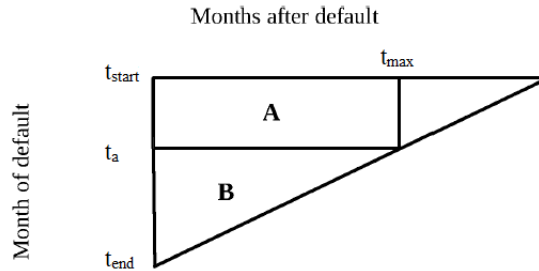


Figure 4: Structure of recovery data.

Area A in *Figure 4* covers all cases where full information about the recovery process of defaulted loans is available while area B covers all cases where only partial information is available, which we refer to as unresolved cases. Excluding cases from area B in the estimation of LGL can cause a significant bias and estimation errors. In addition, the regulator requires to include these unresolved cases ([EBA/GL/2017/16, 2017a](#)). This structure of recovery data is pivotal to understanding when making a decision about the modelling techniques for LGL. This is also the reason we introduce survival analysis as a modelling technique for LGL. One of the advantages of survival analysis is that it can include the partially available information from area B.

3.2 Negative cash flows

The RR should lie in the interval $[0,1]$ because of one of the properties of the survival function, $0 \leq S(t) \leq 1$. Therefore, negative cash flows must be adjusted. In our case, costs for example are negative cash flows. Earlier studies showed different possibilities of adjusting these cash flows. According to [Privara et al. \(2013\)](#), we should use the following formula for all cash flows

$$CF_t = \max(0, \text{repayment}_t - \text{costs}_t) \quad (3)$$

where CF_t , repayment_t and costs_t are the cash flow, repayment and costs at time t . [Witzany et al. \(2010\)](#) suggest to omit negative cash flows and adjust the exposure at default to the cumulative RR in case it exceeds the original EAD. The regulator does not allow to change the EAD after the moment of default ([EBA/GL/2017/16, 2017a](#)). However, all recoveries, including those related to fees capitalized after default, should be included in the calculation of economic loss. Therefore, we adjust negative cash flows by including them into the denominator of the LR, while discounted to the moment of default. The regulator requires to discount the negative cash flows in the same way as the positive cash flows, with Euribor 12-months interest rate + 500 bps ([EBA/GL/2017/16, 2017a](#)).

3.3 Survival analysis

The basic principle of survival analysis is to model the time until an event happens. In our case, the event of interest is a repayment of a monetary unit of a defaulted loan. The following concepts can be found in any statistics book about survival analysis. For a review of survival analysis see [Kleinbaum & Klein \(2005\)](#), [Klein & Moeschberger \(2003\)](#).

The basic principle is as follows. Let T be a non-negative random variable as an instant of exit of a subject, in our case a repayment of a monetary unit of a defaulted loan. The probability density function $f(t)$ and the cumulative distribution function $F(t)$ express the statistical properties of that random variable. The survival function is the complement of the

cumulative distribution function, $S(t) = 1 - F(t)$, and it represents the cumulative probability of the subject still being alive until time t . In our case, this means that the subject in the study, the exposure of the defaulted loan, is not (fully) repaid by the end of the study period. The survival function is therefore a representation of the LR. The hazard rate is defined as follows

$$h(t) = \frac{f(t)}{S(t)} \quad (4)$$

and it represents the rate at which the subjects are exiting exactly at t given survival until t . In our case, this represents the rate at which repayments of monetary units are happening exactly at t . The cumulative of the hazard rate is defined as follows

$$H(t) = \int_0^t h(u)du. \quad (5)$$

The complement of the cumulative hazard function is called the survival function and is derived as follows

$$S(t) = \exp[-H(t)] = \exp\left[-\int_0^t h(u)du\right]. \quad (6)$$

These are the most general concepts of survival analysis methods. The most popular implementation is the Cox Proportional Hazards model.

3.3.1 Cox Proportional Hazards (PH) model

The Cox Proportional Hazards model can be used to model the effects of explanatory variables, in our case risk factors, on the survival function and enables larger flexibility than other types of models. The model gives an expression for the hazard at time t with a given set of explanatory variables, denoted by X . The hazard function for the Cox Proportional Hazards model is defined as follows

$$h(t, X) = h_0(t) \exp\left[\sum_{i=1}^p X_i \beta_i\right] \quad (7)$$

where $h_0(t)$ is the baseline hazard function, $X = (X_1, X_2, \dots, X_p)$ is a vector of the explanatory variables and $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ is the vector we try to estimate. There is no 'error' term in this model as the randomness is implicit to the survival process. The baseline hazard function is assumed to be non-negative, constant over time and independent on the explanatory variables. The baseline hazard function is the hazard function obtained when all explanatory variables are set to zero. In our case, it represents the rate at which repayments of monetary units are happening, without the effects of risk drivers. At the same time, the exponential expression shown here is only dependent on the explanatory variables and does not involve t . The survival function for the Cox Proportional Hazards model is derived from the hazard function. This

function represents the survival at time t for a subject with X explanatory variables. The survival function for the Cox Proportional Hazards model is defined as follows

$$S(t, X) = S_0(t)^{\exp[\sum_{i=1}^p X_i \beta_i]} \quad (8)$$

where

$$S_0(t) = \exp \left[- \int_0^t h_0(u) du \right]. \quad (9)$$

Here $S_0(t)$ is the baseline survival function and also dependent on time only. However, it might be interesting to include time-dependent explanatory variables. Therefore, the Cox Proportional Hazards model can be extended. This means that the explanatory variables for a given subject can change over time. The Extended Cox model that includes both time-dependent and time-independent explanatory variables is defined as follows

$$h(t, X(t)) = h_0(t) \exp \left[\sum_{i=1}^{p_1} X_i \beta_i + \sum_{j=1}^{p_2} \delta_j X_j(t) \right] \quad (10)$$

where $X = (X_1, X_2, \dots, X_{p_1})$ is a vector of the time-independent explanatory variables and $X = (X_1, X_2, \dots, X_{p_2})$ is a vector of the time-dependent explanatory variables. The vector $\delta = (\delta_1, \delta_2, \dots, \delta_{p_2})$ represents the overall effect on the survival time of the time-dependent variables and is not time-dependent. An assumption of this model is that the variable $X_j(t)$ can change over time but the hazard model provides only one value of the coefficients for each time-dependent variable in the model.

3.3.2 Partial likelihood

The estimation of β is done by maximum likelihood methods. The estimates are derived by maximizing a likelihood function, L . The likelihood function is an expression which describes the joint probability of obtaining the data actually observed on the subjects in the study as a function of the unknown parameters in the model being considered. The estimation of β in survival analysis is done by a partial likelihood. This is because the full likelihood function only considers the probabilities for subjects that fail, and does not take the probabilities for censored subjects into account. The advantage of the partial likelihood function in comparison to the full likelihood function is that it can maximize the proportion depending on β alone.

Normally, the partial likelihood is written as the product of several likelihoods, one for each K failure times. At the k th failure time, L_k denotes the likelihood of failing at this time, given survival up to this time. The likelihood function, L , is defined as follows

$$L = L_1 \times L_2 \times L_3 \times \dots \times L_K = \prod_{k=1}^K L_k \quad (11)$$

where L_k is the portion of L for the k th failure time. The partial likelihood function focuses only on subjects who fail but the survival time information prior to censorship is used for those subjects who are censored. This means that a person who is censored after the k th failure time is still part of the risk set used to compute L_k . The censored subjects are in this way included in the analysis.

In order to estimate β the likelihood function needs to be maximized. This is done by maximizing the partial derivatives of the natural log of L with respect to each parameter in the model. In order to obtain this maximization of the natural log of L , the following equation should be solved

$$\frac{\partial \ln L}{\partial \beta_i} = 0 \quad (12)$$

where $i = 1, \dots, p$. *Appendix A* provides a more detailed and mathematical explanation of survival analysis and likelihood methods.

3.4 Censoring

One of the main advantages of survival analysis is that it can handle censored data. Censored data exist when information about the time until an event happens is only known for a certain period of time, as is the case for unresolved defaults. There are different possible censoring types: right censoring, left censoring and interval censoring. Right censoring deals with data where the subject is still alive by the end of the study. Left censoring deals with data where the subject has experienced the event of interest prior to the start of the study. Interval censoring deals with data where the subject has experienced the event in some specific time interval. We only use right censored data in the modelling of LGL. *Figure 5a* shows an example of right censored and uncensored data in ordinary survival analysis (Klein & Moeschberger, 2003).

In the example the specific study period lasts from week 0 until week 12 and the subjects are exposed to some events. Right censoring exists for subject B, C, D and E, since the subjects are either still alive by the end of the study period or an event other than the one of interest happened (in case of subject C and E). The event actually happened with subject A and F.

In case of LGL, censoring exists when (parts of) the loan still needs to be repaid. In our in-default dataset we have resolved and unresolved cases, as discussed earlier an in-default case is considered resolved when the collateral is sold. Censoring can happen in both cases. For simplicity, see the example in *Figure 5b*. In the example, there are four different loans where three are considered resolved (R1, R3 and R4) and one unresolved (U2).

For the resolved cases, loan 1 had a repayment of 80% and loan 4 had a repayment of 100%. Loan 1 still had an amount of 20% and loan 3 still had an amount of 100% which was unpaid

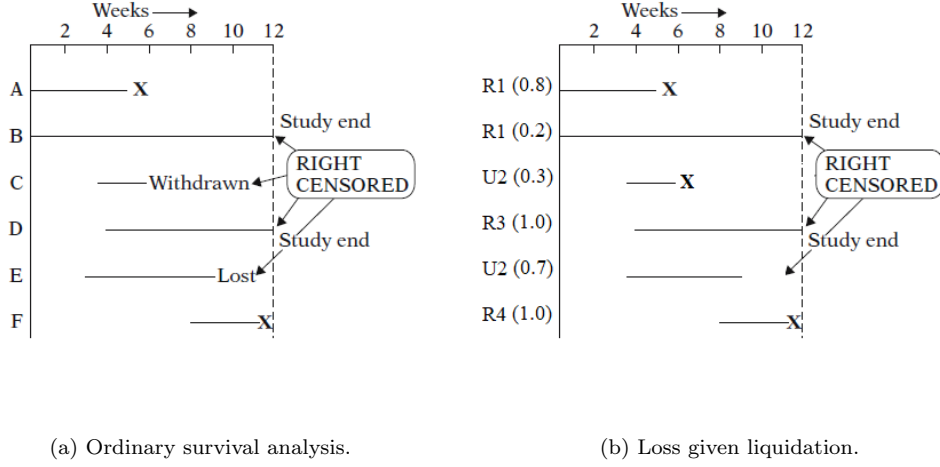


Figure 5: Censored and uncensored data.

by the end of the study period, these parts are considered censored. When looking at the unresolved case, there was one repayment on the loan of 30% but still, 70% of the loan was left unpaid, with no more information available. The part that is still not repaid, in this case, the 70%, is considered censored. Thus, all parts that are unpaid by the end of the study period are considered censored.

We need to transform the original dataset in one suitable for applying survival analysis methods. Therefore, for every payment within the study period, $t \leq t_{max}$, an observation with the weight corresponding to the ratio to exposure at default should be included in the new dataset. Let a be a default case where for every observation of a cash flow $CF_t(a)$ at time t a frequency weight of $d = CF_t(a)$ should be put on the observation corresponding to the ratio to exposure at default. This observation is labeled as not censored. A payment case can however be censored. Therefore, we discuss the different possible censoring methods in the remainder of this section.

3.4.1 Censoring method - recovery rates

According to [Witzany et al. \(2010\)](#), [Privara et al. \(2013\)](#) a payment can be censored for one of the following two reasons.

1. For a complete recovery process (resolved case) where the sum of payments is less than the debt amount, the remainder part of the loan is considered the censored part. So for a complete recovery process where $\sum_{t=1}^{t_{end}} CF_t(a) < EAD(a)$ we include an observation $d = EAD(a) - \sum_{t=1}^{t_{end}} CF_t(a)$ which is censored at t_{max} . Basically, this means that resolved cases will not receive any more payments in any foreseeable future. Therefore, the observation is censored at t_{max} .

2. In case of an incomplete recovery process (unresolved case) where the workout period lasts longer than the study period t_{max} and the payments collected until that moment are less than the EAD, the remainder part of the loan is considered the censored part. So for an incomplete recovery process where $\sum_{t=1}^{t_{end}} CF_t(a) < EAD(a)$ we include an observation $d = EAD(a) - \sum_{t=1}^{t_{end}} CF_t(a)$ which is censored at time t_{end} . This means that the amount of d has not been recovered until the last observation period for each unresolved case. Therefore, the observation is censored at the last known date, t_{end} .

3.4.2 Censoring method - recovery rates with categorization

Another idea by [Privara et al. \(2013\)](#) provides censoring of the data regarding a high-low categorization. Within this method, the loans are seen as single entities. Repayments of monetary units can therefore not be treated separately. The key principle regarding the high-low Cox regression is setting a threshold recovery rate RR_t . For each account where $RR < RR_t$, the recovery process is not finished and it is marked in the dataset as censored. The exit time is set equal to the last study period t_{end} . For all cases where $RR > RR_t$, the recovery process is finished and it is marked in the dataset as not censored. The exit time is then set at the month where the minimum recovery rate RR_t was achieved.

3.4.3 Censoring method - unresolved cases

[Zhang & Thomas \(2012\)](#) address another censoring method. They classify defaults as finished (i.e. written off) or unfinished, which can be compared to resolved and unresolved cases. Observations that are resolved are included in the dataset as uncensored while unresolved are included as censored. The exit time is set equal to the realized RR (so far). Within this method, the concept of time is treated differently than in the above-mentioned methods.

The study of [Privara et al. \(2013\)](#) compared the three proposed methods. Results showed that the first method performed better than the latter two methods. Therefore, we use the first censoring method in this study.

3.5 Explanatory variables

For Rabobank, it is interesting to examine the predicting power of different risk drivers to the LGL estimates. The bank will then be able to act on the knowledge what drives high LGL estimations and what drives low LGL estimations. The vectors $X = (X_1, X_2, \dots, X_{p_1})$ and $X = (X_1, X_2, \dots, X_{p_2})$ in the Cox models indicate these risk drivers. In discussion with Rabobank, we examine the following risk drivers. These are also the ones used in the current cure and liquidation component of the LGD model.

- *Age* - Mean age of all customers connected to the facility.
- *Arrears* - Total arrears relative to the exposure, also considered a change in total arrears in last 3 and last 12 months.
- *Collateral value* - Market (estimated) value of collateral.
- *Confidence level* - Confidence level of the market value of the facility calculated by automated valuation method.
- *Cured before* - Indicates if facility went into default before.
- *Entrepreneur indicator* - Indicates if facility belongs to an "ondernemer in privé".
- *First default trigger* - Type of default trigger at the start of default.
- *Insurance* - Indicates if underlying loan parts are covered by "Nationale Hypotheek Garantie" (NHG).
- *Loan to value* - Percentage of the facility's total exposure amount that is covered by the collateral minus the preclaims.
- *Months in probation* - Number of months the facility is in probation period.
- *Pledged savings amount* - Total amount in a linked savings account, the bank has a legal claim on the value of savings in a linked savings account.
- *Pre-claim* - Total amount of claims from other banks/institutions that have seniority on the recovery from the collateral sale.
- *Product type* - Product type identifier.
- *Time in default* - Amount of months the facility is in default.
- *Valuation type* - Valuation method used - indexation, appraisal, automatic valuation method.

3.6 Variable selection

When developing the LGL model one has to make a decision about which of the above risk drivers to include in the model. There are two main methods for selecting the variables: sequential methods and all-subset methods (Sauerbrei et al., 2007). The latter method is based on fitting all possible models and choose the best model based on agreed criteria. The first method is based on a sequence of tests and looks at whether a variable should be added or removed from the current model. The adding or removing depends on the criteria for

inclusion and normally includes a significance level of $\alpha = 0.05$. The sequential methods are the most popular variable selection methods and also the method we use in this study. Sequential methods can use forward selection, backward selection or a combination of both.

The forward selection starts with no variables in the model and from here the most significant variable is added to the model. This process is repeated until no improvements in the model are exhibited. Backward selection, by contrast, starts with all variables and deleting the least significant one. The model is tested by using a chosen model fit criterion and repeated until no improvements in the model are exhibited. A combination of both (forward and backward) test at each step which variable to be included or excluded (Sauerbrei et al., 2007). We use a combination of both for the selection of risk drivers.

With a combination of both, all risk drivers were tested as explanatory variables to the dependent variable, time to the event, in our case the time to repayment of a monetary unit of a defaulted loan. All risk drivers that exceed the standard significance level of $\alpha = 0.05$ were added to the model where the most significant risk driver was added first. When no improvements in the model are exhibited, the risk driver was removed from the model selection.

In survival analysis, the p -value of the variables is determined by the log-rank test (Kleinbaum & Klein, 2005). The test is used for large samples to provide an overall comparison of the different survival curves. Within this test, the null hypothesis states that there is no overall difference between the survival curves. Under this null hypothesis, the log-rank statistic is approximately chi-square with one degree of freedom. Let $h_i(t)$ be the hazard ratio of group i at time t then

$$H_0 : h_1(t) = h_2(t),$$

$$H_a : h_1(t) = ch_2(t), \quad c \neq 1$$

Thus, the p -value of the variables is determined by from the tables of the chi-square distribution. In case the p -value of the risk driver exceeds the significance level of $\alpha = 0.05$ the risk driver is added in the model.

4 Test methodology

4.1 Data splitting

In order to test a model, one has to use a new dataset, other than the one used to develop the model. Preferably, this dataset should be external, from a group with similar characteristics. This data are often not available and therefore one can use internal data as well. A common approach is splitting the original dataset into two parts: a training set and a test set (Altman et al., 2009).

The training set can then be used for training the model on the data. Afterwards the test set is used to test and evaluate the performance of the model. A negative side of splitting the data is that the data for creating the model are reduced and the observations in the test set are not contributing to the actual model, but only to the testing of the model. The number of samples in the data determines the distribution of the train and test set.

In this study, the dataset originates from Rabobank. Since no other dataset is available, testing with external data is not possible. Therefore, we use internal data for testing the model. The data are randomly split into an 80% training set and a 20% test set.

As described in *Chapter 2*, LR estimations need to be made for performing and non-performing portfolios. Therefore, we discuss the test methodology for both portfolios in more detail.

4.2 Testing of performing models

The test methodology for the performing models is different than for the non-performing models. For performing models, estimations of the LR for time t_{max} should be made at time t_0 . The values of the included risk drivers are the ones observed at the last June before t_0 . *Figure 6a* shows a visualization of this method. The value of risk driver $X1$ varies over time until the last June before t_0 . After that moment the risk driver stays constant and the model includes that value in the LR estimation for time t_{max} made at time t_0 .

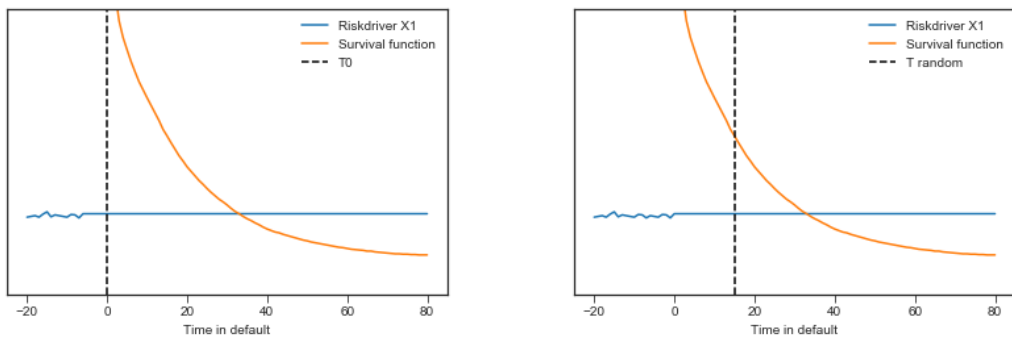
4.3 Testing of non-performing models

Testing of the non-performing models should be done as it is applied to real time data. Therefore, estimations of the LR for time t_{max} should be made at a time t_{random} instead of t_0 . The value of t_{random} is randomly chosen from a discrete uniform distribution between first moment in default, t_0 , and last moment of default, t_{end} , for every defaulted loan.

For the time-independent models, the values of the included risk drivers are the ones observed at time t_0 . This means that, when estimations of the LR for time t_{max} are made at time

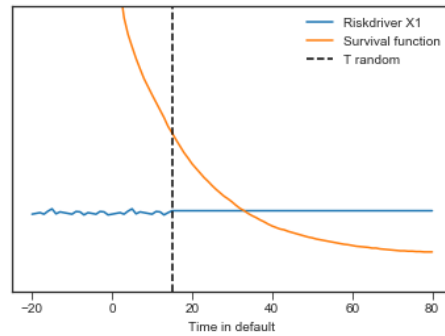
t_{random} , the values of the included risk drivers are the ones observed at time t_0 . *Figure 6b* shows a visualization of this method. The value of risk driver $X1$ varies over time until t_0 . After that moment the risk driver stays constant and the model includes that value in the LR estimation for time t_{max} made at time t_{random} .

For the time-dependent models, the values of the risk drivers change over time until the estimation moment, t_{random} . This means that, when estimations of the LR for time t_{max} are made at time t_{random} , the values of the included risk drivers are the ones observed at time t_{random} . *Figure 6c* shows a visualization of this method.



(a) Performing model.

(b) Non-performing time-independent model.



(c) Non-performing time-dependent model.

Figure 6: Test methodology.

4.3.1 Realized recoveries

Since estimations of the LR for time t_{max} are made at a time t_{random} , recoveries realized until t_{random} should be included in the estimations as well. This can be done by splitting the survival function in two parts and then combine both parts. The first part is the survival before t_{random} , which can be observed from cash flow data. The second part is the survival after t_{random} , which can be estimated with survival analysis methods.

5 Validation methodology

Once a model has been developed, one has to validate it. The validation process should be done periodically to monitor whether the model is still accurate. This is done by backtesting the model and testing its components. There are two fundamental aspects to validate a model: discrimination and calibration. Discrimination refers to the rank order of different estimations. It represents the degree of how well the model is able to differentiate risk. Calibration refers to the accuracy of the estimations on average. In our case, this means that the discriminatory power reflects the ability to assign a correct rank order, while calibration reflects the accuracy of the LGL estimations. In this chapter, we discuss the different tools and metrics used to measure the discriminatory power and calibration.

5.1 Discriminatory power

5.1.1 Loss Capture Ratio

One approach of measuring the discriminatory power of the LGL model is the loss capture (LC) ratio. Its usage is similar to the Gini coefficient. This approach uses three types of curves: the model (rating) LC curve, the ideal (perfect) LC curve, and the random LC curve. *Figure 7* shows an example of these curves (Li et al., 2009). The ideal curve represents the model that has perfect discriminatory power, in our case, this means that all losses are ranked properly. This curve can be constructed by sorting the highest realized LR to the lowest realized LR. It then plots for each fraction the cumulative amount of LR as a percentage relative to the total amount of LR. The rating curve is almost similar to the ideal curve except that the estimated LR instead of realized LR are used for ranking. The random curve is a diagonal line through the origin (0,0) and the point (1,1) and it represents the case where LR are randomly ranked.

From the curves, it is possible to determine the loss capture ratio. This ratio is derived from the area enclosed by the random curve and the model curve (which is area B) divided by the area enclosed by the random curve and the ideal curve (which is area A + B). The loss capture ratio is defined as follows

$$\text{LCR} = \frac{B}{A + B}. \quad (13)$$

The value of the LC ratio will lie between 0 and 1, where a value near 0 means that the model has limited discriminatory power and the closer the value is to 1, the more it represents the discriminatory power of the ideal model. Thus, the higher the LC ratio, the better discriminatory power the model provides

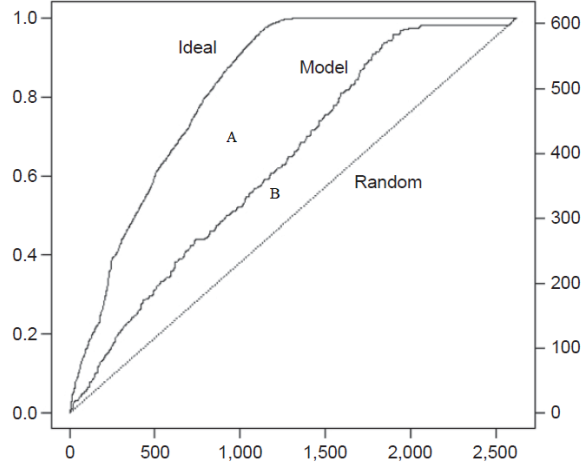


Figure 7: The loss capture curve.

5.1.2 Spearman's rank correlation coefficient

Another approach is the Spearman's rank correlation coefficient (or Spearman's ρ). It measures the monotonicity of the relation between two data sets. This means that it measures the correlation between the rank order, instead of the correlation between the actual observations. The Spearman's rank correlation is a non-parametric test. This means that it does not make any assumption on the underlying distributions of the variables (Zwillinger & Kokoska, 2000).

To understand the concept of Spearman's rank correlation test suppose there are N pairs of observations in two data sets, in our case N pairs of LR. The LR in each dataset should be ranked separately from smallest to largest. Let u_n be the rank of the n^{th} observation in the first dataset, in our case this is the dataset with realized LR. Let v_n then be the rank of the n^{th} observation in the second dataset, in our case this is the dataset with estimated LR. The difference between the respective ranking of a pair is then determined as follows $d_n = u_n - v_n$. This difference should be computed for all N pairs of LR. The Spearman's rank correlation coefficient is defined as follows

$$\rho = 1 - \frac{6 \sum_{n=1}^N d_n^2}{N(N^2 - 1)}. \quad (14)$$

Like other correlation coefficients, this value can vary between -1 and +1 with 0 implying no correlation. In case of a perfect ranking for the realized and the estimated LR, the index score is equal to 1. In case there is a perfect negative correlation this value will be equal to -1. Thus, the closer the Spearman's rank correlation coefficient is to 1, the better discriminator power it provides (Zwillinger & Kokoska, 2000).

5.1.3 Kendall's rank correlation coefficient

Another approach is the Kendall's rank correlation coefficient (or Kendall's τ). This statistic determines the number of concordant and discordant pairs of observations (PSU, 2019). A pair is concordant if it is in the same order and discordant if it is in the reverse order. Consider two data sets, the dataset with the estimated LR, LR^{EST} , and the dataset with the realized LR, LR^{REA} . Two pairs of observations (LR_m^{EST}, LR_m^{REA}) and (LR_r^{EST}, LR_r^{REA}) can take one of the following orders.

1. $LR_m^{EST} < LR_r^{EST}$ and $LR_m^{REA} < LR_r^{REA}$
2. $LR_m^{EST} > LR_r^{EST}$ and $LR_m^{REA} > LR_r^{REA}$
3. $LR_m^{EST} < LR_r^{EST}$ and $LR_m^{REA} > LR_r^{REA}$
4. $LR_m^{EST} > LR_r^{EST}$ and $LR_m^{REA} < LR_r^{REA}$

Two pairs are concordant if they are in the same order, which means that the pairs satisfy Statement 1 or 2. The pairs are discordant if they are in the reverse ordering, in this case the pairs satisfy Statement 3 or 4. It can also happen that two observatories are tied, which means $LR_m^{EST} = LR_r^{EST}$ and/or $LR_m^{REA} = LR_r^{REA}$. The total number of pairs (M) decomposes of five different quantities

$$M = P + Q + X_0 + Y_0 + (XY)_0. \quad (15)$$

Here, P represents the number of concordant pairs, Q represents the number of discordant pairs, X_0 represents the number of tied pairs for LR^{EST} only, Y_0 represents the number of tied pairs for LR^{REA} only and $(XY)_0$ represents the number of tied pairs for both LR^{EST} and LR^{REA} (PSU, 2019). Kendall's rank correlation coefficient is defined as follows

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + X_0)(P + Q + Y_0)}}. \quad (16)$$

Like Spearman's ρ , this value can vary between -1 and +1 with 0 implying no correlation. In case of a perfect ranking for the realized and the estimated LR, the index score is equal to 1. In case there is a perfect negative correlation this value will be equal to -1. Thus, the closer Kendall's tau correlation coefficient is to 1, the better discriminator power it provides.

5.1.4 Concordance index

In survival analysis, the most commonly used metric of discrimination is the concordance probability. It represents a pairwise probability and the usage is equivalent to Kendall's tau (Heller

& Mo, 2016). The concordance probability is a more desirable metric to use than the earlier described metrics, since it can take the actual timing of cash flows and censored observations into account.

The concordance probability for two subjects, in our case two repayments of monetary units, is $P(RR_1^{EST} > RR_2^{EST} | t_1 < t_2)$. Here, RR^{EST} represents the estimated recovery rate and t represents the timing repayment of a monetary unit. Since the RR is the complement of the LR the concordance probability can also be written as $P(LR_1^{EST} < LR_2^{EST} | t_1 < t_2)$. One of the early estimates of the concordance is the C-index (Harrel et al., 1982). The formula for the C-index, in case of LR estimations, is defined as follows

$$\text{C-index} = \frac{\sum_{s_1=1}^S \sum_{s_2=1}^S I(t_{s_1} < t_{s_2}) I(LR_{s_1}^{EST} < LR_{s_2}^{EST})}{\sum_{s_1=1}^S \sum_{s_2=1}^S I(t_{s_1} < t_{s_2})} \quad (17)$$

where S is the total number of cash flows, s_1 is the first subject, s_2 the second subject and I denotes the indicator function. Like the loss capture ratio the value of the C-index can vary between 0 and 1, where the higher the number of the C-index, the better discriminatory power it provides.

5.2 Calibration

Besides measuring the discriminatory power of the model, another way of validating the model is by examining the accuracy of the estimations obtained from the model. One way of doing this is by calculating the loss shortfall (LS).

5.2.1 Loss Shortfall

The loss shortfall (LS) looks at how well the model estimations match the realized values on exposure weighted average. This means that for all facilities in the portfolio the loss rate is multiplied by the exposure at default for that specific facility (Li et al., 2009). The loss shortfall is defined as follows

$$\text{LS} = 1 - \frac{\sum_{f=1}^F \text{EAD}_f \times \text{Expected LR}_f}{\sum_{f=1}^F \text{EAD}_f \times \text{Realized LR}_f} \quad (18)$$

where f represents the specific facility and F is the total number of facilities in the portfolio.

6 Results

The dataset used to build and test the LGL model contains information on two segments of the retail mortgage portfolio of Rabobank. It contains information on the repayments of monetary units of defaulted loans, as well as information on the possible risk drivers (explanatory variables). The information is available for 27,185 defaulted loans and for many of these loans the recovery process has not been completed (unresolved cases).

The workout method described in *Section 2.1* determines the sum of the repayments of monetary units. Based on this, Rabobank can obtain recovery rates (RR) and complementary loss rates (LR). Negative cash flows are treated as described in *Section 3.2*. From the resolved cases 99% was resolved within 80 months. Therefore the maximum workout period as described in *Section 3.4.1*, t_{max} , is set at 80 months.

In this chapter, we discuss the final model selection. First, we discuss the results from the lifetime table in order to get some more insight into the overall survival of the retail mortgage data. Then, we discuss the results of the Cox Proportional Hazards model and Extended Cox model. The final selection of the variables will be given together with their parameter estimates. Lastly, we evaluate the overall performance of the models with the validation methodology described in *Chapter 5*.

6.1 Lifetime data

We transformed the original data into a lifetime data table in order to apply Cox regression techniques. This transformation is done with the modelling methodology described in *Chapter 3*. The lifetime data table provides information, at each time interval, about the number of cash flows that were paid back and censored. Each cash flow has a specific weight relative to the exposure at default. Therefore, the total number of cash flows in each time interval is the sum of the weights of these cash flows.

With this information, we estimated the survival in each time interval with the Kaplan-Meier estimator. This estimator gives the probability of surviving past the previous failure time $t - 1$ multiplied by the conditional probability of surviving past time t , given that the subject has survived to at least time t . *Appendix A.3* explains the mathematical notation of the Kaplan-Meier estimator. In our case, the survival represents the LR at each time interval. We also calculated the probability of failure for each time interval. This probability of failure, also called hazard rate, represents the rate at which subjects are exiting exactly at t given survival until t . In our case, this represents the rate at which repayments of monetary units are happening at a certain time interval.

Time	Number at Risk	Number Event	Number Censored	Prob. of Failure $h(t)$	Survival $S(t)$
1	27184	754	9904	xxxx	xxxx
2	16525	3013	0	xxxx	xxxx
3	13521	1292	0	xxxx	xxxx
4	11467	548	0	xxxx	xxxx
..					
77	1794	8	0	xxxx	xxxx
78	1786	9	0	xxxx	xxxx
79	1777	10	0	xxxx	xxxx
80	1767	6	1761	xxxx	xxxx

Table 1: Lifetime table.

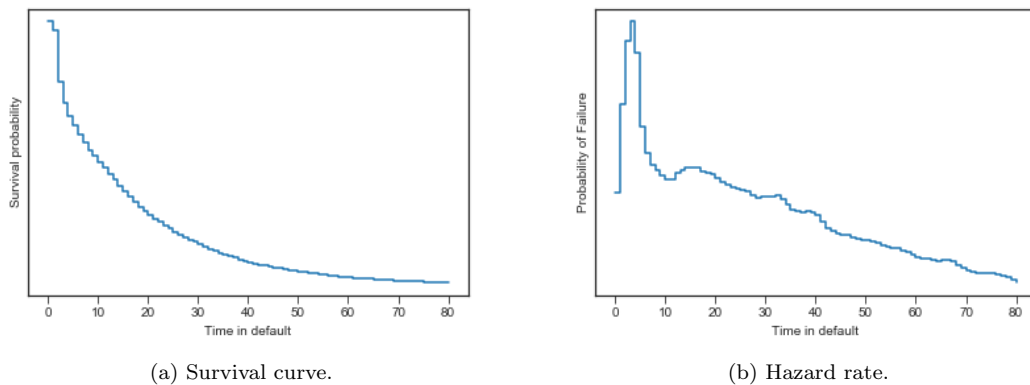


Figure 8: Lifetime table curves.

Table 1 shows the output of the lifetime data.¹ As can be seen, at each time interval the survival, and thus the LR, is estimated. The survival in the last time interval represents the final LR. Based on the survival in each time interval one can plot a survival curve. This curve represents the LR over time. Figure 8a shows a plot of this curve. Here the x-axis represents the time in default and the y-axis represents the survival probability.² The end value of this plot represents the final LR estimate. Figure 8b shows a plot of the hazard rate, which is the rate at which repayments of monetary units are happening exactly at t . It shows that most of the repayments are occurring in the early time intervals.²

¹Values of Prob. of failure and Survival removed due to confidentiality.

²Scales on y-axis removed due to confidentiality.

6.2 Cox Proportional Hazards model

With the Cox Proportional Hazards regression, we tested the effects of explanatory variables on the baseline survival curve. We developed four different Cox Proportional Hazards models based on the two segments of the retail mortgage portfolio of Rabobank. For both segments, we developed a performing model and a non-performing model. The performing model estimates the LR in case a facility goes into default some time in the future. The non-performing model estimates the final LR for facilities that are currently in default. In this section, we discuss the process of the final model selection of one of the models in more detail. It should be noted that we changed some of the variable names because we can not expose them to the public. *Appendix B* shows the final model selection and parameter estimates of the other three models.

6.2.1 Stepwise variable selection

We used a stepwise selection method in order to build the models. This means that one adds the most significant variable to the model first. This process repeats for all explanatory variables that exceed the log-rank test with a standard significance level of $\alpha = 0.05$. When no further improvements in the model exhibit one removes the explanatory variable from the model.

Step	Variable Added	Variable Removed	p -value
1	FirstDefaultTrigger		< 0.00001
2	WeightNHG		< 0.00001
3	LTV		< 0.00001
4	Variable X	Variable X	< 0.00001
5	Variable Y		< 0.00001
6	Variable D	Variable D	< 0.00001
7	Variable A		< 0.00001
8	Variable G	Variable G	< 0.00001
9	Variable C	Variable C	0.00047

Table 2: Stepwise variable selection.

Table 2 shows the results of the stepwise selection method. The first variable added in the model was variable *FirstDefaultTrigger* followed by variable *WeightNHG*, *LTV*, *Variable X*, *Variable Y*, *Variable D*, *Variable A*, *Variable G* and *Variable C*. The third column indicates the variables that were removed from the model because there were no further improvements in the model. The fourth column represents the significant level for each variable obtained from the single variable analysis.

6.2.2 Selected variable estimations

For each selected variable in the model, one can estimate a survival curve for different values of the selected variable. As an example, consider the survival curves for different levels of the variables *FirstDefaultTrigger* and *WeightNHG* in Figure 9.³ The final LR changes whenever a different value of the variable is taken and all the other variables in the model are kept constant. Appendix B shows an overview of the survival curves for different values of the selected variables.

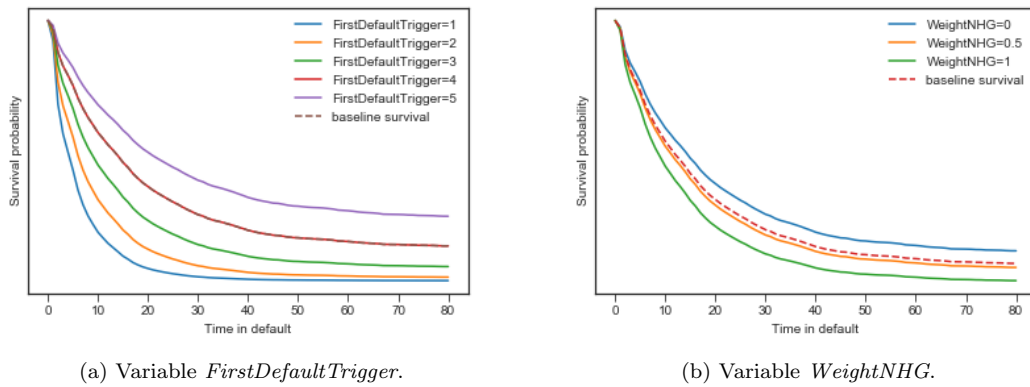


Figure 9: Estimated survival curves for different values of selected variables.

	Coefficient (b)	Standard Error	p	Hazard Ratio	95% Confidence Interval
FirstDefaultTrigger	- 0.351	0.010	< 0.00001	0.703	- (0.373 – 0.330)
WeightNHG	0.407	0.037	< 0.00001	1.503	(0.334 – 0.480)
LTV	- 1.251	0.073	< 0.00001	0.296	- (1.359 – 1.072)
Variable A	0.405	0.070	< 0.00001	1.500	(0.267 – 0.544)
Variable Y	0.152	0.041	0.00021	1.165	(0.071 – 0.233)

Table 3: Selected variable estimations.

The final model includes the variables *FirstDefaultTrigger*, *WeightNHG*, *LTV*, *Variable A* and *Variable Y*. Table 3 shows the estimated regression coefficients, together with their standard errors. The significance level of each coefficient (p) is determined from the tables of the chi-square distribution since, under the null hypothesis, the log-rank statistic is approximately chi-square with one degree of freedom. From the estimated regression coefficient, one can calculate the hazard ratio by taking the exponential of the estimated coefficient. The hazard ratio represents the effect of each variable adjusted for the other variables in the model. For

³Scales on y-axis removed due to confidentiality.

each estimated regression coefficient, one can obtain a confidence interval. *Figure 10* shows the 95% confidence interval for each variable. This figure also shows that the variable *LTV* has the highest predictive power for the LR. The value of the estimated regression coefficients is obtained by maximum likelihood estimation.

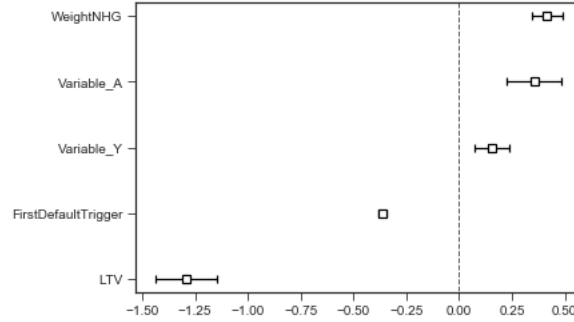


Figure 10: 95% Confidence interval for estimated regression coefficients.

6.3 Extended Cox model

With the Extended Cox model, we tested the effects of time-dependent explanatory variables on the baseline survival curve. We developed two different Extended Cox models based on the same two segments of the retail mortgage portfolio as in the Cox Proportional Hazards models. For both segments, we developed a non-performing model. It is not possible to develop a time-dependent performing model because no information on future risk drivers is available. Within these Extended Cox models, the explanatory variables can change over time. Therefore, the final model selection will be different than the final model selection of the Cox Proportional Hazards model. In this section, we discuss the process of the final model selection of one of the models in more detail. Again, it should be noted that we changed some of the variable names because we can not expose them to the public. *Appendix B* shows the final model selection of the other model.

6.3.1 Stepwise time-dependent variable selection

As with the Cox Proportional Hazards model, we used a stepwise selection method in order to build the models. *Table 4* shows the results of the stepwise selection method. The first variable added in the model was the variable *FirstDefaultTrigger* followed by variable *WeightNHG*, *LTV*, *Variable B*, *Variable X*, *Variable D*, *Variable Y*, *ArrearsChangeSinceMaxPast12Months* and *ArrearsChangeSinceMaxPast3Months*. The third column indicates the variables that were removed from the model because there were no further improvements in the model. The fourth column represents the significant level for each variable obtained from the single variable analysis.

Step	Variable Added	Variable Removed	p -value
1	FirstDefaultTrigger		< 0.00001
2	WeightNHG		< 0.00001
3	LTV		< 0.00001
4	Variable B	Variable B	< 0.00001
5	Variable X	Variable X	< 0.00001
6	Variable D	Variable D	< 0.00001
7	Variable Y	Variable Y	< 0.00001
8	Variable G	Variable G	< 0.00001
9	Variable I	Variable I	0.00311
10	Variable H	Variable H	0.00585

Table 4: Stepwise time-dependent variable selection.

6.3.2 Selected time-dependent variable estimations

The final model includes the variables *FirstDefaultTrigger*, *WeightNHG* and *LTV*. Table 5 shows the estimated regression coefficients, together with their standard errors. As with the Cox Proportional Hazards model, the significance level of each coefficient (p) is determined from the tables of the chi-square distribution since under the null hypothesis, the log-rank statistic is approximately chi-square with one degree of freedom. As with the Cox Proportional Hazards model, the variable *LTV* has the highest predictive power for the LR.

	Coefficient (b)	Standard Error	p	Hazard Ratio	95% Confidence Interval
FirstDefaultTrigger	0.055	0.013	0.00002	1.057	(0.029 – 0.082)
LTV	- 1.165	0.088	< 0.00001	0.311	- (1.338 – 0.993)
WeightNHG	0.278	0.048	< 0.00001	1.320	(0.183 – 0.373)

Table 5: Selected time-dependent variable estimations.

6.4 Model validation

In total, we developed 6 different models based on the two segments of the retail mortgage portfolio of Rabobank. For both segments, we developed a performing, a non-performing time-independent and a non-performing time-dependent model. When one includes time-dependent variables, the final model changes slightly. There are two fundamental aspects to validate the

models: discriminatory power and calibration. Discriminatory power measures the rank order of the different estimations while calibration measures the accuracy of the estimations. In this section, we discuss the performance measurements of the just described models.

	Performance Metric	Time-independent variables	Time-dependent variables
Discriminatory Power	Loss Capture Ratio	xxx %	xxx %
	Spearman's rank	xxx %	xxx %
	Kendall's Tau	xxx %	xxx %
	C-index	xxx %	xxx %
Calibration	Loss Shortfall	-xxx %	-xxx %

Table 6: Performance measurements.

Table 6 shows the performance measurements.⁴ The discriminatory power of the Cox Proportional Hazards model with time-independent variables is higher than the discriminatory power of the Extended Cox model with time-dependent variables. This means that the Cox Proportional Hazards model is better in differentiating risk, and assigning high losses to high realized losses, than the Extended Cox model. However, the loss shortfall of the Extended Cox model is closer to zero than the loss shortfall of the Cox Proportional Hazards model. This means that the Extended Cox model gives more accurate predictions than the Cox Proportional Hazards model. The minus sign in the loss shortfall metric means that the models are currently predicting higher losses than actually observed. This can also be seen in Figure 11, where more observed LR are closer to zero than predicted LR.⁵

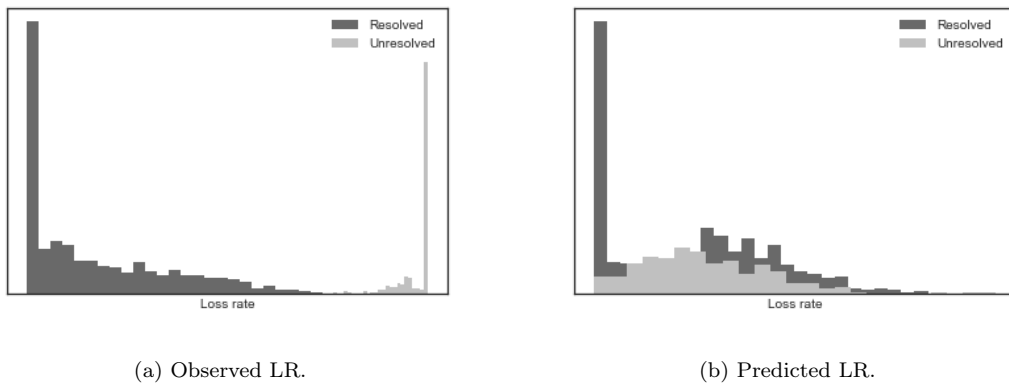


Figure 11: Loss rates.

⁴Performance measurements removed due to confidentiality.

⁵Scales on x-axis and y-axis removed due to confidentiality.

7 Discussion

In this chapter, we discuss the methods and results of this study followed by some limitations and suggestions for further research. First, we give an interpretation and explanation of the results from *Chapter 6*.

The results from the lifetime data table and curves in *Table 1* and *Figure 8* show that the overall survival decreases much faster in the beginning than towards the end. Also, the failure rate, which represents the rate at which repayments of monetary units are happening in a certain time interval, is much higher in the beginning than towards the end. These findings suggest that the more time passes, the less likely it is that a customer will pay back the loan. Since cash flows are discounted to the moment of default, the cash flows received later are less valuable than the ones received earlier in the workout period. The overall survival in the last time interval represents the final LR. This is the amount that Rabobank can expect to lose on average on a defaulted loan.

Compared to the lifetime method, the Cox Proportional Hazards model used a regression method in order to estimate the effects of explanatory variables on the baseline survival curve. The significance level of sixteen variables was tested with a single variable analysis. Only nine of them were included in the stepwise selection of the model. The estimated hazard ratios in *Table 3* explain how much effect a change of one variable has on the baseline survival curve when keeping the values of the other variables constant. The variable *LTV* seems to have the greatest hazard ratio, meaning that it has the highest predictive power for the LR.

The Extended Cox model used a regression method in order to estimate the effects of explanatory variables on the baseline survival curve as well. The method is almost similar to the Cox Proportional Hazards model except that the values of the variables can change over time. As with the Cox Proportional Hazards model, the significance level of sixteen different variables was tested. Only ten of them were included in the stepwise selection of the model. Our results show that the variables *LTV*, *FirstDefaultTrigger* and *WeightNHG* together are good predictors for estimating the LR. These risk drivers are also good predictors in the LGL model of Rabobank currently used.

We measured the performance of the models with different validation metrics. Our results show that the Cox Proportional Hazards model has higher discriminatory power than the Extended Cox model. This means that this model is better at differentiating risk. However, the loss shortfall of the Extended Cox model is closer to zero than the loss shortfall of the Cox Proportional Hazards model. This means that this model is better in making accurate LR predictions than the Cox Proportional Hazards model. However, this comparison is not

completely fair since we used different methods for testing. We explain this limitation in more detail in the next section.

The minus sign in the loss shortfall means that the models are currently predicting higher losses than actually observed. *Figure 11* shows this observation as well. The left histogram shows the loss rates of the resolved and unresolved cases at t_{end} for every defaulted loan in the test dataset. We expect the observed loss rates of the unresolved cases in this histogram to decrease in the future. The right histogram shows the predicted loss rates of the resolved and unresolved cases at t_{max} . This figure gives an interpretation of the distribution of the predicted and realized loss rates of the test dataset. It should not be used as a comparison of the distributions, since the factor of time is not taken into account.

In this study, we used a data splitting approach. This means that the data are randomly split in an 80% train and 20% test set. When reusing the models, the data are split differently. Therefore we calculated a confidence interval for the performance measurements.

The modelling choices made in this study were based on the earlier studies of [Witzany et al. \(2010\)](#), [Zhang & Thomas \(2012\)](#), [Privara et al. \(2013\)](#) and adjusted in a way that satisfies the regulatory requirements. However, earlier studies only examined the effects of time-independent variables, whereas this study also examines the effects of time-dependent variables. Another way of how this study contributes to the literature is by the usage of different validation metrics. The study of [Zhang & Thomas \(2012\)](#) only looked at discriminatory power and goodness of fit, while the studies of [Witzany et al. \(2010\)](#), [Privara et al. \(2013\)](#) did not even examine the discriminatory power of the models. Therefore, this study contributes a new SA method applied to recovery data, and new ways of validating the models, to the literature.

7.1 Limitations

One of the limitations of this study is the distribution of the cash flows of the retail mortgage data. The cash flows are non-evenly distributed over time, where in most cases the most important cash flow happens at the end of the default period. Therefore, our results do not include recoveries realized so far in the non-performing model estimation, in the way we discussed in *Section 4.3*. *Appendix C* shows a more detailed analysis of this limitation.

Another limitation of the retail mortgage data is the limited number of cash flows available for unresolved cases. Incorporating these unresolved cases in the way we discussed in *Section 3.4.1* led to high loss shortfalls. This means that our models were predicting much higher LR than actually observed. Therefore, we tested the models with different workout periods and different censoring times. After testing, we changed the timing of a censored event for unresolved cases,

t_{end} , from last known moment in default to the last known moment of an observed cash flow. Afterwards our results showed better loss shortfalls.

The dataset used to build and test the LGL model contains information on two segments of the retail mortgage portfolio of Rabobank. The information is available for 27,185 defaulted loans. The amount of data available for one segment is five times bigger than the amount available for the other segment. Whenever there are less data, it becomes harder to validate and test the models. We expect that this limitation becomes less important in the future because more data become available.

Another limitation of this study is the treatment of missing values of the risk drivers. We did not completely cover this and therefore some results may be viewed with less confidence. We treated missing values in the most common manner, by replacing them by the median. It might be discussable what would be the best approach and more research needs to be done on this topic.

7.2 Suggestions for further research

This study is primarily focused on survival analysis in the LGL component, used in the LGD model for the retail mortgage portfolio. However, this technique could also be applied in other components of the LGD model, as for example in the loss given cure component. The Cox Proportional Hazards model and the Extended Cox model used in this study only consider one event of interest, liquidation. However, survival analysis does allow to consider more events of interest, this is called competing risk. With competing risk there are at least two events of interest, but only one of such a failure type can actually occur (Kleinbaum & Klein, 2005). As with the retail mortgage portfolio, there are also two events of interest, cure and liquidation. A suggestion for further research would be to examine whether competing risk could be applied for the two different scenarios of the retail mortgage portfolio.

We also suggest to examine alternative survival analysis methods. One of these methods is survival analysis with competing risk, but there are more methods which could be examined, such as for example the accelerated failure time model or the parametric regression models (e.g. Weibull regression) (Kleinbaum & Klein, 2005). In order to get the best performing model for recovery data, it would be worthy to compare alternative survival analysis methods.

This study only looked at the application of the best modelling choices from earlier studies satisfying regulatory requirements. However, in *Chapter 3* we mentioned alternative censoring methods and treatments of negative cash flows. Another suggestion for further research would be to examine whether these methods would result in models performing better. Also, we only

examined the explanatory variables used in the current cure and liquidation model. In order to improve the models, we could consider alternative explanatory variables.

The validation metrics used in this study are already different than the ones used in earlier studies of [Witzany et al. \(2010\)](#), [Zhang & Thomas \(2012\)](#), [Privara et al. \(2013\)](#). However, another suggestion for further research would be to examine alternative validation metrics. Especially for the Cox Proportional Hazards model, there are a lot of different performance metrics ([Austin et al., 2017](#)) which could be used.

Altogether, this study provides a strong foundation of the statistical technique of survival analysis applied in the LGL component, used in the LGD model for retail mortgage portfolios. In extension of this study, alternative techniques and modelling choices could be examined in order to get the model performing best.

8 Conclusion

In this study, we tried to answer the question whether the statistical technique of survival analysis (SA) would be applicable in the modelling of the LGL component, used in the LGD model for retail mortgage portfolios of Rabobank. Therefore, we examined the structure of the recovery dataset of defaulted loans and how to transform it in order to apply SA methods. This dataset contained information on the repayments of monetary units of defaulted loans. We added an observation of every repayment of monetary unit with a weight corresponding to the ratio to exposure at default into a new dataset in order to apply SA methods.

The reason we wanted to examine the application of SA in the LGL component, is the advantage of being able to utilize censored default data in order to make LGL estimations. Rabobank uses the recovery processes of resolved defaulted cases in the model estimation and calibration for LGL. The regulator requires to incorporate incomplete recovery processes of unresolved defaulted cases as well. However, this is more difficult. Earlier studies from [Witzany et al. \(2010\)](#), [Zhang & Thomas \(2012\)](#), [Privara et al. \(2013\)](#) showed the potential of SA for incorporating these incomplete recovery processes. Therefore, we examined which modelling choices, used in earlier studies, were suitable in order to apply SA in the modelling of the LGL component and satisfy regulatory requirements.

We developed six models based on two different survival regression techniques, the Cox Proportional Hazards regression and the Extended Cox regression. With these models, we can examine the predicting power of different risk drivers to the LGL component. This gives insight into what drives high LGL estimations and what drives low LGL estimations. The Extended Cox regression made it possible to include time-dependent risk drivers as well. We measured the performance of the models based on two fundamental aspects: discriminatory power and calibration.

Returning to the main research question, the models scored high on discriminatory power and calibration. Therefore, the results show high potential of applying SA in the modelling of the LGL component, used in the LGD model for retail mortgage portfolios. This results in an alternative to the state of the art LGD model, Rabobank uses, satisfying regulatory requirements.

References

- Altman, D. G., Vergouwe, Y., Royston, P., & Moons, K. G. M. (2009). Prognosis and prognostic research: validating a prognostic model. *BMJ: British Medical Journal*, *338*.
- Austin, P. C., Pencina, M. J., & Steyerberg, E. W. (2017). Predicting accuracy of novel risk factors and markers: A simulation study of the sensitivity of different performance measures for the Cox proportional hazards regression model. *Statistical Methods in Medical Research*, *26*, 1053-1077.
- BCBS. (2017). High-level summary of Basel III reforms.
- EBA/GL/2017/08. (2017). Guidelines on the criteria on how to stipulate the minimum monetary amount of the professional indemnity insurance or other comparable guarantee under Article 5(4) of Directive (EU) 2015/2366 (PSD2).
- EBA/GL/2017/16. (2017a). Guidelines on PD estimation, LGD estimation and treatment defaulted exposures.
- EBA/GL/2017/16. (2017b). Guidelines on the application of the definition of default under Article 178 of Regulation (EU) No 575/2013.
- Harrel, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical test. *Journal of the American Medical association*, *247*, 2543-2546.
- Heller, G., & Mo, Q. (2016). Estimating the Concordance Probability in a Survival Analysis with a Discrete Number of Risk Groups. *Lifetime Data Anal*, *22*, 263-279.
- Klein, J., & Moeschberger, M. (2003). *Survival Analysis Techniques for Censored and Truncated Data*. Springer.
- Kleinbaum, D., & Klein, M. (2005). *Survival Analysis*. Springer.
- Li, D., Bhariok, R., Keenan, S., & Santilli, S. (2009). Validation techniques and performance metrics for loss given default models. *Journal of Risk Model Validation*, *3*, 3-26.
- Privara, S., Kolman, M., & Witzany, J. (2013). Recovery Rates in Consumer Lending: Empirical Evidence and Model Comparison. *Submitted to SSRN*.
- PSU. (2019). *Kendall Tau-b Correlation Coefficient*. Retrieved 2019-04-19, from <https://newonlinecourses.science.psu.edu/stat509/node/158/>
- Rabobank. (2018). Annual Report 2018.
- Rabobank. (2019). *About Rabobank*. Retrieved 2019-02-14, from <https://www.rabobank.nl/bedrijven/english-pages/about-rabobank/>

-
- Sauerbrei, W., Royston, P., & Binder, H. (2007). Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in Medicine*, 26(3), 5512-5528.
- Witzany, J., Rychnovsku, M., & Charamza, P. (2010). Survival Analysis in LGD Modelling. *Financial and Accounting Journal*, 7(1), 6-27.
- Zhang, J., & Thomas, L. (2012). Comparison of single distribution and mixture distribution models for modelling LGD. *International Journal of Forecasting*, 28(1), 204-215.
- Zwillinger, D., & Kokoska, S. (2000). *Standard probability and Statistics tables and formulae*. Chapman & Hall/CRC.

A Survival Analysis

The basic principle of survival analysis is to model the time until an event happens. In our case, the event of interest is a repayment of a monetary unit of a defaulted loan. The following concepts can be found in any statistics book about survival analysis. For a review of survival analysis see [Kleinbaum & Klein \(2005\)](#), [Klein & Moeschberger \(2003\)](#).

A.1 Basic concepts

A.1.1 Cumulative Distribution Function

First, let T be a non-negative random variable. The cumulative distribution function, or CDF, describes the probability that a random variable T will be less than or equal to a value t . In case of survival time analysis, this distribution tells us the probability that the event of interest will happen before time t and is sometimes referred to as the cumulative incidence. The CDF of a random variable T is defined as follows

$$F(t) = P(T \leq t) = \int_0^t f(t)dt. \quad (19)$$

A.1.2 Survival function

In survival analysis, it is more useful to talk about the survival function. The survival function describes the probability of an individual surviving beyond a time t and is defined as follows

$$S(t) = P(T > t) = \int_t^\infty f(t)dt. \quad (20)$$

The survival function is a non-increasing right continuous function of t with $S(0) = 1$ and $\lim_{t \rightarrow \infty} S(t) = 0$. In case of T being a continuous random variable, the survival function is the complement of the cumulative distribution function. That is, $S(t) = 1 - F(t)$, where $F(t) = P(T \leq t)$. Since the survival function is the integral of the probability density function, or PDF, $f(t)$ and the complement of the cumulative distribution function the PDF, CDF and survival function are related as follows

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}. \quad (21)$$

A.1.3 Hazard function

The hazard function is fundamental in survival analysis. It describes the probability that the event happens in an infinitely small step after time t conditional on having survived until time t . It is also known as the hazard rate or failure rate. The hazard rate is defined as follows

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \quad (22)$$

The hazard rate must be non-negative. If T is a continuous random variable, then the hazard rate is related to the survival function as follows

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log S(t). \quad (23)$$

A.1.4 Cumulative hazard function

The cumulative for the hazard function is defined as follows

$$H(t) = \int_0^t h(u) du \quad (24)$$

where the relationship between the survival function and the cumulative hazard function can be derived as follows

$$H(t) = \int_0^t h(u) du = \int_0^t \frac{f(u)}{S(u)} du = \int_0^t -\frac{d}{du} \log S(u) du = -\log S(t). \quad (25)$$

The survival function for continuous lifetimes related to the hazard rate function can then be written as follows

$$S(t) = \exp[-H(t)] = \exp\left[-\int_0^t h(u) du\right]. \quad (26)$$

As can be seen, the cumulative distribution function, the survival function and the hazard function are all defined by the distribution of T .

A.2 Parametric and non-parametric methods

Within survival analysis, there are two main methods that can be used to model a distribution, a non-parametric method and a parametric method. With parametric methods, the underlying distribution assumes to follow a specific distribution, such as for example an exponential or normal distribution. These parametric methods can be used for regression to estimate the effects of explanatory variables. Non-parametric methods are not related to an underlying distribution and are only used for simple standard tests. Next to parametric and non-parametric methods a frequently used method within survival analysis is the semi-parametric method. This method is called the Cox Proportional Hazards model and will be explained in the next section. An advantage of a semi-parametric method is that it does, like parametric methods, handle regression to estimate the effects of explanatory variables but it does not rely on a specific underlying distribution. Semi-parametric methods are an attractive alternative for parametric methods because they can directly relate the effects of explanatory variables on survival time allowing an easier interpretation of the results.

A.3 Kaplan-Meier estimator

Non-parametric methods are not related to an underlying distribution. Normally these non-parametric methods are used for simple standard tests. The mostly used non-parametric method is the Kaplan-Meier estimator, also known as the product limit estimator. The Kaplan-Meier estimator is used to estimate the survival function from lifetime data. The formula for the Kaplan-Meier estimator is defined as follows

$$\hat{S}(t) = \hat{S}(t-1) \times \hat{P}(T > t | T \geq t). \quad (27)$$

This formula gives the probability of surviving past the previous failure time $t-1$ multiplied by the conditional probability of surviving past time t , given that the subject has survived to at least time t . This formula can also be expressed as a product limit. This means that for the survival probability $\hat{S}(t-1)$, the product of all fractions that estimate the conditional probabilities for failure can be taken.

A.4 Cox Proportional Hazards (PH) model

The Cox Proportional Hazards model can be used to model the effects of explanatory variables on the survival function and enables larger flexibility than other types of models. The model gives an expression for the hazard at time t with a given set of explanatory variables, denoted by X . The Cox Proportional Hazards model is defined as follows

$$h(t, X) = h_0(t) \exp\left[\sum_{i=1}^p X_i \beta_i\right] \quad (28)$$

where $h_0(t)$ is called the baseline hazard function, $X = (X_1, X_2, \dots, X_p)$ is a vector of the explanatory variables and $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ is the vector we try to estimate. There is no 'error' term used in this model as the randomness is implicit to the survival process. The baseline hazard function is assumed to be non-negative, constant over time and independent on the explanatory variables. At the same time the exponential expression shown here is only dependent on the explanatory variables and does not involve t . Therefore the X s here are called time-independent variables. Later time-dependent variables will be considered when looking at the Extended Cox model. The survival function of the Cox Proportional Hazards model is defined as follows

$$S(t, X) = S_0(t)^{\exp[\sum_{i=1}^p X_i \beta_i]} \quad (29)$$

where

$$S_0(t) = \exp\left[-\int_0^t h_0(u) du\right]. \quad (30)$$

Here $S_0(t)$ is the baseline survival function and also dependent on time only.

A.4.1 Partial likelihood

Estimation of β is done by using the partial likelihood function. The advantage of the partial likelihood function over the full likelihood function is that it can maximize the proportion depending on β alone. Partial likelihood is only valid when there are no ties in the data. Ties occur when two or more subjects experience the event at the same point in time. The data used in the LGL model can contain tied events. For simplicity, first we explain the basic partial likelihood function. Afterwards, we explain the adjustments to the basic function for dealing with ties by using the Breslow or Efron approximations.

The likelihood function, L , is an expression which describes the joint probability of obtaining the data actually observed on the subjects in the study as a function of the unknown parameters in the model being considered. As described the estimation of the β coefficients is done by a partial likelihood. This is because the likelihood function only considers the probabilities for those subjects who fail, and does not take the probabilities for censored subjects into account. The partial likelihood function can, however, take these probabilities into account. Normally, the partial likelihood is written as the product of several likelihoods, one for each K failure times. At the k th failure time, L_k denotes the likelihood of failing at this time, given survival up to this time. The likelihood function, L , is defined as follows

$$L = L_1 \times L_2 \times L_3 \times \dots \times L_K = \prod_{k=1}^K L_k \quad (31)$$

where L_k = portion of L for the k th failure time. The partial likelihood function focuses only on subjects who fail but the survival time information prior to censorship is used for those subjects who are censored. This means that a person who is censored after the k th failure time is still part of the risk set used to compute L_k . The censored subjects are in this way included in the analysis.

Afterwards, the likelihood function needs to be maximized. This is done by maximizing the partial derivatives of the natural log of L with respect to each parameter in the model. In order to obtain this maximization of the natural log of L , the following equation should be solved

$$\frac{\partial \ln L}{\partial \beta_i} = 0 \quad (32)$$

where $i = 1, \dots, p$.

A.4.2 Hazard ratio

In survival analysis, the effect of one coefficient can be measured with the hazard ratio (HR). In general, the hazard ratio can be calculated by dividing the hazard rate for one individual

with the hazard rate of a different individual. An estimation of the hazard ratio is defined as follows

$$\widehat{\text{HR}} = \frac{\hat{h}(t, X^*)}{\hat{h}(t, X)} \quad (33)$$

where X^* denotes the set of predictors for one individual, and X denotes the set of predictors for another individual. After substituting the Cox model formula into the numerator and denominator of the hazard ratio, the hazard ratio is defined as follows

$$\widehat{\text{HR}} = \exp \left[\sum_{i=1}^p \beta_i (X_i^* - X_i) \right]. \quad (34)$$

A.4.3 Breslow or Efron approximations

Earlier, when using the partial likelihood function, the assumption was made that there are no tied observed survival times in the data. However, it is quite common that tied data exist in survival analysis. Ties occur when two or more subjects experience the event at the same point in time. Therefore, the partial likelihood function as described earlier needs to be reconsidered. Breslow and Efron approximations could be used to adapt the partial likelihood function for tied data.

When using the Breslow's Approximation (1974) and assuming that there are d_k tied survival times at the k th failure time given the risk set $R(t_{(k)})$, then the likelihood function L_k is defined as follows

$$L_k \approx \frac{\exp(\sum_{i \in D_k} X_i \beta_i)}{\left[\sum_{i \in R_k} \exp(X_i \beta_i) \right]^{d_k}} \quad (35)$$

where D_k is the event. The partial likelihood is then defined as follows

$$L = \prod_{k=1}^D L_k \approx \prod_{k=1}^D \frac{\exp(\sum_{i \in D_k} X_i \beta_i)}{\left[\sum_{i \in R_k} \exp(X_i \beta_i) \right]^{d_k}} \quad (36)$$

where D is the total number of events. The Breslow approximation is a good method if d_k is small and/or the number of clients at risk is large. The approximated partial likelihood should then be very close to the exact partial likelihood. However, if these conditions do not hold then the Efron approximation (1977) could better be used.

The partial likelihood of β given by the Efron approximation is defined as follows

$$L = \prod_{k=1}^D L_k \approx \prod_{k=1}^D \frac{\exp(\sum_{i \in D_k} X_i \beta_i)}{\prod_{l=1}^{d_k} \left(\sum_{i \in R_k} \exp(X_i \beta_i) - \frac{l-1}{d_k} \sum_{i \in D_k} \exp(X_i \beta_i) \right)}. \quad (37)$$

A.5 Time-dependent variables

We can extend the Cox Proportional Hazards model to allow time-dependent variables as predictors. Time-dependent variables are variables whose values for a given subject can change over time. The Extended Cox model is defined as follows

$$h(t, X(t)) = h_0(t) \exp \left[\sum_{i=1}^{p_1} X_i \beta_i + \sum_{j=1}^{p_2} \delta_j X_j(t) \right] \quad (38)$$

where $X = (X_1, X_2, \dots, X_{p_1})$ is a vector of the time-independent explanatory variables and $X = (X_1, X_2, \dots, X_{p_2})$ is a vector of the time-dependent explanatory variables. The vector $\delta = (\delta_1, \delta_2, \dots, \delta_{p_2})$ represents the overall effect on the survival time of the time-dependent variables and is not time-dependent. An assumption of this model is that the variable $X_j(t)$ can change over time but the hazard model provides only one value of the coefficients for each time-dependent variable in the model.

A.5.1 Hazard ratio for Extended Cox model

The estimation of the hazard ratio for the Extended Cox model is defined as follows

$$\hat{\text{HR}} = \frac{\hat{h}(t, X^*(t))}{\hat{h}(t, X(t))} \quad (39)$$

which can be rewritten as follows

$$\hat{\text{HR}} = \exp \left[\sum_{i=1}^{p_1} \beta_i (X_i^* - X_i) + \sum_{j=1}^{p_2} \delta_j [X_j^*(t) - X_j(t)] \right]. \quad (40)$$

Here the hazard ratio describes the ratio at a particular time t , for two sets of predictors at time t . $X^*(t)$ and $X(t)$ are both a combined set of predictors containing both the time-dependent and time-independent variables.

B Model selection and parameter estimates

As discussed in *Chapter 6*, we developed four different Cox Proportional Hazards models and two different Extended Cox models based on two segments of the retail mortgage portfolio of Rabobank. For every segment, we developed two Cox Proportional Hazards models (performing and non-performing model) and one Extended model (non-performing model). The final model selection and parameter estimates of one Cox Proportional Hazards model and one Extended Cox model were discussed in more detail in *Chapter 6*. The final model selection and parameters estimates of the other four models can be found in this appendix. It should be noted that we changed some of the variable names because we can not expose them to the public.

B.1 Cox Proportional Hazards model

Step	Variable Added	Variable Removed	p -value
1	WeightNHG		< 0.00001
2	LTV		< 0.00001
3	Variable X	Variable X	< 0.00001
4	Variable D		< 0.00001
5	Variable A		< 0.00001
6	Variable G	Variable G	< 0.00001
7	Variable Y	Variable Y	< 0.00001
8	Variable C	Variable C	0.00755
9	Variable J	Variable J	0.03602

Table B.1: Stepwise variable selection - performing model segment 1.

	Coefficient (b)	Standard Error	p	Hazard Ratio	95% Confidence Interval
WeightNHG	0.640	0.036	< 0.00001	1.898	(0.570 – 0.711)
LTV	- 1.093	0.062	< 0.00001	0.335	- (1.216 – 0.970)
Variable D	0.076	0.014	< 0.00001	1.079	(0.048 – 0.104)
Variable A	0.444	0.050	< 0.00001	1.559	(0.344 – 0.544)

Table B.2: Selected variable estimations - performing model segment 1.

Step	Variable Added	Variable Removed	<i>p</i> -value
1	FirstDefaultTrigger		< 0.00001
2	WeightNHG		< 0.00001
3	Variable X		< 0.00001
4	Variable G	Variable G	< 0.00001
5	Variable D	Variable D	< 0.00001
6	LTV		< 0.00001
7	Variable Z	Variable Z	< 0.00001
8	Variable B		< 0.00001
9	Variable C		< 0.00001
10	Variable F	Variable F	< 0.00001
11	Variable A	Variable A	< 0.00001
12	Variable J	Variable J	0.01358
13	Variable Y	Variable Y	0.04239

Table B.3: Stepwise variable selection - non-performing model segment 2.

	Coefficient (b)	Standard Error	<i>p</i>	Hazard Ratio	95% Confidence Interval
FirstDefaultTrigger	- 0.266	0.005	< 0.00001	0.766	- (0.276 – 0.255)
WeightNHG	0.438	0.017	< 0.00001	1.550	(0.403 – 0.473)
LTV	- 0.737	0.028	< 0.00001	0.478	- (0.793 – 0.681)
Variable B	- 4.265	1.405	0.00240	0.014	- (7.019 – 1.511)
Variable C	- 0.118	0.022	< 0.00001	0.888	- (0.163 – 0.073)
Variable X	0.287	0.016	< 0.00001	1.333	(0.256 – 0.319)

Table B.4: Selected variable estimations - non-performing model segment 2.

Step	Variable Added	Variable Removed	<i>p</i> -value
1	WeightNHG		< 0.00001
2	Variable X		< 0.00001
3	EverCuredBeofre	Variable G	< 0.00001
4	Variable D	Variable D	< 0.00001
5	Variable Z		< 0.00001
6	LTV		< 0.00001
7	Variable A		< 0.00001
8	Variable F	Variable F	< 0.00001
9	Variable C	Variable C	< 0.00001
10	Variable B	Variable B	0.01013
11	Variable Y		0.03426

Table B.5: Stepwise variable selection - performing model segment 2.

	Coefficient (b)	Standard Error	<i>p</i>	Hazard Ratio	95% Confidence Interval
WeightNHG	0.613	0.018	< 0.00001	1.847	(0.577 – 0.650)
Variable X	0.396	0.015	< 0.00001	1.487	(0.366 – 0.426)
Variable Z	1.534	0.0123	< 0.00001	4.638	(1.292 – 1.776)
LTV	- 0.461	0.028	< 0.00001	0.630	- (0.516 – 0.406)
Variable A	0.342	0.022	< 0.00001	1.408	(0.298 – 0.386)
Variable Y	0.042	0.011	0.00033	1.043	(0.019 – 0.065)

Table B.6: Selected variable estimations - performing model segment 2.

B.2 Extended Cox model

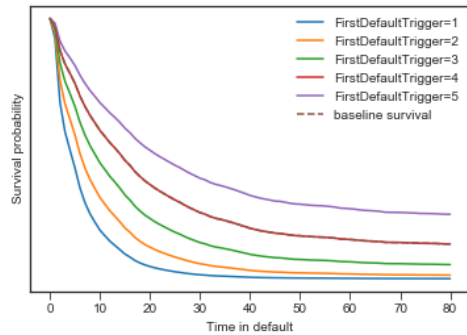
Step	Variable Added	Variable Removed	<i>p</i> -value
1	FirstDefaultTrigger		< 0.00001
2	Variable E		< 0.00001
3	WeightNHG		< 0.00001
4	LTV		< 0.00001
5	Variable A	Variable A	< 0.00001
6	Variable X		< 0.00001
7	Variable D	Variable D	< 0.00001
8	Variable G	Variable G	< 0.00001
9	Variable H	Variable H	< 0.00001
10	Variable C	Variable C	< 0.00001
11	Variable Y	Variable Y	< 0.00001
12	Variable B	Variable B	< 0.00001
13	Variable I	Variable I	0.00156
14	Variable F		0.00193

Table B.7: Stepwise variable selection - non-performing time-dependent model segment 2.

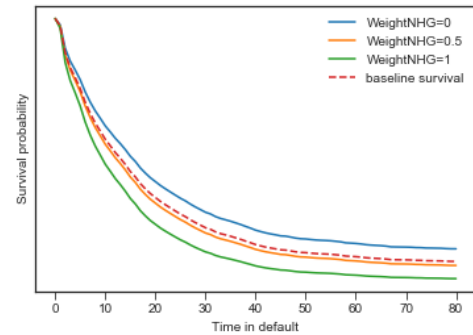
	Coefficient (b)	Standard Error	<i>p</i>	Hazard Ratio	95% Confidence Interval
Variable F	0.000	0.000	0.00083	1.000	(0.000 – 0.000)
Variable X	0.079	0.012	< 0.00001	1.082	(0.054 – 0.104)
FirstDefaultTrigger	0.020	0.007	0.00292	1.021	(0.007 – 0.034)
LTV	- 0.552	0.028	< 0.00001	0.575	- (0.606 – 0.497)
WeightNHG	0.421	0.027	< 0.00001	1.524	(0.368 – 0.474)
Variable E	0.536	0.074	< 0.00001	1.709	(0.390 – 0.682)

Table B.8: Selected variable estimations - non-performing time-dependent model segment 2.

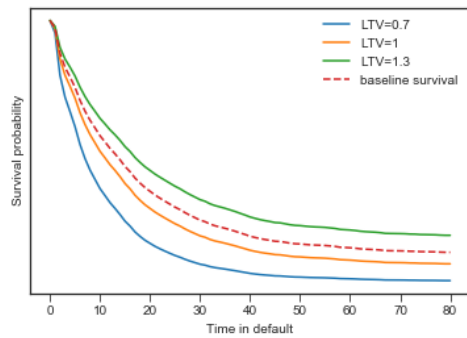
B.3 Survival curves



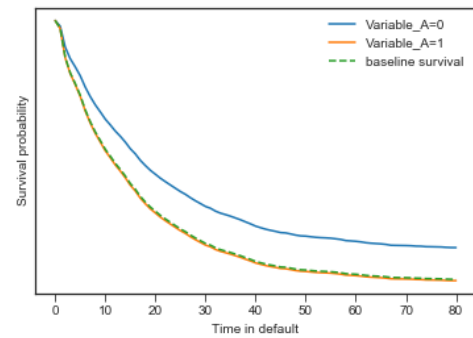
(a) Variable *FirstDefaultTrigger*.



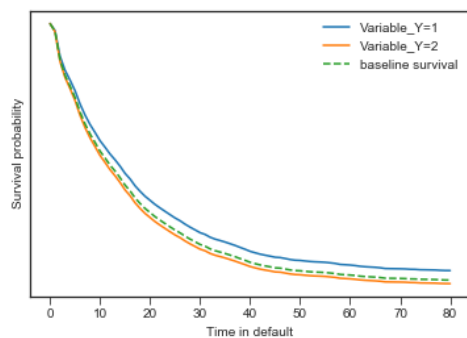
(b) Variable *WeightNHG*.



(c) Variable *LTV*.



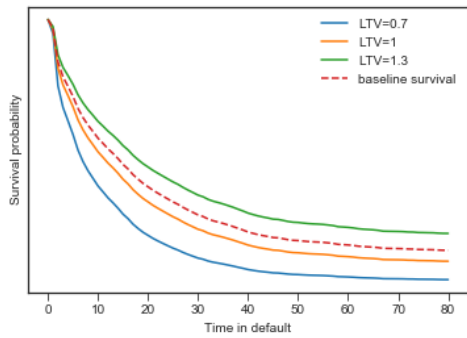
(d) Variable *Variable A*.



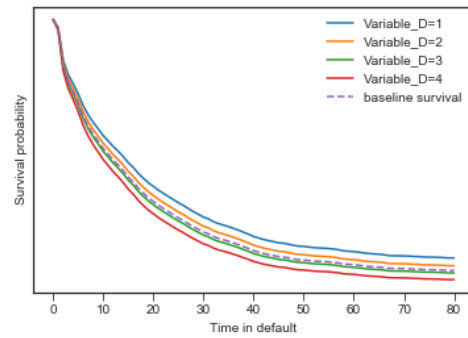
(e) Variable *Variable Y*.

Figure B.1: Estimated survival curves for different values of selected variables - non-performing model segment 1.⁶

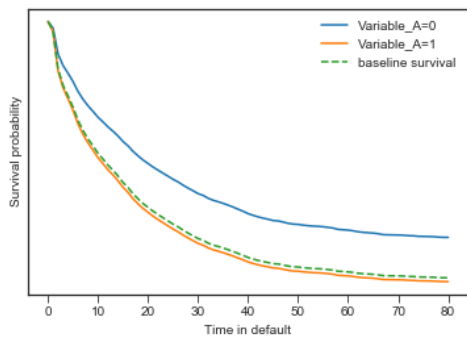
⁶Scales on y-axis removed due to confidentiality.



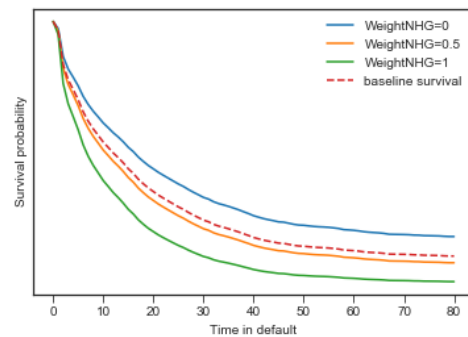
(a) Variable *LTV*.



(b) Variable *Variable D*.



(c) Variable *Variable A*.



(d) Variable *WeightNHG*.

Figure B.2: Estimated survival curves for different values of selected variables - performing model segment 1.⁷

⁷Scales on y-axis removed due to confidentiality.

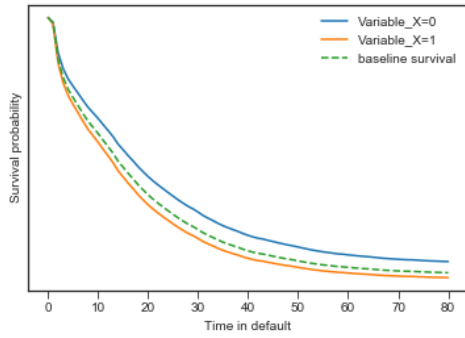
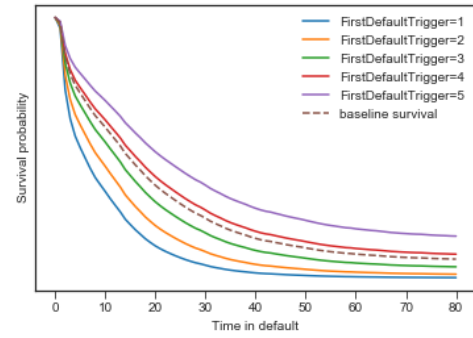
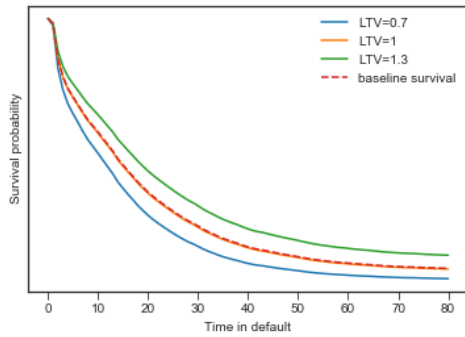
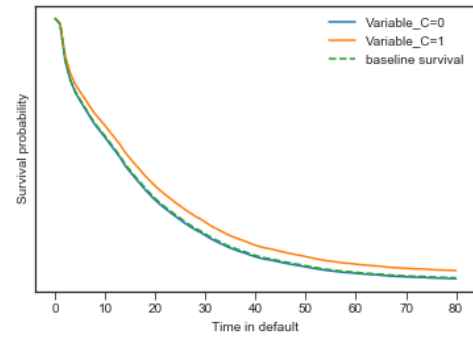
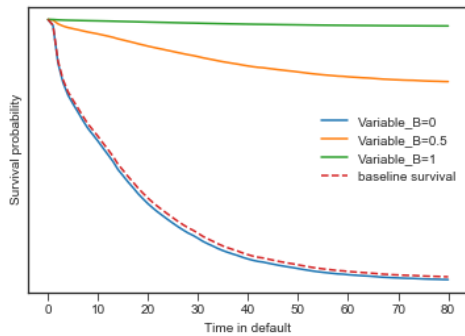
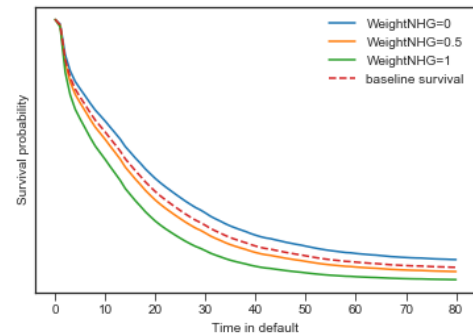
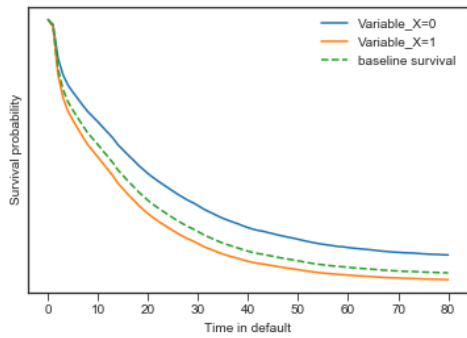
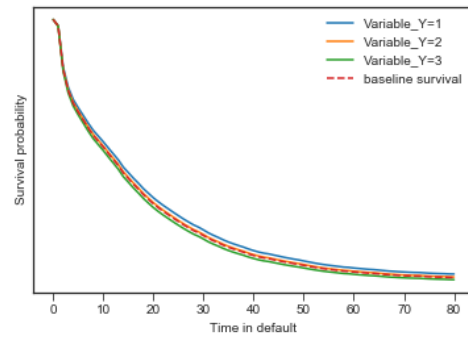
(a) Variable *Variable X*.(b) Variable *FirstDefaultTrigger*.(c) Variable *LTV*.(d) Variable *Variable C*.(e) Variable *Variable B*.(f) Variable *WeightNHG*.

Figure B.3: Estimated survival curves for different values of selected variables - non-performing model segment 2.⁸

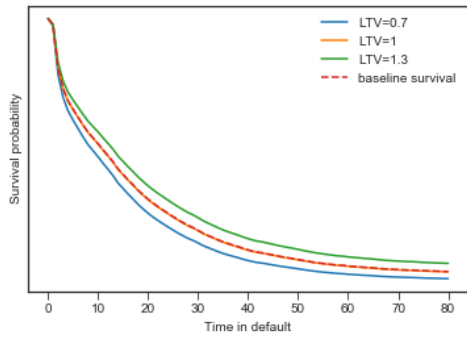
⁸Scales on y-axis removed due to confidentiality.



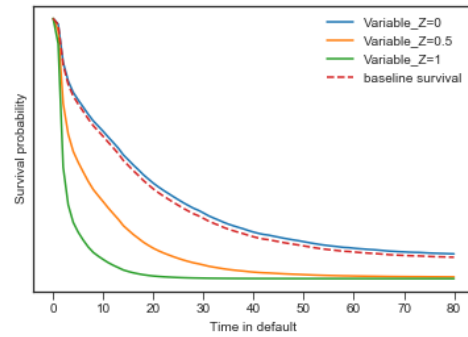
(a) Variable *Variable X*.



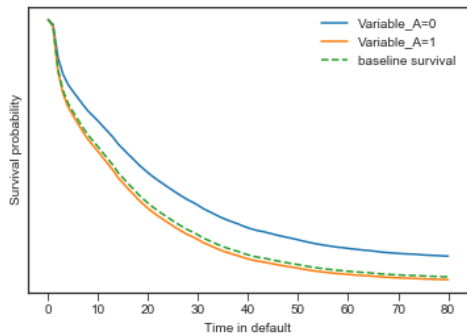
(b) Variable *Variable Y*.



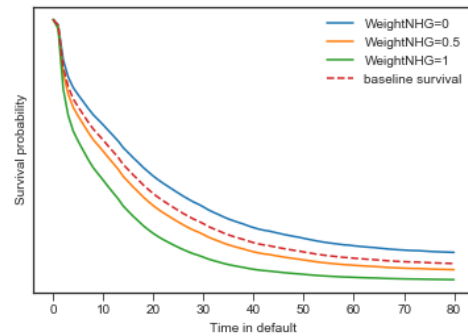
(c) Variable *LTV*.



(d) Variable *Variable Z*.



(e) Variable *Variable A*.



(f) Variable *WeightNHG*.

Figure B.4: Estimated survival curves for different values of selected variables - performing model segment 2.⁹

⁹Scales on y-axis removes due to confidentiality.

C Inclusion of realized recoveries

One of our goals was to incorporate recoveries realized so far in the non-performing models. However, our results show that it is not fair to incorporate recoveries realized so far in the model estimation for mortgage portfolios specific. The results can be seen in [Table C.1](#).¹⁰

Approach	LCR	Spearman	Kendall	C-index	LS
1: time-independent	x_1	x_1	x_1	x_1	x_1
1: time-dependent	x_2	x_2	x_2	x_2	x_2
2: time-independent	$x_1 - 9.6$	$x_1 - 9.3$	$x_1 - 6.3$	$x_1 + 4.3$	$x_1 - 28.0$
2: time-dependent	$x_2 - 1.9$	$x_2 - 5.8$	$x_2 - 4.3$	$x_2 + 28.5$	$x_2 - 61.4$
3: time-independent	$x_1 + 10.6$	$x_1 + 17.9$	$x_1 + 25.6$	$x_1 - 7.7$	$x_1 + 9.3$
3: time-dependent	$x_2 + 24.3$	$x_2 + 25.6$	$x_2 + 34.7$	$x_2 + 16.5$	$x_2 - 10.9$

Table C.1: Performance measurements incorporating recoveries realized so far.

Here, approach 1 represents the approach currently used. Approach 2 represents the approach described in [Section 4.3](#) and approach 3 is the same approach as 2 but, the value of t_{random} is changed to a value randomly chosen between t_0 and t_{80} . As can be seen from the results, including recoveries realized so far led to higher loss shortfalls. This is mainly caused by the non-evenly distribution of the cash flow data, where the most important cash flow (sale of collateral) happens at the end of the default period. Therefore, a lot of the LR observed at time t_{random} are equal to 1, as can be seen in [Figure C.1](#).¹¹

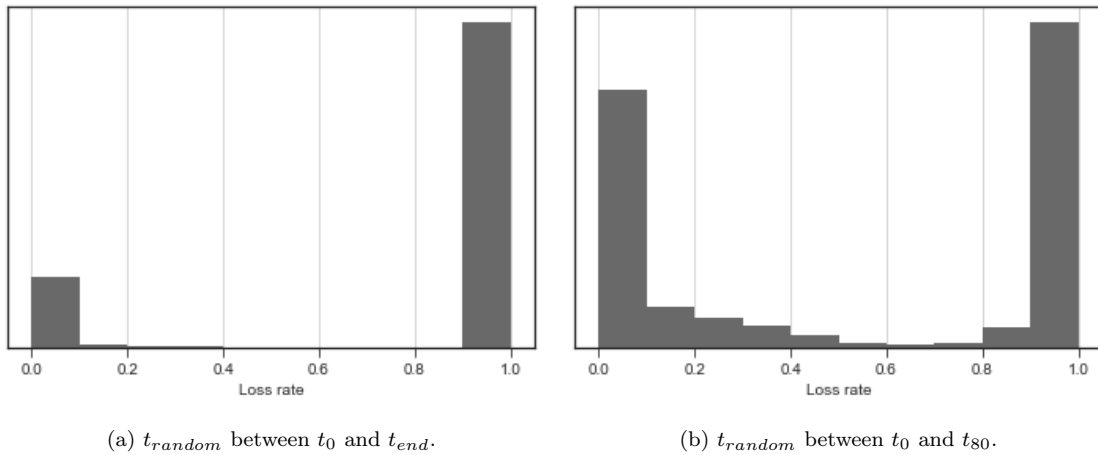


Figure C.1: Distribution of LR at time t_{random} .

¹⁰Performance measurements of first approach removed due to confidentiality.

¹¹Scales on y-axis removed due to confidentiality.