# Predicting Absolute Risk of Mortgage Prepayment using Proportional Hazards and Machine Learning Competing Risks Models in Structured Finance


by

Nick van Gennip (SNR: 2037484)


A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Quantitative Finance and Actuarial Science


Tilburg School of Economics and Management
Tilburg University

Supervised by: Professor Anja de Waegenaere


Date: July 24, 2024

# Acknowledgements

First and foremost, I would like to thank my supervisor Professor Anja de Waegenaere from Tilburg University, who helped me with her remarks and in-depth talks throughout the process. In addition, I would like to thank VB Risk Advisory, and in particular Bernt van Walree, for their guidance throughout the process and for expanding my knowledge in the banking sector. They enabled me to gain experience and knowledge in the backing sector, which motivated me to be thorough in my research. I would also like to thank Professor Thomas Alexander Gerds from the University of Copenhagen, who answered my questions regarding the riskRegression package in R. And finally, I would like to thank my girlfriend and my parents for their support over the last 6 months.

**Abstract**

Accurate predictions of prepayment risk and credit risk over time are of crucial interest for banks, since inaccurate predictions can result in large losses, which one would like to account for in the interest rate set on a loan. If competing risks are disregarded, the probability of prepayment is overestimated, which leads to less competitive interest rates and a weak market position. Using four different survival modelling approaches, from which three are mostly used in medical research, this thesis examines the prediction accuracy of prepayment risk over time while considering default as a competing risk. The four models included in this thesis are: 1) Cox Cause-Specific Hazards model, 2) Fine-Gray Subdistribution Hazard model, 3) Random Survival Forest for Competing Risks, 4) Current model used by VB Risk Advisory at bank X. The most important risk drivers of both prepayment and default are identified using variable selection procedures. The proportionality assumption for both Cox cause-specific hazards model and the Fine-Gray model are evaluated using scaled Schoenfeld residuals and log-log plots. The main goals of this thesis are: i) To improve the current prepayment model used by VB Risk Advisory regarding prediction accuracy, ii) and to clarify the mathematical background of these models used in medical statistics. The results suggest that original interest rate is the most important risk driver of prepayment. Furthermore, the results suggest that the Random Survival Forest for Competing Risk is the best model for predicting the probability of prepayment over time in the same period, meaning that it performs the best in understanding the true underlying relation between variables and prepayment risk. In addition, as the models used in this thesis are neither able to include different prepayment behaviours over time nor yearly effects, the model predictions are also backtested to validate model prediction robustness. It is concluded that the underlying relations between covariates and prepayment behaviour changes over time and that the investigated models are not able to capture these changes.

# Contents

# 1 Introduction

A mortgage is a type of loan between a lender (most often a bank) and a consumer that consumers use to purchase a house. It is agreed to repay in predetermined small, equal, fixed monthly payments over a term (Suknanan, 2023). It is probably the biggest and most important loan an individual has in his life. In the United States there are 84.0 million mortgages at the end of the third quarter of 2023, with a total mortgage debt of $12.14 trillion, which accounts for 70.2% of the consumer debt (Channel, 2023). A total of 1,539,828 residential mortgages were issued in the third quarter of 2023 by banks or other lenders (ATTOM, 2023). A mortgage borrower, also known as mortgagor, is not obligated to comply the initially agreed upon repayment scheme. Such a repayment schedule typically contains the fixed monthly payments, called coupon payments. Depending on the contract details, the mortgagor has the option to partly or fully repay the mortgage earlier, called the prepayment option. Several factors induce prepayment behaviour and prepayment size, such as loan-specific details, economic conditions and personal conditions. Examples of these loan-specific details are loan-to-value ratio, debt-to-income ratio and interest rate on the loan. Examples of economic incentives are shifts in the interest rate and tax benefits. Illustrations of personal circumstances are changes in net income and reception of a heritage.

If the prepayment option is exercised, the mortgage lender, also known as mortgagee, does not collect the anticipated coupon payments and this affects future financial positions. Therefore, it is of high importance that financial institutions, and in particular banks, accurately predict prepayment rates to be able to anticipate on potential prepayments. Banks implement hedging strategies to deal with prepayment risk and in order to be able to do so, banks forecast prepayment behaviour. Fayman and He (2011) showed prepayment risk significantly affects return on loans and return on equity for commercial banks.

Prepayments and defaults affect each other as one cannot occur if the other has occurred. As a consequence, predicting both prepayments and defaults separately may result in overestimation of the risk. Banks need to hold capital for their Risk Weighted Assets (RWA), which is money that a bank is obliged to hold to protect against financial stress and unforeseen losses. This is called the minimum capital requirements and is regulated by the Basel agreements (Bank for International Settlements, 2007). If the risk of prepayment is overestimated, the RWA value is too high compared to the true risks. As a result, the minimum capital requirement is too high. Banks could have invested part of this money.

The main contribution of this thesis is to provide insight in model accuracy for VB Risk Advisory compared to other models. VB Risk Advisory is a quantitative consultancy party, specialised in quantitative financial risk, data analytics and quantitative modelling problems. New proposed models to VB Risk Advisory are requested not to be too complex and computationally costly as clients will not accept models they cannot comprehend. The data used in this thesis is the Single-Family Loan-Level Data set provided by Freddie Mac (2024). It contains fully amortizing 10-, 15-, 20-, 30-, 40-year fixed-rate Single-Family mortgages issued between January 1999 and December 2023.

As this data set contains right-censored data, models were selected that can deal with right-censoring. The models used to compare the VB Risk Advisory model, are the Cox cause-specific hazards model, the Fine-Gray subdistribution hazard model and the Random Survival Forest for competing risks. The first two models are semi-parametric survival models, which allow for easy interpretation. The latter model is a machine learning technique that can give insight into more complex and nonlinear relationships between covariates and prepayment risk, but this comes at a cost of more difficult interpretation.

For VB Risk Advisory it is of crucial interest that they advice their banking clients to apply accurate models to satisfy their needs. Too high predicted risk leads to higher interest rates set on loans, which in turn results in less demand for loans. On the other hand, too low predicted risk leads to potential losses. Therefore, the aim of this thesis is to compare these models regarding prediction accuracy of prepayment risk while considering defaults as competing risk. Prediction accuracy is determined for out-of-sample predictions in the same period as well as for forecasting out-of-sample predictions, i.e. backtesting. Then, it is tested whether the machine learning technique random survival forest outperforms proportional hazards prepayment models. Thereafter, the most suitable approach is advised to VB Risk Advisory.

Furthermore, this paper examines which factors drive prepayment behaviour and if the key assumption, proportionality of the hazards, in both the Cox cause-specific hazards model and the Fine-Gray subdistribution hazard model, is satisfied. In this paper, the focus will be on full prepayments. This thesis adds to the existing literature as it applies the random survival forest for competing risks to prepayment data on mortgages, while it used to be mainly performed on simulated data or medical data. Moreover, the proportionality assumption of the Cox prepayment-specific hazards model and the Fine-Gray subdistribution hazard model is tested thoroughly in this paper and if not satisfied, it is investigated what the effects are on estimated coefficients and prediction accuracy. If deemed necessary, the proportionality assumption is accounted for by using stratification on the non-proportional variables.

The remainder of this thesis is structured as follows. Section 2 provides general background on the mortgage market in the United States and elaborates on related research conducted on prepayment risk. Section 3 provided the methodological background of the Cox cause-specific hazards model, the Fine-Gray subdistribution hazard model, the Random Survival Forest for competing risks, and the model currently used by VB Risk Advisory at bank X. Section 4 describes the data used for the analysis in this thesis and provided non-parametric cumulative incidence curves for the covariates to get insight in the direction of the effect of a covariate on prepayment risk. Section 5 discusses the estimation results and evaluates reliability of the key assumptions. Section 6 evaluates the models based on out-of-sample prediction accuracy. First, another random sample is used as test set to compare prediction accuracy on unseen data of the same period. After, the model predictions are evaluated on unseen data in future time periods to test for forecasting accuracy, which is also known as backtesting. Section 7 concludes, gives advise to VB Risk Advisory and provides limitations of the thesis.

# 2 General background

## 2.1 The mortgage market in the United States

A mortgage market enables people to borrow money to buy a house. In the United States, the mortgage market consists of two parts, which are the primary and secondary market. In the primary market, consumers borrow money from financial institutions for their mortgage. There are several types of financial institutions. These are mortgage brokers, mortgage bankers, commercial banks, credit unions, and savings & loan associations. In the secondary market, mortgage investors like Freddie Mac and Fannie Mae buy mortgages to provide liquidity for financial institutions in the primary market to grant additional loans. The financial institutions that grant mortgages to consumers do not want to wait for the monthly payments to get their money back. Instead, they sell the loans to institutional investors, which enables them to grant more loans. Then, the institutional investors bundle mortgages into so-called mortgage-backed securities. Third-party investors buy the mortgage-backed securities to generate returns on the interest rate payments (Quicken Loans, 2023).

In the United States, the most common mortgage product is the 30-year fixed rate fully amortizing mortgage. It is the only country in the world where this is the dominant home mortgage product. One particular type of this mortgage product is the annuity mortgage, which has a specific type of payment scheme. The monthly amount paid by the borrower is constant over time, but the share of principal payment increases over time while the share of interest payment decreases over time (Kish, 2022).

Prepayment penalties are introduced by financial institutions to have protection against losses. While in The Netherlands prepayments are permitted up to 10% annually and penalties are awarded if more is prepaid, in the United States legislation is different. Prepayment regulation is state dependent as well as financial institution dependent. Most lenders allow borrows to prepay 20% annually (Araj, 2024). One could also add a prepayment premium to the interest rate on a mortgage to cover for potential losses due to prepayment instead of charging a penalty. In this way every borrower pays a little amount to cover the loss a financial institution makes if mortgages are prepaid. This idea is exercised by VB Advisory as requested by a Dutch financial institution, bank X. The downside however, is that this would increase the interest rate for every mortgage, which could weaken their market position.

## 2.2 Previous research

In the literature there are two different types of prepayment rate models. The first class of models, called optimal prepayment models, assume rational behavior of all participants. Assuming that mortgagors prepay their mortgages at the optimal time is unrealistic and therefore the second class of models is mostly used. These are models which take into account covariates such as macroeconomic variables, loan characteristics and mortgagor specific characteristics. These models can be further subdivided into survival models, such as the well-known and most frequently used (Cox) proportional hazards model (Jacobs et al., 2005; Clapp et al., 2002), and binary choice models such

as the logistic regression model (Clapp et al., 2002; Li et al., 2019). Lee et al. (2022) studied the determinants of individual borrowers' prepayment on mortgage loans and concluded that the accelerated failure time model (AFT), which models the effect of covariates on the survival function, potentially outperforms the often used logit model and Cox proportional hazard model, as it has more explanatory power. However, its limitation is the requirement that the survival distribution should be known and specified. As in practice the survival distribution is not known, the AFT model is not used in this thesis. Kau et al. (2009) extended the proportional hazard model by including unobserved heterogeneity when estimating the risk factors for mortgage termination. Although this so-called frailty model is mostly used in medical research (Balan and Putter, 2020; Hougaard, 1995), it can also be applied to prepayment behavior modelling. Through the utilization of the frailty model, the conditional independence of survival times assumption in the Cox model is relaxed. However, as it is computationally costly, hard to interpret and a specific distribution on the frailty term needs to be assumed, the frailty model is not used in this thesis (Huang et al., 2023).

More recently, machine learning techniques, such as neural network and random forest, are investigated and yield promising results (Fu, 2017; Ghatasheh, 2014; van der Star, 2022; Melnyk, 2022). However, their interpretation is challenging, especially for neural networks, and as VB Risk Advisory asked for interpretable models, this thesis does not investigate neural networks. As the random forest yields promising results and is more interpretable than a neural network, this paper investigates the performance of the random forest applied to survival data, called the random survival forest. Ishwaran et al. (2008) introduced the random survival forest and later Ishwaran et al. (2014) expanded it to be able to apply it to competing risks, which in this context means that prepayments and defaults compete as to which will occur first (Li et al., 2023). Their method is fully non-parametric and is used for estimation of the cumulative incidence function, which is defined as the probability of experiencing an event of type $j$ by time $t$. Frydman and Matuszyk (2022) used the random survival forest for competing risks to predict default risk of car loans using prepayment as competing risk, and concluded that the random survival forest for competing risks outperformed the regular random survival forest. In this paper, the focus is on estimating and predicting the cumulative incidence function, which deals with competing risks and hence the random survival forest for competing risks will be used due to its promising results.

Although machine learning techniques get more and more attention in the literature, most research in prepayment rate modelling is conducted using the proportional hazard model, which central concept is the usage of a hazard function. This function then describes the instantaneous risk of the prepayment by an individual, given that the mortgage is not prepaid before. The Cox proportional hazards model is popular due to its easy interpretability, its ability to deal with censored data, and its flexibility. This papers investigates the absolute risk of prepayment, which is the cumulative probability of prepayment over time while considering defaults as competing risk. The estimation of the cumulative probability of prepayment often requires the estimation of the hazard rate of prepayment. Due to its flexibility, the Cox proportional hazards model can be adapted and used in the presence of competing risks for the estimation of the hazard rate. A so-called Cox cause-specific hazards model models a separate hazard function

for each competing event (Austin et al., 2016). This allows one to interpret the effect of covariates on the cause-specific hazard. Another adaptation of a proportional hazards model to be able to include competing risks, is to model the cumulative incidence function (CIF) directly. Here, the CIF is modelled using a subdistribution hazard function, introduced by Fine and Gray (1999). The subdistribution is of the same form as the hazard function in the Cox proportional hazards model, with the adaptation that it models a hazard function derived from a CIF (Columbia University Irving Medical Center, 2023). These two methods are the most used methods for incorporating competing risks and are therefore examined in this thesis. Güneş and Apaydin (2024) used the Cox cause-specific hazards model to model both prepayment-specific hazard and the default-specific hazard, and investigated the prediction accuracy of the hazards, but did not look at cumulative incidence accuracy. Olajubu (2020) concluded that the Cox cause-specific hazards model outperforms the Fine-Gray subdistribution model when predicting cumulative incidences, but noted that predictions of the Fine-Gray model were not reliable as the proportionality assumption was not satisfied.

The Cox cause-specific hazards model and the Fine-Gray model are often compared in medical research. Nolan and Chen (2020) compared the Cox model with the Fine-Gray model when assessing the risk of low-trauma re-fractures and found that the Fine-Gray provides better estimation for the risk of the main outcome of risk in the presence of competing risks. Kim et al. (2023) compared the prediction accuracy of the cause-specific hazards model, the Fine-Gray model and the random survival forest in a chronic kidney disease study. They concluded that in real data analysis all three methods showed similar results in terms of both the significance of the risk factors as well as the prediction accuracy. In case of a simulation study, where a non-linear relation between the covariates and the dependent variable is the true relation, it is observed that the random survival forest is the most robust model. On the other hand, Li et al. (2023) found that the competing risks hazard model outperforms machine learning techniques regarding predictive performance.

The Fine-Gray subdistribution hazard model can be further extended to include (external) time-varying covariates according to some sources (Austin et al., 2019). Austin et al. (2019) looked at all 102 papers that originated from 2015 and the first five months of 2019 that used the Fine-Gray model and investigated which papers included time-varying covariates. They found that only 11 papers used the correct time-varying covariates, from which six did not correctly interpret the results. As noted by Austin et al. (2019), interpretation of results require additional carefulness, which will be done in this thesis. While the Cox cause-specific hazards model allows to interpret the coefficient of a covariate as the effect on the cause-specific hazard, the Fine-Gray model allows to interpret the direction of the effect of a covariate on the cumulative incidence function. If external time-varying covariates are included in the Fine-Gray model, the effect of covariates on the CIF can still be interpreted (Austin et al., 2019). However, after correspondence with the authors, it is decided not to include this in thesis.

Often numerous characteristics are observed and available in data sets for prepayment rates. Such characteristics can be categorized into for example economic, individual-specific, geographical and loan-specific factors. To avoid models to overfit the data,

relevant factors should be identified and then implemented in the model. The literature suggests different approaches for pertinent variable selection, such as LASSO, the Akaike Information Criterion (Meis, 2015) or the nonconcave penalised likelihood approach (Fan and Li, 2002).

Several loan-specific covariates that have been concluded to impact prepayment rates, are examined in this thesis, with the notion that results should be interpreted carefully when dealing with competing events. Jacobs et al. (2005) used the Cox proportional hazards model and found that the type of mortgage, the size of the loan, and the age of the mortgage significantly affects the median duration of the mortgage. Schwartz and Torous (1993) concluded that a high initial Loan-to-Value ratio (LTV) significantly accelerates prepayment. Moreover, prepayments increase when refinancing rates decrease. Kau et al. (2009) concluded that the unobserved Metropolitan Statistical Area level affects the prepayment rate of mortgages. Groot and Lejour (2018) found that the conditions for the allowance of proportional hazard model were violated and therefore investigated heterogeneity in prepayment behavior amongst individuals. They concluded that incentives to prepay can only explain partly the observed prepayment behavior. In addition, they found that prepayment of wealthier households and households with a high net-of-tax interest rate differential can be largely explained by prepayment incentives. Another factor that are hypothesized to impact prepayment rates is the loan size at origination. Wu and Deng (2010) found that mortgages with larger loan amounts at origination are less likely to be prepaid. Quercia et al. (2007) investigated the effect of predatory loan terms on prepayment rates. They also examined the effect of prepayment penalties on prepayment rates. Using a competing risk model, they concluded that extended prepayment penalties increase the probability of prepayment with 20 percent.

Prepayment is also partly driven by macroeconomic factors and therefore these factors are often included in research for estimation of the prepayment hazard. Although macroeconomic factors cannot be incorporated in Cox cause-specific hazards model nor will they be used in this thesis in the Fine-Gray model, it is interesting to keep this in mind as we will perform prediction of prepayment risk on future time periods in this thesis. Underlying macroeconomic factors will change over time and will likely affect prediction accuracy of the models used in future time periods.

Chernov et al. (2018) included multiple macroeconomic factors. They used the lagged growth rate in US personal consumption expenditures, the change in the Conference Board's Consumer Confidence Index, and the change in the unemployment rate. They also included variables which proxy for mortgagors' wealth, since these possibly affect their prepayment behaviour. The proxies are the return on the Barclays US Aggregate Bond Index, the lagged return on the CRSP value-weighted stock index, and the lagged change in the National Association of Home builders Housing Market Index. Quercia (2016) also found that mortgage default and prepayment are sensitive to changes in unemployment rates. Lee et al. (2022) used interest rate spread and house prices as macroeconomic factors and found a significant effect on prepayment rates, which disappeared during the financial crisis and a low interest rate period due to economic stabilization. Green and Shoven (1983) also concluded that market interest rates sig-

nificantly affect prepayment probabilities. They found that the average duration of the mortgage highly depends on interest rates. Bhardwaj and Sengupta (2009) found that the macroeconomic variables Interest Volatility and PV Annualized Rate both have a positive effect on the likelihood of prepayment. These findings are not in line with the expectations. Moreover, they found that unemployment rate increases the probability of prepayment as expected, since homeowners cash-out the benefit from the appreciation in home prices. Li et al. (2019) investigated the impact of GDP growth rate, federal funds, and bankruptcy fillings on prepayment and default rates. They found that the GDP growth rate is negatively correlated with prepayment. It is therefore concluded that a developing economy reduces the chance of prepayment. In contrast, the Federal funds base rate is positively correlated with prepayment, since it increases the cost of refinancing. Moreover, the bankruptcy rate is also positively correlated with the prepayment rate. Yuan and Tao (2023) also considered macroeconomic factors including industrial production, unemployment rate, consumer expectations, and an economic sentiment indicator. They also controlled for credit constraints by using variables such as the Federal funds rate and the cost of borrowing index. They find that mortgagors prepay when the aggregate cost of borrowing is low and the industrial production is rising. Moreover, the aggregate cost of borrowing has a heterogeneous effect on mortgagors with different Fair Isaac Corporation scores.

Section 8 summarizes the most important papers for this thesis.

# 3 Methodology

In this section the models that are used throughout this thesis are described. The Cox cause-specific hazards model, the Fine-Gray subdistribution hazard model, the random survival forest for competing risks, and the current model used by VB Risk Advisory are discussed. The first three models are survival models of which one is a machine learning technique and therefore, to start, it is discussed what survival models and machine learning techniques are in general and how they model prepayment.

## 3.1 Preliminaries

### 3.1.1 Survival function

In this thesis only full prepayments are investigated, so when referring to prepayments, it means full prepayments. The survival function represents the probability that a mortgage is not prepaid before the cut-off date, which is the end-date of the reporting period. Let $n$ be the number of mortgages and let $k \in \{1, \ldots, n\}$. Let $T_k$ be the failure time for loan $k$ and let $C_k$ be the time until censoring for mortgage $k$. Let $\mathcal{S}_k \in \{0, \ldots, H\}$ denote the type of event for mortgage $k$ (Zhang et al., 2011). There are $H$ different causes of failure and $\mathcal{S}_k = 0$ suggests the failure time of mortgage $k$ is censored. In right-censored data, we observe the survival time as:

$$T_k^* = \begin{cases} T_k & \text{if } \mathcal{S}_k \neq 0 \\ C_k & \text{if } \mathcal{S}_k = 0 \end{cases}$$

Here, $T_k$ is the unobserved latent survival time. In other words, $T_k^* = min(T_k, C_k)$. In right-censored data, we observe $\mathcal{S}_k^* = \mathcal{S}_k \mathbb{1}(T_k \leq C_k)$. So, if $\mathcal{S}_k^* = 0$, then this indicates that mortgage $k$ is censored at time $T_k^* = C_k$. It is assumed that the failure causes $1, \ldots, H$ are observable. If we do not consider competing risks, then $\mathcal{S}_k \in \{0, P\}$, where $P$ denotes prepayment and 0 denotes censoring. In the competing risks setting later on, two different causes of failure are considered and hence $\mathcal{S}_k \in \{0, P, D\}$, where $P$ denots prepayment, $D$ denotes default and 0 denotes censoring. Define $F_k(t)$ as the cumulative distribution function of $T_k$. Then the survival function for loan $k$ is:

$$S_k(t) = \mathbb{P}(T_k > t) = 1 - F_k(t), \tag{1}$$

with $S_k(0) = 1$ and $\lim_{t \to \infty} S_k(t) = 0 \; \forall k$ (Chen, 2018). Obviously, the survival function is a monotonic non-increasing function as more prepayments have occurred as time progresses.

### 3.1.2 Hazard function

Let $f_k(t)$ be the probability density function of $T_k$ (Sestelo, 2017). The hazard function is the instantaneous rate at which a prepayment occurs at time $t$ given that the mortgage is not prepaid before time $t$:

$$h_k(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T_k \leq t + \Delta t | T_k > t)}{\Delta t} = \frac{f_k(t)}{S_k(t)}, \tag{2}$$

with $\Delta t > 0$ (Chen, 2018). The cumulative hazard function is the integrated hazard rate over time:

$$H_k(t) = \int_0^t h_k(s)\, ds \tag{3}$$

Using that the survival function $S_k(t) = \mathbb{P}(T_k > t)$ and assuming that the survival function is continuous and the derivative exists at every $t$, gives the following relation between the survival function and the hazard rate (see appendix A.1 for derivation):

$$h_k(t) = -\frac{\partial}{\partial t} log(S_k(t)) = \frac{-S_k'(t)}{S_k(t)} \tag{4}$$

or:

$$S_k(t) = exp(-\int_0^t h_k(s)\, ds) = exp(-H_k(t)) \tag{5}$$

The hazard ratio is defined as the ratio between two hazards for loans $k$ and $l$:

$$\frac{h_k(t)}{h_l(t)} \tag{6}$$

### 3.1.3 Cumulative Incidence Function

The probability of a prematurely finished mortgage $k$ due to prepayment at time $t$ is defined as the incidence of prepayment. The cumulative incidence, or the absolute risk, of mortgage $k$ is the cumulative sum up to time $t$ of the incidence of prepayment up to time $t$. In the case of no competing risks, the cumulative incidence function is defined as (Austin et al., 2016):

$$CIF_k(t) = 1 - S_k(t) \tag{7}$$

## 3.2 Survival models

Survival models model the time until an event occurs, which is here the moment of full prepayment. In practice, we are often more interested in the probability of prepayment within x years instead of the expected time before a prepayment. Survival models contain two key functions, which are the survival function and the hazard function. The hazard function, which is the instantaneous rate at which a prepayment occurs at time t given that the mortgage is not prepaid before time t, provides an intuitive explanation for prepayments. Therefore, survival models are the most used models in the literature (Harrell, 2001). One main advantage of survival models is that they account for censored data. Loan-level data often does not contain information of the mortgage over the full maturity time. Hence, when modelling prepayment rates, it is common to encounter left- or right-censored data. More on censored data follows in the next paragraph. The models discussed in this thesis differ in how they estimate the hazard function as well as the survival function.

#### Censored data

Within survival analysis, a major concern is censored data. This is data in which not the whole duration of all mortgages is observed. Survival analysis mostly has to

deal with right-censored data, which means that the starting date of the mortgage is observed, but the maturity date is not. For prepayment modelling, this indicates that not all prepaid mortgages are classified as prepaid as the prepayment will occur after the final observation in the data set. It also indicates that it is not known whether a mortgage is prepaid or not after the last observation. It is assumed that censored mortgages would have had the same prepayment probability if they were not censored, conditional on the covariates. Hence, censoring is random conditional on the covariates. As this is a key assumption, it should be tested. This can be done by evaluating the proportional hazards assumption as described later. If the proportional hazards do not vary over time, the assumption is valid (Kleinbaum and Klein, 2011).

### 3.2.1 Cox proportional hazards model

The classic Cox proportional hazards model is a semi-parametric model, which takes into account the impact of covariates. It is semi-parametric because the baseline hazard function is not assumed to have a particular shape or distribution, but the hazard rate depends on covariates whose coefficients are estimated (Harrell, 2001). The model is so popular due to its semi-parametric nature. Parametric models typically deliver reliable results only if the specified parametric distribution is correct, whereas the semi-parametric Cox proportional hazards model gives results comparable to those parametric models with a correctly specified distribution. So, there is no need to assume a parametric distribution, which improves reliability and robustness (Kleinbaum and Klein, 2011). Furthermore, the model allows to evaluate the effect of several factors on survival simultaneously.

Let $X_{ki}$ be the the value of time-fixed covariate $i$ for mortgage $k$. Let $h_0(t)$ be the baseline hazard rate, which is the hazard rate if $X_{0i} = 0\ \forall i$. It depends on time, but does not depend on the covariate values. The model can be written as follows:

$$h_k(t) = h_0(t) exp(\sum_{i=1}^{p} \beta_i X_{ki}) \tag{8}$$

$\forall k$ and hence,

$$log(\frac{h_k(t)}{h_0(t)}) = \sum_{i=1}^{p} \beta_i X_{ki}, \tag{9}$$

where *log* is the natural logarithm. The log-hazard is a linear function of the covariates and a population-level baseline hazard. The hazard function as in (8) can be rewritten to obtain the survival function (see appendix A.2 for derivation):

$$S_k(t) = (S_0(t))^{exp(\sum_{i=1}^{p} \beta_i X_{ki})}, \tag{10}$$

where $S_0(t)$ is the baseline survival function. This is defined as the survival function if all covariates are equal to zero and is the same for each mortgage $k$.

Using the relation between the cumulative incidence function and the survival function, when there are no competing risks as described in (7), this gives:

$$1 - CIF_k(t) = (1 - CIF_0(t))^{exp(\sum_{i=1}^{p} \beta_i X_{ki})}, \tag{11}$$

where $CIF_0(t)$ is the baseline cumulative incidence function. It is equal to $1 - S_0(t)$.

The hazard ratio in the Cox proportional hazards model for two mortgages $k$ and $l$ is:

$$\frac{h_k(t)}{h_l(t)} = \frac{h_0(t)exp(\sum_{i=1}^{p} \beta_i X_{ki})}{h_0(t)exp(\sum_{i=1}^{p} \beta_i X_{li})} = exp(\sum_{i=1}^{p} \beta_i (X_{ki} - X_{li})) \tag{12}$$

The hazard rate can be interpreted as the factor with which the baseline hazard is multiplied for that covariate. If the hazard is greater than 1, the baseline hazard is increased proportionally by the hazard. On the other hand, if the hazard is smaller than 1, the baseline hazard is decreased proportionally by the hazard.
As one can see, the hazard ratio is time-independent. This so called proportional hazards assumption is the key assumption in the Cox proportional hazards model. This also indicates that the hazard functions of the two mortgages cannot cross each other (STHDA, 2020).

**Estimation Cox proportional hazards model**

Let $m$ be the number of distinct prepayment times. Let the distinct prepayment times be denoted by

$$t_1 < t_2 < \cdots < t_m \tag{13}$$

And let $t_j \in (t_1, \ldots, t_m)$ be the time that the $j$-th mortgage is prepaid, where $1 \leq j \leq m$. For now, assume that there are no tied events, which has as consequence that $t_j = T_k$ for $j = k$. Following the notation, at time $t_m$, there are $n - m$ mortgages censored. Define the so-called 'risk set' at time $t_j$ for $1 \leq j \leq m$, $R(t_j)$:

$$R(t_j) = \{k \in \{j, \ldots, n\} : T_k \geq t_j\} \tag{14}$$

(Kleinbaum and Klein, 2011)
It is the set of indices of the mortgages that have not yet been prepaid or been censored by time $t_j$ (Cox, 1975). Or stated differently, $R(t_j)$ are the indices of those under observation when the $j$-th mortgage is prepaid (Schoenfeld, 1982). To estimate the coefficients of the Cox proportional hazards model, Cox (1975) introduced the partial likelihood function, which has to be maximised. Cox (1972) showed that conditional on the risk set $R(t_j)$, the likelihood that mortgage $j$ is the prepaid mortgage at time $t_j$:

$$L_j(\beta) = \frac{exp(\sum_{i=1}^{p} \beta_i X_{ji})}{\sum_{l \in R(t_j)} exp(\sum_{i=1}^{p} \beta_i X_{li})} \tag{15}$$

The intuition behind expression (15) is that the likelihood of mortgage $j$ to be prepaid at time $t_j$ is determined by the hazard rate of mortgage $j$ relative to the combined hazard rates of all mortgages at risk at time $t_j$ (Segota, 2023). For the mathematical

derivation, see Appendix A.3. As each prepayment contributes a factor of (15), the partial likelihood then becomes (Cox, 1972):

$$L(\beta) = \prod_j \frac{exp(\sum_{i=1}^p \beta_i X_{ji})}{\sum_{l \in R(t_j)} exp(\sum_{i=1}^p \beta_i X_{li})} \tag{16}$$

So, mortgages that are not prepaid only contribute to the risk set, but do not appear in the numerator. The log partial likelihood is:

$$log(L(\beta)) = \sum_j \sum_{i=1}^p \beta_i X_{ji} - \sum_j log(\sum_{l \in R(t_j)} exp(\sum_{i=1}^p \beta_i X_{li})) \tag{17}$$

Then, to estimate $\beta$, $L(\beta)$ is maximised w.r.t. $\beta$ using the realizations of $X_{ki}$ $\forall k, i$. Or equivalently, $log(L(\beta))$ is maximised over $\beta$. Both result in the same $\beta$ as the natural logarithm is a monotonic transformation, which does not influence the location of the maximum.

$$\hat{\beta} = \arg\max_\beta log(L(\beta)) \tag{18}$$

Interestingly, the baseline hazard function does not have to be known to estimate $\beta$. Estimation of the baseline hazard function is hard, which makes this a desirable property (Chen, 2018).

A disadvantage of the Cox proportional hazards model is that it cannot deal with time-varying covariates as a time-constant hazard ratio is assumed. Macro-economic factors such as the interest rate level and the unemployment rate are in the literature found to affect prepayment rates (Green and Shoven, 1983; Quercia, 2016), and as they are time-varying covariates, it is desirable to include time-varying covariates in the model. This is achieved by adjusting the classical Cox model, which leads to the so called Cox time-varying covariates model, which is discussed in Section 3.2.2.

**Dealing with tied events - Efron's method**

When two or more mortgages are prepaid at the same recorded time, they are said to be tied. In this case, (13) changes to:

$$t_1 \leq t_2 \leq \cdots \leq t_m \tag{19}$$

If time is considered to be a continuous variable, then the probability of two mortgages to be prepaid at the same time is zero. However, as in this thesis, the available data is often observed for discrete time points. The Cox proportional hazards model relies on the assumption that time is continuous, and hence event times are distinct as they are equal with probability zero. So, if we consider two mortgages to be prepaid at the same recorded time in the data set, Cox says that in reality one of the two mortgages was in the risk-set of the other as they did not actually happen at the same time. However, the Cox model cannot determine which mortgage was prepaid first in reality (Xin, 2014). Therefore, Efron (1977) came up with a method to deal with tied events in the Cox model, called the Efron method. It changes the partial likelihood as defined in (16) by changing the contributions of the tied events to the overall partial likelihood. Let $D_j$

be the index set of all mortgages that are prepaid at time $t_j$ and let $d_j$ be the number of loans in $D_j$. The partial likelihood contribution becomes now:

$$L(\beta) = \prod_{j=1}^{m} \frac{exp(\sum_{i=1}^{p} \beta_i X_{ji})}{\prod_{s=1}^{d_j} (\sum_{l \in R(t_j)} exp(\sum_{i=1}^{p} \beta_i X_{li}) - \frac{s-1}{d_j} \sum_{l \in D_j} exp(\sum_{i=1}^{p} \beta_i X_{li}))} \tag{20}$$

Although this may look complicated at first sight, it is a relatively easy correction on the original partial likelihood where tied events were not considered. If we take for example that the first two prepaid mortgages were prepaid at the same time, then we now get that the contribution of these mortgages to the partial likelihood is as follows:

$$\begin{aligned} L_{1,2}(\beta) = & \frac{exp(\sum_{i=1}^{p} \beta_i X_{1i})}{\sum_{l=1}^{n} exp(\sum_{i=1}^{p} \beta_i X_{li})} \\ & * \frac{exp(\sum_{i=1}^{p} \beta_i X_{2i})}{\frac{1}{2} exp(\sum_{i=1}^{p} \beta_i X_{1i}) + \frac{1}{2} exp(\sum_{i=1}^{p} \beta_i X_{2i}) + \sum_{l=3}^{n} exp(\sum_{i=1}^{p} \beta_i X_{li})} \end{aligned} \tag{21}$$

while without tied events it would be (Efron, 1977):

$$L_{1,2}(\beta) = \frac{exp(\sum_{i=1}^{p} \beta_i X_{1i})}{\sum_{l=1}^{n} exp(\sum_{i=1}^{p} \beta_i X_{li})} * \frac{exp(\sum_{i=1}^{p} \beta_i X_{2i})}{\sum_{l=2}^{n} exp(\sum_{i=1}^{p} \beta_i X_{li})} \tag{22}$$

As it is not that straightforward to see how expression (21) follows from (20) for two tied events, it is derived step-by-step in Appendix A.4. The intuition behind the Efron method is that the order in which tied prepayments happen is not known and hence the denominator is reduced with the same fraction. In other words, Efron's approximation assigns fractional weights to tied events. Efron's approximation performs well when the number of tied events is low as well as when the number is high (Rocke, 2021; Borucka, 2014b). Therefore, Efron's method is used in this thesis rather than other approximation methods such as Breslow's method or the Exact method. Note that if the Efron's method is used while there are no tied events, it results in the same partial likelihood as in (16) (Borucka, 2014b).

### 3.2.2 Cox time-varying covariates model

Predictors whose values vary over time can be included in the Cox time-varying covariates model, or also called the extended Cox model. The model contains both time-fixed (denoted by $X$) and time-varying (denoted by $\tilde{X}$) predictors and can be written as:

$$h_k(t) = h_0(t) exp(\sum_{i=1}^{p} \beta_i X_{ki} + \sum_{j=1}^{q} \delta_j \tilde{X}_{kj}(t)) \tag{23}$$

or equivalently,

$$log(\frac{h_k(t)}{h_0(t)}) = \sum_{i=1}^{p} \beta_i X_{ki} + \sum_{j=1}^{q} \delta_j \tilde{X}_{kj}(t), \tag{24}$$

where $log$ is the natural logarithm. $h_0(t)$ is called the baseline hazard, which depends on time, but does not depend on covariate values. There are $p$ time-independent covariates and $q$ time-dependent covariates. The hazard ratio as defined in (6) is constant over time for time-fixed covariates, whereas for time-varying predictors the hazard ratio may vary over time. The model has the following assumptions:

- Survival times between distinct mortgages in the sample are independent conditional on the covariates, which also means that the behaviour of people holding the mortgages are independent conditional on the covariates. So, if one person would have multiple mortgages, his repayment behavior would be independent of the number of mortgages he has.

- The hazard at time $t$ depends on the value of $X_{ki}(t)$ at that same time. In other words, the effect of a time-dependent variable $X_{ki}(t)$ on the survival probability at time $t$ depends on the value of this variable at the same time $t$, and not on the value at an earlier or later time (Kleinbaum and Klein, 2011).

**Testing proportional hazards assumption**

For the Cox proportional hazards model one should test the underlying proportional hazards assumption. As established before, the proportional hazards assumption does not hold for time-dependent covariates. However, the proportional hazards assumption should still hold for the time-independent variables in the Cox time-varying covariates model as it should in the original Cox proportional hazards model. This assumption can be tested using two different approaches. Both testing methods are used. The first method is the graphical method, which uses the evaluation of the log(-log) survival curve. Since $0 \leq S_k(t) \leq 1$, it is that $log(S_k(t)) \leq 0$. Therefore, both sides are multiplied with $-1$ to be able to apply the second logarithm. So, the second logarithm is taken over $-log(S_k(t))$. Applying the log(-log) transformation on the survival curve as stated in (5), we get (see appendix A.5 for derivation):

$$log(-log(S_k(t)) = log(-log(S_0(t))) + \sum_{i=1}^{p} \beta_i X_{ki} \tag{25}$$

The log-log test looks at the difference between the log(-log) survival curves of mortgage $k$ and $l$, which is given by:

$$log(-log(S_k(t)) = log(-log(S_l(t)) + \sum_{i=1}^{p} \beta_i (X_{ki} - X_{li}) \tag{26}$$

Note that in (26), the subindices $k$ and $l$ represent different loans and $i$ represents the different covariates.
If the proportional hazards assumption holds, then the log(-log) survival curves should be parallel to each other. Namely, the difference between the two curves is the linear term of the differences in predictor values, which does not vary over time by the proportional hazards assumption (Sestelo, 2017).

A second approach to test the proportionality assumption is by using the scaled Schoenfeld residuals as introduced by Schoenfeld (1982). It assesses the correlation between scaled Schoenfeld residuals and time. The intuition is that if, for example, a hazard ratio for credit score is greater than unity, than mortgage holders with a high credit score will be overrepresented among the prepaid mortgages. If, the hazard ratio increases with time, the overrepresentation of high credit scores among prepaid mortgages also increases with time. If we then calculate the proportion of high credit scores among

prepayments and the mortgages in the risk set in each short interval of time, we should be able to detect the increasing hazard ratio (Fisher et al., 2004). This testing procedure uses the concept of a risk set, which is already introduced in (14). This approach requires the following steps, which have to be repeated for every covariate $i \in (1, \ldots, p)$ and for all the time points $t_j \in (t_1, \ldots, t_m)$ that a prepayment occurs:

1. Search for an occurrence of prepayment, which the first is named to be at time $t_1$

2. Take the subsample that includes only the data from time $t \geq t_1$. Hence, these are the elements in $R(t_1)$, which is the risk set as defined in (14).

3. Then, the Schoenfeld residual at time $t_1$ for covariate $i$ is defined as:

$$\hat{r}_{1i} = X_{1i} - \frac{\sum_{j \in R(t_1)} X_{ji} exp(\sum_{i=1}^{p} \hat{\beta}_i X_{ji})}{\sum_{j \in R(t_1)} exp(\sum_{i=1}^{p} \hat{\beta}_i X_{ji})}, \tag{27}$$

   which is the difference between the observed value of covariate $i$ on the observed prepaid mortgage 1 and the conditional expectation given $R(t_1)$. Stated differently, it is the difference between covariate $i$ value of mortgage 1 at the time of prepayment of the first mortgage and the corresponding risk-weighted average of covariate values among all the mortgages present in the risk set at time $t_1$.

4. Let $\hat{V}_1$ be the estimated covariance matrix of the covariates at time $t_1$. The dimension of $\hat{V}_1$ is $p \times p$, where $p$ is the number of covariates. On the diagonal are the variance estimates of covariates $i$ for $i = 1, \ldots, p$. Then, $\hat{V}_{1,(i,i)}$ is the variance estimate of covariate $i$ at event time $t_1$. Scale $\hat{r}_{1i}$ by multiplying it with the inverse of a variance estimate of the covariate $i$ for the loans still included in the risk set, which is denoted by $\hat{V}_{1,(i,i)}$. As the risk set becomes smaller over time, the variance of the covariates at each time may change over time. However, Grambsch and Therneau (1994) concluded that this covariance matrix varies slowly over time and is quite stable until the last few event times. Therefore, they used the average covariance matrix of the covariates, which is denoted by $\overline{V} = mean_j\{\hat{V}_j\}$. This yields the scaled Schoenfeld residual $r^*_{1i} = \overline{V}^{-1}_{(i,i)} \hat{r}_{1i}$.

5. This process is repeated for each $t_j$. Grambsch and Therneau (1994) showed that the association from the scaled Schoenfeld residuals with the time-dependence of the Cox proportional hazards regression coefficient. Let $\beta_i(t_j)$ be the value of the time-varying coefficient of covariate $i$ on time $t_j$. Grambsch and Therneau (1994) defined it as:

$$\beta_i(t_j) = \beta_i + \theta_i g_i(t_j), \tag{28}$$

   where $g_i(t_j)$ is a predictable process. So, this intuitively means that a weighted function of time is added to the time-fixed covariate to obtain the time-varying covariate. They showed the following relation:

$$E(r^*_{ji}) + \hat{\beta}_i \approx \beta_i(t_j) \tag{29}$$

   Under the proportionality assumption, the coefficient does not change over time and hence $\beta_i(t_j) = \hat{\beta}_i$, which implies $E(r^*_{ji}) = 0$. So, if we plot $r^*_{ji} + \hat{\beta}_i$ against $t_j$, $r^*_{ji} + \hat{\beta}_i$ should follow a random walk with a constant mean over time in order for the covariate to satisfy the proportional hazards assumption.

With this procedure, a Schoenfeld residual is obtained for every covariate for all points in time where a prepayment occurred. If the Cox proportional hazards assumption holds for a covariate, then the effect of that covariate on the hazard is constant over time. If the scaled Schoenfeld residual is consistently higher/lower over a time interval, then this is evidence that the hazard at that time is higher/lower than implied by the model. This would suggest that the proportional hazards assumption is not valid for that covariate. This can be tested using the zph test developed by Grambsch and Therneau (1994). Using the distribution of $\hat{\theta}$, the asymptotic $\chi^2$ test statistic with $p$ degrees of freedom can be derived and it is tested $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. For further details, see Grambsch and Therneau (1994). As this approach requires the specification of the form of $g_i(t_j)$, next to the zph test, also the graphical displays of the scaled Schoenfeld residuals are evaluated.

## 3.3 Competing risks

There are two main sources of risk when providing a loan. These are the risk of prepayment and the risk of default. These risks are competing events as one cannot occur if the other has occurred. So, both risks compete as to which will occur first (Li et al., 2023). There are two main modelling approaches used in the literature. One is the Cox cause-specific hazards model. In this model, the hazard of each competing event is modelled separately using a Cox proportional hazard model. It treats the competing event, in this case mortgage default, as censored. A second approach is the subdistribution hazard function introduced by Fine and Gray (1999). From now on, $m$ denotes the number of distinct event times (prepayment or default) instead of distinct prepayment times, as we now deal with competing risks. So, $t_j$ denotes now the time the $j$-th mortgage experiences a prepayment or default.

### 3.3.1 Cox cause-specific hazards model

A cause-specific hazards model estimates the hazard function for prepayments and defaults separately, while censoring the other event (Columbia University Irving Medical Center, 2023). This indicates that it is assumed that prepayment and defaults are independent of each other. The cause-specific hazard rate for prepayment is defined as:

$$h_k^P(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T_k \leq t + \Delta t, \mathcal{S} = P | T_k > t)}{\Delta t}, \tag{30}$$

where $\mathcal{S}$ is the reason of mortgage ending and $P$ meaning prepayment. For the discrete time setting, the cause-specific hazard is defined as (Lau et al., 2009):

$$h_k^P(t) = \mathbb{P}(T_k = t, \mathcal{S}_k = P | T_k > t - 1) \tag{31}$$

For default, the cause-specific hazard rate is:

$$h_k^D(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T_k \leq t + \Delta t, \mathcal{S}_k = D | T_k > t)}{\Delta t}, \tag{32}$$

where $D$ is default.
Using the cause-specific hazards for both competing events and the assumption that

prepayments and defaults are independent events, the survival function for both termination events together, defined as $S_k^{CR}(t)$, can be written as:

$$S_k^{CR}(t) = exp(-\int_0^t h_k^P(s)\,ds - \int_0^t h_k^D(s)\,ds) \tag{33}$$

The CIF in a cause-specific hazards model is defined as the probability of prepayment by time $t$ while still at risk of being defaulted (Haushona et al., 2020). If we model the cause-specific hazards separately, as in a cause-specific hazards model, then the cumulative incidence function is defined for each risk. For prepayment, it is calculated as:

$$CIF_k^P(t) = \int_0^t h_k^P(s) S_k^{CR}(s)\,ds, \tag{34}$$

and for defaults it is calculated as:

$$CIF_k^D(t) = \int_0^t h_k^D(s) S_k^{CR}(s)\,ds, \tag{35}$$

Expression (34) is derived in Appendix A.6. So, the survival function contains the cumulative hazard function of both risks. The intuition behind this, is that mortgages should not be prepaid and not be defaulted until time $t$ (Kohl et al., 2015). If the cause-specific hazards rates are modelled using a Cox proportional hazards model as described (8), the Cox cause-specific hazards model is considered. For further details, see Ozenne et al. (2017).

Using the cause-specific cumulative incidence function, we cannot interpret the coefficient as the effect of the covariate on the CIF because of the relations as stated in (34) and (35) (Lambert, 2017). Plugging expression (8) and (33) into (34) can give an effect of a covariate on the cumulative incidence function, but calculations become rather complex.

Since the variable of interest is prepayment, defaults are treated as censored in addition to the usual right-censored observations. This means it is assumed that defaults are non-informative for prepayments. Or in other words, defaults and prepayments are independent of each other. One disadvantage is that this assumption cannot be tested directly. However, using sensitivity analysis, it can be validated whether the assumption is reasonable (Kleinbaum and Klein, 2011).

**Sensitivity analysis in the competing risk model**

It can be assessed what the effect is of the independence of competing risks assumption on the estimated coefficients using sensitivity analysis. Extreme ranges for estimated coefficients in a model can be determined under the violation of the independence assumption. Sensitivity analysis estimates the coefficients by considering worst-case violations of the independence assumption. There are two worst-case situations:

1. All mortgages that are censored due to default are assumed to be prepaid at the time of being censored.

2. All mortgages that are censored due to default are assumed to be prepaid at the largest observed time to event of prepayment

Then, the coefficients of the model in case of these two worst-case situations are compared to the original estimation as described in (30). If results are not significantly different, than we may conclude that at most a small bias can result from the assumption of independence. If there is a significant difference from the original estimation, we learn the extremes to which the results could be biased if the independence assumption is not satisfied (Kleinbaum and Klein, 2011).

**Variable selection in the Cox Cause-specific hazards model**

Best subset selection could be used, which compares all combinations of variables to search which combination results in the best model fit (James et al., 2021). However, this is computational costly and therefore another method is used.

Variable selection in the Cox cause-specific hazards model is performed using a combination of forward and backward selection based on the Akaike Information Criterion (AIC). Using forward or backward selection does not guarantee to find the best possible model out of the $2^p$ possible models. For example, it could be that removing two variables $X_1$ and $X_2$ at once improves the model fit more than removing one other variable $X_3$. Though, a stepwise selection procedure as backward selection would remove $X_3$ (James et al., 2021). Therefore, a combination of forward and backward selection is used. This attempts to mimic more closely best subset selection, while keeping the computational cost low. Let $p$ be the number of parameters and $L(\hat{\beta})$ be the partial log likelihood. The criterion is as follows:

$$AIC(p) = -2log(L(\hat{\beta})) + 2p \tag{36}$$

Another criterion that is often used is the Bayesian Information Criterion (BIC), which is as follows:

$$BIC(p) = -2log(L(\hat{\beta})) + p * log(n)), \tag{37}$$

where $n$ is the number of observations. While the BIC chooses the correct model as $n \to \infty$, the AIC is preferred if the goal is prediction, since it selects a less parsimonious model (Hastie et al., 2009). Therefore, the AIC is used in this thesis to select variables to be included in the model. Both prepayments and defaults have a hazard rate and this variable selection procedure picks the variables that are most informative for the corresponding hazard. So, for each of the two hazards, this procedure is used and hence different variables could be selected. It consists of the following steps as described by Szolnoki (2021):

1. Start with the full model. That is, all covariates are included. Obtain the AIC.

2. Estimate all the models with removing or adding one variable and obtain the AIC for each model. (Note that for the full model, one cannot add a variable).

3. Compare the obtained AICs from 2. with the AIC from 1. and select the model with the lowest AIC value.

4. Stop if the AIC cannot be further reduced by removing or adding a covariate. Use this model as the final model.

One disadvantage of this procedure is that the effect of a covariate on the cumulative incidence function is not clear in the Cox cause-specific hazards model as mentioned before and therefore it could be that selected variables are informative for the cause-specific hazard, but not for the cumulative incidence (Olajubu, 2020).

### 3.3.2 Fine-Gray subdistribution hazard model

Fine and Gray (1999) models the cumulative incidence function with covariates directly by treating the cumulative incidence function (CIF) as a subdistribution function. The subdistribution function models the hazard function derived from a CIF. It is the instantaneous rate of occurrence of prepayment at time $t$ for mortgages that are event free or defaulted before time $t$ (Haushona et al., 2020). The model accounts for this by adjusting the risk set. So, still the estimation procedure is used as in (16), but now the risk set $R(t_j)$ also contains the mortgages that have defaulted before time $t_j$. For prepayment as event, we get:

$$\lambda_k^P(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T_k \leq t + \Delta t, \mathcal{S}_k = P | T_k > t \cup (T_k < t, \mathcal{S}_k \neq P))}{\Delta t} \tag{38}$$

This function estimates the hazard rate for prepayment at time $t$ using the risk set that remains at time $t$ after it accounts for all previous prepayments and defaults. For a discrete time setting, the subdistribution hazard is defined as (Lau et al., 2009):

$$\lambda_k^P(t) = \mathbb{P}(T_k = t, \mathcal{S}_k = P | T_k \geq t \cup (T_k < t, \mathcal{S}_k \neq P)) \tag{39}$$

The risk set is different for the Fine-Gray subdistribution hazard model than for the Cox cause-specific hazards model. Namely, in the Fine-Gray model, a mortgage that defaults, remains in the risk set for prepayment and the mortgages are given a censoring time larger than all event times. On the other hand, the risk set for prepayments in the Cox cause-specific hazards model decreases each time a default occurs (Hinchlie, 2012). While the Cox cause-specific hazards model complicates the interpretation of covariates on the cumulative incidence function, the subdistribution function used by Fine and Gray allows for straightforward interpretation. However, this comes at the cost of the rather counter-intuitive risk set. The model estimates the effect of covariates on this subdistribution hazard function via:

$$\lambda_k^P(t) = \lambda_0^P(t) exp(\sum_{i=1}^{p} \beta_i X_{ki}), \tag{40}$$

where $\lambda_0^P(t)$ denotes the baseline subdistribution hazard function for prepayments (Austin et al., 2019). Note that the model has the same form as the Cox proportional hazards model as defined in (8), but now modelling the subdistribution hazard instead of the hazard. If only time-invariant covariates are included, there is a one-to-one relation between the subdistribution hazard function and the cumulative incidence function. Therefore, the subdistribution hazard model enables one to estimate the direction of the effect of covariates on the CIF. Define the baseline CIF as

$CIF^P_{k0}(t) = \mathbb{P}(T_k \leq t, \mathcal{S} = P | X_{ki} = 0 \ \forall i)$. The relation between the subdistribution hazard function and the CIF is derived in the same way as (10) using the same mathematical steps as in Appendix A.2. It is given by:

$$1 - CIF^P_k(t|X) = (1 - CIF^P_{k0}(t))^{exp(\sum_{i=1}^p \beta_i X_{ki})} \tag{41}$$

If the a covariate is associated with an increase in the subdistribution hazard of prepayment, it is also associated with an increase in the cumulative incidence of prepayment. Consequently, the direction of the effect of a covariate on the CIF can be interpreted if only time-invariant covariates are used (Austin et al., 2019). The proof of this claim can be found in Appendix A.7.

**Time-varying covariates**

First, one should note that there are two types of time-varying covariates, which are the internal and the external time-varying covariates. Internal time-varying covariates are covariates observed over time for each individual mortgage, which is for example marital status. External time-varying covariates consist of two subgroups. One of the subgroups, called the external ancillary time-varying covariates, is our main focus. These covariates are defined as the result of a stochastic process external to the subject. For example, the current 30-year yields from government bonds and the current unemployment rate. The subdistribution hazard model of Fine and Gray can only incorporate time-varying covariates that are externally defined. Since mortgages that defaulted remain in the risk set for prepayments, time-varying covariates have to be observed for the mortgage after the default happened. In the data set, internal time-varying covariates of mortgages that defaulted are not observed after default and as a result cannot be considered in the risk set of prepayment (Austin et al., 2019). Since only external time-varying covariates are considered, the relationship between the subdistribution hazard and the cumulative incidence function still holds. Let $\Lambda_k(t) = \int_0^t \lambda^P_k(s)$ be the cumulative subdistribution hazard function. The relationship between the subdistribution hazard and the CIF now becomes:

$$CIF^P_k(t|X(s), s \leq t) = 1 - exp(-\Lambda_k(t|X(s), s \leq t)) \tag{42}$$

$$= 1 - exp(-\int_0^t \lambda^P_0(s) exp(\sum_{i=1}^p \beta_i X_{ki}(s))\, ds) \tag{43}$$

Since one cannot bring $exp(\sum_{i=1}^p \beta_i X_{ki}(s))$ outside of the integral, one cannot interpret anymore that a covariate that has an effect on the subdistribution hazard of prepayments, has an effect in the same direction on the CIF of prepayments.

REMARK: Austin et al. (2019) reviewed the interpretation of time-varying covariates in the Fine-Gray model of papers published in 2015 and in the first five months of 2019. They concluded that of the 11 studies all but 2 wrongfully used at least one internal time-varying covariate and that all papers did not clearly explain how they used time-varying covariates. Moreover, they found that 6 out of the 11 suggested that the time-varying covariate was associated with the risk of an event. Also, the papers in the literature contradict each other about whether or not (external) time-varying covariates are allowed in the Fine-Gray model, and hence I question the use of time-varying

covariates. Moreover, for prediction of the CIF with time-varying covariates, predicting the evolution of the macroeconomic variables would be required. After reaching out to the author of one of the reports (Therneau et al., 2024) that suggested that it is allowed to use external time-varying covariates and he told me that he questioned his own paper and needed to review it, I decided to not progress with external time-varying covariates.

**Variable selection in the Fine-Gray model**

In the Fine-Gray subdistribution hazard model, variables are selected using backward selection. While one could use the AIC or BIC as criterion, in this paper, the BICcr criterion is used as proposed by Kuk and Varadhan (2013). In this criterion, the penalty uses the total number of prepayments instead of the total number of mortgages as in AIC and BIC. Their approach is based on Volinsky and Raftery (2000), who used the number of uncensored events instead of the number of observations in the BIC for variable selection in the Cox proportional hazards model. Their intuition behind this was that there are only as many terms in the partial likelihood function as there are uncensored events, which can be seen in (16). So, only the mortgages that are prepaid contribute to the partial likelihood of the Fine-Gray model. They concluded that using the number of uncensored events in the BIC improved the criterion. Moreover, the asymptotic properties are still valid. Kuk and Varadhan (2013) translated this idea to the competing risk setting. Define $n^*$ as the total number of full prepayments. The BICcr is then defined as follows:

$$BICcr = -2log(L(\hat{\beta})) + p * log(n^*) \tag{44}$$

The BICcr is a more stringent penalty than the AIC, which results in a more parsimonious model than using AIC. On the other hand, the BICcr has a less stringent penalty than BIC. So, the number of parameters selected by BICcr, will be between the number of parameters selected by AIC and BIC (Kuk and Varadhan, 2013).

## 3.4 Machine Learning: Tree based methods

To understand tree based methods, this section briefly explains the relatively easy classification tree model that predicts the classification of mortgages. That is, it predicts whether a mortgages is prepaid or not rather than predicting a cumulative incidence function. The prediction of the cumulative incidence function is described in Section 3.4.1.

Tree based methods stratify the predictor space into a finite number of non-overlapping regions. For observations that fall into the same region, the same prediction is made. That is, the mean of prepayment rate for the training observations in that region. A tree is grown starting at the top of the tree, which is called the root node. This root node is split into two daughter nodes, the left-hand node and the right-hand node. The splitting is based on recursive binary splitting. This process is repeated for each node until the number of observations in the last node, called the terminal node, would drop below a threshold (Frydman and Matuszyk, 2022). So, the data is subdivided into regions based on a splitting criteria. Tree based methods make use of recursive binary splitting rather than finding the optimal regions as this would be too computational

costly. As a result, classification trees often fail to make good predictions. However, combining multiple trees using a so-called random forest, significantly improves predictions (James et al., 2021). Therefore, a random forest will be used in this thesis, which will be discussed now.

**Bootstrapping and Random Forest**

If we would just split the data into one training set and one test set, the splitting would directly affect the estimation and prediction. By increasing the number of training samples, the effect of the split reduces. As we generally have one data set, the widely used method called bootstrapping can be used to obtain multiple training sets. A bootstrap sample is a random sample of the original data set with replacement. The size of the bootstrap sample is equal to the size of the original training data set. Or in other words, one observation of the original data set can occur multiple times in the bootstrap sample. Sampling with replacement leads to larger random samples which reduces bias. Moreover, sampling with replacement also reduces the variance of the Random Forest and prevents overfitting (Lyu, 2021). For a large sample, on average 63% of the original observations will occur one or more times in the bootstrap sample, whereas the other 37% of the observations will not occur (Weathers, 2017). The proof of this statement is stated in Appendix A.8. As these are random samples with replacement of the original data set, the trees grown on the bootstrap samples are likely to be very correlated. The random forest mitigates this by taking a random subsample of the predictors at each split when building a tree. As a rule of thumb, this subsample includes $m = \lfloor \sqrt{p} \rfloor$ predictors for classification trees and $m = \lfloor \frac{p}{3} \rfloor$ for regression trees as concluded by Breiman (2001). Taking a random split of the predictors doesn't allow the strong predictors to always be the first split and hence less strong predictors get a chance. Each tree makes a prediction and as last step in the random forest, the average of the predictions of all grown trees is taken to get the final prediction. The trees are grown deep and are not pruned, which results in high variance. Averaging trees results in lower variance of the prediction while keeping the bias of the prediction low (James et al., 2021).

### 3.4.1 Random survival forest for competing risks

The random survival forest (RSF) is an extension to the random forest discussed in the previous section. While the random forest is used to make predictions for classification and regression, the random survival forest for competing risks can predict the survival curve as well as the cumulative incidence function in the presence of competing risks. The idea of the RSF is still to build a tree on each bootstrap sample and then decorrelate the trees by randomly selecting $\lfloor \sqrt{p} \rfloor$ predictors at each split. However, there are several choices of splitting rules and they involve splitting based on hazard functions or cumulative incidence functions instead of survival functions. The mostly used split criterion for the RSF is the log-rank test statistic. At each node, the $\lfloor \sqrt{p} \rfloor$ randomly selected predictors are considered and the best split is selected as the maximum of the log-rank statistic over all possible split points over all $\lfloor \sqrt{p} \rfloor$ predictors (Wright et al., 2017). Important remark is that although $\lfloor \sqrt{p} \rfloor$ predictors are randomly selected at each node, the eventual split is performed on one of these variables and

not on multiple variables simultaneously. A second splitting rule used in the presence of competing risks, is a modification of Gray's test (Gray, 1988). Both splitting rules are described and used in this thesis, but first, in the next section, it is explained how the CIF is estimated in the RSF for competing risks and how tied events are dealt with.

**Estimation of the ensemble CIF**

Let $B$ denote the number of bootstrap samples, which thus also equals the number of trees. Let $c_{k,b}$ be the number of times mortgage $k$ occurs in bootstrap sample $b$. Observe the covariates of mortgage $k$, $X_k$, and follow the splitting criteria through the tree to end up in a terminal node and to be able to define the CIF of mortgage $k$ for the $b$-th tree. Define $h_b(X_k)$ as the indices of the mortgages that are in the same terminal node as mortgage $k$ in the bootstrap training sample $b$. Let the node-specific number of prepayments at time $t_j$ in bootstrap sample $b$ be denoted by $N_b^P(t_j|X_k) = \sum_{k\in h_b(X_k)} c_{k,b}\mathbb{1}_{\{T_k=t_j,\mathcal{S}=P\}}$ and let the number of mortgages at risk at time $t_j$ in the terminal node be denoted by $Y_b(t_j|X_k) = \sum_{k\in h_b(X_k)} c_{k,b}\mathbb{1}_{\{T_k\geq t_j\}}$. Let the survival function of all mortgages that are in the same terminal node as mortgage $k$ be $\hat{S}_b(t_j|X_k) = \prod_{u\leq t_j}(1 - \sum_{s\in\{P,D\}} \frac{N_b^s(u|X_k)}{Y_b(u|X_k)})$, which is the Kaplan-Meier estimator. Let the CIF estimate at time $t_j$ of the terminal node containing mortgage $k$ then be defined as:

$$\widehat{CIF}_b^P(t_j|X_k) = \sum_{u=1}^{j} \hat{S}_b(t_{u-1}|X_k)Y_b(t_u|X_k)^{-1}N_b^P(t_u|X_k), \tag{45}$$

which is the Aalen-Johansen estimator of the cumulative incidence function. While Ishwaran et al. (2014) used notation for continuous time, (45) is a discrete representation.

Thus, at the terminal node we look how many mortgages experience cause $s \in \mathcal{S}$, which is divided by the total number of mortgages at risk in that terminal node. This is multiplied with the survival function. Intuitively this can be motivated by that mortgages should have survived until that point in time, or in other words, mortgages should be event-free until that point in time.

The ensemble estimate of the CIF is then obtained by averaging over the trees and hence, for each $j = 1, \ldots, m$, the ensemble estimate of the CIF at time $t_j$ is:

$$\overline{CIF}^P(t_j|X_k) = \frac{1}{B}\sum_{b=1}^{B} \widehat{CIF}_b^P(t_j|X_k), \tag{46}$$

where $B$ is the number of trees. So, given the covariates of the mortgage $k = 1, \ldots, n$, the CIF estimate of mortgage $k$ within each bootstrap sample is averaged to get the ensemble estimate of the CIF for mortgage $k$. The cause-$\mathcal{S}$ termination is equal to:

$$\overline{M}^{\mathcal{S}}(\tau|x) = \int_0^\tau \overline{CIF}^{\mathcal{S}}(t|x)\,dt = \frac{1}{B}\sum_{b=1}^{B} \hat{M}_b^{\mathcal{S}}(\tau|x) \tag{47}$$

(Ishwaran et al., 2014)

### Log-rank splitting criteria

While the log-rank test statistic was originally used for two-sample testing with survival data, it can be used as splitting criteria in a random survival forest with competing risks. In this section, the notation of Ishwaran et al. (2021) is followed. First, consider a node to be split. Let $X_i$ be the value of covariate $i$. A splitting then looks like $L : X_i \leq c$ and $R : X_i > c$ if $X_i$ is a continuous predictor, where $L$ is the left-hand branch and $R$ denotes the right-hand branch. If $X_i$ is a binary predictor, then a splitting looks like $L : X_i = 1$ and $R : X_i = 0$. Let the distinct event times be denoted by $t_1 < t_2 < \cdots < t_m$ as defined before. Let $d_{L,i}(t_j)$ and $d_{R,i}(t_j)$ equal the number of prepayments at time $t_j$ in branches $L$ and $R$ respectively based on splitting on covariate $i$. Let $Y_{L,i}(t_j)$ and $Y_{R,i}(t_j)$ be the number mortgages at risk at time $t_j$ in branches $L$ and $R$ respectively based on splitting on covariate $i$. Then:

$$Y_{L,i}(t_j) = \sum_{k=1}^{n} \mathbb{1}_{\{T_k \geq t_j, X_{ki} \leq c\}} \tag{48}$$

and

$$Y_{R,i}(t_j) = \sum_{k=1}^{n} \mathbb{1}_{\{T_k \geq t_j, X_{ki} > c\}} \tag{49}$$

Let $Y(t_j) = Y_{L,i}(t_j) + Y_{R,i}(t_j)$ and $d(t_j) = d_{L,i}(t_j) + d_{R,i}(t_j)$. The log-rank split-statistic value for the split is then:

$$L(i,c) = \frac{\sum_{j=1}^{m}(d_{L,i}(t_j) - Y_{L,i}(t_j)\frac{d(t_j)}{Y(t_j)})}{\sqrt{\sum_{j=1}^{m}\frac{Y_{L,i}(t_j)}{Y(t_j)}(1-\frac{Y_{L,i}(t_j)}{Y(t_j)})(\frac{Y(t_j)-d(t_j)}{Y(t_j)-1})d(t_j)}} \tag{50}$$

We then get the best split $(i,c)^*$ for the node by taking the maximum over $i$ and $c$ of the absolute value of the log-rank split-statistic:

$$(i,c)^* = \arg\max_{i,c} |L(i,c)| \tag{51}$$

Intuitively, this means that the obtained $(i,c)^*$ are the covariate $i$ and the split that maximize the difference between the cause-specific Nelson-Aalen estimates in the resulting branches of the node (James et al., 2021). This splitting procedure is particularly useful if the main purpose is to detect variables that affect the cause-specific hazard of prepayments (Ishwaran et al., 2014; Weathers, 2017).

### Modified Gray's splitting rule

The splitting rule as described in (50) may not be optimal if the purpose is to predict cumulative event probabilities as the splitting rule is based on maximizing the difference in cause-specific hazards rather than maximizing the difference in cumulative incidence. If the purpose is prediction of cumulative incidences, it may be better to apply a splitting rule that select variables based on their direct effect on the CIF. The modified Gray's splitting rule does this. The test modifies the risk set used in the

log-rank splitting criteria. Now, $Y_j$ is replaced by:

$$Y_j^* = \sum_{k=1}^{n} \mathbb{1}_{\{T_k \geq t_j \cup (T_k < t_j \cap \mathcal{S}=D)\}} \tag{52}$$

So, the risk set consists of mortgages that are not prepaid or defaulted before time $t_j$ and mortgages that are defaulted before time $t_j$, but which are not censored (Ishwaran et al., 2014).

### 3.4.2 Out-of-bag estimate of prediction error

Out-of-bag (OOB) observations are the loans that are left out due to bootstrap sampling with replacement. In bag observations are the loans that are included in the bootstrap sample and are used for growing the tree. One can evaluate the grown forest performance using prediction on the OOB loans. To be able to define the out-of-bag estimate of the prediction error, we first have to define the event-specific cumulative hazard function. Let $t_{1,X_k} < t_{2,X_k} < \cdots < t_{m(X_k),X_k}$ be the unique event times in the terminal node that contains $X_k$. The number of event times $m$ depends on the node, which is denoted by $m(X_k)$. The prepayment-specific cumulative hazard function in the terminal node of $X_k$ in tree $b$ is defined as:

$$H_b^P(t_j|X_k) = \sum_{t_{u,X_k} \leq t_j} \frac{N_b^P(t_u|X_k)}{Y_b(t_u|X_k)} \tag{53}$$

So, it is the cumulative sum of the hazards until time $t_j$. Let $O_k$ be trees where loan $k$ is OOB. Then, for $k = 1, \ldots, n$, the OOB ensemble estimator of the cumulative hazard of prepayment is:

$$\overline{H}_k^{P,OOB}(t_j) = \frac{1}{|O_k|} \sum_{b \in O_k} H_b^P(t_j|X_k) \tag{54}$$

So this intuitively means that loan $k$ is dropped down the trees that did not use mortgage $k$ when they were grown, and average the estimated cumulative hazard from these trees. Then, for $k = 1, \ldots, n$, the OOB ensemble prepayment cause termination is equal to:

$$\overline{M}_k^{P,OOB} = \sum_{j=1}^{m} \overline{H}_k^{P,OOB}(t_j) \tag{55}$$

Then, mortgage $k$ is said to have a worse outcome than mortgage $k'$ if:

$$\overline{M}_k^{P,OOB} > \overline{M}_{k'}^{P,OOB} \tag{56}$$

Harrell's C-index (Harrell, 2001) is used as the OOB estimate of the prediction error and is calculated using the following steps as described in Ishwaran et al. (2021):

1. Take all pairs of loans over all the data
2. Omit pairs where shorter event time is censored and also if both loans have the same time until censoring. The set of the remaining loans is defined as $\mathcal{L}$

3. If $T_k \neq T_{k'}$, count 1 for each $l \in \mathcal{L}$ in which the shorter time had the worse predicted outcome

4. If $T_k \neq T_{k'}$, count 0.5 for each $l \in \mathcal{L}$ in which $\overline{M}_k^{P,OOB} = \overline{M}_{k'}^{P,OOB}$

5. If $T_k = T_{k'}$, count 1 for each $l \in \mathcal{L}$ in which $\overline{M}_k^{P,OOB} = \overline{M}_{k'}^{P,OOB}$

6. If $T_k = T_{k'}$, count 0.5 for each $l \in \mathcal{L}$ in which $\overline{M}_k^{P,OOB} \neq \overline{M}_{k'}^{P,OOB}$

7. Let concordance index C be:

$$C = \frac{\textit{Sum over all counts}}{|\mathcal{L}|} \tag{57}$$

8. The OOB prediction error is equal to $1 - C$

If the prediction error equals 0.5, then we do no better than random guessing. If the prediction error equals 0, this indicates perfect prediction (Ishwaran et al., 2021).

### 3.4.3 Variable selection in Random Survival Forest for competing risks

Variable are selected based on their variable importance (VIMP) and minimal depth. Both methods measure the importance of variables in the model. A combination of both will be used in this thesis.

#### Variable importance (VIMP)

The VIMP measures the increase in prediction error when a variable is randomly "noised-up". Ishwaran and Lu (2018) introduced a procedure to estimate confidence intervals for VIMP, which used the calculation of VIMP as described in Ishwaran (2007). It is calculated using the following steps, which are repeated for each covariate $X_i$:

1. Drop each OOB loan down the in-bag tree. If the covariate $X_i$ is splitted, assign a daughter node randomly (so with equal probability one of the daughter nodes is selected). Then, at every split afterwards, assign a daughter node with equal probability until a terminal node is reached.

2. The prepayment specific CIF is calculated with the "noised-up" variable as being the value at the terminal node

3. The VIMP for covariate $X_i$ is the difference between the prediction error for the original prepayment specific CIF (the OOB observations are dropped down the in bag tree without randomization) and the prediction error for the new ensemble obtained using randomizing $X_i$ assignments.

4. This process is repeated for each tree to obtain a distribution of VIMP across the trees. The ensemble VIMP is the mean VIMP across the trees.

The idea behind the VIMP is that the difference in prediction performance is directly effected by the location in the tree of the "noised-up" variable. If the variable contains more information about prepayment risk, then it is higher op the tree and consequently, the more perturbed the noised-up prediction becomes relative to the original tree-specific prediction. Predictors with a high VIMP value have predictive ability, whereas predictors with zero or negative VIMP have no predicative ability (Ishwaran, 2007).

#### Minimal depth

Minimal depth as described by Ishwaran et al. (2010) assesses the forest construction to rank variables. Variables with high impact on predictions are assumed to be more frequently near to root node. Minimal depth numbers the node levels based on their distance to the root node. Important to mention is that if a tree is split multiple times into two branches based on the same variable, the minimal depth is calculated for the split of that variable that is nearest to the root node of the tree. The depth of the first split for each variable is averaged over all trees. So, variables with low minimal depth are variables with high predictive ability.

#### Variable selection criterion

Ishwaran and Lu (2018) used a subsampling approach to derive standard errors and confidence intervals for VIMP in the random survival forest. If for a variable the VIMP equal to zero is in the 95% confidence interval, then this implies there is no evidence that the variable contributes to the prediction accuracy of the random survival forest and consequently, the variable is removed from the data. However, if the minimal depth of the variable is low, then the variable will not be removed.

### 3.4.4 Hyperparameter tuning

When the random survival forest for competing risks is used, there are parameters that need to be set when growing the trees. These parameters are called hyperparameters. The hyperparameters that need to be chosen are the number of trees to build, the minimum number of observations in a terminal node, the number of chosen variables at each split, the number of splits at each node and the maximum depth to which a tree is grown. If for example the maximum depth of a tree is too high, the model will be overfitted leading to high variance. On the other hand, if a tree is not grown deep enough, the model is not fitted well leading to high bias (GeeksforGeeks.org, 2022). As the values of the hyperparameters need to be chosen before fitting the model, it is up to the researcher to assess what these values should be.

To reduce the effect of the hyperparameter value choice of the researcher, hyperparameter tuning is used. Different methods are available to select the set of hyperparameters. One of these methods introduces an additional sample, called validation sample. Then, a large grid for each hyperparameter is taken. Using an optimization function, the combination of hyperparameters is chosen that has the lowest prediction error on the validation sample. Then, this set of hyperparameters is used to fit the random sur-

vival forest for competing risks on the training set and evaluate the performance on the test sample. One cannot choose the hyperparameters that results in the best predictions on the test set as this would lead to fitting the model to the test set (Jordan, 2017).

In this thesis a modification of the previously mentioned strategy is used to obtain the values of the hyperparameters. A modification is performed since the described strategy is computationally costly and can be inefficient. Rather than introducing a validation sample, the hyperparameters are chosen based on model performance on the OOB loans. This is possible since the OOB loans are independent of the training set. The procedure used in this thesis to obtain the hyperparameters is the following:

1. Set the number of splits at each node equal to 2 to avoid biasing splits towards continuous variables (Loh and Shih, 1997; Ishwaran et al., 2010). Set the number of variable chosen at each split be equal to $\lceil \sqrt{p} \rceil$ as this is the universally accepted rule of thumb (James et al., 2021). Note that we now round up the number to the nearest integer greater than the square root of $p$ instead of rounding down to the nearest integer as described before. This is decided because of the low number of variables available in the data set.

2. Specify a small grid for the minimum number of observations in a terminal node and the number of trees grown. This small grid is based on the current literature and lowers the computational effort.

3. Set the value of minimum number of observations equal to the default value of 15 and minimize the OOB prediction error, defined as 1 - C, where C is defined in (57), over the grid of number of trees.

4. Fix the number of trees to the obtained number of trees as calculated in the previous step and minimize the OOB prediction error over the grid of minimum number of observations in the terminal node.

## 3.5 Current VB Advisory prepayment model

Currently, VB works with a model that neither takes into account covariates nor competing risks. They multiply their anticipated interest rate income with the cumulative incidence of prepayment to correct for interest rate income lost due to early repayments. Prepayment is here defined as the percentage of total portfolio value that is prepaid in a specific period. This also includes partial prepayments, but since partial prepayments are very rare in practice and as it is relatively low amount in euros compared to the full prepayments, this does not heavily influence the percentage. It is assumed that time of prepayment $T_k$ is exponentially distributed with parameter $\gamma$, where $\gamma$ is the last 12-month average percentage of prepayments. So, the parameter $\gamma$ can be seen as a 12-month moving average of prepayment percentage. Then, their model for the cumulative incidence function (CIF) looks as follows:

$$CIF_k^P(t) = \mathbb{P}(T_k \leq t, \mathcal{S} = P) = 1 - exp(-\gamma * t) \tag{58}$$

## 3.6 Prediction performance evaluation methods

To be able to evaluate which model performs better, different evaluation methods can be used. As the goal of the thesis is to predict the cumulative incidence of prepayment over time, two evaluation methods are introduced that evaluate the prediction performance of the models. This enables to compare the different models regarding prediction accuracy. The methods used in this thesis are the integrated time-dependent Brier score and the time-dependent area under the ROC curve. For both methods, the score() function in the riskRegression package of R was used.

### 3.6.1 Integrated time-dependent Brier score

A first evaluation method for the prediction performance is the integrated time-dependent Brier score (IBS). It is the mean squared error between the predicted cumulative incidence function and the indicator that the mortgage is prepaid, integrated over time.

Gerds and Schumacher (2006) introduced a consistent estimation of the Brier score with right-censored data by using Inverse Probability of Censoring Weighting to weight observations, which is applicable if competing risks are present. The idea is the give observations that are not censored more weight than observations that are censored. First, we start by defining the estimate of the censoring distribution to be the Kaplan-Meier estimate. Let $r_{j'}$ be the number of censored observations at time $t_{j'}$. Then the estimate of the censoring distribution is $\hat{G}(t) = \prod_{j':t_{j'} \leq t}(1 - \frac{r_{j'}}{|R(t_{j'})|})$. And as we will later, in (60), plug in time $t_j$ into $\hat{G}(t)$, $j'$ is used in the definition of $\hat{G}(t)$ to make a clear distinction between indices and input. This gives a better view how $\hat{G}(t)$ is calculated if $t_j$ is used as input. Competing events are treated as censored. Inverse Probability of Censoring Weighting (IPCW) for each mortgage $k$ at time $t$ is defined as (Ishwaran et al., 2014):

$$\hat{W}_k(t) = \frac{\mathbb{1}\{T_k \leq t, \mathcal{S} \neq 0\}}{\hat{G}(T_k)} + \frac{\mathbb{1}\{T_k > t\}}{\hat{G}(t)} \tag{59}$$

Then, for $j = 1, \ldots, m$, the Brier score at time $t_j$ for mortgage $k$ is defined as:

$$BS_k^P(t_j) = \begin{cases} (\widehat{CIF_k}^P(t_j) - 1)^2 & \text{if } T_k \leq t_j \text{ and event is prepayment} \\ (\widehat{CIF_k}^P(t_j))^2 & \text{if } T_k > t_j \text{ or event is of other type} \end{cases}$$

Intuitively this means that loans that are prepaid by time $t_j$ have a prediction error equal to their squared distance of cumulative incidence prediction and 1. On the other hand, we would like the model to predict a cumulative incidence equal to zero at time $t_j$ for loans that are not prepaid at or before time $t_j$. So, if loans are not prepaid at or before time $t_j$, they have a prediction error equal to the squared distance between their predicted cumulative incidence at time $t_j$ and zero.

For $j = 1, \ldots, m$, the Brier score at time $t_j$ is calculated by using the weighted average with weights being the IPCW (Ishwaran et al., 2014; Frydman and Matuszyk, 2022):

$$BS^P(t_j) = \frac{1}{n} \sum_{k=1}^{n} \hat{W}_k(t_j)((\mathbb{1}\{T_k \leq t_j, \mathcal{S} = P\} - \widehat{CIF_k}^P(t_j))^2 \tag{60}$$

IPCW increases the weights of the observations that have an observed prepayment. If for example we estimate that 1/4 of the mortgages have censoring times greater than 12 months and mortgage $k$ is prepaid at $t = 12$. Then, the IPC weight of mortgage $k$ is $\hat{W}_k(12) = 4$. This can be interpreted as the mortgages representing 4 mortgages. Three mortgages censored before $t = 12$ and itself (Vock et al., 2016).

By integrating over time, it accounts for changes in prediction accuracy over time. For prepayment, it is defined as (Ishwaran et al., 2014):

$$IBS^P(\tau) = \frac{1}{\tau}\int_0^\tau BS^P(t)\,dt, \tag{61}$$

where $\tau \leq t_m$ is a point in time until which point the integrated Brier score is evaluated. Often this is set is equal to $t_m$, the last observed event time (Ishwaran et al., 2014). Expression (61) is essentially a time-averaged Brier score. A lower value of the IBS is considered a better prediction.

### 3.6.2 Time-dependent area under the ROC curve (AUC)

A second evaluation method for prediction performance is the time-dependent AUC. Before it is described how the time-dependent AUC is calculated, the True Positive Rate, the False Positive Rate and the ROC curve have to be defined.

The True Positive Rate, or 1−Type II error, is defined as the percentage of prepaid loans that were correctly predicted as prepaid. The False Positive Rate, or Type I error, is the percentage of non-prepaid loans that were predicted as prepaid. The Receiver Operating Characteristics (ROC) curve is constructed by comparing the True Positive Rate with the False Positive Rate at different classification thresholds (James et al., 2021). First, fix a point in time and define a threshold $\xi$. For each loan, if the predicted cumulative incidence value lies above this threshold, it is predicted that the loan will be prepaid by that point in time. Then calculate the inverse probability censoring weights, which estimate the probability of being uncensored, using the Kaplan-Meier estimate to account for the bias introduced by censoring. These weights are used to calculate the True Positive Rate (TPR) (Kamarudin et al., 2017). Let $\delta_k = \mathbb{1}(T_k \leq C_k)$ and let $T_k^*$ as defined in Section 3.1.1. Hung and Chiang (2010) defined the estimates of the weighted TPR and FPR. This can be adjusted to incorporate competing risks. Then the estimates of the weighted TPR and FPR are defined as:

$$\widehat{TPR}(\xi, t) = \frac{\sum_{k=1}^n \mathbb{1}(\widehat{CIF}_k(t) > \xi, T_k^* \leq t, \mathcal{S} = P)\frac{\delta_k}{n\hat{G}(T_k^*)}}{\sum_{k=1}^n \mathbb{1}(T_k^* \leq t, \mathcal{S} = P)\frac{\delta_k}{n\hat{G}(T_k^*)}} \tag{62}$$

$$\widehat{FPR}(\xi, t) = \frac{\sum_{k=1}^n \mathbb{1}((\widehat{CIF}_k(t) > \xi, T_k^* > t) \cup (\widehat{CIF}_k(t) > \xi, T_k^* \leq t, \mathcal{S} = D)}{\sum_{k=1}^n \mathbb{1}((T_k^* > t) \cup (T_k^* \leq t, \mathcal{S} = D))} \tag{63}$$

For intuition, the fraction $\frac{\delta_k}{n\hat{G}(T_k^*)}$ in (62), corrects for censoring. It is the fraction of non-censored loans at time $T_k$ divided by the probability of being observed at time $T_k$. It gives an estimate what the number of loans would be if there would be no censoring.

The numerator in (63) counts the loans that are predicted to be prepaid by time $t$, but which are not due either a time to event larger than $t$, or due to a default before or at time $t$. The denominator counts the loans that are not prepaid by time $t$. These are loans that experience an event after time $t$ or have experienced a default before or at time $t$. Note that the False Positive Rate does not have a term that corrects for censoring, because we assume that censoring time and the CIF are independent of each other, which gives the same $\frac{1}{n\hat{G}(t)}$ term for each loan in the numerator and denominator, which then cancel each other out (Blanche et al., 2013).

Setting the threshold too low results in a high True Positive Rate, but also in a high False Positive Rate. On the other hand, setting the threshold too high, results in a low False Positive Rate, but also in a low True Positive Rate. So, by looking at a range of threshold, this results in a graph of the True Positive Rate against the False Positive Rate for different thresholds, which is the ROC curve. From this, the optimal threshold can be determined. As this could result in overfitting, the area under the ROC curve (AUC) is calculated to obtain a consistent evaluation criterion. The AUC is calculated using the score() function in the riskRegression package that uses the trapezoidal method of calculating the area under the ROC curve. This method divides the area in smaller subareas, which creates smaller trapezoids and calculates the sum of the areas of these smaller trapezoids to approach the AUC (Yeh, 2002). The higher the AUC, the better the prediction. These steps are repeated at each event time to get an AUC for each event time. The time-dependent AUC is plotted over time to see the prediction performance of the model over time (Saha and Haegerty, 2015).

# 4 Data

The data used is the Single-Family Loan-Level Data set provided by Freddie Mac (2024), which is supervised and regulated by the Federal Housing Finance Agency. It contains fully amortizing 10-, 15-, 20-, 30-, and 40-year fixed-rate Single-Family mortgages. A random sample of 50,000 loans is available for each origination year between 1999 and 2022. For the year 2023, there is a random sample of 37,500 loans available. These random samples are updated every quarter to include the most recent monthly observed factors. Hence, the random sample of year 1999 consists of monthly observations from January 1999 until December 2023 from loans that were issued in 1999. Additionally, supplementary data from the year of issuance is available for each loan. In total, there are now 1,237,500 loans observed each with a unique Loan Sequence Number. All mortgages are either prematurely terminated by prepayment or default, matured or right-censored. There is no left-censored data.

- First, mortgages whose Metropolitan Statistical Area (MSA) are not observed, are removed from the data set. This is because the original idea was to add macroeconomic factors that are observed for each MSA to the data set. So, it is assumed that missing data about the MSA is independent of prepayment and default rates, as this would otherwise bias results. As it was later decided not to proceed with the inclusion of macroeconomic factors, this step is a limitation of the research as mentioned in Section 7. Then, 964,999 mortgages remain in the data set.

- Next, mortgages for which it is not known whether it is the first property bought by the mortgagor are removed. This leaves 964,424 mortgages.

- Moreover, mortgages for which the Number of Units is not known are removed. This leaves 964,357 mortgages.

- If the Original Loan-to-Value Ratio is not known, the mortgages are removed from the data set. This leaves 964,327 mortgages.

- If the Original Debt-to-Income Ratio is not known, the mortgages are removed from the data set. This leaves 901,872 mortgages.

- Next, if a mortgage does not have a prepayment penalty it is removed from the data. This is because this could potentially affect prepayment behaviour, which would be desirable to include, but due to limited number of mortgages with a prepayment penalty, it is excluded (only 623 mortgages in the whole data set). Now, 901,249 mortgages remain.

- If the type of property is not known, then the mortgage is removed. This leaves 901,225 mortgages.

- If the credit score is not known, the mortgages are removed from the data set. This leaves 899,918 mortgages.

- It was also decided to remove mortgages for which the occupancy status is not known or the amortization type is ARM. However, there were no mortgages in the data set that had this problem, so no mortgages were removed in this step.

So, in the end, 1,237,500 mortgages were reduced to 899,918 mortgages.

To reduce computation time, a random samples of 20,000 mortgages is taken, which is drawn without replacement. To check whether the sample drawn is indeed representative for all mortgages, a second sample is drawn and checked if the characteristics are similar. This second sample is also used for out-of-sample predictions as later described in Section 6.1. The original random sample of 20,000 mortgages is also split into two samples to check for persistence of prepayment drivers and prediction accuracy using backtesting as described in Section 6.2.

The prepayment rate is derived from the Zero Balance Code, the Zero Balance Effective Date and the Maturity Date, which are all provided in the data set. A Zero Balance Code equal to 01 indicates the loan is prepaid or matured. If the Zero Balance Effective Date, i.e. the date on which the event triggering the Zero Balance Code took place, is not the same date as the maturity date, then this implies a prematurely payment. Combining these insights, it indicates a prepayment.

The definition of default provided by the European Central Bank Regulation (European Central Bank, 2024) is that a default on a loan can be considered to have occurred when either or both of the following have taken place:

1. The bank considers that the obligor is unlikely to pay its credit obligations in full to the bank, without recourse by the bank to actions such as realising security. This is the "unlikeness to pay" criterion.

2. The obligor is more than 90 consecutive days past due on any material credit obligation to the bank. This is the "days past due" criterion.

So, in this thesis, a loan is classified as defaulted if the delinquency status exceeds 90 days. Some of these loans are observed after these 90 days. These observations of the defaulted loans are removed from the data set to be able to correctly count the time to event using the table() function in R. Moreover, mortgages that are prematurely ended due to short sale or Charge Off, REO (Real Estate Owned) disposition, or reperforming loan securitizations are also classified as defaulted. These are events that typically occur if the borrower is in financially distress. With a short sale, the obligor sells the house before the lender seize it in foreclosure, because the borrower is unable to make the interest payments (Chen, 2023). REO disposition refers to the situation where the house is owned by the lender because it failed to sell in a foreclosure auction after the borrower defaulted on their mortgage (Chen, 2024). Reperforming loans are loans that defaulted, but came out of default as the borrower resumed payments (Kenton, 2022). There are three loan in the data set and all three defaulted again. Therefore, the loans are considered to have defaulted when they first went into default.

The data set also consists of 28 loans that are "defect prior to other termination event". These causes are not specified in more detail. One of these loans was already classified as defaulted by the definition of default taken. Since this thesis focuses on the modelling of prepayment rates while incorporating competing events, the other 27 loans are also classified as defaults.

The loan with Loan Sequence Number F04Q20559684 is removed from the data due to data quality issues. Data was missing for several months.

Available loan-specific covariates are presented in Table 1. Data on the continuous covariates are presented in Table 2. In Appendix C, bar charts are provided for the categorical variables in Figures 27-32. The measurement of *number of borrowers* has changed from the second quarter of 2018 onwards. Before, it was only known if it was one borrower or more borrowers. After, the exact number of borrowers was known. As this leads to measurement differences, the data on the number of borrowers from Q2 2018 onwards is transformed to the same format as before.

The *channel* variable can take the values "Retail", "Broker", "Correspondent, and "Third Party Origination not specified" (which occurs 3613 times in the sample). This last value indicates that it is not known whether a broker or a correspondent was involved in the origination of the mortgage. It is only known that a third party was involved. Therefore, it is merged with the values "Broker" and "Correspondent" into a newly created value "Third Party". So, the variable *channel* is subdivided into two dummy variables called *channel: Retail* and *channel: Third Party*.

As one can observe in the Appendix C Figures 28-30, several categories contain very few observations. To avoid this, these categories are bundled. So, the *number of units* will be a dummy variable that equals one if the mortgage consists of one unit and zero if the mortgage consists of more than one unit. *Occupancy status* is changed into a dummy variable that equals one if the house is used as primary residence and zero if not. The *property type* dummy variable equals one if the property type secured by the mortgage is a single-family home and zero otherwise.

As Table 22 in Appendix B shows, the number of loans per state are unbalanced. In other words, there are states with few observations, which can cause problems with estimation. Therefore, the *states* are assigned to four *regions*, as described by United States Census Bureau (2022). The *regions* are determined based on geographical, historical and economic characteristics. Although outside the United States there are often five *regions* used, the US government uses four *regions*. These *regions* are called Northeast, Midwest, South, and West. The number of observations for each *region* are tabulated in the Appendix B Table 23. Table 24 in Appendix B shows the selected covariates for estimation of the models and their reasons to be selected.

Table 1: Loan-specific variables (Freddie Mac, 2024)

| Variable Name | Description |
|---|---|
| Channel | Indicates whether a broker, correspondent or a third party which is not specified was involved in the issuing of the loan. If a mortgage does not fit in one of the three categories, it indicates retail. |
| Credit score | Score to indicate likelihood that borrower will timely repay future obligations. |
| First Payment Date | The agreed date of the first scheduled payment |

| | |
|---|---|
| First Time Homebuyer Flag | Indicates whether the borrower will reside in the mortgaged property as a primary residence and had no ownership interest in a residential property during the three-year period preceding the purchase date. |
| Interest-only Indicator | Indicator that denotes whether the loan only requires interest payments for a specified period. |
| Loan Purpose | Indicates whether the mortgage loan is a Cash-out Refinance mortgage, No Cash-out Refinance mortgage, or a Purchase mortgage. |
| Loan Sequence Number | Unique identifier assigned to each loan. |
| Maturity Date | The month in which the final monthly payment on the mortgage is scheduled. |
| Mortgage Insurance Percentage | The percentage of loss coverage on the loan in case of default. |
| MSA | Metropolitan Statistical Area |
| Number of borrowers | The number of borrowers who are obligated to repay the mortgage |
| Number of units | Denotes whether the mortgage is a one-, two-, three- or four-unit property. |
| Occupancy Status | Denotes whether the mortgage type is owner occupied, second home, or investment property. |
| Original DTI Ratio | Original Debt-to-Income Ratio is the sum of the borrower's monthly debt payments divided by the total monthly income used to underwrite the loan as of the origination date. |
| Original Loan Term | Duration of the loan in months. |
| Original LTV | Original Loan-to-Value is the original mortgage loan amount on the note date divided by the lesser of the mortgaged property's appraised value on the note date or its purchase price. |
| Original UPB | Unpaid Principal Balance of the mortgage on the note date. |
| Property State | State in which mortgaged property is situated. |
| Property Type | Denotes whether the proporty type secured by the mortgage is a condominium, leasehold, planned unit development, cooperative share, manufactured home, or Single-Family home. |
| Current Actual UPB | Reflects the mortgage ending balance. |
| Estimated LTV | Current Loan-to-Value ratio estimated through Freddie Mac's Automated Valuation Model. |
| Interest Rate | The interest rate for each month of the mortgage |
| Loan Age | The number of scheduled payments from the time the loan was originated or modified up to and including the performance cutoff date. |
| Monthly Reporting Period | As-of month for loan information contained in the loan record. |
| Payment Deferral | A flag indicating a loan has been granted a payment deferral in the current or prior period. |

| | |
|---|---|
| Remaining Months to Legal Maturity | The remaining number of months to maturity. |
| Zero Balance Code | A code indicating the reason the loan's balance was reduced to zero. |
| Zero Balance Effective Date | The date on which the loan's balance was reduced to zero. |
| Zero Balance Removal UPB | The amount of total UPB remaining on the loan immediately prior to the application of the Zero Balance Code |

Table 2: Summary statistics continuous variables

| Variable name | Mean | Std. Dev | Min. | Max. |
|---|---|---|---|---|
| Credit Score | 741.95 | 50.58 | 300.00 | 836.00 |
| Mortgage Insurance Percentage | 5.44 | 10.71 | 0.00 | 40.00 |
| Original Debt-to-Income Ratio | 34.47 | 10.66 | 1.00 | 65.00 |
| Original UPB | 225526.28 | 128855.08 | 17000.00 | 1114000.00 |
| Original Loan-to-Value | 72.07 | 17.15 | 6.00 | 100.00 |
| Interest Rate | 5.06 | 1.46 | 1.75 | 10.75 |
| Original Loan Term | 329.57 | 66.38 | 120.00 | 480.00 |

Figure 1 shows the number of issued loans per month from January 1999 until December 2023. After the financial crisis of 2007-2009, the number of issued loans reduced and from that time onwards, the number of issued loans rose again. Figure 2 shows the total number of outstanding loans in the random sample. Note that this figure is not just the cumulative function of Figure 1 as prepayments and defaults cause the dips in the graph. The prepayment and default rate over time are presented in Figures 3 and 4 respectively. The prepayment rate has spikes in the periods 2001-2003, 2011-2013, and 2020-2022. In the mid-2000s, the prepayment rate increases drastically. Several factors could have contributed to this observation. One explanation could be that 30-year fixed interest rates on mortgages were relatively low compared to the period before this time. Prepayment rates are namely influenced by market rate fluctuations, which in research is discovered to be one of the main factors. If market rates drop, this may trigger the prepayment option (even though a penalty might neutralize the effect). Homeowners have an incentive to refinance their mortgages to secure a lower interest rate on their mortgage. However, as past research suggests, and also assumed in this paper, people do not act rationally. A second explanation is the housing boom that occurred in the mid-2000s. During this period, home prices rose and demand for housing was high. Lee et al. (2022) concluded that an increase in house price significantly increases prepayment rates, which can be seen during the housing boom in the early 2000s. Homeowners are given an incentive to sell their home or to refinance it. During the COVID-19 pandemic, the interest rate was kept low to support the economy, which could lead to people refinancing their loans, which results in more prepayments. It is observed that there are spikes in the default rate during the financial crisis starting in 2009 and also a relatively high spike at the start of the COVID-19 pandemic.

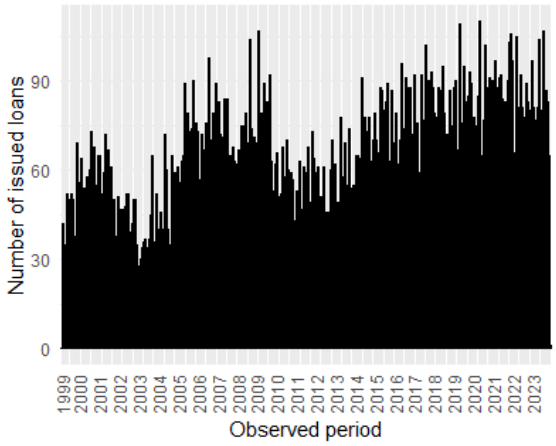


Figure 1: Number of newly issued loans over time

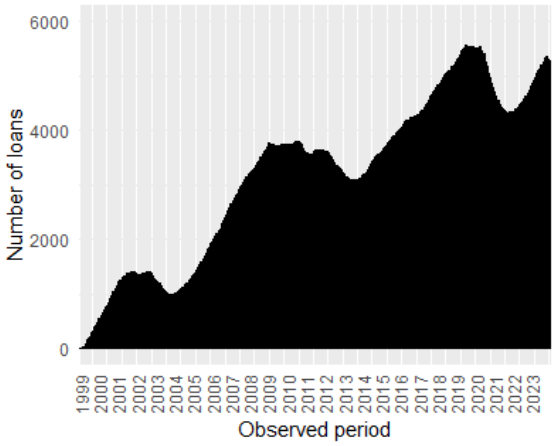


Figure 2: Total number of outstanding loans over time

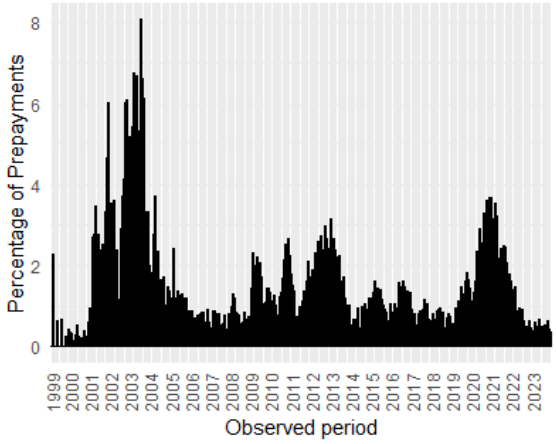


Figure 3: Prepayment rate over time

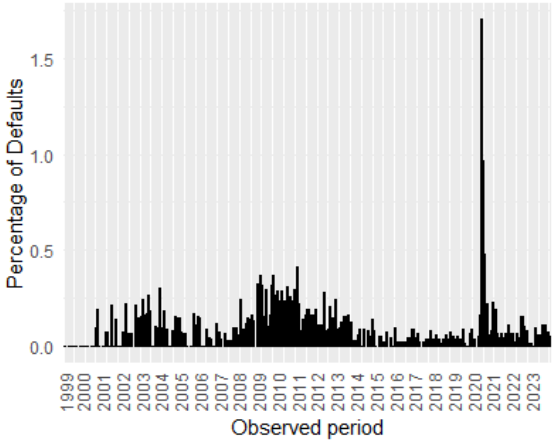


Figure 4: Default rate over time

Table 3 shows the number of loans that are still active on 31 December 2023, prepaid or defaulted, or that are matured. With $\frac{13{,}690}{19{,}999} * 100\% = 68.5\%$ of the mortgages being prepaid, prepayment is the termination cause that is most observed.

Table 3: Overview loan status in training set

| State of mortgage | Number of observations |
|---|---|
| Still active on 31 December 2023 | 5,270 |
| Prepaid | 13,690 |
| Default | 1,024 |
| Matured | 15 |
| Total | 19,999 |

## 4.1 Driver analysis

Figure 5 and Figure 6 show the relationship of the loan size and the loan term with the cumulative incidence respectively. It uses the non-parametric Aalen-Johansen estimator

of the cumulative incidence. The Aalen-Johansen estimator estimates the cumulative incidence function in the presence of competing risks non-parametrically. Let $dN^P(t_j)$ be the number of mortgages prepaid at time $t_j$ given that they did not experience an event before. This condition can be observed as the probability is calculated using only loans that are still in the risk set at time $t_j$. Let $|R(t_j)|$ be the total number of mortgages at risk at time $t_j$. Let $p^P(t_j) = \frac{dN^P(t_j)}{|R(t_j)|}$ be the non-parametric estimate of the probability of a prepayment at time $t_j$. Then, for distinct event times $j = 1, \ldots, m$, the Aalen-Johansen estimate of the cumulative incidence function is given by (Stegherr et al., 2020):

$$CIF^{AJ}(t_j) = \sum_{s \leq t_j} (\prod_{\tau < s} (1 - p^P(\tau)) * p^P(s)) \tag{64}$$

So, intuitively, it suggests that we estimate the probability that a loan is prepaid by time $t_j$ by multiplying the probability that the loan is not prepaid until time $t_j$ with the probability it is prepaid at time $t_j$ given that it did not experience an event before. If we subdivide the data based on a variable value, then we can estimate the Aalen-Johansen cumulative incidence non-parametrically for each variable outcome. Since continuous variables have a lot of unique values, this would result in an unclear graph with multiple lines. Therefore, for continuous variables, the data is separated based on quartiles of a covariate.

In Figure 5, the cumulative incidence is plotted for the four quartiles of *original unpaid balance* to show the effect of *original unpaid balance* on the cumulative probability of prepayment. The figure shows that higher loan sizes are associated with a higher cumulative incidence. So, higher loan amounts appear to be more likely to be prepaid. For *loan term*, three categories are used instead of quartiles as the distribution of loan terms is not very smooth, which then would result in non-informative curves. Namely, 16,305 out of the 19,999 loans have a term of 360 months, which suggests that quartiles are not informative. Figure 6 suggests that shorter *loan terms* are less likely to be prepaid in the first 150 months. An explanation for this could be that it is less beneficial to payoff short-term loans than long-term loans if market rates drop as one is already closer to the maturity date. One interesting thing we see from Figure 6, is that there is an increase in cumulative incidence just before the time hits 180 months for loans with a *loan term* between 120 and 180 months. This has been investigated in the data, since this increase is relatively large and looks random. The data shows that there are 22 loans that are prepaid after 179 months, which had a loan term of 180 months. This may suggest that people want to save that little bit of extra interest they would pay if they stick to the payment schedule. As the remaining amount to be paid one month before maturity is relatively low, people often can afford to pay these two months at once. We do not see this spike for the other two *loan term* categories, because the end of the *loan term* is not reached in the data, as the data only covers 24 years. This probably will cause *original loan term* to be non-proportional for prepayments.

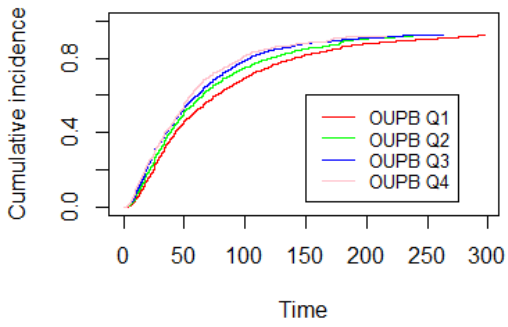


Figure 5: Cumulative incidence of prepayment for the four quartiles of Original Unpaid Principle Balance

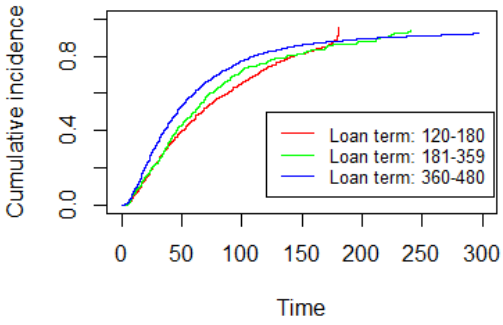


Figure 6: Cumulative incidence of prepayment for loan terms of the three categories 120-180 months, 181-359 months, and 360-480 months

The Aalen-Johansen estimators of the cumulative incidence for the other covariates are presented in Appendix C. Mortgages with interest rates in the fourth quartile seem to have a higher probability of being prepaid at any point in time than mortgages with lower interest rates as can be seen in Figure 33. Moreover, mortgages with interest rates in the lowest 25% quantile have a lower probability of being prepaid. Remarkably, mortgages with interest rates in the third quartile are less likely to be prepaid than mortgages with interest rates in the second quartile. Intuitively, interest rates on mortgages and prepayment are related through the current market rate. If market rates drop, borrowers with mortgages with high interest rates are intuitively more likely to refinance their loans, which would result in a prepayment.

Figure 34 shows the cumulative incidence curve for mortgages that were issued to *first-time homebuyers* compared to mortgages that were issued to non-first-time homebuyers. Intuitively, it is not clear in which direction *first-time homebuyer* affects the likelihood of prepayment. Namely, people who take on a mortgage for the first time are for example more likely to be young people. This could be an indicator for less financial stability. For example, they are more likely to be fired from their jobs as they are less likely to have a permanent contract, which would increase the likelihood of default, but decrease the likelihood of prepayment. This relation should mainly be visible for short loan age. On the other hand, young people are more likely to experience a change in household composition or more likely to get divorced. Both could see prepayment rates to be higher for first-time homebuyers than for non-first-time homebuyers. In Figure 34, it appears that up to 50 months, mortgages held by *first-time homebuyers* are less likely to be prepaid. The same holds from 100 months onwards. Between 50 and 100 months, it is less evident what the relation is. Moreover, since the cumulative incidence curves are relatively close to each other, the effect of a *first time homebuyer* is suggested not to be large.

Figure 35 suggests mortgages on a one-unit property are more likely to be prepaid than mortgages on a two-, three-, or four-unit property. A reason for this could be

that if the underlying property consists of more than one unit, it has a more complex financial structure, which could lower the probability of prepayment. Figure 36 suggests a time-varying or non-linear effect of *credit score* on prepayment. Namely, from this graph, one cannot conclude in which direction credit scores and prepayments are related. As a result, the proportionality assumption of the proportional hazard models will likely not be satisfied. It looks like low credit scores are associated with lower prepayments, but for the second, third and fourth quartile, the cumulative incidence curves cross each other. Figure 37 indicates that the effect of the channel is not evident or there might even not be an effect at all. For *occupancy status* there seems to be an effect, as observed in Figure 38. If the collateral of the mortgage is used as primary residence, the probability of prepayment appears to be higher than if the mortgage is used for other purposes. In Figure 39, there appears to be an effect of *original debt-to-income ratio* on the probability of prepayment from year 10 onwards. Loans with low *original DTI* appear to have a higher probability of prepayment than loans with a high *original DTI*. Intuitively, this makes sense. If the debt is low compared to income, this suggests more financial flexibility for the borrower. Consequently, this enables one to prepay their loan. Note however, that income will change over time, which changes the debt-to-income ratio over time. If income increases, the debt-to-income ratio decreases compared to the *original DTI ratio*, which would give even more financial flexibility. On the other hand, if income decreased after the issuing date, this would lead to a reduction in prepayment rate not only through less financial flexibility, but also due to an increase in the number of defaults. Figure 40 does not show a stable answer to the question in which direction the effect of *mortgage insurance percentage* is on the probability of prepayment is. Moreover, it suggests that up until month 100 there is no effect of *mortgage insurance percentage* on the probability of prepayment. For *original loan-to-value*, there also appears not to be a clear effect on probability of prepayment, observable in Figure 41. In Figure 42, it is observed that loans with underlying properties situated in the Western region seem to have a higher probability of being prepaid than loans with underlying properties situated in other regions. As we cannot be sure at this stage whether the effects observed are statistically significant, these figures should be interpreted carefully.

# 5 Model estimation

In this section, the estimates of the Cox cause-specific hazards model, the Fine-Gray subdistribution hazard model and the random survival forest for competing risks are presented and discussed. First, the estimation results of the Cox cause-specific hazards model are discussed followed by the discussion of the Fine-Gray subdistribution hazard model estimates. Lastly, the results of estimation of the random survival forest are presented. The current VB Advisory model does not require estimation of parameters and hence, only its prediction accuracy is compared to the other three models in Section 6.

## 5.1 Cox cause-specific hazards model

The proportionality assumption of the Cox prepayment-specific hazard and the Cox default-specific hazard should be tested. If this assumption is not satisfied, it is tested using sensitively analysis what the effect is on estimated coefficients and prediction accuracy. If there is a significant difference in prediction accuracy, one could deal with non-proportionality using two method, which are both examined. One is the inclusion of time-varying effects. Interactions between the non-proportional variables and time are included to incorporate the time-varying effect of the variables on the hazard. Based on the pattern of the Schoenfeld residuals over time, one can also decide to use interactions between the non-proportional variable and subsets of time. Moreover, one can also incorporate interactions of non-proportional variables with functions applied to the time to event, for example the natural logarithm (Bellera et al., 2010).

**Stratification as solution to non-proportionality**

A second approach is to stratify non-proportional variables. The idea is to split the sample into subgroups based on the non-proportional variable and then estimate a different baseline hazard function for each subgroup (Borucka, 2014a). Each stratum has a different partial likelihood function (see (16)), but by multiplying these different partial likelihood functions, one single partial likelihood function is obtained. As a result, the estimated coefficients of the variables will be the same for each stratum (Kleinbaum and Klein, 2011). An important remark is that the number of observations per subgroup should not be too low, because this leads to an unreliable estimate for the baseline hazard in a subgroup and as a result will result in bad predictions. Stratifying a continuous variable can lead to a large number of subgroups if for each value a subgroup is created, which leads to a low number of observations per subgroup. Therefore, it is decided to stratify on the quartiles of the continuous variable in that case.

### 5.1.1 Variable selection

But before testing the proportional hazards assumption, variable selection is performed. Figures 28 and 29 show little variation in the *number of units* and *occupancy status* respectively in the data. As this suggests these two variables will not be able to contribute to explaining the variation in the cumulative incidences for different mortgages, these two variables were removed. Using the combination of forward and backward selection

method based on the AIC as described in Section 3.3.1, the variables *channel*, *first time homebuyer flag*, and *original loan-to-value (LTV)* are excluded for the estimation of the prepayment-specific hazard rate. The step() function in R is used and the details of the selection process are stated in Appendix B Table 25. The details on how to interpret the table are given above Table 25 in Appendix B. For the Cox default-specific hazard rate, the variables *original loan term*, *mortgage insurance percentage*, *first time homebuyer flag*, and *region* are excluded. The details of the selection procedure are tabulated in Appendix B Table 26. For both cause-specific hazards, the selection procedure did not include a variable again after it was removed, and therefore using only backwards selection using AIC would have resulted in the same selected variables.

### 5.1.2 Testing proportionality assumption

The next step is to test the proportional hazards assumption for both cause-specific hazards, which is obtained using the cox.zph function in R. First, for each variable, the sum of the scaled Schoenfeld residual and the estimated time-fixed coefficient is plotted against time, as described in formula (29) in the Schoenfeld procedure. This is exercised to get a first indication whether the proportionality assumption is satisfied for each variable. Figures 7, 8 and 9 show the sum of the scaled Schoenfeld residuals and the estimated time-fixed coefficient over time for each covariate. So, Beta(t) on the y-axis refers to $r^*_{ji} + \hat{\beta}_i$ as described in formula (29) of the Schoenfeld procedure. It suggests that *mortgage insurance percentage* could potentially violate the proportionality assumption as there seems to be a slight positive trend over time of the scaled Schoenfeld residuals. Namely, the coefficient is constant under the proportionality assumption and hence deviations over time are caused by changes in scaled Schoenfeld residuals. Also, the *region* of the underlying property of the mortgage seems to have non-constant effect over time on the cause-specific hazard of prepayment. Another variable that seems to violate the proportionality assumption is *original interest rate*.

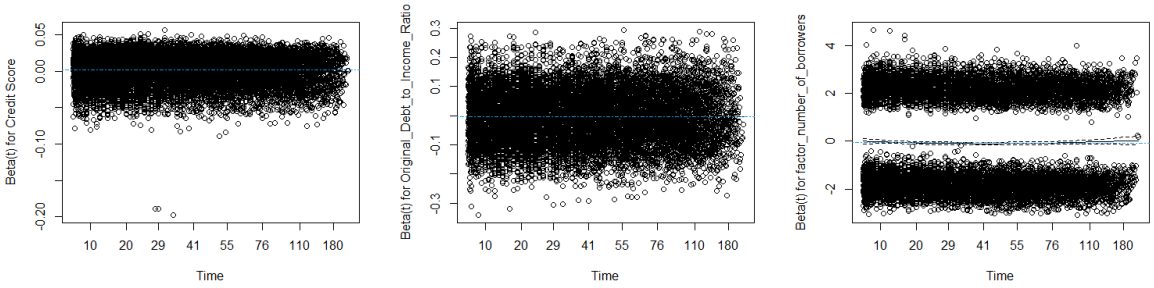


Figure 7: Scaled Schoenfeld residuals Credit Score, Original DTI and Number of Borrowers respectively in Cox prepayment-specific hazards model

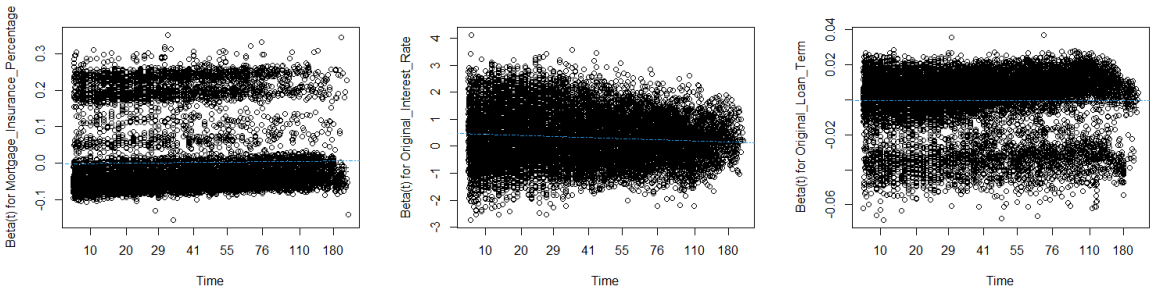


Figure 8: Scaled Schoenfeld residuals MI percentage, Original Interest Rate and Original Loan Term in Cox prepayment-specific hazards model

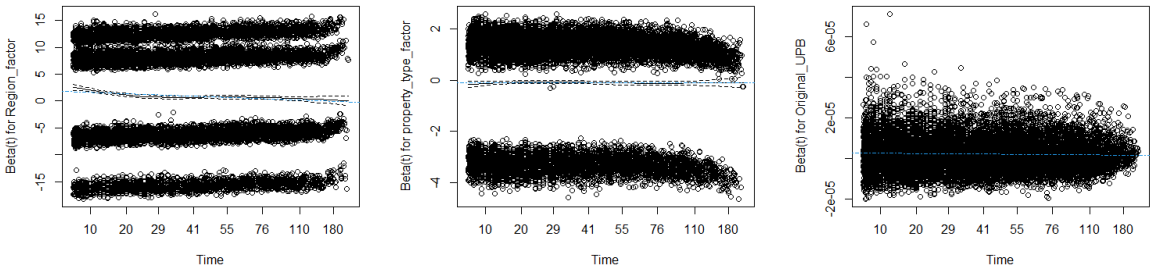


Figure 9: Scaled Schoenfeld residuals Region, Property Type and Original Unpaid Principle Balance in Cox prepayment-specific hazards model

For the three variables that seem to violate the proportionality assumption, these hypotheses are also tested using the log-log survival curve as described in (26). As plotting the log-log survival curve for continuous variables will lead to unclear pictures, the continuous variable *original interest rate* is subdivided into four categories, which are the quartiles. *Mortgage insurance percentage* is subdivided into two categories: no insurance (MI percentage = 0) and insurance (MI percentage greater than zero). The plots are visible in Figures 10 and 11. One can see that the log-log survival curve of no mortgage insurance crosses the log-log survival curve of mortgages with insurance and therefore the proportionality assumption is not satisfied. For the *original interest rate*, the log-log survival curves of the second and third quartile also cross each other, so it looks like the proportionality assumption is not satisfied following (26). We see that the log-log survival curve for the lowest quartile of *original interest rate* stops earlier than for the other quartiles. This is because mortgages with lower loan terms typically have lower interest rates because there is less uncertainty in interest rate fluctuations. Note also that there is relatively large spike in the log-log survival curve for the second quartile of *original interest rate* at the end, which is, as described earlier in Section 4, because people tend to prepay their mortgage close to the maturity date. This spike is exactly at 179 months. So apparently, those mortgages had an *original interest rate* between 3.88% and 4.88%. For mortgages in different *regions*, the proportionality assumptions does not seem to hold for later time points although the curves are relatively close to each other.

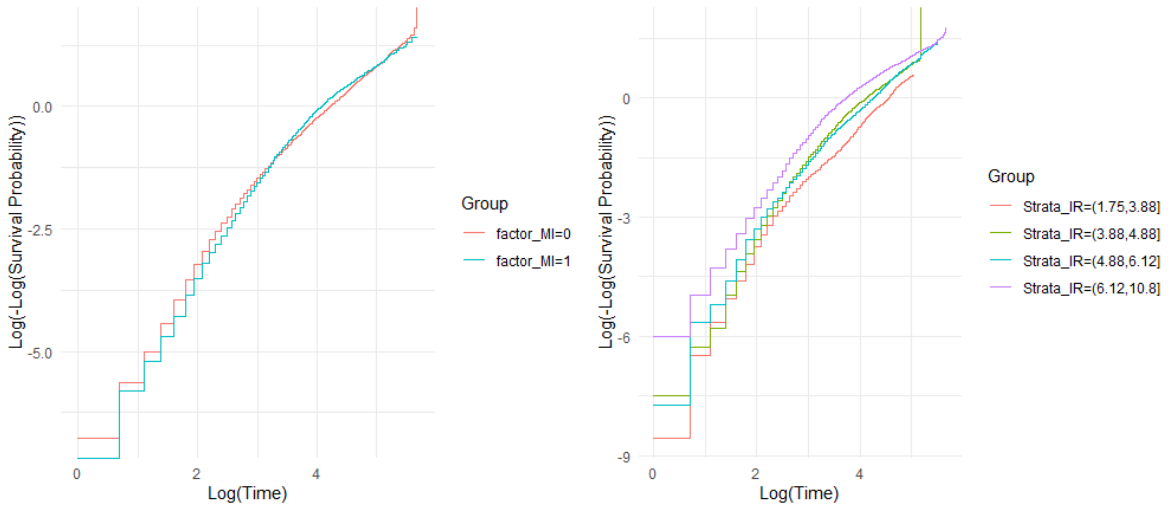


Figure 10: Log-log survival curves of MI factor and Interest Rate quartiles in Cox prepayment-specific hazards model

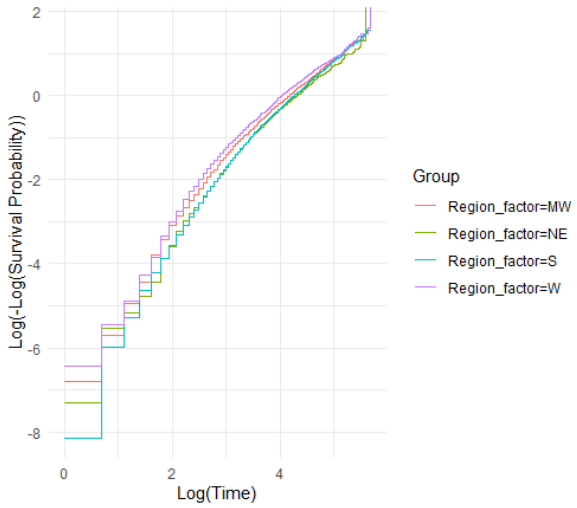


Figure 11: Log-log survival curves of the Regions in Cox prepayment-specific hazards model

Third, the proportionality assumption is tested using the Chi-squared test statistic. If the p-value is smaller than 5%, this suggests that the proportionality assumption is not satisfied for that variable. The results are tabulated in Table 4. As one can observe, the proportionality assumption is not satisfied for the variables *credit score*, *original loan term*, *original interest rate* and *region*. First, we should deal with the variable that most violates the proportionality assumption first, which is *original interest rate*, and after, test again the proportionality assumption for each variable. Interactions between covariates and (functions of) time were incorporated based on the plots 7, 8 and 9 of the scaled Schoenfeld residuals. This resulted in non-proportionality of even more variables, no matter which combination of time and non-proportional variable interaction was used. This "guessing" of the right interactions between variables and time is not

the right statistical strategy. In addition, applying a function of time has no theoretical reason. For both these reasons, it was concluded stratification is better suited for this data. Stratifying on the quartiles of *original interest rate*, resulted in Table 27, which is shown in Appendix B to keep this section organised. Note that both *credit score* and original loan term changed from non-proportional to proportional. Next, stratification on the *region* variable is performed, and the proportionality test is shown in Table 28 in Appendix B.

Thus, at this stage, 16 subgroups are created. Only the *mortgage insurance percentage* remains as non-proportional variable. Using too many stratification variables leads to a small number of observations in the subgroups. Using *mortgage insurance percentage* as additional stratification variable led to a small number of observations in some subgroups, even when the *mortgage insurance percentage* was changed into a dummy variable indicating whether the insurance on the mortgage was 0%. In addition, Table 25 suggests that *mortgage insurance percentage* is the variable included that improves model fit the least. While excluding this variable leads to omitted variable bias as it has some explanatory power, it is decided to remove the variable from the model of the Cox prepayment-specific hazard. The impact of this on the estimated coefficients is tested in Section 5.1.3. Then, the remaining variables all satisfy the proportionality assumption, as visible in Table 5. Note that both *original loan term* and *credit score* seemed to violate the proportionality assumption based on the Chi-squared test statistic in Table 4, but did not according to Table 5 after correcting for non-proportionality of other covariates.

Table 4: Test of proportional hazard assumption for prepayment-specific hazard

| Variable | chisq | df | p-value |
|---|---|---|---|
| Credit Score | 5.4589 | 1 | 0.019 |
| Original DTI | 0.7686 | 1 | 0.381 |
| Original UPB | 2.5162 | 1 | 0.113 |
| Original Loan Term | 12.6119 | 1 | 3.8e-04 |
| Property Type | 2.1970 | 1 | 0.138 |
| Number of Borrowers | 0.0519 | 1 | 0.820 |
| Region | 51.5200 | 3 | 3.8e-11 |
| Original IR | 139.0247 | 1 | 2e-16 |
| Mortgage Insurance Per. | 3.7016 | 1 | 0.054 |
| Global | 206.9986 | 11 | 2e-16 |

Table 5: Final test of proportional hazard assumption for prepayment-specific hazard after stratification

| Variable | chisq | df | p-value |
|---|---|---|---|
| Credit Score | 0.2252 | 1 | 0.64 |
| Original DTI | 0.0762 | 1 | 0.78 |
| Original UPB | 1.1554 | 1 | 0.28 |
| Original Loan Term | 0.7805 | 1 | 0.38 |
| Property Type | 0.1721 | 1 | 0.68 |
| Number of Borrowers | 0.2907 | 1 | 0.59 |
| Global | 2.9658 | 6 | 0.81 |

For the Cox default-specific hazard, the proportionality assumption should also be tested as it is used to estimate the cumulative incidence function of prepayment as stated in (33) and (34). The output of the cox.zph function in the first two steps is tabulated in Appendix B Tables 29 and 30. Subgroups are created based on the quartiles of loan-to-value and the number of borrowers. This leads to 8 subgroups. In total, there are currently $16 * 8 = 128$ subgroups, which is a relatively large number. This has consequences for prediction, but this will be addressed in Section 6. The final output, after stratification, of the proportionality assumption evaluation is presented in Table 6.

Table 6: Test of proportional hazard assumption for default-specific hazard using LTV and Number of Borrowers as stratifying variables

| Variable | chisq | df | p-value |
|---|---|---|---|
| Credit Score | 1.086 | 1 | 0.297 |
| Original DTI | 0.934 | 1 | 0.334 |
| Original UPB | 0.873 | 1 | 0.350 |
| Original Interest Rate | 3.340 | 1 | 0.068 |
| Channel | 0.306 | 1 | 0.580 |
| Property Type | 0.984 | 1 | 0.321 |
| Global | 9.781 | 6 | 0.134 |

The scaled Schoenfeld residuals for prepayment are plotted again in Appendix C Figures 43 and 44 for the variables that satisfy the proportionality assumption after controlling for non-proportionality in *original interest rate* and *region*. The patterns are almost the same as observed in the Figures 7, 8 and 9. The scaled Schoenfeld residuals for variables when estimating the effect on the Cox default-specific hazard are plotted in the Appendix C Figures 45 and 46. We see that the figures also suggest that the proportionality assumption is satisfied for these variables.

### 5.1.3 Estimation results using stratification on non-proportional variables

Now, the final model can be estimated accounted for non-proportionality. The results are tabulated in Table 7. All covariates but *original loan term* are statistically signifi-

cant at a 1% significance level. *Credit score* and *original unpaid principle balance* have a positive effect on the hazard rate for prepayment. The hazard rate can be interpreted as the factor with which the baseline hazard is multiplied for that covariate. If the hazard is greater than 1, the baseline hazard is increased proportionally by the hazard. On the other hand, if the hazard is smaller than 1, the baseline hazard is decreased proportionally by the hazard. For the categorical variables, the interpretation is straightforward. If a mortgage is issued to a *single borrower*, this decreases the hazard with 6.7% ceteris paribus. One could also say that a *single borrower* is associated with a 0.0069 decrease in log hazard rate of prepayment ceteris paribus. If the *property* underlying the loan is used as Single Family home, then the hazard of prepayment decreases with 7.8% ceteris paribus. A one-unit increase in *original debt-to-income ratio* reduces the hazard of prepayment with 0.2% ceteris paribus. This is the expected sign, as higher debt compared to income leads to less financial freedom compared to a low debt-to-income ratio, which would intuitively reduce the probability of prepayment at any time. Note that the covariates have an effect on the cumulative incidence of prepayment that is different than the estimated coefficients and that we cannot compare these results with the Aalen-Johansen curves (64) as the effect of variables on the Cox cause-specific hazard does not need to be in the same direction as the effect of the variable on the cumulative incidence (34). This was explained in Section 3.3.1 and is the result of the independence assumption of prepayment and default.

Table 7: Estimation results Cox prepayment-specific hazard using stratification

| | Strata: quartiles IR and Region | | | |
|---|---|---|---|---|
| Variable name | coef | Hazard rate (exp(coef)) | s.e. | p-value |
| Credit Score | 1.384e-03 | 1.001 | 1.815e-04 | 0.000*** |
| Original DTI | -2.263e-03 | 0.998 | 7.986e-04 | 0.005** |
| Original UPB | 2.008e-06 | 1.000 | 8.026e-08 | 0.000*** |
| Original Loan Term | 3.203e-06 | 1.000 | 1.482e-04 | 0.983 |
| Property Type = SF | -8.082e-02 | 0.922 | 1.920e-02 | 0.000*** |
| Num of Borrowers = 1 | -6.932e-02 | 0.933 | 1.766e-02 | 0.000*** |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

It is examined what the effect is on the estimated hazard ratios if the non-proportional variable *mortgage insurance percentage* is included. The results are tabulated in Appendix B Table 31. It is observed that the hazard ratios do not differ much compared to the estimated hazard ratios found in Table 7. The same variables are statistically significant at a 5% significance level. Note that the *mortgage insurance percentage* variable is not statistically significant at a 5% significance level and hence, it is decided to exclude this variable for further analysis.

The estimation results of the Cox default-specific hazard estimation is presented in Table 8. All covariates but *property type* are statistically significant at a 5% significance level. The variables intuitively have the expected signs. Higher *credit score* are

associated with a decrease in the hazard of default, which ultimately the *credit score* is designed for. It gives a score to people based on how likely it is that that person defaults, with higher scores suggesting higher reliability. A higher *original debt-to-income ratio* implies more debt compared to the income one receives, which makes a person more likely to default as one has less financial stability. *Original interest rate* also positively affects the hazard of default. So, mortgages with a high *original interest rate* have a higher default hazard than mortgages with a low *original interest rate*. Note that *original unpaid principle balance*, although statistically significant, has a relatively small effect on both prepayment hazard rate as on default hazard rate.

Table 8: Estimation results Cox default-specific hazard using stratification

| | Strata: quartiles LTV and Num of Borrowers | | | |
|---|---|---|---|---|
| Variable name | coef | Hazard rate (exp(coef)) | s.e. | p-value |
| Credit Score | -1.154e-02 | 0.989 | 5.141e-04 | 0.000*** |
| Original DTI | 2.482e-02 | 1.025 | 3.045e-03 | 0.000*** |
| Original UPB | 2.854e-06 | 1.000 | 2.899e-07 | 0.000*** |
| Original IR | 2.469e-01 | 1.280 | 2.842e-02 | 0.000*** |
| Channel = R | -1.535e-01 | 0.858 | 6.451e-02 | 0.017* |
| Property Type = SF | -1.247e-01 | 0.883 | 6.894e-02 | 0.071 |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

### 5.1.4 Testing independence assumption prepayment and default

As described in Section 3.3.1, using sensitivity analysis by considering the worst-case violations of the independence assumption, it can be assessed whether the independence assumption of prepayment and default is reasonable. Tables 9 and 10 show the estimation results of the two worst-case violations of the independence assumption of the competing risks. For Table 9, the censored observations due to default were assumed to all be prepaid at that time. So, the time to event remained the same, but the event changed from default to prepaid. If we compare the results to the estimated coefficient values in Table 7, we see that *credit score*, *original DTI*, *original loan term* and *number of borrowers* are all statistically insignificant at a 5% significance level in the worst-case violation, whereas they were statistically significant at a 5% significance level before. For Table 10, the censored observations due to default were assumed to be prepaid at the last observed time to prepayment in the data. The last observed time to prepayment in the data is 292 months. The table shows that the same variables are statistically significant at a 5% significance level. These results suggest that if the independence assumptions is violated, the concluded effects of some variables could be different. However, this approach cannot determine whether the independence assumption of competing risks is satisfied or not. It only gives insight how the results would change if we did not assume independence (Kleinbaum and Klein, 2011).

Table 9: Estimation results for worst-case violation 1) of independence assumption of competing risks

| | Strata: quartiles LTV and Num of Borrowers | | | |
|---|---|---|---|---|
| Variable name | coef | Hazard rate (exp(coef)) | s.e. | p-value |
| Credit Score | 2.460e-04 | 1.000 | 1.717e-04 | 0.152 |
| Original DTI | -6.598e-04 | 0.999 | 7.704e-04 | 0.392 |
| Original UPB | 2.080e-06 | 1.000 | 7.752e-08 | 0.000*** |
| Original Loan Term | 6.618e-05 | 1.000 | 1.443e-04 | 0.647 |
| Property Type = SF | -8.545e-02 | 0.918 | 1.853e-02 | 0.000*** |
| Num of Borrowers = 1 | -2.269e-02 | 0.978 | 1.699e-02 | 0.182 |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 10: Estimation results for worst-case violation 2) of independence assumption of competing risks

| | Strata: quartiles LTV and Num of Borrowers | | | |
|---|---|---|---|---|
| Variable name | coef | Hazard rate (exp(coef)) | s.e. | p-value |
| Credit Score | 2.665e-03 | 1.003 | 1.729e-04 | 0.000*** |
| Original DTI | -4.392e-03 | 0.996 | 7.790e-04 | 0.000*** |
| Original UPB | 1.327e-06 | 1.000 | 7.728e-08 | 0.000*** |
| Original Loan Term | -2.166e-04 | 1.000 | 1.441e-04 | 0.133 |
| Property Type = SF | -4.837e-02 | 0.953 | 1.854e-02 | 0.009* |
| Num of Borrowers = 1 | -1.308e-02 | 0.877 | 1.699e-02 | 0.000*** |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

### 5.1.5 Example in-sample prediction cumulative incidence

Figure 12 shows an example output of in-sample prediction of the cumulative incidence of two selected loans. For this, the estimated coefficients of the variables of Tables 7 and 8 are used to calculate the in-sample prediction of the cumulative incidence of prepayment as in expression (34). As there are multiple continuous variables, it is not possible to keep other covariates the same and look at the effect of one covariate on the curve of the two loans. Namely, there are no loans which have identical *original loan-to-value ratios*, *debt-to-income ratios*, etc. and have one covariate value that differs. This figure is intended just as illustration how the cumulative incidence functions are estimated in the Cox cause-specific hazards model. One can see that loan 1 has a higher probability of being prepaid up until any point in time compared to loan 2.

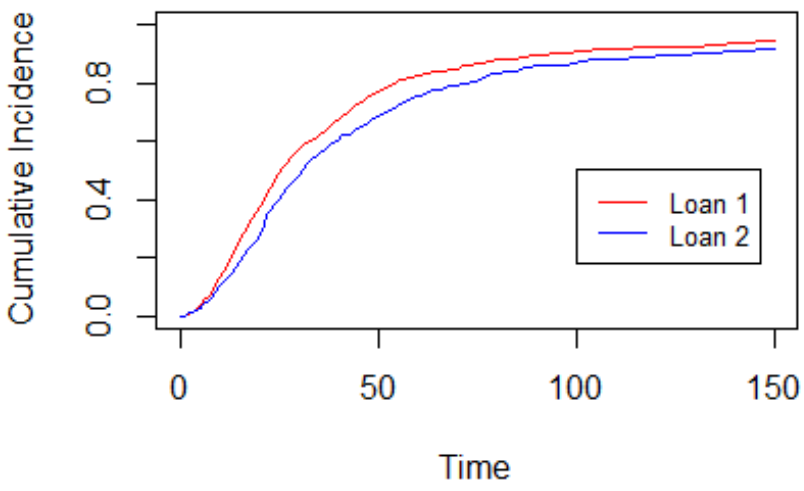


Figure 12: In-sample predicted cumulative incidence curves for two randomly selected loans in the Cox cause-specific hazards model

### 5.1.6 Estimation results without dealing with non-proportionality

As concluded from Section 5.1.2, the variables *original interest rate*, *region* and *mortgage insurance percentage* violate the proportional hazards assumption of the Cox prepayment-specific hazard. This section shows the differences in estimated coefficients of the variables if it is assumed all variables satisfy the proportional hazards assumption. The reason for this is that in the literature the focus shifts more to sensitivity analysis if model assumptions are not satisfied to check the impact of this violation on the estimated and predicted results.

Table 11 shows the results of the Cox prepayment-specific hazard if non-proportionality is not dealt with and hence, no stratification is used. Comparing these results with the results of Table 7, it is observed that the hazard rates of the coefficients change. In particular, the hazard rate of *property type* = *SF* changed from 0.922 to 0.904, which is the largest change observed. Also, the variable *original loan term* is now statistically significant at a 5% significance level, while in Table 7 it had a p-value of nearly 1. In Table 11, all coefficients are statistically significant at a 5% significance level apart from *region W*.

Table 11: Estimation results Cox prepayment-specific hazard if no stratification is used

| Variable name | coef | Hazard rate (exp(coef)) | s.e. | p-value |
|---|---|---|---|---|
| Credit Score | 1.948-03 | 1.002 | 1.822e-04 | 0.000*** |
| Original DTI | -2.382e-03 | 0.998 | 7.921e-04 | 0.003** |
| Original UPB | 2.311e-06 | 1.000 | 8.047e-08 | 0.000*** |
| Original Loan Term | -4.922e-04 | 1.000 | 1.467e-04 | 0.000*** |
| Property Type = SF | -1.008e-01 | 0.904 | 1.915e-02 | 0.000*** |
| Num of Borrowers = 1 | -6.511e-02 | 0.937 | 1.762e-02 | 0.000*** |
| Region NE | -2.742e-01 | 0.760 | 2.933e-02 | 0.000*** |
| Region S | -1.842e-01 | 0.832 | 2.373e-02 | 0.000*** |
| Region W | -4.561e-02 | 0.955 | 2.477e-02 | 0.066 |
| Original IR | 3.375e-01 | 1.401 | 7.903e-03 | 0.000*** |
| MI percentage | 2.110e-03 | 1.002 | 8.580e-04 | 0.01* |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 12 shows the results of the Cox default-specific hazard if no stratification is used. Comparing these results with Table 8, one can conclude the estimated coefficients do not change much. The same variables are statistically significant at a 5% significance level. Interestingly, the variables *original loan-to-value* and *number of borrowers = 1* are both statistically significant at a 5% significance level. The variable *number of borrowers = 1* has the largest effect on the Cox default-specific hazard. If a loan is issued to one borrower, the hazard of default increases with 78.3% ceteris paribus.

Table 12: Estimation results Cox default-specific hazard if no stratification is used

| Variable name | coef | Hazard rate (exp(coef)) | s.e. | p-value |
|---|---|---|---|---|
| Credit Score | -1.150e-02 | 0.987 | 5.033e-04 | 0.000*** |
| Original DTI | 2.474e-02 | 1.025 | 3.047e-03 | 0.000*** |
| Original UPB | 2.871e-06 | 1.000 | 2.894e-07 | 0.000*** |
| Original IR | 2.430e-01 | 1.275 | 2.836e-02 | 0.000*** |
| Channel = R | -1.491e-01 | 0.862 | 6.441e-02 | 0.021* |
| Property Type = SF | -1.007e-01 | 0.904 | 6.878e-02 | 0.143 |
| Original LTV | 1.847e-02 | 1.019 | 2.256e-03 | 0.000*** |
| Num of Borrowers = 1 | 5.785e-01 | 1.783 | 6.487e-02 | 0.000*** |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

As the main purpose of this thesis is to predict absolute prepayment risk, the prediction of the Cox cause-specific hazards model using stratification is compared to the Cox cause-specific hazards model but without stratification in Section 6.1.

## 5.2 Fine-Gray subdistribution hazard model

First, variable selection is performed using backward selection based on the BICcr as described in Section 3.3.2. The detailed steps are provided in Appendix B Table 32. The variables *mortgage insurance percentage*, *original loan-to-value ratio*, *channel* and *first time homebuyer flag* are removed. Then, the proportionality assumption is tested for each variable. Figures 47, 48, 49 and 50 in Appendix C show the scaled Schoenfeld residuals plotted over time for each covariate. For the variables *number of borrowers = 1*, *original unpaid principle balance*, *region South*, *region North-East*, *region West* and *property type = SF* the proportionality assumption seems to be satisfied. For the variables *original interest rate*, *credit score*, *original debt-to-income ratio* and *original loan term*, there seems to be a violation of the proportionality assumption in the first 100 months. This violation is relatively small. After month 100, the proportionality assumption seems to be a reasonable assumption. Assuming that the proportionality holds for each variable, gives the results as tabulated in Table 13. These results show the estimated coefficients of the subdistribution hazards model as described in (40). All coefficients are statistically significant at a 5% significance level except the *region West* indicator. The coefficients of the variables can now be interpreted as the effect they have on the subdistribution hazard. The sign of the coefficient can be interpreted as the direction of the effect the variable has on the cumulative incidence. So, if the property underlying the mortgage is located in the North-Eastern region, this reduces the subdistribution hazard of prepayment with 20.5% ceteris paribus. Moreover, it indicates that an underlying property located in the North-Eastern region is associated with a lower cumulative incidence of prepayment ceteris paribus. The results in Table 13 show that *original unpaid principle balance*, *credit score* and *original interest rate* are positively associated with the cumulative incidence of prepayment. The signs of *original interest rate* and *original unpaid principle balance* are not as expected according to the reasoning in Table 24 in Appendix B.

Table 13: Estimation results Fine-Gray subdistribution hazard model

| Variable name | coef | Hazard rate (exp(coef)) | s.e. | p-value |
|---|---|---|---|---|
| Original DTI | -4.18e-03 | 0.996 | 7.85e-04 | 0.000*** |
| Original UPB | 1.80e-06 | 1.000 | 8.37e-08 | 0.000*** |
| Credit Score | 3.21e-03 | 1.003 | 1.87e-04 | 0.000*** |
| OLT | -5.06e-04 | 0.999 | 1.37e-04 | 0.000*** |
| Num of Borrowers = 1 | -1.25e-01 | 0.883 | 1.78e-02 | 0.000*** |
| Property Type = SF | -8.04e-02 | 0.923 | 1.92e-02 | 0.000*** |
| Region NE | -2.29e-01 | 0.796 | 2.86e-02 | 0.000*** |
| Region S | -1.75e-01 | 0.839 | 2.35e-02 | 0.000*** |
| Region W | -2.28e-02 | 0.977 | 2.51e-02 | 0.360 |
| Original IR | 3.10e-01 | 1.363 | 8.05e-03 | 0.000*** |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

## 5.3 Random survival forest for competing risks

In this section the random survival forest for competing risks is estimated using both Gray's splitting rule and the log-rank splitting rule. For both splitting rules, variable selection and hyperparameter tuning are performed.

### 5.3.1 Gray's splitting rule - Variable selection

First, variable selection in the random survival forest for competing risks is performed, for which the minimal depth and the VIMP are used. The VIMP is event-specific and hence the VIMP documented in this section is the VIMP for prepayments. The combination of hyperparameters could impact the variables selection, which is undesirable (Binder et al., 2020). Therefore, the variable selection procedure, as described in Section 3.4.3, is performed for two different sets of hyperparameters. First, the set of default hyperparameters are 100 trees, 4 variables chosen at each split, 2 number of splits at each node, and 15 number of observations in a terminal node, which are the default hyperparameter values of the rfsrc() function in the RandomForestSRC package. The results are shown in Table 14 and Figure 13. The confidence intervals of the VIMP for each variable are derived using the subsample.rfrsc() function in the RandomForestSRC package. The *channel* variable has VIMP equal to zero in its 95% confidence interval, which implies there is no evidence channel improves prediction accuracy of the random survival forest for competing risks as described in Section 3.4.3. As the Minimal Depth of the *channel* variable is also relatively large compared to the other variables, it is concluded to remove the *channel* variable. It is checked with another hyperparameter combination if the variable selection results in the same variable removed. The combination used is 200 trees, 4 variables chosen at each split, 2 splits at each node, and a minimum of 25 observations in each terminal node. The variable *channel* is still removed due to the same reasons, so it is concluded to remove the variable *channel*. The *original interest rate* on a loan appears to be the most important driver of prepayment risk as this variable has the lowest minimal depth and the highest VIMP.

Table 14: Minimal depth and VIMP for each variable in the random survival forest for competing risks using Gray's splitting rule

| Variable name | Minimal Depth | VIMP |
|---|---|---|
| Original Interest Rate | 1.26 | 0.172 |
| Credit Score | 2.00 | 0.030 |
| Num of Borrowers = 1 | 2.01 | 0.008 |
| Region | 2.03 | 0.011 |
| Original UPB | 2.31 | 0.064 |
| Original Loan Term | 2.36 | 0.037 |
| Occupancy Status | 3.35 | 0.032 |
| Num of Units = 1 | 3.39 | 0.039 |
| Original DTI | 3.43 | 0.018 |
| MI Percentage | 3.59 | 0.028 |
| Original LTV | 3.68 | 0.015 |
| Property Type | 4.04 | 0.005 |
| First Time Homebuyer | 4.24 | 0.010 |
| Channel | 4.33 | 0.001 |

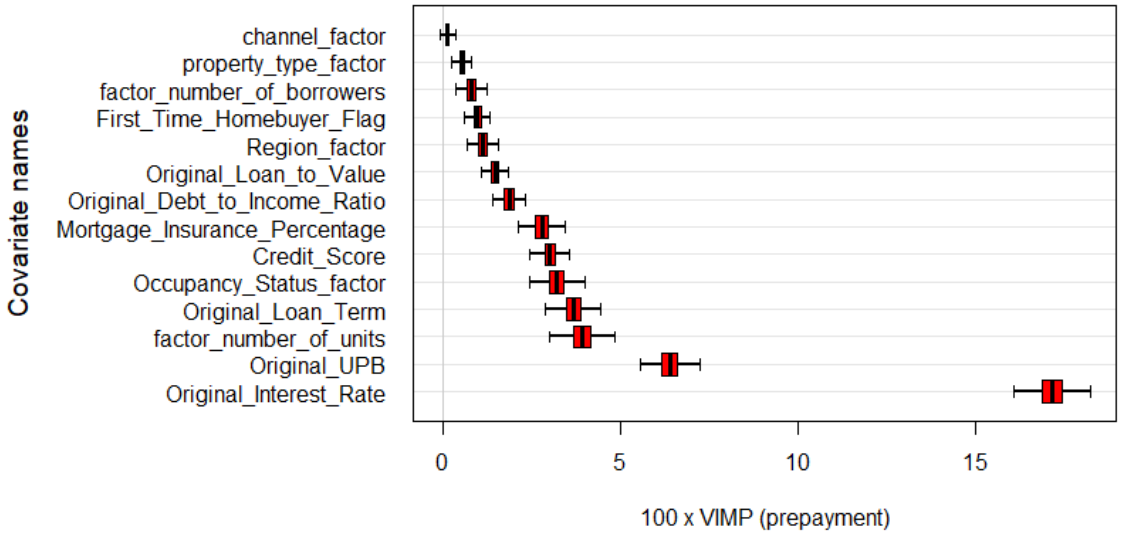


Figure 13: Confidence intervals for VIMP of covariates using Gray's splitting rule

### 5.3.2 Gray's splitting rule - Hyperparameter tuning

The hyperparameter values are obtained using the procedure described in Section 3.4.4. The number of splits at each node is set equal to 2 and the number of variables chosen at each split to 4 (the square root of the total number of variables rounded up to the nearest integer). First, the default setting of 15 observations at each terminal node is

used and the obtained OOB error is plotted against the number of trees used. This can be observed in Figure 14. The OOB error is calculated as $1 - C$ with $C$ being the C-index as in formula (57), as described in Section 3.4.2. For this figure, the OOB error was calculated for the number of trees equal to 5, 10, 25, 50, 100, 200 and 500. While the OOB error is decreasing in the number of trees, the number of trees is chosen to be set equal to 200 as growing a tree larger than 200 does not result in a significant decrease in the OOB error. This also lowers computational effort compared to growing for example 500 or even 1000 trees. Figure 15 shows the OOB error plotted against the minimum number of observations in the terminal node. The grid taken on the minimum number of observations in a terminal node is 10, 15, 20, 25, 35, 50, 65, 75, 85. It is concluded that 75 observations should be taken in the terminal node. So, to conclude, the hyperparameter values are taken as follows:

- Number of trees: 200
- Number of variables chosen at each split: 4
- Number of splits at each node: 2
- Minimum number of observations in a terminal node: 75

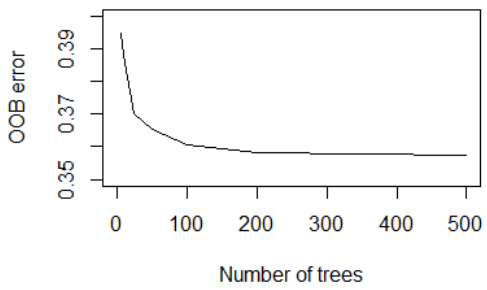


Figure 14: Out-of-bag error plotted against the number of grown trees for Gray's modified splitting rule

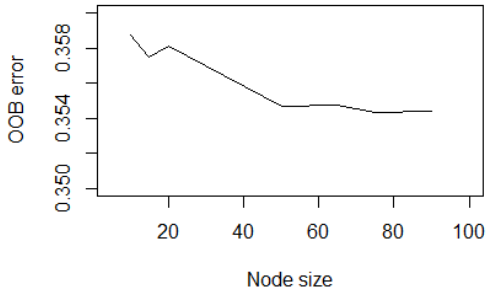


Figure 15: Out-of-bag error plotted against the minimum number observations in the terminal node for Gray's modified splitting rule

### 5.3.3 Log-rank splitting rule - Variable selection

For the log-rank splitting rule, as described in (50) and (51), the same steps were performed as for the modified Gray's splitting rule. The set of default hyperparameters are again 100 trees, 4 variables chosen at each split, 2 number of splits at each node, and 15 number of observations in a terminal node. Table 15 shows the minimal depth and VIMP for each variable using the log-rank splitting rule. The order of the variables based on minimal depth is very similar compared to Table 14, where the Gray's modified splitting rule was used. *Original interest rate*, *original loan term* and *original UPB* appear to be the most important drivers of prepayment according to their Minimal Depth and VIMP. It is concluded from Figure 16 that the variable *channel* should be

removed from the data as zero is in the 95% confidence interval of the VIMP. Moreover, Table 15 shows that the Minimal Depth is also relatively large compared to the other Minimal Depths. It is checked with another hyperparameter combination if the variable selection results in the same variable removed. The combination used is 200 trees, 4 variables chosen at each split, 2 splits at each node, and a minimum of 25 observations in each terminal node. The variable *channel* is still removed due to the same reasons, so it is concluded to remove the variable *channel*. The *original interest rate* on a loan appears to be the most important driver of prepayment risk as this variable has the lowest minimal depth and the highest VIMP.

Table 15: Minimal depth and VIMP for each variable in the random survival forest for competing risks using the log-rank splitting rule

| Variable name | Minimal Depth | VIMP |
|---|---|---|
| Original Interest Rate | 1.28 | 0.180 |
| Original UPB | 1.77 | 0.075 |
| Original Loan Term | 2.05 | 0.037 |
| Region | 2.19 | 0.010 |
| Num of Borrowers = 1 | 2.79 | 0.004 |
| Occupancy Status | 2.89 | 0.032 |
| Num of Units = 1 | 3.16 | 0.044 |
| Original LTV | 3.36 | 0.019 |
| Credit Score | 3.53 | 0.022 |
| MI Percentage | 3.67 | 0.026 |
| Property Type | 3.67 | 0.005 |
| Original DTI | 3.78 | 0.016 |
| Channel | 3.85 | 0.001 |
| First Time Homebuyer | 4.39 | 0.009 |

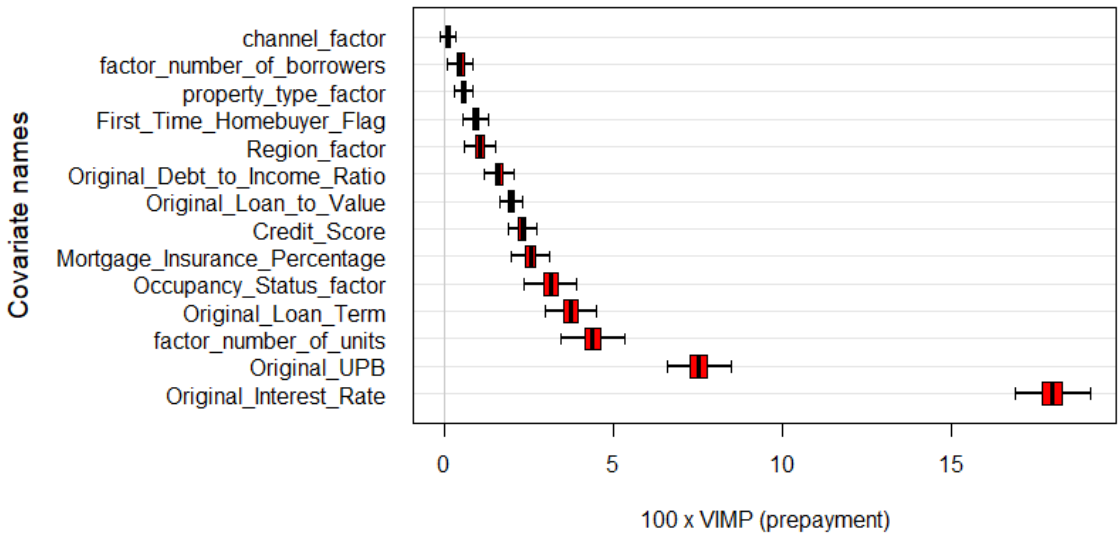


Figure 16: Confidence intervals for VIMP of covariates using the log-rank splitting rule

### 5.3.4 Log-rank splitting rule - Hyperparameter tuning

It was again concluded, by Figure 17, to use 200 trees in order to obtain a relatively low OOB error while keeping computational cost sufficiently low. Now, the lowest OOB error is obtained when the minimum node size is set equal to 65 as visible in Figure 18. Using 65 as the minimum number of observations in the terminal node, the OOB error for the log-rank splitting rule equals 0.3492. On the other hand, using 75 as the minimum number of observations in the terminal node, for the Gray's modified splitting rule, the OOB error equals 0.3543. This suggests that the log-rank splitting rule outperforms the Gray's splitting rule regarding prediction accuracy, which is not expected as Gray's splitting rule is more designed for prediction of the cumulative incidence (Ishwaran et al., 2014).

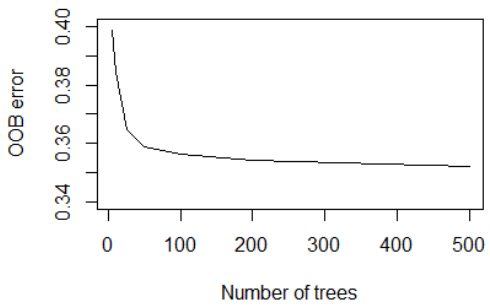


Figure 17: Out-of-bag error plotted against the number of grown trees for log-rank splitting rule

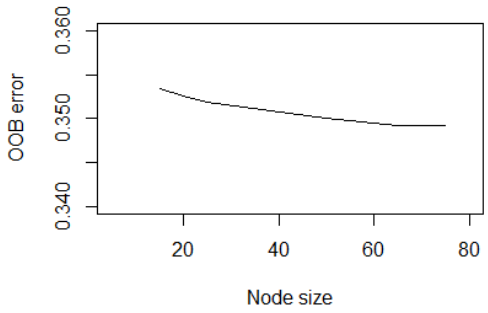


Figure 18: Out-of-bag error plotted against the minimum number observations in the terminal node for log-rank splitting rule

So to conclude the hyperparameter values are taken as follows:

- Number of trees: 200
- Number of variables chosen at each split: 4
- Number of splits at each node: 2
- Minimum number of observations in a terminal node: 65

While the Gray's modified splitting rule is designed for prediction purposes, the OOB error is lower if the log-rank splitting rule is used. Therefore, the log-rank splitting rule is used for model comparison with the Cox cause-specific hazards model, the Fine-Gray subdistribution hazard model and the current VB Advisory model.

# 6 Model evaluation

This section shows the prediction accuracy measured by the time-dependent Brier score and the time-dependent Area under the ROC curve as described in Sections 3.6.1 and 3.6.2 respectively. The time-dependent Brier score is calculated as stated in (60) and the time-dependent AUC is calculated using (62) and (63). First, in Section 6.1, the prediction accuracy is compared between the models on an unseen sample of the data using the same time-period. Then, in Section 6.2, backtesting is performed.

## 6.1 Prediction accuracy models on unseen data of same time period

The prediction accuracy of the models is evaluated on a test set. This test set contains mortgages originated between January 1999 and December 2023, which are not in the training set used to estimate the coefficients of the model as tabulated in Table 7 and Table 8. The characteristics of the test set are tabulated in Table 16. The distribution of the status of the mortgages is similar to the distribution in the training set as observable in Table 3.

Table 16: Overview loan status in test set

| State of mortgage | Number of observations |
|---|---|
| Currently active | 5,213 |
| Prepaid | 13,746 |
| Default | 1,017 |
| Matured | 24 |
| Total | 20,000 |

### 6.1.1 Comparison of stratification versus no stratification in Cox cause-specific hazards model

First, the prediction accuracy in the Cox cause-specific hazards model is evaluated when we control for non-proportionality of covariates versus when we do not control for non-proportionality. Figures 19 and 20 show the comparison of prediction accuracy using time-dependent Brier score and the time-dependent AUC respectively. The Cox cause-specific hazards model, using stratification to deal with non-proportionality of the cause-specific hazards, is compared with the Cox cause-specific hazards model if we do not deal with non-proportionality of the hazards. In Figure 19, it is observed that both Cox cause-specific hazards models outperform the Aalen-Johansen estimator for cumulative incidence prediction of the first approximately 125 months. After approximately 125 months, the Cox cause-specific hazards model without using stratification performs worse than the Aalen-Johansen model. Note also that the Cox cause-specific hazards model without stratification slightly performs better for predictions in the first approximately 60 months. Figure 20 shows again that the Cox cause-specific hazards model that does not correct for non-proportionality performs better at first, but as time progresses, the model performance drops below the Cox cause-specific hazards model with stratification. So, it is concluded that the model should be chosen based

on preference of short-term model prediction accuracy or long-term model prediction accuracy. The integrated Brier score is tabulated in Table 17, which is calculated as stated in (61) using $\tau = 150$. It is observed that the Cox cause-specific hazards model using stratification slightly performs better than the Cox cause-specific hazards model without using stratification. For model comparison in the remainder of this thesis, stratification is used if a relatively large violation of the proportionality assumption is observed.

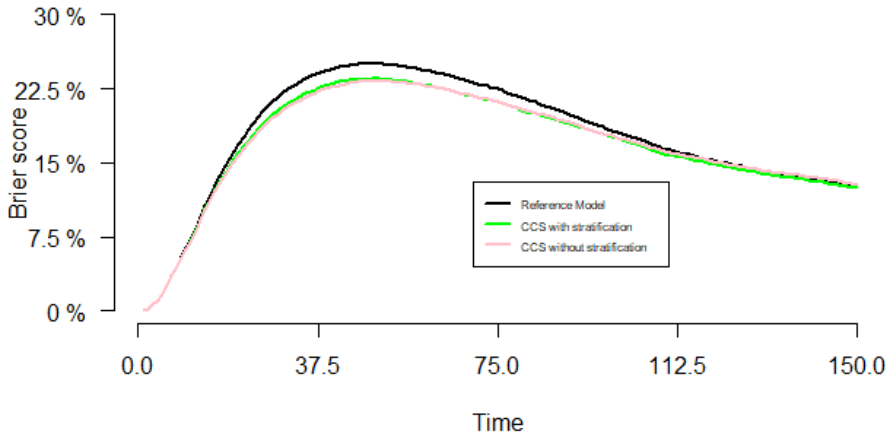


Figure 19: Brier score plotted over time for the Cox cause-specific hazards model using stratification, Cox cause-specific hazards model without stratification, and the Aalen-Johansen prediction model as reference model

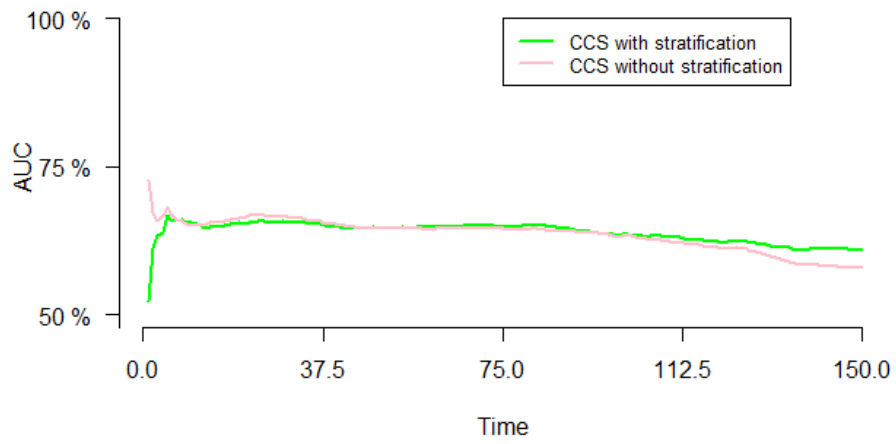


Figure 20: AUC plotted over time for the Cox cause-specific hazards model using stratification and the Cox cause-specific hazards model without stratification

Table 17: Integrated Brier score for Aalen-Johansen estimator and Cox cause-specific hazards model with and without stratification until 150 months

| Model | Integrated Brier score |
|---|---|
| Aalen-Johansen | 0.1775 |
| Cox cause-specific with stratification | 0.1692 |
| Cox cause-specific without stratification | 0.1695 |

### 6.1.2 Comparison of Cox cause-specific hazards model, Fine-Gray subdistribution hazards model and random survival forest for competing risks

Figure 21 shows the Brier score over time for the models of Cox cause-specific hazards, Fine-Gray and random survival forest for competing risks. Reference curve is the non-parametric Aalen-Johansen prediction model, where no covariates are being used. Interestingly, the Brier score curve is decreasing after approximately 45 months. An explanation for this is that it could be that for earlier times, to cumulative incidence is underestimated while already most of prepayments happened. Table 18 shows the integrated Brier score for the models. It shows that the random survival forest for competing risks has the lowest integrated Brier score and therefore is concluded to be the best model to capture the relations between variables and the cumulative incidence of prepayment.

Figure 22 shows the ROC curve for the models at month 150. The True Positive Rate, as described in (62), is plotted against the false positive rate, as described in (63). It shows that at month 150, the AUC is the highest for the random survival forest

for competing risks and therefore at month 150 the random survival forest for competing risks performs the best across the models considered regarding AUC. This figure is calculated for each time between 0 and 150, which results in Figure 23. Figure 23 shows the time-dependent AUC over time for the models. Predictions are only made up to 150 months, because they become less reliable due to limited number of observations with time to event greater than 150. From both prediction metrics, it is concluded that the random survival forest for competing risks predicts the cumulative incidence of prepayment of a mortgages the best amongst the evaluated models. The Brier score over time shows that the random survival forest for competing risks especially performs better for time to events between 35 and 75 approximately. On the other hand, according to the time-dependent AUC, the random survival forest consistently outperforms the Cox prepayment-specific hazards model and the Fine-Gray subdistribution hazard model, except in the first few months. Note that for VB Risk Advisory the most important evaluation metric is the Brier score as this evaluates absolute risk predictions, while the AUC evaluates risk predictions of loans relative to each other. Thus, for future pricing purposes, the Brier score is the most relevant. The Cox prepayment-specific hazards model and the Fine-Gray subdistribution hazard model seem to perform relatively similar to each other with respect to prediction accuracy. However, for months further in the future, the prediction performance of the Fine-Gray model drops as visible in both Figure 21 and Figure 23. The Brier score of the Fine-Gray model even exceeds the Brier score of the Aalen-Johansen prediction model (as described in (64)) for times greater than 120 approximately.

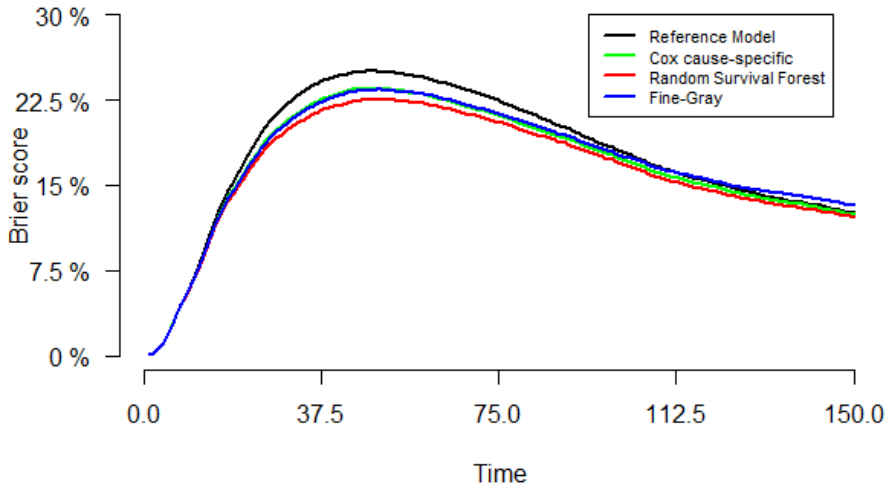


Figure 21: Brier score plotted over time for the Cox cause-specific model, the Random Survival Forest, the Fine-Gray subdistribution hazard model and the Aalen-Johansen prediction model as reference model

Table 18: Integrated Brier score for Aalen-Johansen estimator, Cox cause-specific hazards model with stratification, Fine-Gray subdistribution hazard model and random survival forest for competing risks until 150 months

| Model | Integrated Brier score |
|---|---|
| Aalen-Johansen | 0.1775 |
| Cox cause-specific with stratification | 0.1692 |
| Fine-Gray subdistribution hazard model | 0.1712 |
| Random survival forest for competing risks | 0.1640 |

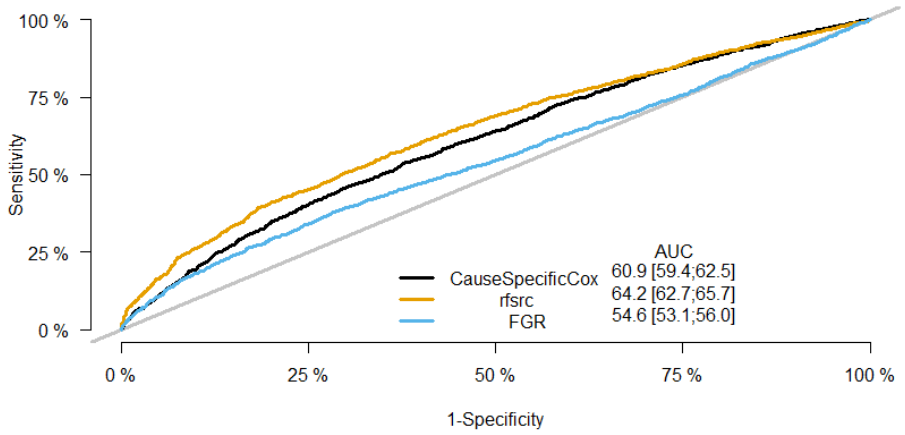


Figure 22: ROC curve for Cox cause-specific model (CauseSpecificCox), the Random Survival Forest for competing risks (rfsrc) and the Fine-Gray subdistribution hazard model (FGR) at t = 150

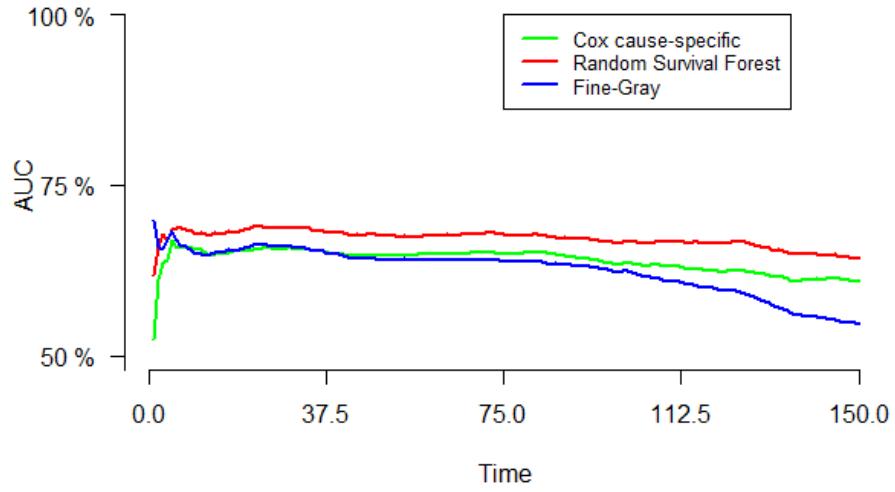


Figure 23: Time-dependent AUC plotted over time for Cox cause-specific model, the Random Survival Forest and the Fine-Gray sub-distribution model

## 6.2 Backtesting

All models discussed previously only consider time to event and not seasonal patterns. In this section, it is investigated how these models perform on unseen future data. This is relevant for translating the prepayment risk into pricing. In practice, the predicted cumulative incidence curve can be used for pricing purposes. Therefore, it is also necessary to investigate prediction performance in the future. In the literature, the data is often randomly split into training data and testing data, and prediction accuracy is measured on the testing data. A similar idea was conducted in Section 6.1. This procedure is however less relevant if our goal is predicting the absolute risk of prepayment in the future. For this, we need backtesting. The relation between variables and the cumulative incidence could change over time. In this section, the models are estimated again using the same steps as in Section 5, but now on a subset of the training sample. The subset contains the loans that originated between January 1999 and December 2010. Then, prediction accuracy is measured on the subset of loans that originated between January 2011 and December 2023. So, loans that are in the first subset do not occur in the second subset. The tables and figures of the intermediate steps in the estimation process, such as variable selection, proportionality checking, and hyperparameter tuning are not included in this thesis in order to maintain clarity.

The details of the data on 1999-2010 and 2011-2023 are tabulated in Table 19 and Table 20 respectively. In the 1999-2010 data set 53.6% of the loans is fully prepaid. In the 2011-2023 data set 52.7% of the mortgages is fully prepaid.

Table 19: Overview loan status training set 1999-2010

| State of mortgage | Number of observations |
|---|---|
| Currently active | 3,539 |
| Prepaid | 4,521 |
| Default | 376 |
| Matured | 0 |
| Total | 8,436 |

Table 20: Overview loan status test set 2011-2023

| State of mortgage | Number of observations |
|---|---|
| Currently active | 5,074 |
| Prepaid | 6,094 |
| Default | 394 |
| Matured | 1 |
| Total | 11,563 |

In the Cox cause-specific hazards model, the variables *property type*, *mortgage insurance percentage*, *first time homebuyer flag* and *channel* were removed for the prepayment-specific hazard using the variable selection procedure as described in Section 3.3.1. For the default-specific hazard, the variables *mortgage insurance percentage*, *first time homebuyer flag*, *property type* and *original loan term* were removed. After assessing the scaled Schoenfeld residuals, stratification was performed on the *unpaid principle balance* variable for the prepayment-specific hazard and on the *number of borrowers* for the default-specific hazard. Although the statistical test suggests that *region* also is non-proportional for the prepayment-specific hazard, it is decided to not stratify on this variable for several reasons. First, the coefficients of the other variables do not change much. More specifically, the same variables are statistically significant at a 5% significance level and the largest change in hazard ratio is 0.3 basis points, while for the other 5 out of 6 variables, the difference in hazard ratio is in the order of 0.001 basis point. In addition, the scaled Schoenfeld residuals plot, as presented in Figure 24, shows that the proportionality assumption is reasonable, or at least not drastically violated. And lastly, as there are now only 11,563 mortgages in the data set, using too many stratified variable leads to a low number of observations within each stratum.

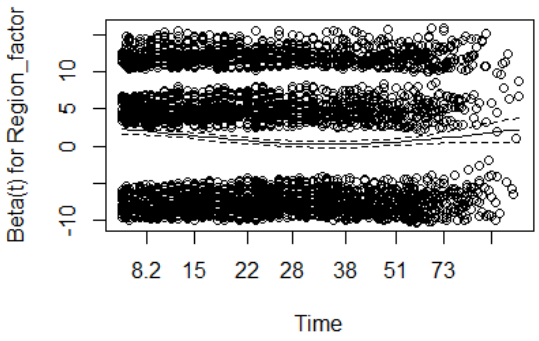


Figure 24: Scaled Schoenfeld residuals for Region on prepayment-specific hazard

In the Fine-Gray subdistribution hazard model, the variables *first time homebuyer flag*, *channel*, *property type*, *mortgage insurance percentage*, *original loan term*, *original debt-to-income ratio* were removed from the data after performing variable selection as described in Section 3.3.2. The proportionality assumption was evaluated using the plot of the scaled Schoenfeld residuals for each variable. The proportionality assumption seemed reasonable for each variable, so no stratification was performed.

In the random survival forest, variable selection based on minimal depth and VIMP resulted in the removal of the variables *first time homebuyer flag* and *property type*. The following hyperparameter values were used after hyperparameter tuning:

- Number of trees: 200
- Number of variables chosen at each split: 4
- Number of splits at each node: 2
- Minimum number of observations in a terminal node: 25

As there is less data available than in Section 6.1, it makes sense that the tree is grown deeper, which explains that the minimum number of observations in a terminal node is now lower. Note that for the variable selection and hyperparameter tuning in the random survival forest, the OOB error was used, which is not the data of the 2011-2023 test set. If we would do this, we would fit the model on the test data, which likely underestimates the prediction error.

Figures 25 and 26 show the Brier score and AUC over time of the models on 2011-2023 data. The non-parametric Aalen-Johansen estimator is used as reference model. The non-parametric Aalen-Johansen estimator is estimated on 1999-2010 data and those predicted risks are used for prediction on 2011-2023. If the models have any predictive power, they should perform better than the Aalen-Johansen estimator. As one can see in Figure 25, the considered models all perform worse than the Aalen-Johansen estimator from month 50 onwards, which suggests none of the considered models predict

absolute prepayment risk better than the non-parametric estimator. A potential explanation for this is that the relation between covariates and prepayment (in particular the CIF) change over time, or too few data is used to correctly estimate the underlying relation between the covariates and prepayment and/or default. One could potentially update the model parameters before issuing the loan to use the most recent model parameters. In this thesis, it is assumed all loans that are issued between 2011 and 2023 use to model parameters as estimated on the first of January 2011. In practice, the strategy taken is often to use data available at loan origination to predict prepayment risk. The focus of the models is on predicting short-term risk. After some time, newly available data on the loans become available and are used to evaluate current risk positions. If this proves that current risk positions are unsatisfactory, recalibration of the prepayment model is performed.

Figure 26 shows that the AUC of the Aalen-Johansen estimator is 0.5, which is true by construction, because for each loan the same risk is predicted as no covariates are used. Interestingly, the VB model performs worse than the Aalen-Johansen estimator, which indicates that it performs bad in predicting relative risks of loans. The VB performs worse than randomly guessing which loan has more prepayment risk compared to another loan. This is likely the result of changing prepayments over time. In the VB Risk Advisory model, newly issued loans are predicted to have a high probability of prepayment if there were a relatively large number of prepayments in the last 12 months. Due to changes over time of prepayment, this could lead to bad predictions of absolute and relative prepayment risk. The Cox prepayment-specific hazards model and the Fine-Gray subdistribution hazard show similar results and both outperform the random survival forest for competing risks if we look at the AUC.

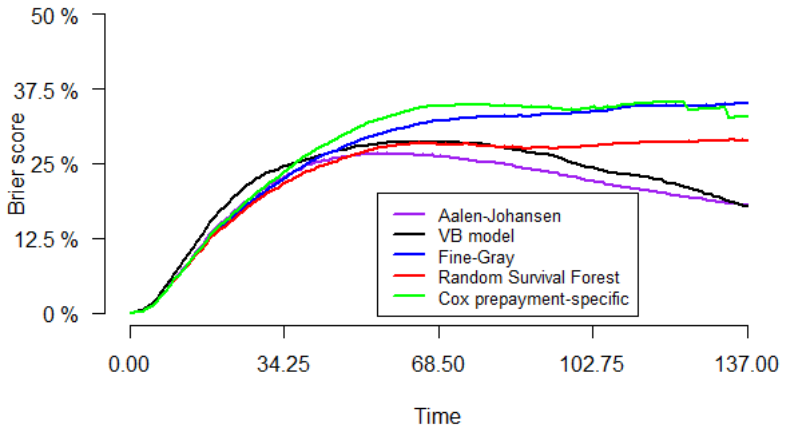


Figure 25: Brier score over time of predictions on future unseen data (2011-2023)

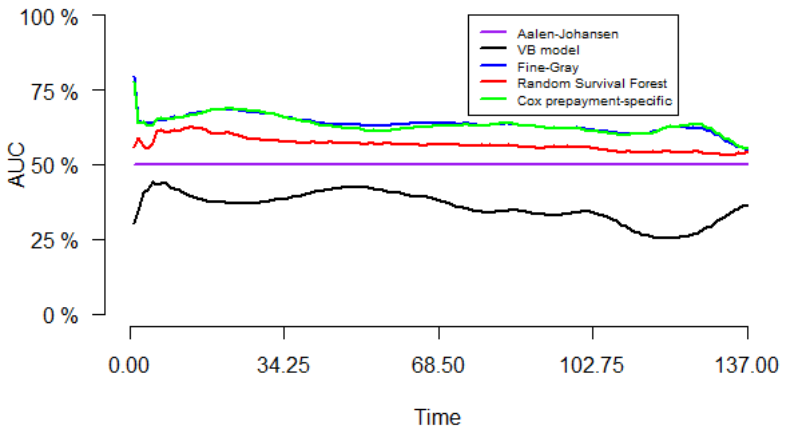


Figure 26: AUC over time of predictions on future unseen data (2011-2023)

# 7 Conclusion

This thesis examined prediction accuracy of the Cox cause-specific hazards model, the Fine-Gray subdistribution hazard model, the random survival forest for competing risks, and the current model used by VB Risk Advisory at bank X on full prepayment risk using default as competing risk. The goal of this thesis was to improve the current VB Risk Advisory model by using interpretable and computationally friendly (survival) models. Unfortunately, there is no evidence that the investigated models in this thesis significantly outperform the current model used by VB. Data was collected from the Single-Family Loan-Level data set provided by Freddie Mac on the period January 1999 until December 2023.

Variable selection was performed on all survival models. For the Cox cause-specific hazards model, the relevant variables were selected based on the Akaike Information Criterion using forward and backward selection. For the Fine-Gray subdistribution hazard model, the variables were selected based on the Bayesian Information Criterion for competing risks (BICcr) using backward selection. Lastly, in the random survival forest for competing risks, the features were selected using a selection procedure based on the Variable Importance and the Minimal Depth. Using a grid search algorithm, the hyperparameters were selected in the random survival forest for competing risks.

Using log-log survival curves and scaled Schoenfeld residuals, it appeared that the proportionality assumption of the Cox cause-specific hazards model and the Fine-Gray subdistribution hazard model was violated for several variables. Using sensitivity analysis, the effect of the violation of the proportional hazards assumption on the estimated coefficients and prediction accuracy was examined. If this violation was interpreted as significant, then stratification was used on non-proportional variables to obtain reliable estimation and prediction results.

Another goal of this thesis was to identify the risk drivers of prepayment. If the proportional hazards assumption is assumed to be true, then the Cox cause-specific hazards model suggests that *original interest rate* has the largest positive effect on the prepayment-specific hazard. The Fine-Gray subdistribution hazard model concluded that *original interest rate* has the largest effect on the subdistribution hazard for prepayment and positively affects the cumulative incidence of prepayment. The random survival forest for competing risks showed that *original interest rate* is the most important driver of prepayment risk.

To evaluate prediction accuracy of the models, the Brier score over time and the time-dependent Area under the ROC curve were used, both with Inverse Probability of Censoring Weighting. Using a test sample on data from 1999-2023, the results suggest that the random survival forest for competing risks is the best model regarding prediction accuracy over time. It can better capture the underlying relation between the covariates and the risk of prepayment than the other models.

Where most research only evaluates the models on data from the same period, this thesis also examined the prediction accuracy of the models on future unseen time peri-

ods as this is most relevant for pricing purposes. It appeared that none of the models consistently outperforms the non-parametric Aalen-Johansen prediction regarding time-dependent Brier score. Regarding time-dependent AUC, the current model of VB is the only model that performs worse than the Aalen-Johansen prediction.

This thesis has a number of limitations, from which some of them could be improved in future research. Firstly, for the prediction accuracy using backtesting, the time frame is limited both for estimation as well as prediction. This results in less available information for estimation, which affects prediction accuracy. Also, the split of the data could potentially affect estimated coefficients and predictions as prepayments and defaults are not constant over time.

In addition, the current models do not incorporate macroeconomic variables that could influence prepayment behaviour over time, such as *remaining number of months until maturity*. Although this is not possible for the Cox cause-specific hazards model and the Fine-Gray subdistribution hazard model, the random survival forest for competing risks can be extended to include time-varying covariates.

Moreover, future research could also look at prediction accuracy if the model parameters are updated each month to better incorporate variations of prepayment behaviour and relations between variables and prepayment over time. In this thesis, the same coefficients were used for mortgages that were issued for example in 2011 and 2023 to predict cumulative incidences.

Another limitation is that variable selection used in this thesis for the Cox cause-specific hazards model assesses the effect of the variable on the cause-specific hazard rather than on the cumulative incidence function. It could be that variables that do not affect the cause-specific hazard do have a significant effect on the cumulative incidence.

Furthermore, the mortgage market in the United States has different characteristics than the mortgage market in The Netherlands and hence VB Risk Advisory should test the models using data from the Dutch mortgage market before they implement the models in practice.

And lastly, loans for which the Metropolitan Statistical Area is not available were removed from the data, because at first the idea was to use this variable to select macroeconomic variables. After it was concluded it was not possible to include this in the thesis, the data set could not be updated to incorporate also the loans that had missing values of the MSA. The reason for this was that my laptop repeatedly crashed due to the large size of the data and the R file.

# 8 Closely linked papers

| Paper | Data set | Method | Result |
|---|---|---|---|
| Meis (2015) | Freddie Mac, Single-Family Loan-Level data set. 01-01-1999 until 31-09-2013. Fixed rate fully amortizing mortgages with 30 years maturity. Random sample of 10.000 taken | Comparison between Multinomial logit model, competing risk model (Kaplan-Meier), Markov models | Multinomial logit most effective |
| Szolnoki (2021) | Fannie Mae, Single-Family Fixed Rate Mortgage data set. 01-2000 until 12-2019. Fixed-term fully amortizing that are fully documented. Random sample of 425,722 mortgages. Unemployment, House Price Index, Personal Income, CPI, Number of new residential sales, Yield curve were added as macro-economic variables | Extended Cox model including time-varying covariates, discrete time logistic model, relative risk forest, conditional inference forest. Variable selection based on AIC | The two machine learning techniques outperform the traditional approaches with respect to predictive performance. Consumer Price Index, Interest rate incentive, Loan Delinquency and Loan Amount are the most important drivers for prepayments. A parallel downward shift in interest rates and a positive macroeconomic scenario increases prepayment rates. |
| Frydman and Matuszyk (2022) | Data on car lease contract status from a Polish financial institution | Random survival forest for competing risks and Fine-Gray model using prepayment as competing risk | Random survival forest for competing risks outperforms Fine-Gray model |

| | | | |
|---|---|---|---|
| Ishwaran et al. (2014) | Competing risks data on HIV/AIDS patients and simulated data | Random survival forest with competing risks, Cox cause-specific hazards model, Cox-likelihood based boosting, and Fine-Gray model | In low-dimensional linear simulations, RSF performs the worst. In low-dimensional quadratic and interaction models, RSF outperforms the other methods |
| Olajubu (2020) | Freddie Mac single family loan-level credit performance on fully amortizing fixed-rate mortgages | Cox cause-specific hazards and Fine-Gray. Hazard rates are estimated for both defaults and prepayments using competing risks | Cox cause-specific hazards model outperforms Fine-Gray model |
| Kau et al. (2009) | 30-year fixed rate single-family residential mortgages from a large financial service institution. 734,721 loans from 1976 until 2004. | Cox proportional hazard model with unobserved heterogeneity (frailty model). Choice of covariates is intuitive. LTV ratio, Original loan size and points paid at origination, interest rate spread, housing price dynamics, unemployment rate as indicator of economic conditions, FICO score | It is important to account for the within-group correlation in Metropolitan Statistical Area (MSA) among individual mortgages and hence to use frailty model. |

# References


Araj, V. (2024). Prepayment penalty: What it is and how to avoid it. https://www.rocketmortgage.com/learn/prepayment-penalty#:~:text=Most%20mortgage%20lenders%20allow%20borrowers,a%20loan%20at%20a%20time. [accessed: 16-04-2024].

ATTOM (2023). Home-mortgage lending declines again across u.s. during third quarter as mortgage rates climb. *Mortgage Origination, Real Estate News*.

Austin, C., Latouche, A., and Fine, J. P. (2019). A review of the use of time-varying covariates in the fine-gray subdistribution hazard competing risk regression model. *Statistics in Medicine*, 39(2):103–113.

Austin, P. C., Lee, D. S., and Fine, J. P. (2016). Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609.

Balan, T. A. and Putter, H. (2020). A tutorial on frailty models. *Statistical Methods in Medical Research*, 29(11):3424–3454.

Bank for International Settlements (2007). Part 2: The first pillar - minimum capital requirements. https://www.bis.org/publ/bcbs128b.pdf[accessed: 02-07-2024].

Bellera, C. A., MacGrogan, G., Debled, M., de Lara, C. T., Brouste, V., and Mathoulin-Pélissier, S. (2010). Variables with time-varying effects and the cox model: Some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Medical REsearch Methodology*, 10:1–12.

Bhardwaj, G. and Sengupta, R. (2009). Did prepayments sustain the subprime market? *CentER Discussion Paper*, 38 S.

Binder, M., Moosbauer, J., Thomas, J., and Bischl, B. (2020). Multi-objective hyperparameter tuning and feature selection using filter ensembles. In *Proceedings of the 2020 genetic and evoluationary computation conference*, pages 471–479.

Blanche, P., Latouche, A., and Viallon, V. (2013). Time-dependent auc with right-censored data: a survey. *Risk assessment and evaluation of predictions*, pages 239–251.

Borucka, J. (2014a). Extensions of cox model for non-proportional hazards purpose. *Ekonometria*, 3(45):85–101.

Borucka, J. (2014b). Methods for handling tied events in the cox proportional hazard model. *Studia Oeceonomica Posnaniensia*, 2(2):91–105.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Channel, J. (2023). Mortgage statistics: 2024. *Lendingtree*.

Chen, J. (2023). What Is a Short Sale on a House? Process, Alternatives, and Mistakes to Avoid. https://www.investopedia.com/terms/r/real-estate-short-sale.asp[accessed: 01-05-2024].

Chen, J. (2024). Real Estate Owned (REO) Definition, Advantages, and Disadvantages. https://www.investopedia.com/terms/r/realestateowned.asp[accessed: 01-05-2024].

Chen, Y.-C. (2018). *STAT 425: Introduction to Nonparametric Statistics. Lecture 5: Survival Analysis*. Washington University in St. Louis.

Chernov, M., Dunn, B. R., and Longstaff, F. A. (2018). Macroeconomic-driven prepayment risk and the valuation of mortgage-backed securities. *The Review of Financial Studies*, 31(3):1132–1183.

Clapp, J. M., Goldberg, G. M., Harding, J. P., and LaCour-Little, M. (2002). Movers and shuckers: Interdependent prepayment decisions. *Real Estate Economics*, 29(3):411–450.

Columbia University Irving Medical Center (2023). Competing risk analysis. https://www.publichealth.columbia.edu/research/population-health-methods/competing-risk-analysis[accessed: 09-04-2024].

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

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

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

European Central Bank (2024). ECB guide to internal models - Definition of default. pages 79–80.

Fan, J. and Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30(1):74–99.

Fayman, A. and He, L. T. (2011). Prepayment risk and bank performance. *The Journal of Risk Finance*, 12(1):26–40.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Association*, 94(446):496–509.

Fisher, L. D., van Belle, G., Heagerty, P. J., and Lumley, T. (2004). *Biostatistics: A Methodology for the Health Sciences*. John Wiley  Sons, second edition.

Freddie Mac (2024). Single Family Loan-Level Dataset. https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset[accessed: 23-04-2024].

Frydman, H. and Matuszyk, A. (2022). Random survival forest for competing credit risks. *Journal of the Operational Research Society*, 73(1):15–25.

Fu, Y. (2017). Combination of random forests and neural networks in social lending. *Journal of Financial Risk Management*, 6:418–426.

GeeksforGeeks.org (2022). Random forest hyperparameter tuning in python. https://www.geeksforgeeks.org/random-forest-hyperparameter-tuning-in-python/[accessed: 16-05-2024].

Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040.

Ghatasheh, N. (2014). Business analytics using random forest trees for credit risk prediction: A comparison study. *International Journal of Advanced Science and Technology*, 72:19–30.

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

Gray, R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, 16(3):1141–1154.

Green, J. and Shoven, J. B. (1983). The effects of interest rates on mortgage prepayments. *National Bureau of Economic Research*.

Groot, S. P. and Lejour, A. M. (2018). Financial incentives for mortgage prepayment behavior: Evidence from dutch micro data. *Journal of Housing Economics*, 41:237–250.

Güneş, T. and Apaydin, A. (2024). Prepayment and default risks of mortgage-backed security collateral pools. *Ege Academic Review*, 24(1):21–42.

Harrell, F. E. (2001). *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition.

Haushona, N., Esterhuizen, T. M., Thabane, L., and Machekano, R. (2020). An empirical comparison of time-to-event models to analyse a composite outcome in the presence of death as a competing risk. *Contemporary Clinical Trials Communications*, 19.

Hinchlie, S. R. (2012). Survival Analysis for Junior Researchers, Department of Health Sciences, University of Leicester.

Hosmer, D. W. and Lemeshow, S. (1999). *Applied Survival Analysis Regression Modeling of Time to Event Data*. John Wiley Sons.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis*, 1:255–273.

Huang, S. and Deng, H. (2021). *Data Analytics: A Small Data Approach*. Chapman and Hall/CRC, first edition.

Huang, X., Xu, J., and Zhou, Y. (2023). Exploring complex survival data through frailty modeling and regularization. *Mathematics*, 11(21):4440.

Hung, H. and Chiang, C.-T. (2010). Estimation methods for time-dependent auc models with survival data. *Canadian Journal of Statistics*, 38(1):8–26.

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1:519–537.

Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J., and Lau, B. M. (2014). Random survival forests for competing risks. *Biostatistics*, 15(4):757–773.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860.

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., and Lauer, M. S. (2010). High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105(489):205–217.

Ishwaran, H., Lauer, M. S., Blackstone, E. H., Lu, M., and Kogalur, U. B. (2021). randomForestSRC: random survival forests vignette. http://randomforestsrc.org/articles/survival.html[accessed: 03-04-2024].

Ishwaran, H. and Lu, M. (2018). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in Medicine*, 38(3):558–582.

Jacobs, J., Koning, R., and Sterken, E. (2005). Modelling prepayment risk. *University of Groningen*.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R*. Springer, second edition.

Jordan, J. (2017). Hyperparameter tuning for machine learning models. https://www.jeremyjordan.me/hyperparameter-tuning/[accessed: 16-05-2024].

Kamarudin, A. N., Cox, T., and Kolamunnage-Dona, R. (2017). Time-dependent roc curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology*, 17(53).

Kau, J. B., Keenan, D. C., and Li, X. (2009). An analysis of mortgage termination risks: A shared frailty approach with msa-level random effects. *The Journal of Real Estate Finance and Economics*, 42:51–67.

Kenton, W. (2022). Reperforming Loan (RPL): What it Means, How it Works. https://www.investopedia.com/terms/r/reperforming-loan.asp[accessed: 01-05-2024].

Kim, J., Lee, S., Ji Hye Kim AND, D. W. I., Lee, D., and Oh, K.-H. (2023). Comparing predictions among competing risks models with rare events: application to know-ckd study - a multicentre cohort study of chronic kidney disease. *Scientific Reports*, 13(13315).

Kish, R. J. (2022). The dominance of the u.s. 30-year fixed rate residential mortgage. *Journal of Real Estate Practice and Education*, 24(1):1–16.

Kleinbaum, D. G. and Klein, M. (2011). *Surivival Analysis. A Self-Learning Text*. Springer, third edition.

Kohl, M., Plischke, M., Leffondré, K., and Heinze, G. (2015). Pshreg: A sas macro for proportional and nonproportional subdistribution hazards regression. *Computer Methods and Programs in Biomedicine*, 118(2):218–233.

Kuk, D. and Varadhan, R. (2013). Model selection in competing risks regression. *Statistics in Medicine*, 32(18):3077–3088.

Lambert, P. C. (2017). The estimation and modeling of cause-specific cumulative incidence functions using time-dependent weights. *The Stata Journal*, 17(1):181–207.

Lau, B., Cole, S. R., and Gange, S. J. (2009). Competing risk regression models for epidemiologic data. *American Journal of Epidemiology*, 170(2):244–256.

Lee, S. W., Ryu, S. B., Kim, T. Y., and Jeon, J. Q. (2022). A comparative study on determinants of housing mortgage prepayment of individual borrowers. *Journal of Derivatives and Quantitative Studies*, 30(4):278–295.

Li, Z., Li, A., Bellotti, A., and Yao, X. (2023). The profitability of online loans: A competing risks analysis on default and prepayment. *European Journal of Operational Research*, 306(2):968–985.

Li, Z., Li, K., Yao, X., and Wen, Q. (2019). Predicting prepayment and default risks of unsecured consumer loans in online lending. *Emerging Markets Finance and Trade*, 55:118–132.

Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, 7(4):815–840.

Lyu, Q. (2021). Implementing Random Forest. https://towardsdatascience.com/implementing-random-forest-26dd3e4f55c3[accessed: 14-05-2024].

Meis, J. (2015). Modelling prepayment risk in residential mortgages.

Melnyk, R. (2022). Modelling prepayment risk in residential mortgages.

Nolan, E. K. and Chen, H.-Y. (2020). A comparison of the cox model to the fine-gray model for survival analyses of re-fracture rates. *Archives of Osteoporosis*, 15(86).

Olajubu, O. J. (2020). Competing risks of default and prepayment of mortgage market.

Ozenne, B., Sorensen, A. L., Scheike, T., Torp-Pedersen, C., and Gerds, T. A. (2017). riskregression: Predicting the risk of an event using cox regression models. *The R Journal*, 9:440–460.

Quercia, R. G. (2016). Differential impacts of structural and cyclical unemployment on mortgage default and prepayment. *The Journal of Real Estate Finance and Economics*, 53:346–367.

Quercia, R. G., Stegman, M. A., and Davis, W. R. (2007). The impact of predatory loan terms on subprime foreclosures: The special case of prepayment penalties and balloon payments. *Housing Policy Debate*, 18(2):311–346.

Quicken Loans (2023). A guide to the mortgage market. https://www.quickenloans.com/learn/how-does-the-mortgage-market-work[accessed:16-04-2024].

Robert, D. (2021). The mathematical relationship between the survival function and the hazard function. https://towardsdatascience.com/the-mathematical-relationship-between-the-survival-function-and-hazard-function-74559bb6cc34[accessed: 16-04-2024].

Rocke, D. M. (2021). Survival regression models. https://dmrocke.ucdavis.edu/Class/EPI204-Spring-2021/Lecture11SurvivalRegression.pdf[accessed: 08-04-2024].

Saha, P. and Haegerty, P. (2015). Time-dependent prediction accuracy in the presence of competing risks. *Biometrics*, 66(4):999–1011.

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

Schwartz, E. S. and Torous, W. N. (1993). Mortgage prepayment and default decisions: A poisson regression approach. *Journal of the American Real Estate and Urban Economics Association*, (4):431–449.

Segota, I. (2023). Unbox the cox: Intuitive guide to cox regressions. *Towards Data Science*.

Sestelo, M. (2017). *A short course on Survival Analysis applied to the Financial Industry*. Bookdown.org.

Stegherr, R., Allignol, A., Meister, R., Schaefer, C., and Beyersmann, J. (2020). Estimating cumulative incidence functions in competing risks data with dependent left-truncation. *Statistics in Medicine*, 39(4):481–493.

STHDA (2020). Cox proportional-hazards model. http://www.sthda.com/english/wiki/cox-proportional-hazards-model[accessed: 12-03-2024].

Suknanan, J. (2023). What is a mortgage and how does it work? *CNBC*.

Szolnoki, P. (2021). Modelling mortgage prepayments in the united states - a comparison of different methods.

Therneau, T., Crowson, C., and Atkinson, E. (2024). Multi-state models and competing risks. https://www.vps.fmvz.usp.br/CRAN/web/packages/survival/vignettes/compete.pdf[accessed: 14-05-2024].

United States Census Bureau (2022). Census regions and divisions of the united states. https://www2.census.gov/programs-surveys/economic-census/2022/geographies/reference-maps/2022-ec-regions.pdf[accessed: 03-06-2024].

University of Michigan (2015). *Modeling of Survival Data*. https://public.websites.umich.edu/~yili/lect4notes.pdf[accessed: 26-03-2024].

van der Star, T. (2022). Forecasting mortgage prepayment.

Vock, D. M., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P. E., Vazquez-Beneitez, G., and O'Connor, P. J. (2016). Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *Journal of Biomedical Informatics*, 61:119–131.

Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics*, 56(1):256–262.

Weathers, B. (2017). Comparison of survival curves between cox proportional hazards, random forests, and conditional inference forests in survival analysis. *All Graduate Plan B and other Reports*, 927.

Wright, M. N., Dankowski, T., and Ziegler, A. (2017). Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Statistics in Medicine*, 36(8):1272–1284.

Wu, H.-M. and Deng, C. (2010). A study of prepayment risks in china's mortgage-backed securitization. *China Economic Journal*, 3(3):313–326.

Xin, X. (2014). *Ties Between Event Times and Covariate Change Times in Cox Models*. PhD thesis, The University of Guelph.

Yeh, S.-T. (2002). Using trapezoidal rule for the area under a curve calculation. *Proceedings of the 27th Annulal SAS® User Group International (SUGI'02)*, pages 1–5.

Yuan, Y. and Tao, R. (2023). Prepayment and credit utilization in peer-to-peer lending. *Managerial Finance*, 49(12):1849–1864.

Zhang, X., Zhang, M.-J., and Fine, J. (2011). A proportional hazards regression model for the subdistribution with right-censored and left-truncated competing risks data. *Statistics in Medicine*, 30(16):1933–1951.

# A Proofs

## A.1 Proof relation survival function and hazard function

We want to prove that the hazard function is related to the survival function in the following way:

$$h_k(t) = -\frac{\partial}{\partial t} log(S_k(t))$$

*Proof.* As starting point, take $h_k(t) = \frac{f_k(t)}{S_k(t)}$ as in (2) and $S_k(t) = 1 - F_k(t)$ as in (1). Moreover, note the following:

$$f_k(t) = \frac{d}{dt} F_k(t)$$

If we now substitute A.1 into (2), and then substitute $F_k(t) = 1 - S_k(t)$, we get:

$$\begin{aligned} h_k(t) &= \frac{f_k(t)}{S_k(t)} \\ &= \frac{\frac{\partial}{\partial t} F_k(t)}{S_k(t)} \\ &= \frac{\partial}{\partial t}(1 - S_k(t)) * \frac{1}{S_k(t)} \\ &= -\frac{\partial}{\partial t} log(S_k(t)), \end{aligned}$$

where the last equation follows from the chain rule applied to a log function (Robert, 2021). □

## A.2 Proof survival function Cox proportional hazards model

We want to prove that the survival function in the Cox proportional hazards model is equal to:

$$S_k(t) = S_0(t)^{exp(\sum_{i=1}^{p} \beta_i X_{ki})}$$

*Proof.* As starting point, take the hazard function in the Cox proportional hazards model, which is:

$$h_k(t) = h_0(t) exp(\sum_{i=1}^{p} \beta_i X_{ki})$$

integrating both sides:

$$\begin{aligned} \int_0^t h_k(s)\, ds &= \int_0^t h_0(s) exp(\sum_{i=1}^{p} \beta_i X_{ki})\, ds \\ &= exp(\sum_{i=1}^{p} \beta_i X_{ki}) \int_0^t h_0(s)\, ds \end{aligned}$$

Using that the cumulative hazards function $H_k(t))$ is the integrated hazard rate over time $H_k(t) = \int_0^t h_k(s)\, ds$, gives

$$H_k(t) = exp(\sum_{i=1}^{p} \beta_i X_{ki}) H_0(t)$$

Using that $S_k(t) = exp(-H_k(t))$ or equivalently $H_k(t) = -log(S_k(t))$, results in:

$$-log(S_k(t)) = exp(\sum_{i=1}^{p} \beta_i X_{ki})(-log(S_0(t)))$$

and since $exp(\sum_{i=1}^{p} \beta_i X_{ki})$ is a scalar, we get:

$$log(S_k(t)) = log(S_0(t)^{exp(\sum_{i=1}^{p} \beta_i X_{ki})})$$

Applying the exponential function on both sides, yields:

$$S_k(t) = S_0(t)^{exp(\sum_{i=1}^{p} \beta_i X_{ki})}$$

(Hosmer and Lemeshow, 1999) □

## A.3 Proof of conditional probability in Cox model

To prove:

$$L_j(\beta) = \frac{exp(\sum_{i=1}^{p} \beta_i X_{ji})}{\sum_{l \in R(t_j)} exp(\sum_{i=1}^{p} \beta_i X_{li})}$$

*Proof.* Let $R(t_j)$ be the risk set at time $t_j$. As the conditional probability of mortgage $j$ is the contribution of mortgage $j$ to the partial likelihood, the conditional probability is first denoted by $L_j(\beta)$.

$$\begin{aligned} L_j(\beta) &= \mathbb{P}(\textit{mortgage } j \textit{ is prepaid at } t_j \,|1 \textit{ prepayment from } R(t_j)) \\ &= \frac{\mathbb{P}(\textit{mortgage } j \textit{ is prepaid at } t_j \,|\textit{at risk at } t_j)}{\sum_{l \in R(t_j)} \mathbb{P}(\textit{mortgage } l \textit{ is prepaid} | \textit{at risk at } t_j)} \\ &= \frac{h_0(t) exp(\sum_{i=1}^{p} \beta_i X_{ji})}{\sum_{l \in R(t_j)} h_0(t) exp(\sum_{i=1}^{p} \beta_i X_{li})} \\ &= \frac{exp(\sum_{i=1}^{p} \beta_i X_{ji})}{\sum_{l \in R(t_j)} exp(\sum_{i=1}^{p} \beta_i X_{li})} \end{aligned}$$

(University of Michigan, 2015) □

## A.4 Proof example Efron's approximation

To prove: if mortgages 1 and 2 are both prepaid at time $t_1$, then it follows that their contribution to the partial likelihood is as described in (21).

*Proof.* Let mortgage $j = 1$ and $j = 2$ both be prepaid at time $t_1$. Then, $D_1 = \{1, 2\}$ and $d_1 = 2$. Define $L_{1,2}(\beta)$ as the partial likelihood contribution of the mortgages 1

and 2. Note that at $t_1$ all $n$ mortgages are at risk and hence $R(t_1) = \{1, \ldots, n\}$ Then,

$$
\begin{aligned}
L_{1,2}(\beta) &= \prod_{j=1}^{2} \frac{exp(\sum_{i=1}^{p} \beta_i X_{ji})}{\prod_{s=1}^{2}(\sum_{l \in R(t_1)} exp(\sum_{i=1}^{p} \beta_i X_{li}) - \frac{s-1}{2} \sum_{l \in 1,2} exp(\sum_{i=1}^{p} \beta_i X_{li}))} \\
&= \prod_{j=1}^{2} \frac{exp(\sum_{i=1}^{p} \beta_i X_{ji})}{\sum_{l=1}^{n} exp(\sum_{i=1}^{p} \beta_i X_{li}) * (\sum_{l=1}^{n} exp(\sum_{i=1}^{p} \beta_i X_{li}) - \frac{1}{2} \sum_{l=1}^{2} exp(\sum_{i=1}^{p} \beta_i X_{li})))} \\
&= \frac{exp(\sum_{i=1}^{p} \beta_i X_{1i})}{\sum_{l=1}^{n} exp(\sum_{i=1}^{p} \beta_i X_{li})} * \frac{exp(\sum_{i=1}^{p} \beta_i X_{2i})}{\sum_{l=1}^{n} exp(\sum_{i=1}^{p} \beta_i X_{li}) - \frac{1}{2} \sum_{l=1}^{2} exp(\sum_{i=1}^{p} \beta_i X_{li}))} \\
&= \frac{exp(\sum_{i=1}^{p} \beta_i X_{1i})}{\sum_{l=1}^{n} exp(\sum_{i=1}^{p} \beta_i X_{li})} \\
&* \frac{exp(\sum_{i=1}^{p} \beta_i X_{2i})}{\frac{1}{2} exp(\sum_{i=1}^{p} \beta_i X_{1i}) + \frac{1}{2} exp(\sum_{i=1}^{p} \beta_i X_{2i}) + \sum_{l=3}^{n} exp(\sum_{i=1}^{p} \beta_i X_{li})}
\end{aligned}
$$

□

## A.5 Proof of log-log survival curve

We want to proof that the log-log survival curve in the Cox proportional hazards model is:

$$log(-log(S_k(t))) = log(-log(S_0(t))) + \sum_{i=1}^{p} \beta_i X_{ki}$$

*Proof.* Let $p$ be the number of covariates, let $k$ be the index for the mortgage and let $X_{ki}$ be the value of the $i'th$ covariate for mortgage $k$. Let the survival be as stated in A.2:

$$S_k(t) = S_0(t)^{e^{\sum_{i=1}^{p} \beta_i X_{ki}}}$$

Taking the natural logarithm on both sides gives:

$$log(S_k(t)) = log(S_0(t)) * exp(\sum_{i=1}^{p} \beta_i X_{ki})$$

Since $0 \leq S_k(t) \leq 1$, $log(S_k(t)) \leq 0$, both sides are multiplied with $-1$ to be able to take second natural logarithm:

$$-log(S_k(t)) = -log(S_0(t)) * exp(\sum_{i=1}^{p} \beta_i X_{ki})$$

Now, taking again the natural logarithm on both sides, gives:

$$
\begin{aligned}
log(-log(S_k(t))) &= log(-log(S_0(t)) * exp(\sum_{i=1}^{p} \beta_i X_{ki})) \\
&= log(-log(S_0(t))) + log(exp(\sum_{i=1}^{p} \beta_i X_{ki})) \\
&= log(-log(S_0(t))) + \sum_{i=1}^{p} \beta_i X_{ki}
\end{aligned}
$$

(Kleinbaum and Klein, 2011) □

## A.6 Proof of cause-specific cumulative incidence

To prove:

$$CIF_k^P(t) = \int_0^t h_k^P(s) S_k^{CR}(s)\, ds$$

*Proof.* Let $k \in \{1, \ldots, n\}$, $h_k^P(t)$ be the prepayment hazard function of mortgage $k$, $S_k^{CR}(t)$ be the survival function of both competing events together. Start with expression (30). Rewriting gives:

$$\begin{aligned} h_k^P(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T_k \leq t + \Delta t, \mathcal{S} = P | T_k > t)}{\Delta t} \\ &= \frac{1}{\mathbb{P}(T_k > t)} \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T_k \leq t + \Delta t, \mathcal{S} = P)}{\Delta t} \\ &= \frac{1}{\mathbb{P}(T_k > t)} f_k^P(t) \qquad \textit{if } f_k^P(t) \textit{ exists} \end{aligned}$$

Rewriting gives:

$$f_k^P(t) = h_k^P(t)\mathbb{P}(T_k > t) = h_k^P(t) S_k^{CR}(t)$$

Using that $CIF_k^P(t) = \int_0^t f_k^P(s)\, ds$, gives that:

$$CIF_k^P = F_k^P(t) = \int_0^t h_k^P(s) S_k^{CR}(s)\, ds$$

(Frydman and Matuszyk, 2022) □

## A.7 Proof that the effect of a covariate on the subdistribution hazard function is in the same direction as the effect of the covariate on the cumulative incidence

To prove, for each $k = 1, \ldots, n$:

$$1 - CIF_k^P(t|X) = (1 - CIF_{k0}^P(t))^{exp(\sum_{i=1}^p \beta_i X_{ki})} \tag{65}$$

suggests that a positive $\beta_i$ indicates an increase in $CIF_k^P(t|X)$.

*Proof.* Rewriting (65) gives:

$$CIF_k^P(t|X) = 1 - (1 - CIF_{k0}^P(t))^{exp(\sum_{i=1}^p \beta_i X_{ki})}$$

Take the derivative if $CIF_k(t|X)$ w.r.t. $\beta_i$. This gives:

$$\frac{\partial CIF_k^P(t|X)}{\partial \beta_i} = -(1 - CIF_{k0}^P(t))^{exp(\sum_{i=1}^p \beta_i X_{ki})} * log(1 - CIF_{k0}^P(t)) * exp(\sum_{i=1}^p \beta_i X_{ki}) X_{ki},$$

where log is the natural logarithm. Let us consider the terms on the right hand side of the equation seperately.

- As $0 < CIF_{k0}^P(t) < 1$, $log(1 - CIF_{k0}^P(t)) < 0$.

- $exp(\sum_{i=1}^{p} \beta_i X_{ki}) > 0.$
- $-(1 - CIF_{k0}^{P}(t))^{exp(\sum_{i=1}^{p} \beta_i X_{ki})} < 0.$

So, the right-hand side of the equation is positive if $X_{ki}$ is positive. Since we prove that the effect of $\beta_i$ on $CIF_k^P(t|X)$ is in the same direction as the effect of $\beta_i$ on $X_{ki}$, we can now conclude that if $\beta_i$ is positively associated with $X_{ki}$, then $\beta_i$ is also positively associated with $CIF_k^P(t|X)$. In addition, if $\beta_i$ is negatively associated with $X_{ki}$, then $\beta_i$ is also negatively associated with $CIF_k^P(t|X)$. And hence, $\frac{\partial CIF_k^P(t|X)}{\partial \beta_i} > 0$ □

## A.8 Proof of left-out observations percentage Random Forest

To prove: approximately 37% of the loans are left-out by using bootstrap sampling with replacement in the Random Forest.

*Proof.* Let $n$ be the number of loans in the data set. As we take a random sample with replacement, the probability that a loan is selected is $\frac{1}{n}$. The probability that a loan is not selected is then $1 - \frac{1}{n}$. Note that $n$ observations are taken. Then, the probability that a loan is not selected out of the $n$ drawn loans:

$$\lim_{n \to \infty} (1 - \frac{1}{n})^n = \frac{1}{e} \approx 37\%$$

(Huang and Deng, 2021) □

# B Tables

Table 22: Number of observations per state

| Variable name | Number of loans |
|---|---|
| Property State_AK | 31 |
| Property State_AL | 192 |
| Property State_AR | 108 |
| Property State_AZ | 624 |
| Property State_CA | 2665 |
| Property State_CO | 558 |
| Property State_CT | 194 |
| Property State_DC | 48 |
| Property State_DE | 63 |
| Property State_FL | 1366 |
| Property State_GA | 644 |
| Property State_HI | 50 |
| Property State_IA | 151 |
| Property State_ID | 137 |
| Property State_IL | 972 |
| Property State_IN | 425 |
| Property State_KS | 181 |
| Property State_KY | 262 |
| Property State_LA | 176 |
| Property State_MA | 565 |
| Property State_MD | 452 |
| Property State_ME | 67 |
| Property State_MI | 641 |
| Property State_MN | 491 |
| Property State_MO | 409 |
| Property State_MS | 61 |
| Property State_MT | 29 |
| Property State_NC | 616 |
| Property State_ND | 31 |
| Property State_NE | 94 |
| Property State_NH | 81 |
| Property State_NJ | 578 |
| Property State_NM | 97 |
| Property State_NV | 228 |
| Property State_NY | 730 |
| Property State_OH | 669 |
| Property State_OK | 155 |
| Property State_OR | 348 |
| Property State_PA | 698 |
| Property State_PR | 4 |
| Property State_RI | 68 |
| Property State_SC | 318 |

| | |
|---|---|
| Property State_SD | 28 |
| Property State_TN | 295 |
| Property State_TX | 1421 |
| Property State_UT | 281 |
| Property State_VA | 619 |
| Property State_VT | 26 |
| Property State_WA | 609 |
| Property State_WI | 377 |
| Property State_WV | 49 |
| Property State_WY | 17 |

Table 23: Number of observations per U.S. region

| Region of mortgage | Number of observations |
|---|---|
| Midwest | 4469 |
| Northeast | 3007 |
| South | 6849 |
| West | 5674 |

| Variable Name | Reason of selection |
|---|---|
| Channel | Third parties receive a commission for each loan they facilitate. They have thus an incentive to prepayment of their clients as they can facilitate more often, so it is assumed that loans originated by third parties positively affect the prepayment rate |
| Credit Score | A higher credit score is associated with a lower probability of default. Therefore, a higher credit score is associated with a higher probability of prepayment |
| First Time Homebuyer Flag | People who buy their first home are likely to be young and have less financial stability |
| Interest Rate | A higher interest rate is associated with higher risk and higher payments, which suggests a negative correlation with the rate of prepayment and a positive correlation with default |
| Mortgage Insurance Percentage | Theoretically, mortgage insurance is negatively related to default rates. The lower the percentage of the mortgage that is covered by insurance, the higher the probability of default (Meis, 2015). |
| Number of Borrowers | If there are more than 1 people obligated to repay the mortgage, then it intuitively suggests that this will positively affect the probability of prepayment as there likely would be more financial stability |
| Number of Units | If the underlying property consists of more than 1 unit, it has a more complex financial structure, which is assumed to lower the probability of prepayment |
| Occupancy Status | If the underlying house is the primary residence, then occupancy status is positively associated with prepayment rate in the literature (Olajubu, 2020) |

| | |
|---|---|
| Original DTI Ratio | A higher DTI Ratio results in higher risk to meet the scheduled payments and therefore it is theoretically positively associated with the rate of default and negatively related with the rate of prepayment |
| Original Loan Term | The term of the mortgage is assumed to impact prepayment behaviour, while the direction is not immediately clear |
| Original LTV | In previous research, Original Loan-to-Value ratio is found to positively affect prepayment rates (Schwartz and Torous, 1993), but in other research it is found to negatively affect prepayment rates van der Star (2022) |
| Original UPB | In previous research, Original Unpaid Principle Balance is often found to negatively affect prepayment rates (Wu and Deng, 2010) |
| Property Region | Prepayment rates are observed to differ amongst states and regions in the United States (Kau et al., 2009) |
| Property Type | Previous literature suggests that Single-Family homes are less likely to be prepaid than other property types (Meis, 2015) |

Table 24: Selected covariates from data

In Table 25 the steps of the forward and backward selection procedure using AIC are tabulated. One should interpret the table as follows:

1. In step 1 "Mod. 0" indicates the full model if no variables are removed. The column shows the AIC of the model if the variable in the "Var." column is removed from the model. A lower AIC suggests a better model fit and hence, if removing a variable leads to lower AIC then the "Mod. 0", this variable is removed. So, from Table 25, it is concluded *first time homebuyer flag* is removed.

2. As the *first time homebuyer flag* variable was removed in Step 1, the reference model has one variable less than originally. This reference model is called "Mod. 1" and it is the full model without the *first time homebuyer flag* variable. Now, one of the options is the add the FTHF as denoted by +FTHF. Step 2 suggests that removing *original loan-to-value ratio* results in a better model fit than only removing FTHF.

3. These steps are repeated until doing nothing, so not removing or adding another variable, results in the lowest AIC. This can be seen in step 4. Now, "Mod. 3" has the lowest AIC. This indicates that the variables *first time homebuyer flag*, *original loan-to-value ratio* and *channel* are removed in the variable selection procedure.

Table 25: Forward and backward selection prepayment hazard rate using AIC

| Step 1 | | Step 2 | | Step ... | Step 4 | |
|---|---|---|---|---|---|---|
| Var. | AIC | Var. | AIC | ... | Var. | AIC |
| -FTHF | 242969 | -OLTV | 242969 | ... | **Mod. 3** | 242968 |
| -OLTV | 242970 | -Chan. | 242969 | ... | +Chan. | 242969 |
| -Chan. | 242970 | -MI | 242970 | ... | +OLTV | 242969 |
| **Mod. 0** | 242970 | **Mod. 1** | 242970 | ... | +FTHF | 242970 |
| -MI | 242971 | +FTHF | 242971 | ... | -MI | 242972 |
| -ODTI | 242977 | -ODTI | 242977 | ... | -ODTI | 242975 |
| -OLT | 242979 | -#Bor | 242981 | ... | -OLT | 242978 |
| -#Bor | 242982 | -OLT | 242979 | ... | -#Bor | 242980 |
| -Prop. | 242996 | -Prop. | 242995 | ... | -Prop. | 242994 |
| -CS | 243086 | -CS | 243086 | ... | -CS | 243083 |
| -Reg | 243096 | -Reg | 243096 | ... | -Reg | 243095 |
| -OUPB | 243710 | -OUPB | 243710 | ... | -OUPB | 243726 |
| -OIR | 244742 | -OIR | 244749 | ... | -OIR | 244754 |

Abbreviations a) Chan.: Channel, b) CS: Credit Score, c) FTHF: First Time Homebuyer Flag, d) MI: Mortgage Insurance percentage, e) ODTI: Original Debt-to-Income ratio, f) OIR: Original Interest Rate, g) OLT: Original Loan Term, h) OLTV: Original Loan-to-Value, i) OUPB: Original Unpaid Principle Balance, j) Prop.: Property type, k) Reg: Region, l) #Bor: Number of borrowers.

Table 26: Forward and backward selection default hazard rate using AIC

| Step 1 | | Step 2 | | Step ... | Step 5 | |
|---|---|---|---|---|---|---|
| Var. | AIC | Var. | AIC | ... | Var. | AIC |
| -Reg | 17174 | -OLT | 17173 | ... | **Mod. 4** | 17170 |
| -OLT | 17177 | -MI | 17173 | ... | -Prop. | 17170 |
| -MI | 17177 | -FTHF | 17173 | ... | +FTHF | 17171 |
| -FTHF | 17177 | **Mod. 1** | 17174 | ... | +MI | 17172 |
| -Prop. | 17178 | -Prop. | 17175 | ... | +OLT | 17172 |
| **Mod. 0** | 17178 | -Chan. | 17178 | ... | -Chan. | 17174 |
| -Chan. | 17181 | +Reg | 17178 | ... | +Reg | 17174 |
| -OLTV | 17207 | -OLTV | 17207 | ... | -ODTI | 17235 |
| -OIR | 17237 | -OIR | 17232 | ... | -OLTV | 17241 |
| -ODTI | 17244 | -ODTI | 17239 | ... | -OIR | 17241 |
| -#Bor | 17257 | -#Bor | 17254 | ... | -#Bor | 17250 |
| -OUPB | 17259 | -OUPB | 17256 | ... | -OUPB | 17257 |
| -CS | 17585 | -CS | 17585 | ... | -CS | 17582 |

Abbreviations a) Chan.: Channel, b) CS: Credit Score, c) FTHF: First Time Homebuyer Flag, d) MI: Mortgage Insurance percentage, e) ODTI: Original Debt-to-Income ratio, f) OIR: Original Interest Rate, g) OLT: Original Loan Term, h) OLTV: Original Loan-to-Value, i) OUPB: Original Unpaid Principle Balance, j) Prop.: Property type, k) Reg: Region, l) #Bor: Number of borrowers.

Table 27: Test of proportional hazard assumption for prepayment-specific hazard after stratification of Original Interest Rate in Cox cause-specific hazards model

| Variable | chisq | df | p-value |
|---|---|---|---|
| Credit Score | 0.8364 | 1 | 0.3604 |
| Original DTI | 0.0406 | 1 | 0.8404 |
| Original UPB | 2.9132 | 1 | 0.0879 |
| Original Loan Term | 0.4638 | 1 | 0.4959 |
| Property Type | 1.0766 | 1 | 0.2995 |
| Number of Borrowers | 0.8211 | 1 | 0.3648 |
| Region | 58.3018 | 3 | 1.4e-12 |
| Mortgage Insurance Per. | 8.5674 | 1 | 0.0034 |
| Global | 65.5937 | 10 | 3.1e-10 |

Table 28: Test of proportional hazard assumption for prepayment-specific hazard after stratification of Original Interest Rate and Region in Cox cause-specific hazards model

| Variable | chisq | df | p-value |
|---|---|---|---|
| Credit Score | 0.243 | 1 | 0.622 |
| Original DTI | 0.102 | 1 | 0.749 |
| Original UPB | 1.027 | 1 | 0.311 |
| Original Loan Term | 0.865 | 1 | 0.352 |
| Property Type | 0.189 | 1 | 0.664 |
| Number of Borrowers | 0.298 | 1 | 0.585 |
| Mortgage Insurance Per. | 5.905 | 1 | 0.015 |
| Global | 8.100 | 7 | 0.324 |

Table 29: Test of proportional hazard assumption for default-specific hazard in Cox cause-specific hazards model

| Variable | chisq | df | p-value |
|---|---|---|---|
| Credit Score | 1.082 | 1 | 0.298 |
| Original DTI | 0.150 | 1 | 0.699 |
| Original UPB | 0.497 | 1 | 0.481 |
| Original Interest Rate | 2.395 | 1 | 0.122 |
| Channel | 1.086 | 1 | 0.297 |
| Property Type | 2.594 | 1 | 0.107 |
| Original Loan-to-Value | 9.850 | 1 | 0.002 |
| Number of Borrowers | 8.738 | 1 | 0.003 |
| Global | 28.625 | 8 | 0.000 |

Table 30: Test of proportional hazard assumption for default-specific hazard after stratifying OLTV in Cox cause-specific hazards model

| Variable | chisq | df | p-value |
|---|---|---|---|
| Credit Score | 1.143 | 1 | 0.285 |
| Original DTI | 0.634 | 1 | 0.426 |
| Original UPB | 0.349 | 1 | 0.555 |
| Original Interest Rate | 3.111 | 1 | 0.078 |
| Channel | 0.598 | 1 | 0.440 |
| Property Type | 1.548 | 1 | 0.214 |
| Number of Borrowers | 7.680 | 1 | 0.006 |
| Global | 18.316 | 7 | 0.011 |

Table 31: Estimation results Cox prepayment-specific hazard model with Mortgage Insurance Percentage included

| | Strata: quartiles IR and Region | | | |
|---|---|---|---|---|
| Variable name | coef | Hazard rate (exp(coef)) | s.e. | p-value |
| Credit Score | 1.411e-03 | 1.001 | 1.823e-04 | 0.000*** |
| Original DTI | -2.324e-03 | 0.998 | 8.000e-04 | 0.004** |
| Original UPB | 2.004e-06 | 1.000 | 8.032e-08 | 0.000*** |
| Original Loan Term | -3.401e-04 | 1.000 | 1.499e-04 | 0.821 |
| Property Type = SF | -7.934e-02 | 0.924 | 1.922e-02 | 0.000*** |
| Num of Borrowers = 1 | -7.044e-02 | 0.932 | 1.768e-02 | 0.000*** |
| MI Percentage | 1.431e-03 | 1.001 | 8.658e-04 | 0.098 |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 32: Backward selection prepayment subdistribution hazard rate using BICcr for Fine-Gray subdistribution hazard model

| Step 1 | | Step 2 | | Step ... | Step 5 | |
|---|---|---|---|---|---|---|
| Var. | BICcr | Var. | BICcr | ... | Var. | BICcr |
| -MI | 246199.7 | -OLTV | 246190.2 | ... | **Mod. 4** | 246175.2 |
| -OLTV | 246199.7 | -FTHF | 246190.6 | ... | -OLT | 246177.7 |
| -FTHF | 246200.1 | -Chan. | 246193.7 | ... | -Prop. | 246183.2 |
| -Chan. | 246203.2 | **Mod. 1** | 246199.7 | ... | -ODTI | 246193.3 |
| **Mod. 0** | 246209.1 | -OLT | 246200.1 | ... | -#Bor | 246216.4 |
| -OLT | 246209.6 | -Prop. | 246208.2 | ... | -Region | 246259.8 |
| -Prop. | 246217.6 | -ODTI | 246216.3 | ... | -CS | 246488.8 |
| -ODTI | 246225.8 | -#Bor | 246238.7 | ... | -OUPB | 246638.6 |
| -#Bor | 246248.2 | -Region | 246282.9 | ... | -OIR | 247729.8 |
| -Region | 246292.3 | -CS | 246506.7 | ... | | |
| -CS | 246516.1 | -OUPB | 246659.4 | ... | | |
| -OUPB | 246668.6 | OIR | 247746.2 | ... | | |
| -OIR | 247754.1 | | | ... | | |

Abbreviations a) MI: Mortgage Insurance percentage, b) OLTV: Original Loan-to-Value, c) FTHF: First Time Homebuyer Flag, d) Chan.: Channel, e) OLT: Original Loan Term, f) Prop.: Property type, g) ODTI: Original Debt-to-Income ratio, h) #Bor: Number of Borrowers, i) CS: Credit Score, j) OUPB: Original Unpaid Principle Balance, k) OIR: Original Interest Rate.

# C Figures

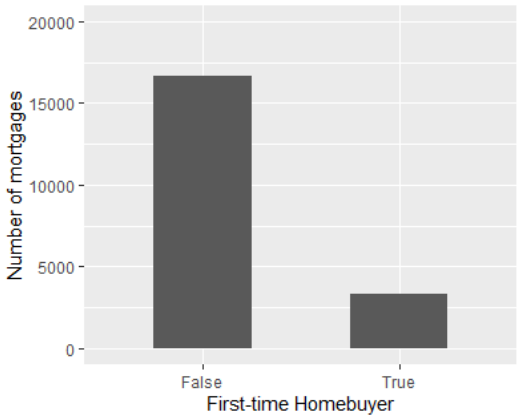


Figure 27: Number of loans with first-time homebuyers

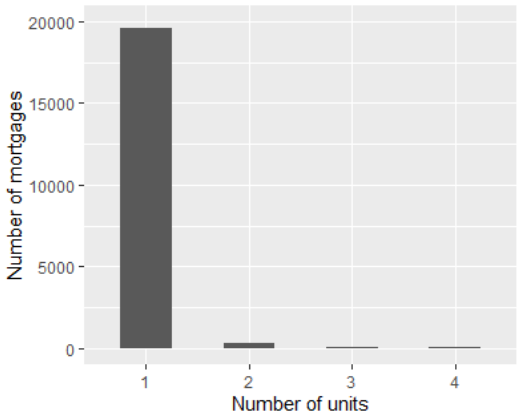


Figure 28: Number of units for mortgages

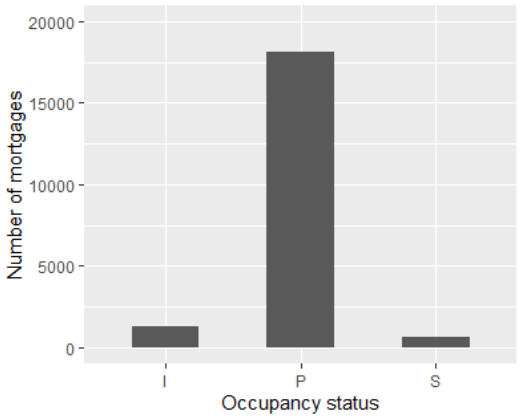


Figure 29: Distribution of occupancy status

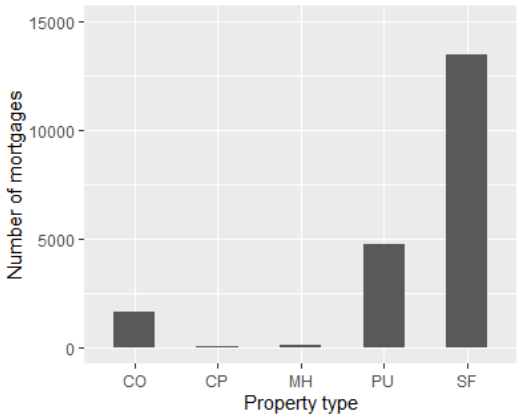


Figure 30: Distribution of property types

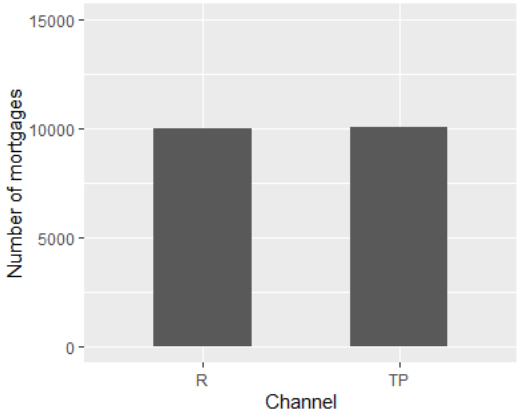


Figure 31: Distribution of channel types

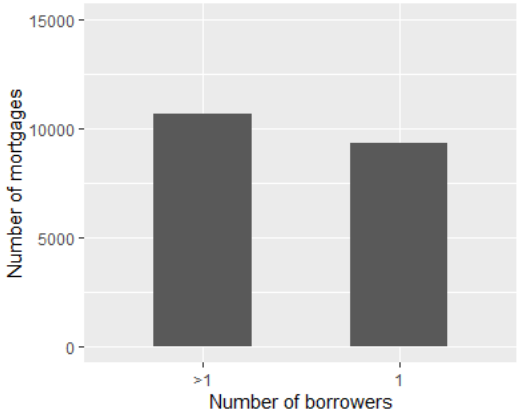


Figure 32: Distribution number of borrowers

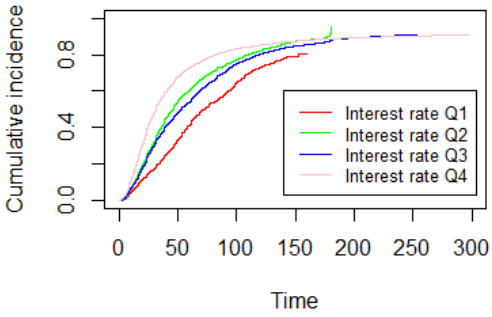


Figure 33: Cumulative incidence of prepayment for the four quartiles of interest rate

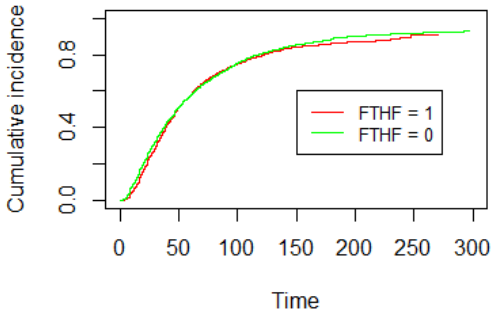


Figure 34: Cumulative incidence of prepayment for first-time homebuyer flag

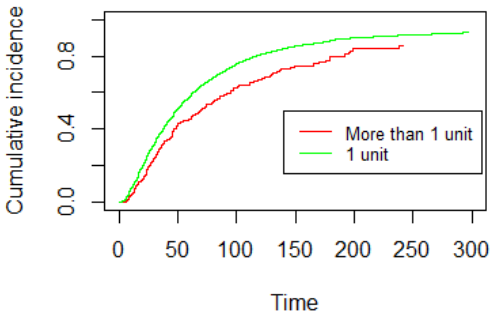


Figure 35: Cumulative incidence of prepayment for the number of units

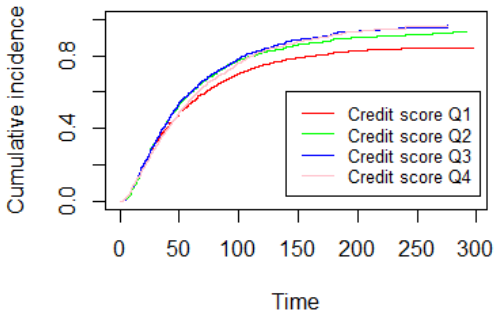


Figure 36: Cumulative incidence of prepayment for the four quartiles of credit score

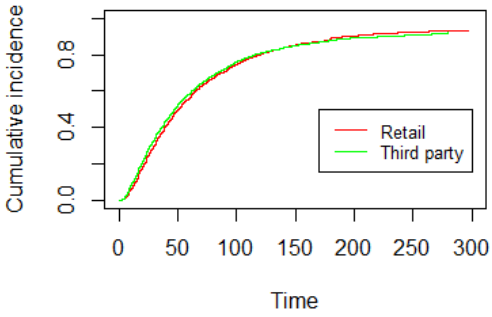


Figure 37: Cumulative incidence of prepayment for the different channels

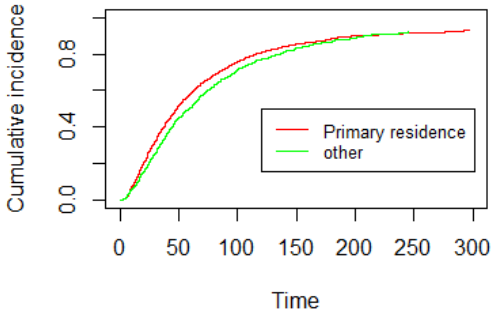


Figure 38: Cumulative incidence of prepayment for the different Occupancy Status

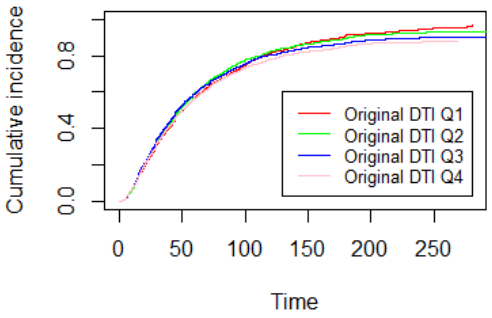


Figure 39: Cumulative incidence of prepayment for the four quartiles of Original Debt-to-Income Ratio

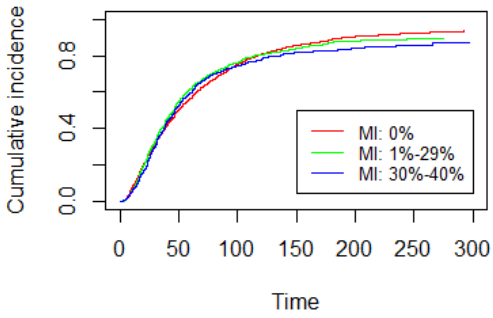


Figure 40: Cumulative incidence of prepayment for the four quartiles of Mortgage Insurance Percentage

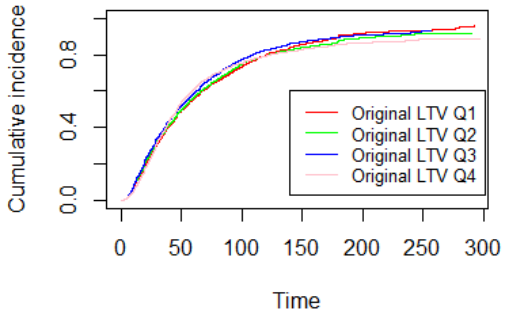


Figure 41: Cumulative incidence of prepayment for the four quartiles of Original Loan-to-Value Ratio

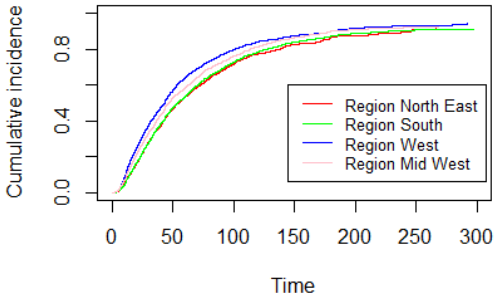


Figure 42: Cumulative incidence of prepayment for the four regions in the US

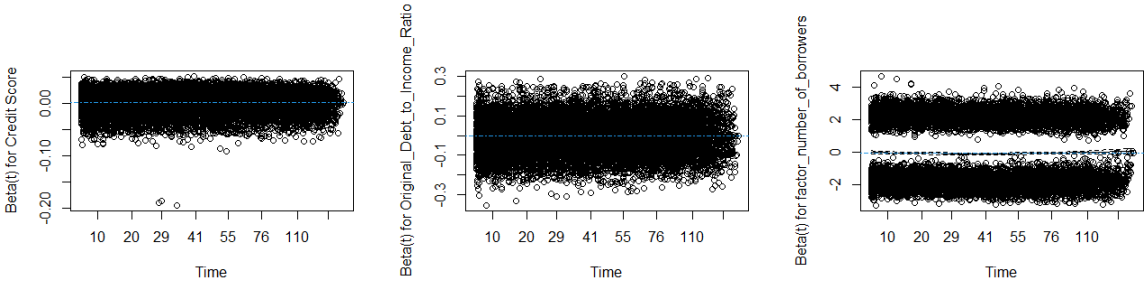


Figure 43: Schoenfeld residuals of Credit Score, Original Debt-to-Income Ratio and Number of Borrowers after controlling for non-proportional effect on Cox prepayment-specific hazard

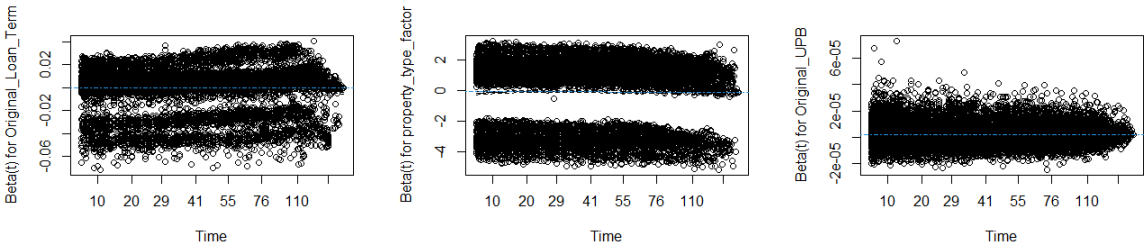


Figure 44: Scaled Schoenfeld residuals of Original Loan Term, Property Type and Original Unpaid Principle Balance after controlling for non-proportional effect on Cox prepayment-specific hazard

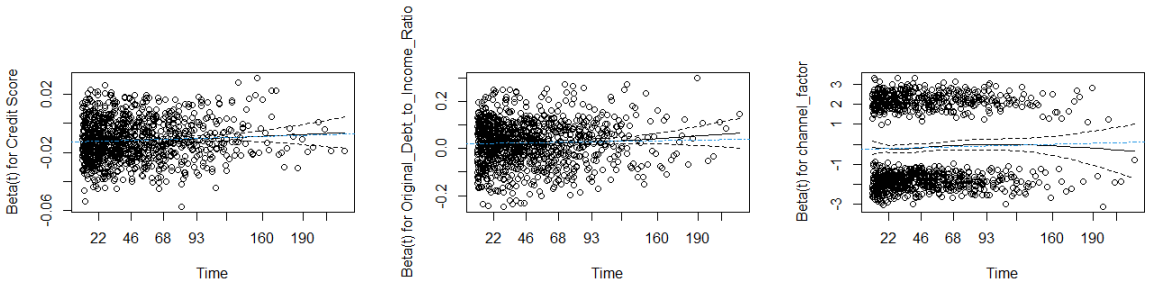


Figure 45: Scaled Schoenfeld residuals of Credit Score, Original Debt-to-Income Ratio and Channel after controlling for non-proportional effect on Cox default-specific hazard

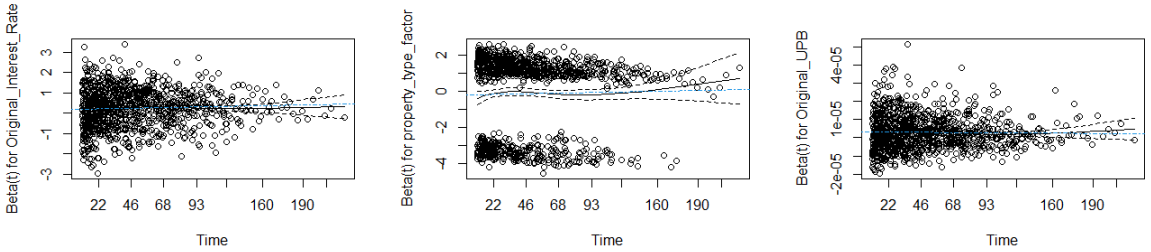


Figure 46: Scaled Schoenfeld residuals of Original Interest Rate, Property Type and Original Unpaid Principle Balance after controlling for non-proportional effects on Cox default-specific hazard

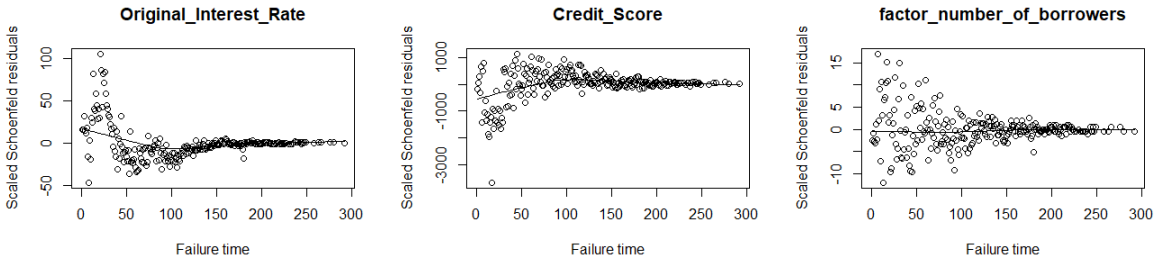


Figure 47: Scaled Schoenfeld residuals of Original Interest Rate, Credit Score and Number of borrowers in Fine-Gray subdistribution hazard model

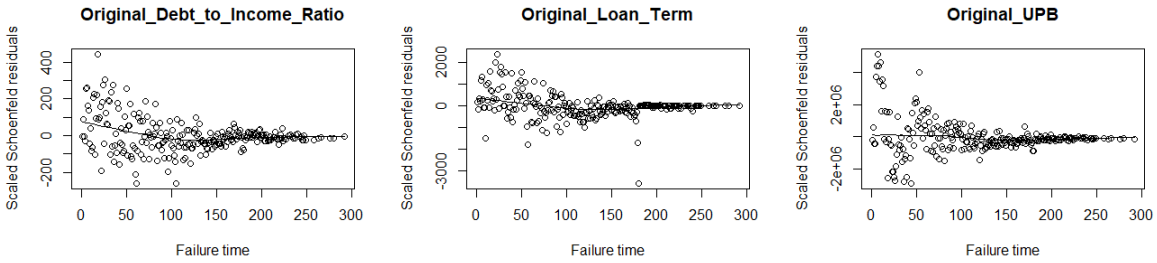


Figure 48: Scaled Schoenfeld residuals of Original Debt-to-Income Ratio, Original Loan Term and Original Unpaid Principle Balance in Fine-Gray subdistribution hazard model

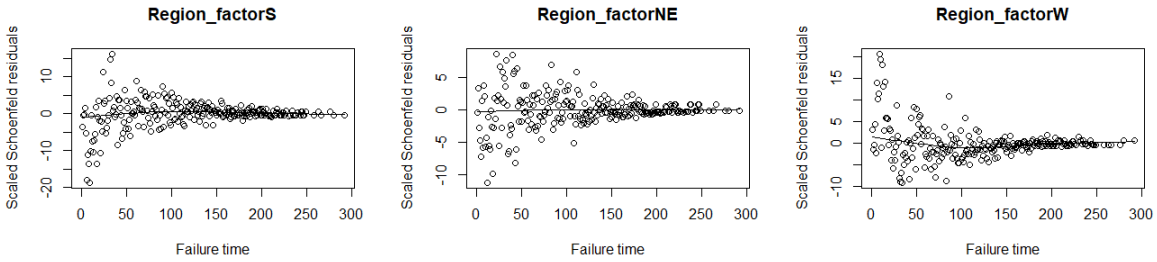


Figure 49: Scaled Schoenfeld residuals of Regions South, North-East and West in Fine-Gray subdistribution hazard model

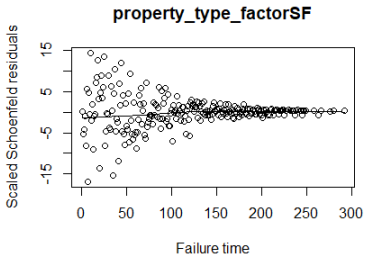


Figure 50: Scaled Schoenfeld residuals of Property Type = Single-Family in Fine-Gray subdistribution hazard model