Modelling mortgage prepayments in the United States - A comparison of different methods

by
Peter Szolnoki 614284
Msc. Tilburg University 2021

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Econometrics and Mathematical Economics

Tilburg School of Economics and Management
Tilburg University

Supervisor: dr. Martin Salm

Date: 2021 June

Acknowledgements

I would like to express my gratitude to my supervisor, dr. Martin Salm, who helped me with his remarks and useful critiques on numerous occasions. Furthermore, I am very grateful to my colleagues at Achmea Bank from whom I learned a lot and from where my interest in prepayment modelling comes from. Last, but not least: the encouragement of my family and friends was also invaluable, thank you!

Abstract

This thesis compares different modelling approaches to model early repayments (prepayments) of mortgages in the context of the US mortgage market. The relevance of the topic is attributed to the non-negligible interest rate and liquidity risk that prepayment poses for financial institutions. The approaches used in the thesis make it possible to handle censoring and to make lifetime predictions on each individual loan in a mortgage portfolio given a macroeconomic forecast. Loan-level and highly granular macroeconomic data is combined to identify the key factors behind the prepayment decision. Four models are estimated: i) a semiparameteric Extended Cox model, ii) a discrete time logistic model, iii) a relative riks forest and iv) a conditional inference forest model. The results indicate that Consumer Price Index, Interest rate incentive, Loan Delinquency and Loan Amount are the most important drivers for prepayments. The predictive performance of the models are evaluated based on different metrics and by all of these the two machine learning forest-based models outperform traditional approaches. Furthermore, sensitivity analyses indicate that a parallel downward shift in interest rates and a positive macroeconomic scenario increases prepayment rates substantially.

Table of contents

List of Figures

List of Tables

1. Introduction

From the perspective of a mortgage lender (or mortgagee) the timely payment of the scheduled principal amount in a given payment period is of great importance. If the borrower (or mortgagor) does not pay back on the loan as agreed upon in the mortgage contract, arrears can be interpreted as a signal for financial difficulty faced by the client and are associated with an increased probability of default. Historical arrears records are indeed one of the most common elements of applied credit risk models, especially for Probability of Default (PD) models. The need for the appropriate accounting of the expected credit loss faced by financial institutions became more prevalent after the Great Recession (2007-2009) and the new regulatory background was laid out in International Financial Reporting Standard number 9 (IFRS9) and in the Current Expected Credit Loss (CECL) Methodology. The calculated expected credit loss is important from a practical perspective in that it determines the capital requirement of the institution.

Another aspect of the payment schedules of the mortgages are related to early repayments or prepayments. Prepayment happens when the client pays more at a time than contractually agreed in a given payment period. This can mean a full prepayment or a partial prepayment (curtailment). The possibility of prepayment on mortgages can be viewed as an embedded American call option in a mortgage contract that creates interest rate risk for a financial institution since when a prepayment occurs, interest is no longer paid by the mortgagor on the prepaid principal amount based on the scheduled payment scheme agreed upon at loan origination. Furthermore, liquidity is also affected by prepayments through uncertainties in cash-flows created by prepayments. The prepayment option hence complicates the valuation of mortgages offered by financial institutions. For this reason, understanding driving factors behind prepayment and creating models that predict future prepayments as accurately as possible based on these factors can create monetary value by reducing risks associated with prepayments. This can contribute to more effective hedging strategies and Asset and Liability Management (ALM) and to a more efficient financial market.

Compared to credit risk modelling, the regulation for prepayment modelling for fixed-rate mortgages is usually less standardized. However, prepayment risk has also proved to be significant: most of the times in the 2015-2020 period annualized prepayment rates exceeded 10% in the US and it even reached 60% in 2004 (Federal Housing Finance Agency, 2020). Although the expected loss on prepayment is usually lower than the loss in the event of default, it is not negligible, especially when the prepayment option is priced incorrectly and when there is no prepayment penalty that can (partially) compensate for lost interest income.

Accounting for the effect of interest rates on prepayments based on option-theoretic grounds is typically expected to be an element of most of the existing prepayment models. Interest rate incentive or refinancing incentive can be defined as the difference between the contractually agreed interest rate at the start of the interest period and the prevailing actual market reference rate or refinancing rate in any given point in time. The contract rate is constant over time while the reference rate is dynamically changing over time. The bigger the difference between the fixed contract rate and the market rate,

the bigger the incentive to refinance the mortgage at a lower available rate. In reality it can also be observed that many clients do not prepay in spite of being in the money for the prepayment option. Hence, the relationship between incentive and prepayment is stochastic in practice and many other factors can have an effect on how interest rate incentive influences prepayments.

One such obvious factor is the penalty that the client is required to pay in the event of early repayment. In the US it is not as common to pay penalty for prepayment as in other countries. For example, in The Netherlands the client is usually allowed to prepay between 10 and 20% of the mortgage in a given year. Other factors such as inflation, unemployment or loan- and client-specific characteristics can also influence the exercise of the prepayment option. Understanding these factors and accounting for them in prepayment models is a challenging task that can be approached differently. With a greater availability of loan-level mortgage data, an approach that combines the option-theoretic approach and insights of data analytics can be beneficial for learning about these mechanisms and to predict and manage the risks stemming from prepayments more effectively.

The goal of this thesis is twofold. On the one hand, the first research objective is to identify how different factors (including macroeconomic variables) affect prepayments in the US. With combining many potentially relevant variables and highly granular macroeconomic and loan data, it is expected that this thesis can add new insights into the existing literature. On the other hand, the thesis also aims to compare different traditional and machine learning (ML) models based on their predictive performance.

The thesis focuses on a time-to-event analysis approach. This choice was made since mortgage data usually has a (right-)censored nature, which means that for a number of mortgages the exact survival time is unknown because the end of the duration time is not observed. With other techniques, such as linear regression or multinomial choice, it is more difficult to handle censoring. Not accounting for censoring can be problematic, because if one ignores it, the sample statistics obtained through the estimators are not representative of the parameters that relate to the survival distribution (Hosmer & Lemeshow, 1999). Furthermore, the estimation of survival curves takes into account all the available information at a given point in time to estimate the time of occurrence of a prepayment event given that it did not happened before (Banasik et al., 1999). This feature can be very valuable for practical implementation because it makes it possible to predict the evolution of a portfolio over time.

One of the goals of this thesis is to compare the predictive performance of traditional and more recent survival analytical approaches. Traditional methods include nonparametric methods such as the empirical Kaplan-Meier (KM) or product-limit survival function estimator, semiparametric methods based on Cox (1972) where the baseline hazard function is not modelled explicitly. In these types of models, the baseline hazard is multiplied by a hazard ratio which is the function of the covariates. Thus, if the hazard ratio is greater than 1, the overall hazard increases and the overall hazard decreases if the ratio is below 1. Continuous time parametric methods (such as the Weibull model or the Gompertz model) require that survival times follow a specific probability distribution which may not be fulfilled in practice. However, they have the advantage that these models have a smaller number of parameters to estimate and known distributional properties. Further-

more, in a discrete time setting, the hazard function describes a conditional probability which makes it possible to model hazard by traditional binary response models.

Another class of survival analytic method relates to machine learning techniques. One of the most commonly used ML techniques is random forest (RF). Random forest is a machine learning ensemble method which is constructed based on a collection of decision trees. In the context of survival analysis, 2 relatively new methods are examined in this thesis: conditional inference forests and random relative risk forests with time-varying covariates. These methods can add more insights into complex and nonlinear relationships between variables explaining prepayments, however it is more difficult to interpret the results of these models compared to traditional methods.

This thesis focuses on the fixed-rate mortgage-backed market in the USA. This choice was made since most of the easily publicly available mortgage data and the biggest mortgage market is in the US. The total mortgage debt outstanding exhibits an increasing trend since 2013 amounting to 16.56 trillion USD in the third quarter of 2020 (Statista, 2021a). The biggest part of this mortgage debt has been securitized and sold on the secondary market. At the time of writing this thesis the majority of MBSs consists of so-called agency securities that are guaranteed by the US government through Ginnie Mae or through government sponsored enterprises (GSEs): Fannie Mae and Freddie Mac. The main goal of these agencies is to channel liquidity into the primary mortgage market by buying mortgages from lenders through the issuance of MBSs. For this thesis, the Single-Family Fixed Rate Mortgage dataset of the Federal National Mortgage Association (FNMA or Fannie Mae) is used. Currently Fannie Mae is the largest issuer of MBSs with 3.6 trillion USD worth of MBS as of February 2021. The dataset used in the thesis contains fully-amortizing penalty-free fixed-rate mortgages originated between 2000 and 2019. The `R` (R Core Team, 2017) programming language is used for the analysis throughout the thesis.

The remainder of the thesis is structured as follows. In Section 2 a literature review of prepayment modelling is provided. In Section 3, the methodological background of survival analysis including traditional approaches and machine learning approaches, namely relative risk forest and conditional inference forests are discussed. In Section 4 the loan level and macroeconomic data is described and also descriptive analysis is performed regarding the relevant drivers of prepayments. Section 5 summarizes the main results for the different models. Section 6 is dedicated to the evaluation of the models based on different metrics and sensitivity analyses. Section 7 concludes with the discussion of the main results and an overview of limitations and directions for possible future research.

2. General context of the thesis

2.1. Brief overview of the US mortgage market

The size of the mortgage market in the United States is 16.56 trillion USD as of 2020-Q3 (Statista, 2021a). The total volume of mortgage-backed securities (MBS) was 10.9 trillion USD at this time (SIFMA, 2021). Single-family residential mortgages make up the biggest proportion of the total mortgage outstanding amount by 11.51 trillion USD. The remaining part consists of 1.72 trillion USD worth of mortgages on multifamily, 277.9 billion USD on farm property and 3.09 trillion USD of mortgages on nonresidential and non-farm properties (Statista, 2021b, 2021c, 2021d).

The distribution of mortgages by holder type shifted a lot throughout history (see Figure 1). The holder of the mortgage can be different from the mortgage originator if the mortgage was sold, but The Board of Governors of the Federal Reserve System (Fed) records this information. From the 1970s many mortgage pools and trusts entered the market and started to hold mortgages and issue private-label mortgage-backed securities. After the financial crisis, mortgages held by these entities decreased as the federal government took over a large portion of their portfolios. (FRED, 2018). Furthermore, the proportion of MBSs issued by agencies grew significantly, reaching 92% in 2020 (SIFMA, 2021). An agency MBS means that it is issued by Fannie Mae, Freddie Mac or Ginnie Mae and have an implicit credit guarantee by the federal government.

The total mortgage debt outstanding exhibited a great increase in the second half of the 20th century: in 1960-Q1 the total mortgage debt outstanding amounted to 51% of the total personal income, by 1980-Q1 this figure had risen to 64% while at the end of the century it was already 77%. In the first 7 years of the 21st century this ratio increased even more rapidly, reaching its peak value of 121% by 2007-October (see Figure 2a). The origination of new mortgages slowed down substantially after the financial crisis resulting a decrease in the total mortgage outstanding amount after a very long time. From 2014 onward the outstanding amount started to increase again, however to a more moderate pace than before the 2007-2009 financial crisis. These trends are summarized in Figure 2b.

The majority of the mortgages in the market are 30-year fixed-rate, amortizing mortgages. Before the Great Depression (1929-1933) most of the mortgages were adjustable-rate mortgages with balloon payments, where the client had to pay back the original principal at maturity. However, this resulted in a disaster during the crisis, since home owners did not refinance when house prices decreased drastically and the loan value exceeded the collateral value. To avoid this risk, several regulatory and institutional changes were introduced including the establishment of the Federal Housing Agency in 1936 and Fannie Mae in 1938 (R. K. Green & Wachter, 2005).

After the Great Depression, amortizing and fixed-rate mortgages became dominant. Amortizing mortgage refers to a payment scheme where the client pays back part of the principal amount in each month, so the outstanding principal decreases over time. The most common type of amortizing mortgage is called an annuity mortgage which follows a payment scheme where in each period the client pays the same amount. This

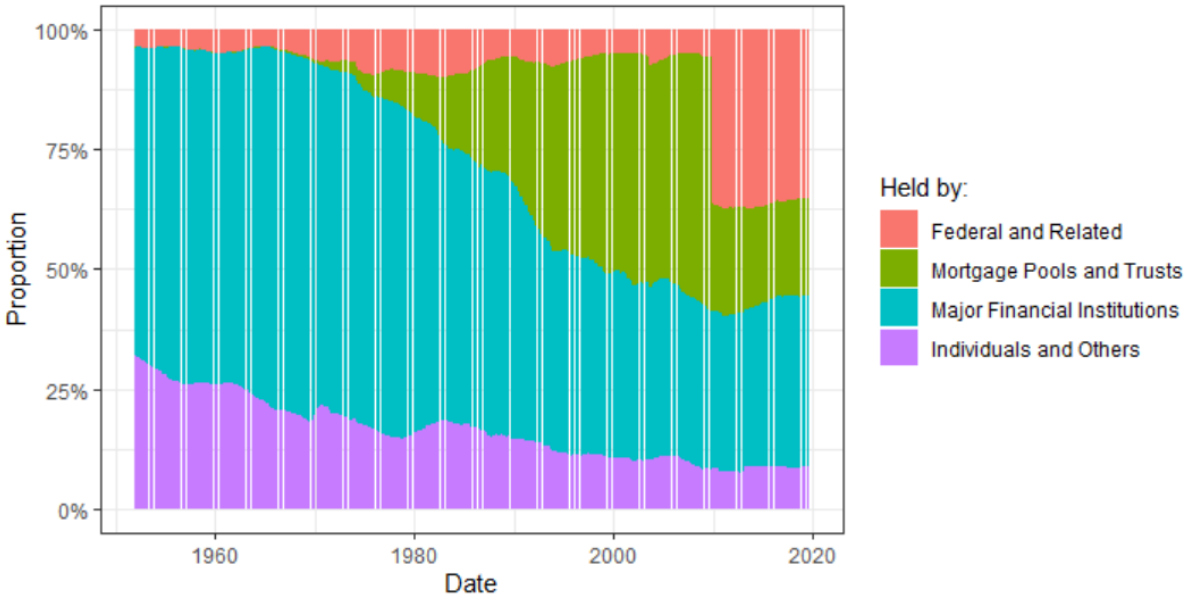


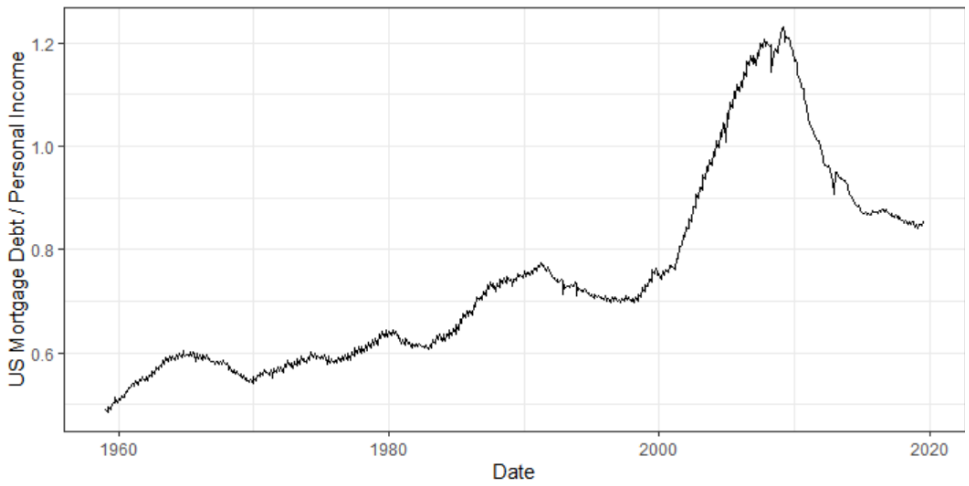
Figure 1: Mortgage Debt Outstanding by Type of Holder in the US over time, based on (FRED, 2018).

involves that the share of interest payment is decreasing while the share of principal payment for each installment is increasing with time towards the maturity date. Another, less typical kind of amortizing mortgage is the linear mortgage where the same principal amount is paid back in each payment period. However, this means that the client pays more in the earlier periods since the interest is also higher at this time.

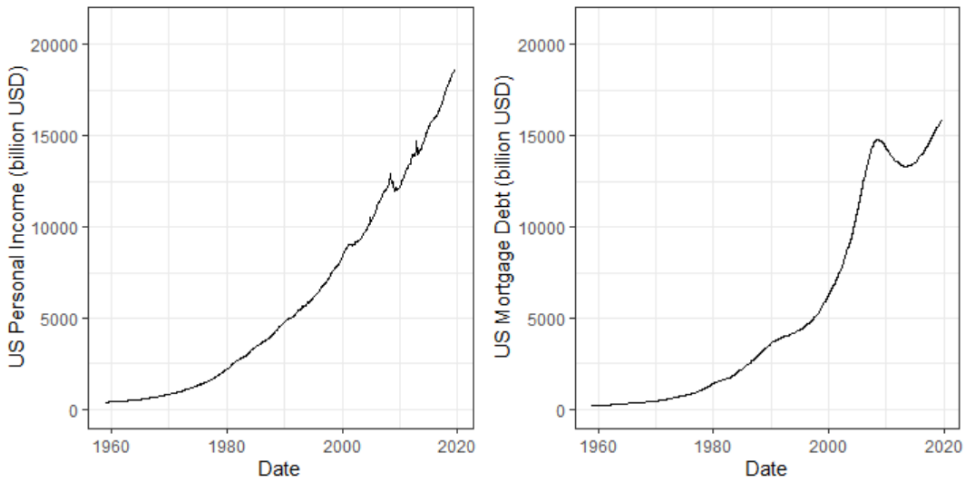
In contrast to the market in The Netherlands, most mortgages do not have a pre-specified interest reset period and prepayment penalties. Reset period is shorter than the loan term and it involves that the interest rates are adjusted from time to time. For example, a 30-year mortgage with 3 10-year reset periods means that there are possible adjustments after 10 and 20 years. Since there are usually no prepayment penalties in the US, it is expected that full prepayments are more common compared to The Netherlands.

One of the most important aspects of the mortgage market in the US is securitization. From the mid-1960s the mortgage market experienced a shortage of funds as a consequence of increasing interest rates. Congress responded to the liquidity problem by splitting Fannie Mae 1968 into a privatized "new" Fannie Mae and Ginnie Mae and creating Freddie Mac in 1970 with the goal to channel funds to the secondary mortgage market through securitization (R. K. Green & Wachter, 2005). Securitization is facilitated by creating a special purpose vehicle (SPV) which purchases mortgages from lenders, pools them and issues marketable liquid securities. The first MBS was issued by Ginnie Mae in 1970, while Freddie Mac issued its first MBS one year later. Fannie Mac issued its first MBS in 1981 (Alink, 2002). Securitization gained traction in the 1980s.

While securitization provided more liquidity in the mortgage market, it was also associated with laxer lending standards in the pre-crisis period and a surge in sub-prime mortgages. Kara et al. (2016) find that banks that were more active at issuing MBSs were more aggressive in loan pricing which may indicate that some organizations overestimated the possible risk-adjusted returns associated with trading MBSs. After the



(a) Ratio of Mortgage Debt Outstanding and Personal Income over time in the US.



(b) Personal Income and Mortgage Debt Outstanding over time in the US.

Figure 2: Mortgage market trends in the US.

financial crisis as part of the first and third quantitative easing (QE) program the Fed bought a great part of the MBSs. Furthermore, several regulatory changes were adopted to control securitization in reaction to the crisis as part of the Dodd-Frank: Title XIV - Mortgage Reform and Anti-Predatory Lending Act.

2.2. Previous research on mortgage prepayments

Prepayment models can be divided into first-generational option-theoretic models that are based on the evaluation of the optimal exercise of the call option embedded in the mortgage contract and second-generational empirical econometric approaches that rely on data analysis with the aim to identify variables that can explain prepayments. The distinction between these approaches is less pronounced nowadays since most of the empirical models include interest-rate-related variables. One exception is the Public Securities Association (PSA) prepayment benchmark model which was developed in 1985 and it simply assumes that prepayment percentage increase linearly with loan age by 0.2% until the 30th month after which a 6% prepayment rate is assumed (Schultz, 2016)). The rationale behind this model is that the number of refinancing and relocations is increasing with time. However, this model was mostly used as a benchmark and it is no longer commonly used in practice by financial institutions.

One can also make a distinction based on what level of aggregation is used in the analysis. The first prepayment models looked at mortgage pools or groups of mortgages that were assumed to be similar. The common argument behind this approach was its practical convenience and the belief that mortgage pools diversify risks associated with individual mortgages (Geanakoplos et al., 2012). This approach, however, may not be optimal since big part of the heterogeneities remain hidden within a certain pool of mortgages. With the availability of more data and computing power it became common to model prepayment using loan level data. Loan level data is frequently matched with collateral data and non-confidential client data in practice. Furthermore, most of the prepayment models also include a set of macroeconomic variables.

Dunn and McConnell (1981) were the first who developed a pricing model for mortgages based on the prepayment option for GNMA (Ginnie Mae) securities. In line with this work, first generational models aimed at expressing the option-theoretic value of the mortgage by calculating the expected present value of a mortgage by solving a partial differential equation with backward induction (Kau et al., 1992). Kau and Keenan (1995) derived the value of a mortgage with prepayment option which can include the most important characteristics as state parameters: coupon rate, outstanding balance, loan age, interest rates and property value. The last two of these parameters can be modelled with stochastic processes. Based on this result, the mortgagor should prepay on the mortgage when it is possible to refinance the mortgage for the remaining term with a lower coupon rate then for the current mortgage. This solution however does not account for costs and a presumes a "ruthless" exercise of the prepayment option. Schwartz and Torous (1989) derived a combination of option-theoretic mortgage valuation model and an estimated prepayment function. This model requires the estimation of interest rates processes since it includes 2 state variables for interest rates: the instantaneous riskless interest rate and the yield on the default-free consol bond.

Deng et al. (2005) proposes an extension of this model by adding time-varying covariates for defaults and prepayments that treats transaction costs as part of the unobserved heterogeneity. They also distinguish between option-related (e.g. interest rates, property values) and non-option-related variables (e.g. unemployment and divorce events). These extensions of option-theoretic models reflected on the fact that pure option-based models

were not able to capture prepayment behavior precisely. Hence, there has been a gradual transition toward a more econometric approach and empirical models became more dominant by the beginning of the 21st century.

Zenios and Kang (1993) added 3 variables to the refinance incentive explained by option-theoretic approaches: seasonality, seasoning and burnout. Seasonality can reflect the peak in house sales in the summer and other different seasonal patterns in different settings which may be attributed to fiscal reasons. For example, in the case of The Netherlands, a year-end peak in prepayments is commonly observed (Alink, 2002). Seasoning is related to the increasing probability of refinancing and relocation with loan age. Burnout means that after some time it becomes less likely that a client would prepay since clients with older mortgages have a bigger chance that they already had a favorable opportunity to prepay but they did not exercise this option.

Numerous other explanatory variables were included in other prepayment models which can be categorized as macroeconomic, client-related, loan-related or property-related variables. The non-exhaustive list of some commonly used variables is the following:

- Macroeconomic variables: interest rates, unemployment, GDP, inflation, yield curve steepness, house prices, housing turnover rate
- Client-related variables: client age, gender, income, credit history, family size, marital status
- Loan-related variables: contract rate, principal amount, insurance amount, redemption type, loan age, loan purpose, first loan indicator, arrears amount (delinquency)
- Property-related variables: geographical location, collateral value, property type, energy label

Some explanatory variables combine these categories. Interest rate incentive can be defined as the percentage point difference between the reference mortgage rate and the contractual rate while Loan to Value (LTV) is defined as the loan amount relative to the collateral value. The debt to income (DTI) ratio is another variable that can be used for modelling. If GDP is not available at a granular level, personal income can be a proxy for it. Furthermore, if marital status is not available, geographical divorce rate can be used, if available.

For some of the explanatory variables the potential impact on prepayment is clearer while for other variables it is more ambiguous. It is expected that unemployment is associated with worsening economic conditions hence the disposable income of some households decreases. Sirignano et al. (2016) confirms this and also finds that unemployment is a very strong explanatory variable. However, other authors (Deng et al., 2005; Foote et al., 2010) report no significant relationship. GDP by the same logic is expected to be positively correlated with prepayments, however the relevant literature is limited and Ho and Su (2006) report a negative association.

Alink (2002) suggests that the trend in the interest rates and the steepness of the yield curve also have an effect on prepayments. He suggests that less risk-averse clients may

refinance their long-term mortgages with short-term loans when the difference between the rates is considerable. Furthermore, if the yield curve is steep, it can be more attractive to take a loan with shorter maturity with the expectation that longer term interest rates will drop in the future. By doing this, clients lose some protection if interest rates begin to rise. For this reason, Alink (2002) assumes a small positive effect of yield curve steepness on prepayment in the Netherlands. Since most mortgages in the US have a maturity of 15-or 30-years, this effect may be also limited in the current setting.

When house prices increase, the client has more ability to refinance, since the loan value relative to the collateral value (LTV) decreases. This is because lower LTV can make it possible to refinance at lower rates. When house prices decline, the borrower becomes "property-constrained" since the value of the mortgage increases compared to the collateral value in relative terms and the client may be reluctant to relocate (J. R. Green & Shoven, 1983). If the LTV value exceeds 100% the property goes under water so the client may not even have the ability to change home. By similar token, when consumer prices increase, the client can become more financially constrained (if wage inflation does not keep up with CPI), so prepayment is expected to be lower. The number of home sales at a time involves more relocations hence more full prepayments (Hayre, 2003). It can be however difficult to find granular housing turnover data or identify prepayments where relocation is the main trigger. Hence variables such as client age, loan age, property type, loan term and seasonality can be proxies for relocations (Jacobs et al., 2005).

Hayre et al. (2000) observe that apartment-dwellers prepay more often. They suggest that prepayment with mortgages on apartments are more likely, since there is a bigger probability that the client will change home in the future due to upward social mobility (upgrading). This is also in line with the findings of Yiwen (2007); namely that high-income and younger single clients prepay more often in China. Younger individuals typically have an increasing expected income with age and an increasing number of children (until a certain age) which is consistent with relocations. Charlier and Van Bussel (2003) also find that apartment owners opt for prepayment more often in The Netherlands. One could also expect for these reasons that first mortgages are more likely to prepay, however, clients with mortgages in the past may have already refinanced their first mortgages so they can also have a higher propensity to prepay.

Wu and Deng (2010) find that females and clients with bigger outstanding are more likely to prepay. Abrahams (1997) suggests that since larger loans usually finance more expensive homes, relocations are less common between large homes and hence lower prepayment rates are expected. On the other hand, Alink (2002) suggest that refinancing is more beneficial financially for larger loans with the same interest rate difference between the coupon rate and the refinance rate, so a positive relationship can be expected. Literature on how gender plays a role in prepayments is scarce and the intuitive reasons are less clear. A different relationship can be found between credit scores and prepayments before and after the financial crisis. Clients with lower credit scores were more likely to prepay before the crisis than after which may be attributed to stricter mortgage underwriting standards after the crisis (Amar, 2020). The prepayment rates of delinquent clients (clients with arrears) are expected be lower since arrears is an indicator of financial difficulties.

Charlier and Van Bussel (2003) distinguished between mortgages based on redemption

types and formulate different equations to define the option-theoretical refinance incentives. Alink (2002) finds that interest-only mortgages have a higher prepayment rate and these mortgages also prepay earlier. State and regional characteristics can also affect prepayments through geography, degree urbanization and socioeconomic differences. Furthermore, better energy labels can indicate that the client invests in the house and is less likely to relocate, so prepayments are expected to be smaller.

From a modelling perspective, the first empirical models were using either ordinary least-squares (OLS) or nonlinear least-squares estimation methodologies (Chau et al., 2000; Zenios & Kang, 1993)) by defining the functional forms of the explanatory variables first separately. Nowadays the two most common econometric approaches used for prepayment modelling are (nested) multinomial or binary choice regression (An et al., 2010; Hall & Maingi, 2019; Wu & Deng, 2010) and Cox proportional hazard survival models (Charlier & Van Bussel, 2003; Ho & Su, 2006).

Recently Deep Learning and Artificial Neural Network approaches also became popular for prepayment modelling (Amar, 2020; Saito, 2018; Sirignano et al., 2016). Neural networks mimic the human brain and have proven to be extremely successful in recognizing patterns and also nonlinear relationships. One such network consists of neurons which can be grouped into 3 type of layers: input layers, usually several hidden layers and the output layer. Deep learning approaches use multiple hidden layers in a neural network to improve performance. The first hidden layer reads data from the input layers and then transforms it by aggregating the input weights and adding biases and transforming them with an activation function. The optimal weights and biases are assigned with an optimization algorithm. The predictive performance of these approaches is usually higher than for traditional survival analysis techniques since they can account for nonlinear effects and interactions. However, with these approaches it is more difficult to identify the impact of different factors on prepayment.

Another novel class of prepayment model uses Bayesian methods that provide description of probabilistic inferences dynamically. George et al. (2008) was the first who studied prepayments with applying Bayesian analysis by using a Markov Chain Monte Carlo approach. Bhattacharya et al. (2019) uses a Bayesian competing risk proportional hazards model to model default and prepayment behavior simultaneously and also develop posterior and predictive inferences. They estimate marginal posterior distributions of the baseline default and prepayment rates and also for other covariates. Based on these they can estimate predictive posterior distributions for the time of prepayment and defaults.

3. Methodological context of the thesis

The history of survival analysis dates back to the first life table that was created by John Graunt in 1662. However, the biggest advancement in the field started from the second half of the 20th century. The estimation of survival probabilities and hazard rates by Kaplan and Meier (1958) was a great breakthrough in the history of survival analysis. The development of a proportional hazard model by Cox (1972) was another big milestone in the field. With the exponentially increasing computing capacity, different machine learning (ML) techniques have been also developed to create more efficient and accurate survival models. Section 3.1 explains some preliminary concepts of survival analysis. In Section 3.2 the focus is on traditional techniques, while in Section 3.3 some machine learning approaches are examined.

3.1. Preliminaries of survival analysis

Survival analysis aims to answer questions related to how much time it takes until a certain event occurs. This time is denoted by T and any observed time is denoted by t. The event is also frequently called as failure in the literature[1]. When this event is death, the term survival analysis can be taken literally since the time until death corresponds with survival time. However, there are many other examples where the outcome of interest is not death, such as time until response by a medical treatment, time until a customer is paying subscription fee for a service or product or full prepayment in our case. When more events are in focus of the same analysis, a competing risk approach can be used (see e.g. Bhattacharya et al. (2019)). Partial prepayments (or curtailments) are out of scope of this analysis since it is not as common as full prepayment and the partially prepaid amount is significantly smaller than in case of full prepayments, as it will be seen in Figure 9 later. Furthermore, a different approach should be used to account for partial prepayments, such as recurrent time-to-event analysis or a different modelling technique.

Survival data is different from other common data type in two important aspects: a) survival times cannot be negative and b) some observations are censored i.e. the exact survival times are not observed.

3.1.1. Censoring

The most important characteristics of survival data that distinguishes it from other data types is that it contains censored observations. The most common type of censoring is right-censoring. An observation is right-censored when only the start of the spell is known but not the end, so $T>t$ (Jenkins, 2005). This can mean censoring related to a cutoff date, where the event of interest for a given subject occurs after the end of the

[1]Most of the definitions in the field of survival analysis originate from the biostatistics literature which may not be intuitive in other applications. Yet the naming conventions of survival analysis are commonly used and this thesis also follows them.

observation period, for example a subject prepays in 2020 but data is only available until 2019. Alternatively, it can also happen that a subject is not observed after a given time (but before the end of the observation period) and the event had not occurred. One such example is when survival measures prepayment and the subject defaults in 2018 and not observed later or the subject is lost for some unknown reason. An example for censoring related to cutoff date is Subject 2 in Figure 3. In this case the survival time is the length of the observed time since information after the end of the study/end date of the database is not known. An example for censoring due to other event or losing the observation is Subject 3 in Figure 3. Subject 1 is not censored and prepays/dies after 40 weeks, while Subject 4 is only observed from week 20 and not observed after week 60, so the censored survival time is 40 weeks in this case.

Left-censoring happens when it is only known that the survival time exceeds some value because the start date of the spell is not observed[2] ($t>T$). This is not very common in practice and it is also not present in this thesis since loans are observed from origination. Considering only right-censoring, the observed survival time can be expressed as:

$$T_i = \begin{cases} T_i^* & \text{if } d_i = 1 \\ C_i & \text{if } d_i = 0 \end{cases} \tag{1}$$

where T_i^* is the unobserved latent survival time, d_i is a censoring indicator which, by convention, takes a value of 0 in case of censored observations and C_i is the time of censoring. It is easy to see that $T_i = min(T_i^*, C_i)$.

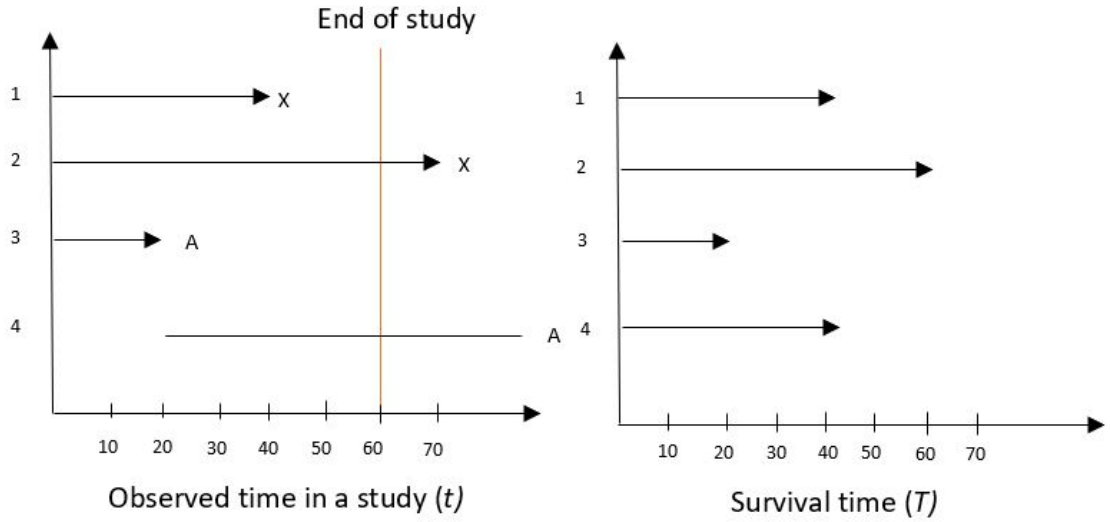


Figure 3: Examples for survival times with censored and uncensored observations. Event is denoted by X and if the subject survived it is denoted by A (own construction).

[2]This can happen for example when a patient gets enrolled in a study with symptoms and dies in 5 years. Then it is only known that the survival time is more than 5 years and survival time relates to the length of being ill.

3.1.2. Survival and hazard functions

The survival function expresses the probability of failure (the event occurring) as a function of survival time:

$$S(t) = Pr(T > t) \tag{2}$$

This function is decreasing monotonically (but not necessarily strictly) and $S(0) = 1$ and $S(\infty) = 0$. The hazard function is in deterministic relationship with the survival function as survival can be also expressed as the probability of death as a function of survival time:

$$S(t) = 1 - Pr(T < t) = 1 - F(t) \tag{3}$$

where $Pr(T < t)$ is the non-decreasing cumulative distribution function of T.

The hazard function can be defined as the instantaneous risk of experiencing an event, given that the event had not occurred beforehand:

$$h(t) = \lim_{\Delta \to 0} \frac{P(t < T < t + \Delta t | T > t)}{\Delta t} \tag{4}$$

where $\Delta > 0$. The hazard function is always greater than 0, but it cannot be interpreted as a probability since it depends on the unit of measurement of time and it can be larger than 1.

Let $f(t)$ is the probability density function (PDF) of survival time that can be expressed $f(t) = -\frac{d}{dt}S(t)$. Since the expected survival time is defined as:

$$\mathbb{E}(T) = \int_0^\infty t f(t)\, dt \tag{5}$$

,after integrating by parts the expected survival time is equivalent with the area under the survival curve:

$$\mathbb{E}(T) = \int_0^\infty S(t)\, dt. \tag{6}$$

The hazard function (or conditional failure rate) can be also expressed from the PDF and the survival function:

$$h(t) = \frac{f(t)}{S(t)} \tag{7}$$

The cumulative hazard function is the integrated hazard rate over time:

$$H(t) = \int_0^t h(x)\, dx \tag{8}$$

and based on the derivation of Jenkins (2005), the survival function can be also written as:

$$S(t) = exp\left(-\int_0^t h(x)\,dx\right) = exp(-H(t)). \tag{9}$$

3.2. Traditional survival analysis methods

In this section, non-machine learning survival analysis methods are studied. Section 3.2.1 discusses nonparametric methods such as the Kaplan-Meier (KA) and Nelson-Aalen (NA) estimators. Section 3.2.2 explains parametrics methods including the discrete logistic model and Section 3.2.3 presents different aspects of the Cox-type semiparametric method.

3.2.1. Nonparametric methods

If the distribution of survival and censoring times in a given dataset of interest follows distributions with well-known mathematical properties, it would be possible to estimate the survival curves relatively easily. However, in practice this is often not the case and the empirical distribution can add insights to the analysis. Kaplan and Meier (1958) developed an estimator for the empirical survival curves that is also called the product-limit estimator. This estimator is widely used in practical applications nowadays as well.

Assuming a discrete time setting, it is intuitively easy to see that the probability of being alive in a given time is the product of all conditional survival probabilities up to that time. Alternatively the probability of survival at time $\bar{T}$ is the probability of surviving until the previous time $T-1$ multiplied by the conditional probability of surviving after $\bar{T}$ given survival until $T-1$ (Kleinbaum & Klein, 2010). Formally we can write:

$$S(\bar{T}) = \prod_{i=1}^{\bar{T}} p_i \tag{10}$$

where p_i is the probability of surviving at time i. The Kaplan-Meier method estimates p_i by comparing the the proportion of the survived subjects at a given time to all subjects:

$$\widehat{p_i} = \frac{n_i - d_i}{n_i}, \tag{11}$$

where d_i is the number of events at at that time, and n_i is the number of individuals at risk before time that time. It is important to note that censored subjects at a given time

are not at risk. From (10) and (11) the KM estimator can be defined as:

$$\widehat{S}(t) = \prod_{i:t_i \leq t} \widehat{p}_i. \tag{12}$$

The most common standard error of the estimator can be obtained by transforming (10):

$$\ln S(\bar{T}) = \sum_{i=1}^{\bar{T}} \ln p_i \tag{13}$$

and assuming that the number of individuals surviving through time follow a binomial distribution (Collett, 2015). The estimator for the KM standard error that is obtained with this assumption is called the *Greenwood formula* based on Greenwood et al. (1926):

$$\widehat{S.E.}(\widehat{S}(t)) = \widehat{S}(t) \sqrt{\sum_{i:t_i \leq t} \left(\frac{d_i}{n_i(n_i - d_i)} \right)}. \tag{14}$$

The confidence interval around the estimated survival curves can be obtained with this formula after specifying the desired confidence level.

One may be interested whether two or more survival curves are significantly different. This can also be an important step in discovering the relevant explanatory variables for a given problem see for example Alink (2002). Although visual representation of survival curves can be indicative to answer this question, it is useful to perform a formal test. The formal test starts by formulating the hypotheses. The *null hypothesis* is that there is no significant difference between the survival curves while the *alternative hypothesis* is that there exist such a difference. The most common formal tests are the log-rank test and the Wilcoxon test.

These tests are χ^2-type where the observed and expected number of events are used as a basis and the degrees of freedom is the number of comparison groups less one. These test statistics based on Hosmer and Lemeshow (1999) can be generally defined as:

$$\frac{[\sum_{i=1}^{\bar{T}} w_i(d_{1i} - \widehat{e}_{1i})]^2}{\sum_{i=1}^{\bar{T}} w_i^2 \widehat{v}_{1i}} \tag{15}$$

where w_i is the weighting vector, $\widehat{v}$ is the estimator of the hypergeometric distribution and $\widehat{e}_{1i}$ is the expected number of events at a given time calculated based on the contingency table for group and survival status.

These tests assume noninformative censoring which means that censoring is unrelated to the outcome of interest. The log-rank and the Wilcoxon tests are similar, but while the Wilcoxon test weights the test statistic by the number of observations, the log-rank test weights the test-statistic at each time equally.

Apart from the KM estimator one another relatively common nonparametric estimator of survival curves in the Nelson-Aalen (NA) estimator:

$$\widehat{S}(t) = \prod_{i=1}^{\bar{T}} \exp(-d_i/n_i) \tag{16}$$

For small d_i/n_i the KM and NA estimators are similar.

While nonparametric methods are relatively easy to construct and can be very useful in modelling the empirical relationship behind survival times, it is more difficult to measure how certain explanatory variable affect survival times. For this reason, very often semiparametric methohds are used, see Section 3.2.3.

3.2.2. Parametric methods

It can be appealing to assume that hazard functions can be defined by distributions with well-known properties. This assumption also involves parametric survival function modelling since there is a deterministic relationship between the hazard function and the survival function as seen in (7). If this approach is possible, the parameters can be estimated by Maximum Likelihood:

$$L(\theta) = \prod_{i=1}^{n} f(t_i, \theta)^{d_i} \, S(t_i, \theta)^{1-d_i} \tag{17}$$

where θ represents the parameters of interest and d_i is the right-censoring indicator. Adjusting the likelihood contribution for censoring is necessary since for these cases only the starting time of the spell is known.

This formula can be transformed to *log-likelihood* as:

$$\ell(\theta) = \sum_{i=1}^{n} d_i \log f(t_i, \theta) + \sum_{i=1}^{n} (1 - d_i) \log S(t_i, \theta) \tag{18}$$

Often in practice the observed hazard rates do not follow any well defined distributions. Common types of continuous time parametric methods include the Weibull and Gompertz models. Details of continuous time parametric models are not discussed further, since they will not be used in this thesis. For more details, see e.g. Jenkins (2005).

However, as Berger and Schmid (2018) showed, discrete time parametric specifications make it possible to derive the hazard function from binary response regressions. To see this, one can write the discrete equivalent of the hazard and survivor functions:

$$h(t, X_i) = P(T_i = T | T_i > T, X_i) \tag{19}$$

$$S_t(t, X_i) = \prod_{i=1}^{\bar{T}} (1 - h(i, X_i)) \tag{20}$$

In this discrete specification the hazard function can be defined as a distribution function with time-varying intercepts and a coefficient vector:

$$h(t, X_i) = h(\alpha_t + \mathbf{x}_i^\top \boldsymbol{\beta}) \tag{21}$$

If we impose a logistic distribution on the hazard function we can write:

$$h(t, X_i) = \frac{exp(\alpha_t + \mathbf{x}_i^\top \boldsymbol{\beta})}{exp(1 + \alpha_t + \mathbf{x}_i^\top \boldsymbol{\beta})} \tag{22}$$

which is called the logistic hazard model. It is possible to interpret the above equation as the probability of an event happening at a given time divided by the probability that the event happens later than that time.

3.2.3. Semiparametric survival methods

Cox proportional hazards method

In order to identify the impact of different covariates on survival time, the use of semiparametric methods, and especially Cox-type models are very popular. The basic Cox-model is called the proportional hazards model by Cox (1972). The model can be written based on the hazard function:

$$h(t, \mathbf{x}) = h_0(t)\, e^{\mathbf{x}^\top \beta} \tag{23}$$

As one can see, the hazard function can be divided into two parts: $h_0(t)$ is called the baseline hazard that is not directly modelled and only depends on time. The $e^{\mathbf{x}^\top \beta}$ part on the other hand does not depend on time, but depends on the linear combination of covariates. Hence, this version of the model assumes that covariates are essentially time-independent. The proportionality refers to the property that the hazard ratios of two individuals are proportional:

$$\frac{h(t, x_1)}{h(t, x_2)} = \exp(\beta \cdot (x_1 - x_2)) \tag{24}$$

To determine the values of the coefficients of the model Cox (1975) suggested to use partial maximum likelihood estimation (PMLE). Due to proportionality it is possible to

write the probability of failure of an individual at a time as:

$$1 - p_i = \frac{exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{\sum_{j \in R_i} exp(\mathbf{x}_i^\top \boldsymbol{\beta})}. \tag{25}$$

where R_i refers to the risk set of individuals at a given time.

The Cox partial likelihood can be expressed as multiplying this expression for all times:

$$L(\beta) = \prod_{i \in K} \frac{exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{\sum_{j \in R_i} exp(\mathbf{x}_j^\top \boldsymbol{\beta})}. \tag{26}$$

where K is the number of observed event times. This expression is called partial likelihood since it does not contain censored observations. Cox (1972) argued that by maximizing this expression the estimated parameters have the same distributional properties as maximum likelihood estimators have. Tsiatis et al. (1981) formally proved the consistency and asymptotic normality properties and large sample properties of the estimator.

The log partial likelihood can be also expressed from the above expression:

$$\ell(\beta) = \sum_{i \in K} \mathbf{x}_i^\top \boldsymbol{\beta} - \sum_{i \in K} log \left[\sum_{j \in R_i} exp(\mathbf{x}_j^\top \boldsymbol{\beta}) \right]. \tag{27}$$

This estimator only works properly if there are no event ties in the data. In practice, however, ties in survival times occur frequently. Cox (1975) and Kalbfleisch and Prentice (2011) developed extensions to account for ties, however these methods can be computationally very intensive with big data, hence commonly approximations are used. The most commonly used approximations are the Breslow (1974) and Efron (1977) approximations.

Proportionality in the hazard rates is an assumption that is important to test in practice, since if the assumption is violated, the model can also be invalid. To assess proportionality, usually Schoenfeld residuals are used. These residuals can be derived from the log partial likelihood expression in (27). First, we take the derivative of the log partial likelihoods:

$$\ell'(\beta) = \sum_{i \in K} \left[\mathbf{x}_i^\top - \sum_{j \in R_i} \mathbf{x}_j \cdot \frac{exp(\mathbf{x}_j^\top \boldsymbol{\beta})}{\sum_{k \in R_i} exp(\mathbf{x}_k^\top \boldsymbol{\beta})} \right]. \tag{28}$$

We can express the individual Schoenfeld residual as:

$$\hat{r_i} = \mathbf{x}_i^\top - \sum_{j \in R_i} \mathbf{x}_j \cdot \frac{exp(\mathbf{x}_j^\top \hat{\boldsymbol{\beta}})}{\sum_{k \in R_i} exp(\mathbf{x}_k^\top \hat{\boldsymbol{\beta}})} = \mathbf{x}_i^\top - \bar{\mathbf{x}}_i^\top(t_i) \tag{29}$$

.

Under the proportional hazards assumption $\mathbb{E}\left[\bar{\mathbf{x}}_i^{\top}(t_i)\right] = 0$.

Extended Cox model: adding time-varying covariates and coefficients

Often covariates of interest (such as macroeconomic variables or income) varies over time. Luckily, the Cox model can be extended to account for time-varying covariates. We can write:

$$h(t, \mathbf{x}(t)) = h_0(t)\, e^{\mathbf{x}(t)^{\top}\beta} \tag{30}$$

The above equation allows variables to change with time. Naturally, the above expression also permits that a variable is time-constant, hence the model is usually called extended Cox model (Kleinbaum & Klein, 2010). The parameters can be determined by partial maximum likelihood estimation in this case as well:

$$L(\beta) = \prod_{k=1}^{K} \frac{exp(\mathbf{x}(\mathbf{t})_i^{\top}\boldsymbol{\beta})}{\sum_{j \in R_i} exp(\mathbf{x}(\mathbf{t})_i^{\top}\boldsymbol{\beta})}. \tag{31}$$

However, in this case both the numerator and the denominator is time-dependent. For this reason, the hazard ratio is no longer proportional in this model either:

$$\frac{h(t, x_1^*(t)}{h(t, x_2(t)} = \exp(\beta \cdot (x_1^*(t) - x_2(t))) \tag{32}$$

where $x_1^*(t)$ and $x_2(t)$ are two sets of predictors at time t for two subjects. One can see that the hazard ratio depends on time, so it is no longer constant.

The addition of time-varying covariates can be useful in practice and it also ensures that one does not use covariates values after t for predicting the outcome at t. In practice when an extended Cox model is used, the data is usually transformed into a *start-stop* (counting process) format which can result in a significantly bigger dataset if there are many time periods. This method is also often referred to as the Andersen and Gill (1982) method.

Other possible extensions of Cox-type semiparametric models include the addition of time-varying coefficients (instead of covariates), by specifying the model as:

$$h(t, \mathbf{x}) = h_0(t)\, e^{\mathbf{x}^{\top}\beta(t)} \tag{33}$$

Adding time-varying coefficients can be a good approach to use when the proportional hazards assumption is violated, however it makes interpretation more difficult.

Regularized Cox models

When the data is high-dimensional and there are more predictors, then observations ($p>n$), the partial maximum likelihood estimation fails. For this reason, it is a common technique to add a penalty term to the partial likelihood maximization equation when this problem occurs. Tibshirani (1997) proposed a LASSO approach to the Cox model by which some parameters are set to 0. An alternative solution combining lasso and ridge penalties was proposed by Park and Hastie (2007) by using the following elastic net penalty:

$$\lambda \sum_{k=1}^{p} |\beta_k| + (1-\lambda) \sum_{k=1}^{p} {\beta_k}^2 \tag{34}$$

where λ is a tuning parameter. For this approach Simon et al. (2011) developed an efficient algorithm that uses cyclical coordinate descent.

3.3. Machine learning survival analysis methods

In this section two forest-based methods are introduced with time-varying covariates: the relative risk forest and the conditional inference forest.

Both of these method are tree-based which means that the outcome can be viewed as an aggregation of numerous trees. Decision and regression tree approach rely on partitioning the feature space into a number of regions which consist of tree nodes. The node that does not have any child nodes is called the terminal node. The generation of new nodes depends on a splitting criterion and the partitioning process ends when a stopping criterion is met. Splitting is based on a measure that maximizes purity within the newly generated nodes by some metric. The outcome from one single tree can be very sensitive to the training data and can result in overfitting. However, as Breiman (2001) showed, ensemble tree methods can be attractive since they can preserve low bias while reducing variance.

The most notable ensemble methods for survival analysis purposes are random survival forest (Ishwaran et al., 2008) , relative risk forest (Ishwaran et al., 2004) and conditional inference forest (Hothorn et al., 2006). For a random forest, variance reduction is achieved by growing trees based on different bootstrapped samples from the original data and by selecting a subset of possible variables randomly at each node randomly. This results in a lower correlation between the trees which reduces the global generalization error of the ensemble (Breiman, 2001). For a conditional inference forest, multiple hypotheses are tested based on conditional distributions of candidate variables.

The originally proposed models, however, did not account for time-varying covariates. Yao et al. (2020) recently proposed extensions of conditional inference forests and relative risk forests to make these models compatible with time-dependent covariates by modifying the splitting criteria. The detailed mathematical description of these methods are out or scope of this thesis, for more details please see to the cited literature.

Both the relative risk forest and the conditional inference forest algorithms have 3 important steps:

1. Transform the data to a *start-stop* format.
2. Apply the relevant forest algorithm to the transformed data.
3. Estimate survival functions based on the output of the forest algorithms.

3.3.1. Relative risk forest

Relative risk forest relies on the methods of relative risk trees (LeBlanc & Crowley, 1992) and random forests (Breiman, 2001). Relative risk trees aim to create tree structures that represent the parametric part of the Cox-type hazard function see (23) (LeBlanc & Crowley, 1992). [3] The main steps of growing a relative risk tree can be summarized as follows (based on (Ishwaran et al., 2004)):

1. Estimate the Aalen (1978) estimator for the baseline cumulative hazard function.
2. Grow a Classification and Regression Tree (CART) by recursive partitioning by maximizing a Poisson tree likelihood using the censoring indicator as the outcome.
3. Calculate a one-step maximum likelihood estimate for each terminal node.
4. Iteratively update this estimate until meeting the stopping criterion.
5. To determine the individual hazard rate (relative to the baseline), place the subject within a terminal node. This hazard is called the relative risk.

The problem with a single tree is that it can exhibit large variance. By aggregating many relative risk trees into a relative risk forest, it is possible to reduce variance while maintaining low bias. The relative risk tree algorithm has the following steps:

1. Apply bootrapping to the data.
2. Select a prespecified number of covariates randomly from the pool of covariates.
3. Grow a full tree based on the bootrapped data and the randomly selected covariates (see relative risk tree above).
4. Repeat the process for each tree and obtain the ensemble relative risk measure for each covariate.

[3] Note that this part does not necessarily have to be a log-linear function of the parameters.

3.3.2. Conditional inference forest

The motivation behind conditional inference forest is to separate the steps of variable selection and splitting while growing the tree (Das et al., 2009). The algorithm for growing a conditional inference tree has been developed by Hothorn et al. (2006):

1. Define a generic function with nonnegative integer valued case weights.
2. At each node test the global null hypothesis whether the distribution of the right-censored survival time is independent of the conditional distribution of a given covariate is tested. Finish the algorithm if the null cannot be rejected at a prespecified level of α.
3. Select the covariate with the strongest association with the outcome.
4. Split the sample space into two based on the selected variable by a splitting criterion.
5. Modify the case weights based on the size of the nodes.
6. Repeat steps 2-6) until the stopping criterion is met.

To create a conditional inference forest, the conditional inference trees are aggregated.

3.4. Evaluation metrics for model performance

In this section commonly used performance metrics for survival analysis are described: the concordance-index or Harrel's C-index, the (integrated) Brier score and the time-dependent Area Under the (ROC) Curve (AUC).

3.4.1. Concordance index

The concordance index measures the discriminative ability of a model by comparing the order of realized and predicted survival times and measuring the ratio of concordant pairs to the total number of pairs. Censored observations are excluded since their survival times cannot be effectively compared. The value of an index ranges from 0 to 1 where 1 means perfect concordance and 0.5 shows that the discriminative ability of the model is not better than random. The C-index can be written formally as:

$$C = \frac{\sum_{j:d_i=1} \sum_{j:y_i<y_j} 1[(\mathbf{x}_i^\top \boldsymbol{\beta}) > (\mathbf{x}_j^\top \boldsymbol{\beta})]}{P} \tag{35}$$

where P is the total number of comparable pairs. One can possibly also account for tied survival times by adding $0.5[(\mathbf{x}_i^\top \boldsymbol{\beta}) = (\mathbf{x}_j^\top \boldsymbol{\beta})]$ to the numerator.

One possible shortcoming of this measure is that it only compares the survival times of subjects relative to each other and not the absolute survival times.

3.4.2. Brier score

An alternative to concordance index is to directly compare survival probabilities. The most commonly used such method is called the Brier score. The Brier score is the mean squared difference between the realized outcome and the predicted survival probability:

$$BS(t*) = \frac{1}{N}\sum_{i=1}^{N}(y - \widehat{S}(t*))^2 \tag{36}$$

where $t*$ is the time point of interest. The best predictive accuracy is reached when the score is 0, while a noninformative model would result in a 0.25 score. As one ˙can see, the Brier score is time dependent. For this reason, an integrated Brier score is commonly used which takes into consideration all the time periods of interest:

$$IBS(max(t)) = \frac{1}{max(t)}\int_0^{max(t)} BS(t)dt \tag{37}$$

3.4.3. Time-dependent AUC

Another possible way to evaluate survival model is the time-dependent Area Under the Receiver Operating Characteristics (ROC) curve. This method can be viewed as an extension to binary classification evaluation based on this metric. The ROC curve is constructed by comparing the True Positive Rate to the False Positive Rate at different classification thresholds. The AUC measures the area under this curve and its range is between 0 and 1. A 45 degree diagonal line corresponds to a 0.5 AUC value which is equivalent with a noninformative random guessing. An AUC close to one means a better classification performance while a score below 0.5 would mean that the inverse of the classifier is better than random classification.

Extending the AUC method to a survival analysis setting is possible by defining the cumulative ROC curve at each time. The area under the cumulative ROC curve measures how well the model can differentiate between subjects who die before that time and subjects who die after that time. Based on Hung and Chiang (2010), the dynamic AUC can be defined as:

$$AUC(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I(y_j > t)I(y_i \leq t)w_i I(\hat{r}(x_j) \leq \hat{r}(x_i))}{(\sum_{i=1}^{n} I(y_i > t))(\sum_{i=1}^{n} I(y_i \leq t)w_i)} \tag{38}$$

where $\hat{r}$ is the individual hazard rate and w_i denotes the inverse probability of censoring weights which are estimated separately (e.g. with a KM estimator). It is also possible to define the average AUC measure over time in one figure in a similar manner than in case of the integrated Brier score.

4. Data description and exploratory analysis

In this section the data that is used during the analysis is described. First, details on the Single-Family Fixed Rate Mortgage dataset is laid out in Section 4.1, then the relevant macroeconomic variables are described in Section 4.2. This is followed by Section 4.3 where variable transformations are explained. Finally, in Section 4.4 single-variable patterns are studied.

4.1. The Single-Family Fixed Rate Mortgage dataset

The Single-Family Fixed Rate Mortgage dataset by Fannie Mae is the main data source that is used for the analysis. This dataset is publicly available [4] and contains loan level data with monthly frequency between 2000 and 2020 for single-family loans. For the thesis the period between 2000 January and 2019 December is used. Separate data files and possible historical corrections are published quarterly after 4 months of the relevant quarter. The dataset contains acquisition and loan performance data. This means that some variables are related to loan origination and are static while other variables are dynamic and change monthly. The loan performance data for a loan until the latest available date is included in the file that is related to origination date. Hence the structure of the dataset can be represented as on Figure 4.

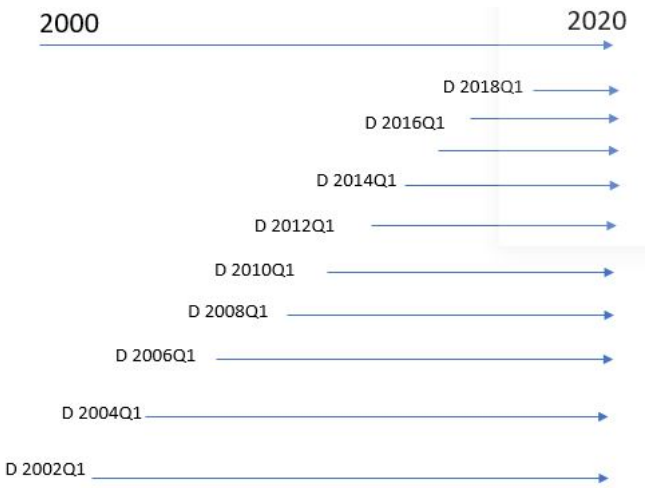


Figure 4: Single-Family Fixed Rate Mortgage data files structure.

The dataset contains fixed-term fully amortizing, single-family mortgages acquired by Fannie Mae for which full documentation is available and are originated after 1999. Mortgage loans with prepayment penalties, government insured loans, Home Affordable Refinance Program (HARP) mortgage loans, Refi Plus mortgage loans, or non-standard mortgage loans are excluded. Furthermore, loans with LTV larger than 97%, with bi-weekly payment schedule, loans subject to long-term standby commitments or third-party

[4] https://capitalmarkets.fanniemae.com/credit-risk-transfer/single-family-credit-risk-transfer/fannie-mae-single-family-loan-performance-data

risk-sharing arrangements and loans acquired on a negotiated bulk basis are excluded (Fannie Mae, 2021a). The dataset contains mortgages from all the 50 US states and Washington D.C. and also some mortgages from unincorporated territories: Puerto Rico, Guam and the United States Virgin Islands.

There are altogether 44,489,082 unique loans in the total dataset which corresponds to 552 gigabytes of data after unzipping the files and saving them to an external drive. Since most of the mortgages are observed for more than 48 months (which are different observations), it is not feasible to analyse the total dataset due to memory constraints. Furthermore, some quarterly *.csv* files are bigger than 10 gigabytes which make them difficult to read. For these reasons the following strategy is followed:

1. Read only the loan IDs from each *.csv* file create a vector of the unique loan IDs from each file (since in one file 3 months are included).
2. Merge these IDs and take a random sample and obtain 1% of the loan IDs without replacement.
3. Convert the *.csv* files into SQLite databases.
4. Query all relevant fields from the SQLite databases for the selected loan IDs.
5. Bind together all the tables obtained in the previous step.

After following these steps, an unbalanced sample panel of 425,722 mortgages is obtained with 21,715,489 observations.

In the raw dataset the following relevant variables are available:

Table 1: Variable description (Fannie Mae, 2021b).

Variable name	**Description**
Monthly Reporting Period	The month and year of the period.
Seller Name	The name of the seller of the mortgage.
Original Interest Rate	The original interest rate on a mortgage
Current Interest Rate	The rate of interest in effect for the periodic installment due.
Original UPB	The dollar amount of the loan as stated on the note at the time the loan was originated.
Current Actual UPB	The current actual outstanding unpaid principal balance of a mortgage loan.
Original Loan Term	The number of months in which regularly scheduled borrower payments are due defined at origination.
Origination Date	The date of each individual note.
Loan Age	The number of calendar months since the mortgage loan's origination date.
Remaining Months to Legal Maturity	The number of calendar months remaining until the mortgage loan is due to be paid in full based on the maturity date as defined in the mortgage documents.

Remaining Months To Maturity	The number of calendar months remaining until the outstanding unpaid principal balance of the mortgage loan amortizes to a zero balance, taking into account any additional prepayments.
Maturity Date	The month and year in which a mortgage loan is scheduled to be paid in full.
Original Loan to Value Ratio (LTV)	The ratio, expressed as a percentage, obtained by dividing the amount of the loan at origination by the value of the property.
Number of Borrowers	The number of individuals obligated to repay the mortgage loan.
Debt To Income (DTI)	The ratio obtained by dividing the total monthly debt expense by the total monthly income of the borrower.
Borrower Credit Score at Origination	A numerical value used by the financial services industry to evaluate the quality of borrower credit.
First Time Home Buyer Indicator	An indicator that denotes if the borrower or co-borrower qualifies as a first-time homebuyer.
Loan Purpose	An indicator that denotes whether the mortgage loan is either a refinance mortgage or a purchase money mortgage.
Property Type	An indicator that denotes whether the property type secured by the mortgage loan is a condominium, co-operative, planned urban development (PUD), manufactured home, or single-family home.
Occupancy Status	The classification describing the property occupancy status at the time the loan was originated.
Property State	A two-letter abbreviation indicating the state of the property.
Metropolitan Statistical Area (MSA)	The numeric Metropolitan Statistical Area Code for the property.
Mortgage Insurance Percentage	The original percentage of mortgage insurance coverage obtained for an insured mortgage loan.
Current Loan Delinquency Status	The number of months the obligor is delinquent as determined by the governing mortgage documents.
Loan Payment History	The coded string of values that describes the payment performance of the loan over the most recent 24 months.
Modification Flag	An indicator that denotes if the mortgage loan has been modified.
Zero Balance Code	A code indicating the reason the loan's balance was reduced to zero or experienced a credit event, if applicable.
Zero Balance Effective Date	Date when the loan balance was reduced to zero.
UPB at the Time of Removal	The unpaid principal balance of the loan at the time of removal.

Foreclosure Date	The date on which the completion of the legal action of foreclosure occurred.
Borrower Assistance Plan	An indicator that denotes the type of assistance plan that the borrower is enrolled.

The number of observations and the total principal amounts in the 1% sample and the number of newly originated loans and outstanding amounts are presented in Figure 5. It can be seen that the size of the sample most often grew over time and we can observe a very big increase both in the number of loans and total outstanding amount in 2002-2003. Furthermore, in most of the months there were more than 1000 new loans originated usually above 200 million USD in each month, but also several spikes can be observed.

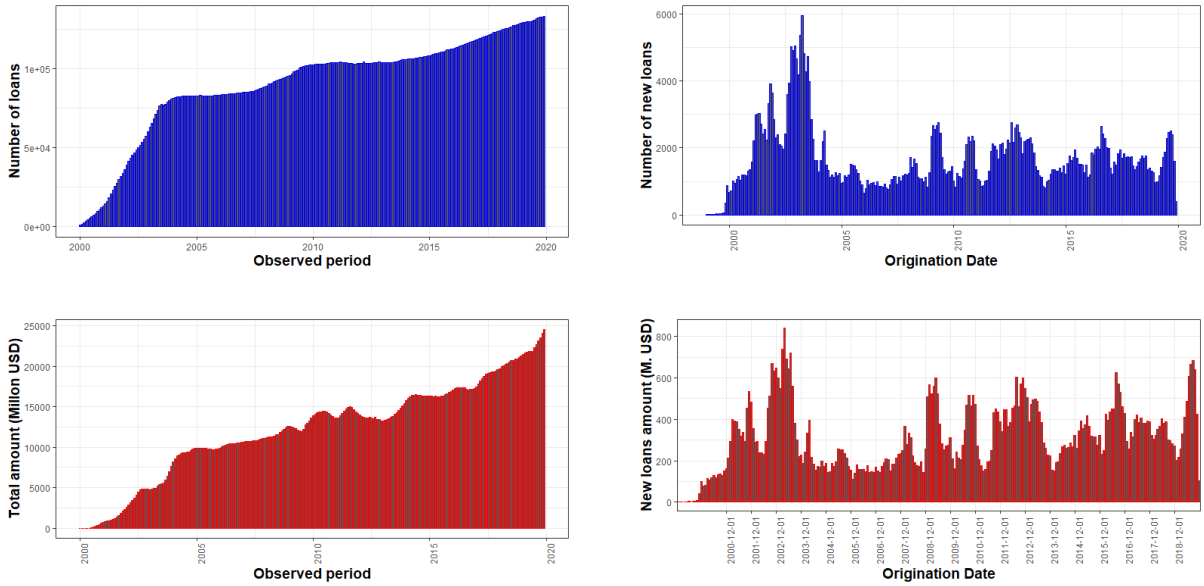


Figure 5: Total and new number of loans and total and new outstanding principal amounts.

The full prepayment indicator outcome variable can be derived from the Zero Balance code variable which is non-missing in the month when a loan balance is reduced to 0. Matured and prepaid loans are categorized together in the data, so first loans are recoded as matured when the smallest adjusted remaining interest period is smaller than or equal to 3 months. At the latest date (December, 2019) the final states of the loans are summarized in Table 2. One can see that the proportion of prepaid loans nearly reaches 66% and less than 2% of the loans matured. This is not surprising, since most of the loan terms are 30 years and the time period is 20 years. Other terminations are less frequent; none of these termination events happen significantly more than 1% of the loans. Among these, Real-Estate-Owned (REO) disposition is the most common which refers to the completion of a liquidation process.

The proportion of the fully prepaid loans in each month is presented in Figure 6. This figure is annualized using the $CPR_{annualized} = 1-(1-CPR_{monthly})^{12}$ formula. There was a big spike in prepayments in 2002-2003 which also corresponded with a great decrease in mortgage rates and increased refinancing incentive (see Figure 10). In the recent years, annualized prepayment rates remained around 10%, but in the end of 2019 a somewhat bigger spike can be observed.

The 10 biggest sellers to Fannie Mae in terms of the number of loans are listed in Table 3. The geographical distribution of the mortgages largely follows the population of the states and territories. California contributed to the total number of mortgages by more than 14%, the second and the third largest states are Texas and Florida around 5% as it can be seen in Figure 7.

Additional details and summary statistics about the Single-Family Fixed Rate Mortgage dataset are presented in Appendix A.

Table 2: Latest state of loans as of 2019 December.

State	Count	Proportion
Active	117,015	30.9%
Prepaid	279,296	65.6%
Matured	7,268	1.7%
Third Party Sale	497	0.1%
Short Sale	1,063	0.3%
Repurchased	715	0.2%
Deed-in-Lieu; REO Disposition	4,294	1%
Notes Sales	278	0.1%
Reperforming Loan Sale	638	0.1%

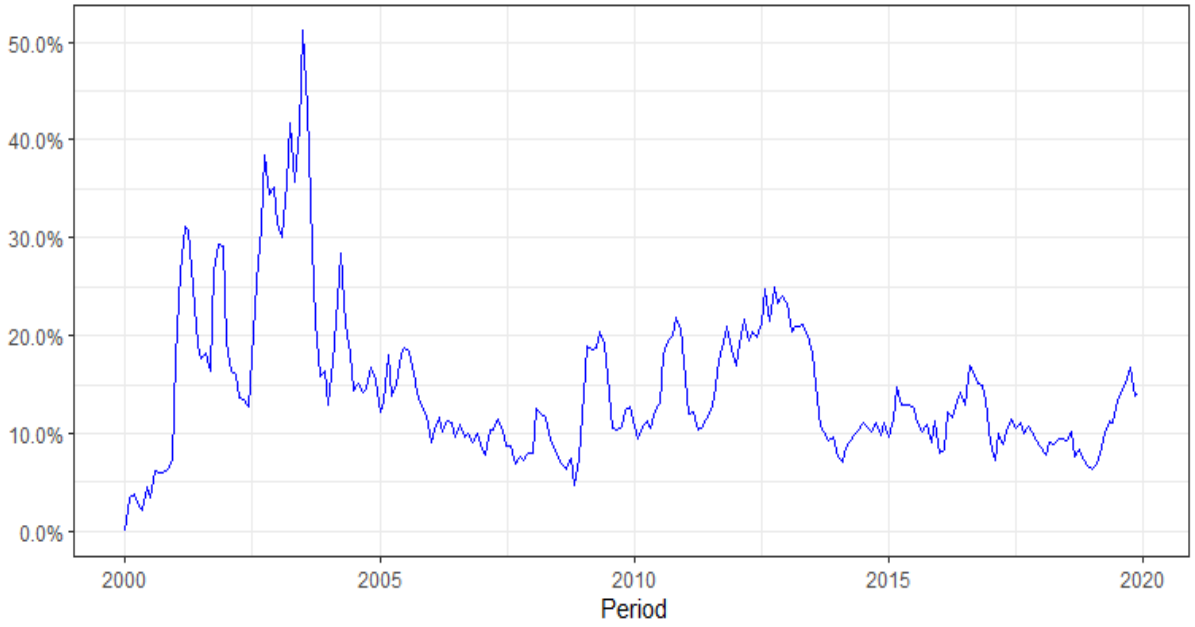


Figure 6: Fraction of Fully Prepaid Loans (annualized).

4.2. Macroeconomic data

Although the loan dataset contains a number of variables related to individual loans that can influence the timing of prepayments, it is also useful to include macroeconomic variables in prepayment models. These variables are time-varying and can have a great

Table 3: Largest mortgage sellers, 2000-2019 December.

Number of mortgages	
J.P. Morgan Chase Bank	51,360
Wells Fargo Bank	48,193
Bank of America	45,289
Citimortgage, Inc	17,646
Gmac Mortgage, Llc.	12,220
Flagstar Bank, Fsb.	10,033
Suntrust Mortgage Inc.	8,583
Quicken Loans Inc.	6,927
Amtrust Bank	5,266
First Tennessee Bank National Association	4,746

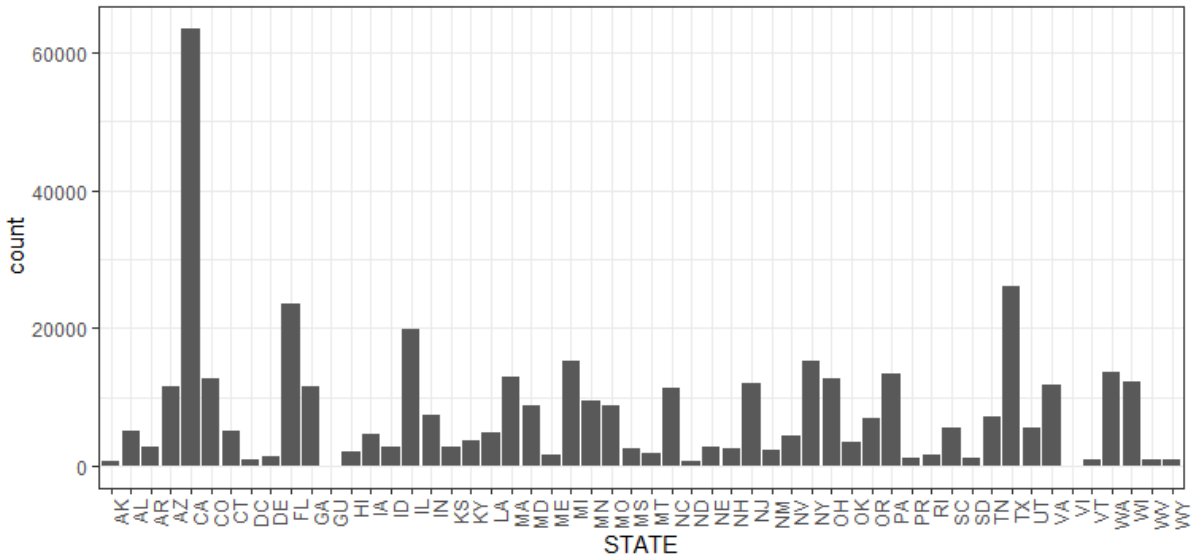


Figure 7: Number of mortgages by states and territories 2000-2019 December.

effect on exercising the prepayment option. From the option-theoretic literature one key variable is interest rate. However, it was chosen in this thesis to construct the interest rate incentive variable internally as described in the next section. The reason for this is that there is no publicly available historical interest rate information for Fannie Mae that also accounts for Loan to Value differences. Differentiating based on Loan to Value is important, since a new loan with higher LTV is expected to have a higher interest rate as well.

Apart from interest rates, other macroeconomic variables were also considered. These variables differ in terms of granularity and periodicity. The most granular level is MSA (Meteropolitan Statistical Area) with 406 categories, which is followed by State (50 states and Washington D.C. and 3 unincorporated territories) and Region (West, South, Midwest, Northeast) levels. The macroeconomic data was joined to the loan level data based on geography and date. For the yield curves the daily values were converted to monthly levels in such a way that values as of the 15^{th} of the months (or the closest available dates) were taken. Not all mortgages have collaterals in MSAs, in these cases, the state average

value is used for unemployment and HPI. There were several changes in metropolitan statistical area codes over time due to unifications, divisions, or renaming, especially in 2013. These changes were taken into account and the MSA-level macroeconomic data was adjusted accordingly (see details in Federal Housing Finance Agency (2021b)). The ratio of population and new residential properties is used as a proxy for housing turnover. The considered variables and their features are presented in Table 4.

After joining the macroeconomic data to the loan table and taking the average values over time, one can observe the macroeconomic trends in Figure 8. There are considerable geographical differences in the macroeconomic variables (except for yield curves). The regional heterogeneities are presented in Appendix B.

Table 4: Considered macroeconomic variables and their features.

Variable	**Granularity**	**Periodicity**	**Source**
Unemployment	MSA	Monthly	U.S. Bureau of Labor Statistics (2021b)
House Price Index	MSA	Monthly	Federal Housing Finance Agency (2021a)
Personal Income Per Capita & Population	State	Quarterly	Bureau of Economic Analysis (2021)
Consumer Price Index	Region	Monthly	U.S. Bureau of Labor Statistics (2021a)
Number of new residential sales	Region	Monthly	Census Bureau (2021)
Yield curve	Country	Daily	U.S. Department of Treasury (2021)

4.3. Variable transformations

As it can be seen in Table 14, the proportion of missing values is not very big. The Zero balance code variable is available only for the last months of the loans, while Loan Age, Remaining months, Adjusted Remaining months, Current interest rate and the Modification flag is not available for terminating loan, hence these values are replaced with 0. Insurance percentage is replaced with 0 as well when there are no available values. For the remaining variables with missing values first the last available value for a given loan is used as replacement and if it is not available, the portfolio level median at a given month is used as a replacement value.

Furthermore, due to borrower privacy considerations, the first few (most commonly 6) months of the unpaid principal amounts are set to 0 in the data (Fannie Mae, 2021a). For this reason, this value is replaced. First, the number of months with unknown principal months is derived. Then, P_0, the difference between the original principal balance and the first known principal balance is calculated. Next, the payable total amount amount is derived from the annuity formula where the number of periods is the number of unknown

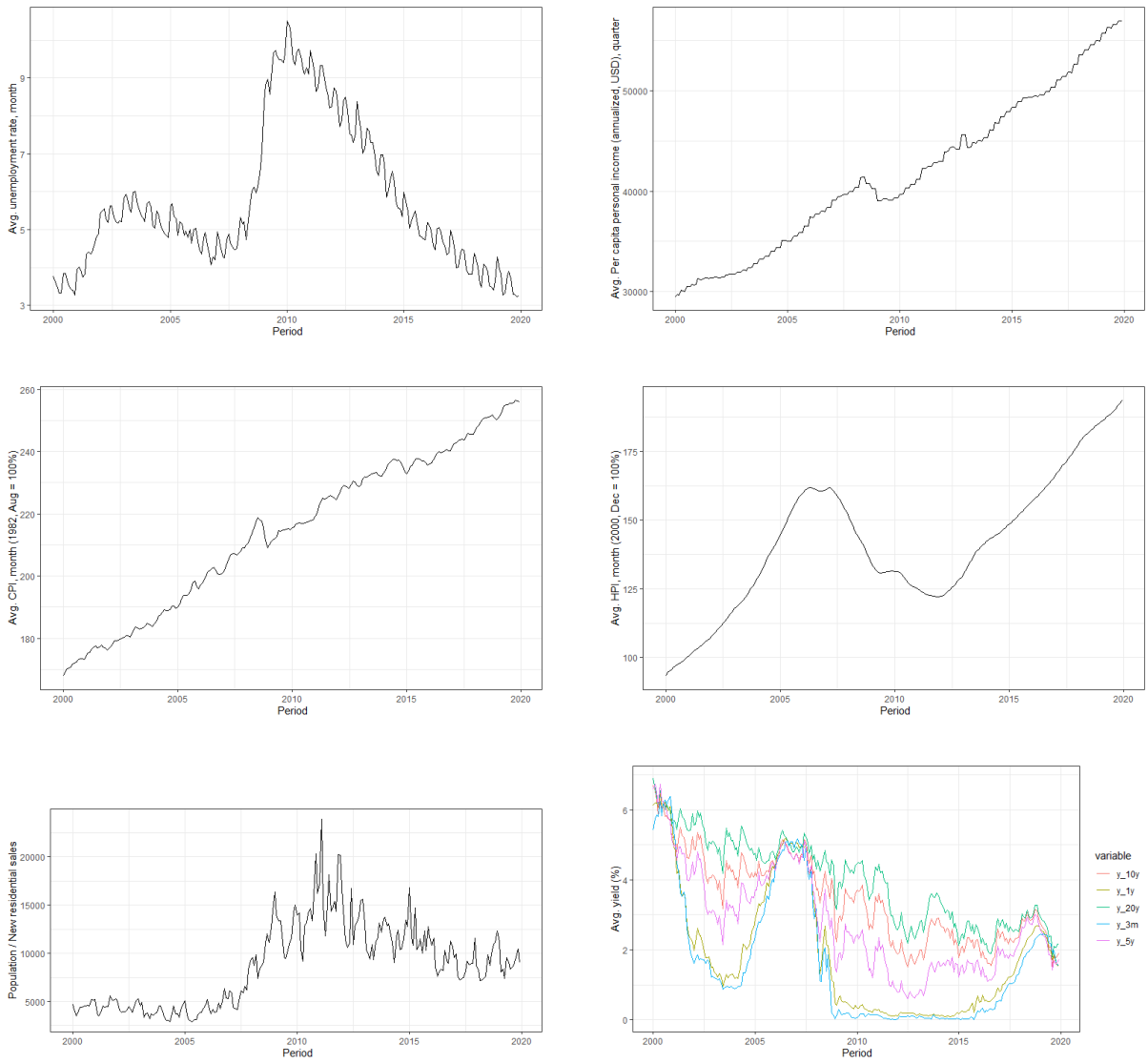


Figure 8: Average values of macroeconomic variables over time.

month and the principal amount is the total unknown principal amount:

$$Total\ Payment_i = P_0 \frac{r(1+r_i)^{ni}}{(1+r_i)^{ni}-1} \tag{39}$$

where r is the actual mortgage rate and n is the number of payment periods. The total redemption payment can be determined by subtracting the interest amount:

$$R_{it} = Total\ Payment_i - P_{t-1,i} \cdot r_i \tag{40}$$

Finally the cumulative sum of the redemption amounts is subtracted from the original amount to substitute the unknown principal balance values.

Based on the annuity formula it is also possible to investigate how often partial prepayments happen. However, sometimes uneven payments between months make this

calculation more difficult: it frequently happens that a client does not pay in one month but pays twice as much in the next next month. For this reason, in this thesis partial prepayment is defined as a paid amount in a month that is at least three times higher than the calculated redemption with the annuity formula. After creating this variable it is possible to compare the proportion of partial and full prepayments in terms of dollar amount. As it can be seen in Figure 9, the proportion of partial prepayments is relatively low, the average proportion is 2.8%. It is considered to be a small amount and given the relative difficulty of modelling partial prepayments with survival methods, this is out of scope of this thesis. However, when applying the model one should note this distinction.

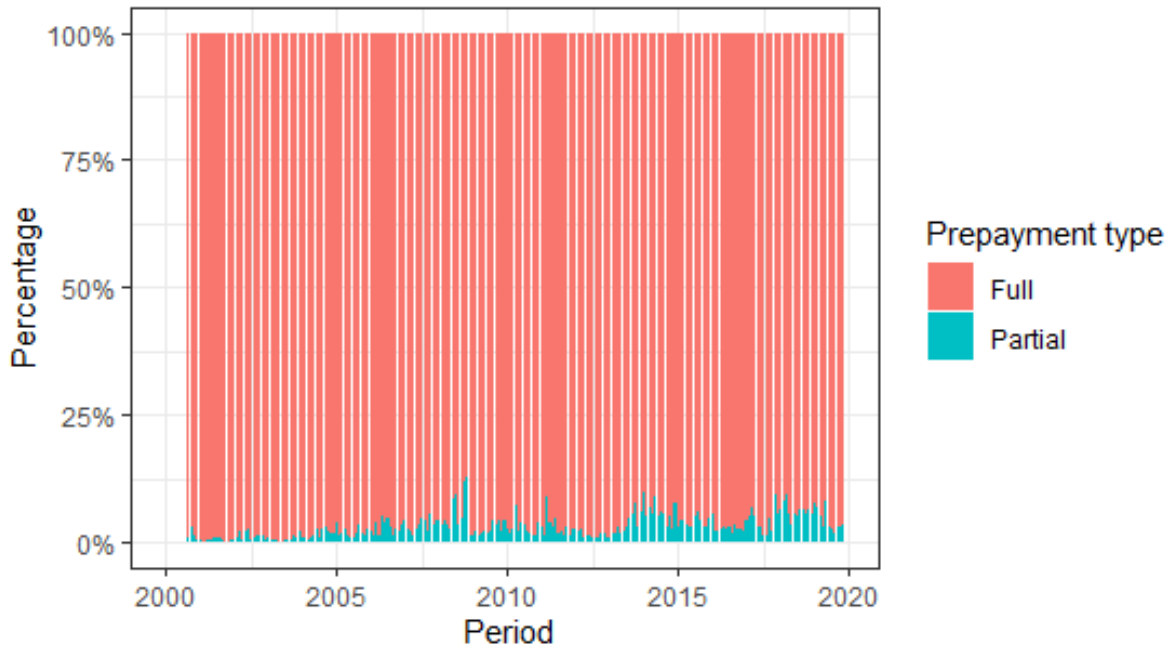


Figure 9: Proportion of full and partial prepayment amounts over time.

Another variable that needs to be transformed is the Loan to Value since it is only available at origination. From the first LTV and principal balance it is possible to derive the collateral value at origination. Next, a dynamic LTV can be obtained as:

$$Dynamic\,LTV_{ti} = Principal_{ti}/(Collateral\,Value_i^{t0} \cdot \frac{HPI^t{}_i}{HPI_i^{t0}}) \tag{41}$$

Next, the interest rate incentive variable is obtained internally. First, the reference rates or refinancing rates should be determined. This is determined based on the newly offered interest rates for loans with similar LTV and loan terms at each quarter. Hence, LTV and loan term categories are created. This is necessary since there may be no newly offered loans with exactly the same characteristics as for the loan of interest. Hence, the buckets should be sufficiently wide, yet not too wide to reach the goal of distinguishing interest rates based on loan terms and (dynamic) LTV. Quarters are used instead of months for the same reason. Based on these considerations, quintiles were used to categorize LTV and the loan term was determined based on 10, 15, 20, 25 years, hence altogether there are 5 categories here as well. After constructing the reference rates, the interest

rate incentive is defined as the percentage point difference between the actual mortgage rate and the reference rate. The average reference rates and refinancing incentives are displayed in Figure 10.

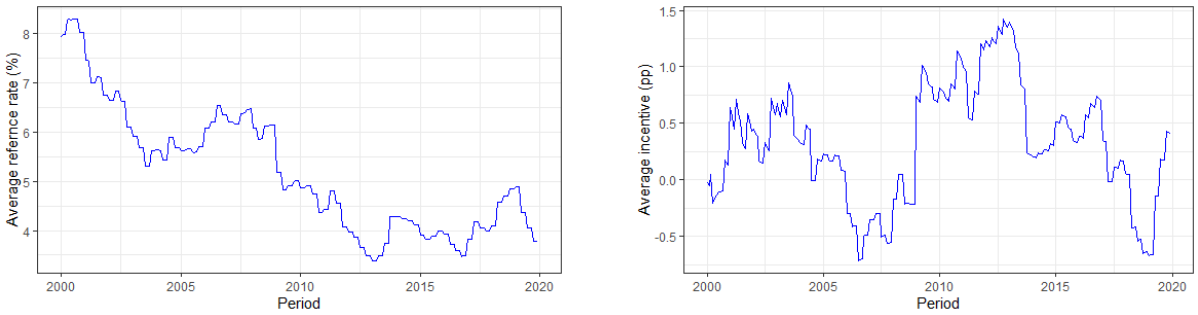


Figure 10: Average reference rates (left) and incentive (right) over time.

Finally, some other variables are also constructed. A binary delinquency indicator is created which takes a value of 1 if the client has arrears that are past due by more than 30 months. The yield curve steepness variable can be defined as the relationship between the long term and the short term yield. This is a simplification, since the yield curve is a broken curve based on different maturities, however for simplification, the is defined based on 2 points. For the short and long maturities, 1 and 20 years were chosen, so these points define the steepness in each month. Moreover, lagged macroeconomic variables were also created. By including a lagged incentive variable, it is possible to capture the burnout effect, namely that is is less likely that the client would opt for refinancing for similar incentive after some time has passed. Additionally, lagged unemployment variables were also constructed since unemployment tends to have a longer effect and unemployment status is positively correlated over time for an individual. The lagged variables were constructed for each quarter for incentive up to one year (since incentive is defined quarterly) to account for the burnout effect. For unemployment 1, 3, 6 and 12 months were included since 6 months is the maximum duration of unemployment insurance in most of the states and 1 year can capture long-term unemployment effects.

4.4. Driver analysis

In this section the association of certain variables with prepayment times are studied. As Hosmer and Lemeshow (1999) recommends, all survival analysis should start with bivariate analysis of potential drivers. Two techniques are used: the prepayment rates are plotted for certain values or buckets for the selected variables and then nonparametric survival curve estimation and testing is performed. Furthermore, in each survival plot the p-value of the log-rank tests are presented. If this value is smaller than 5%, the null hypothesis of identical survival curves for the certain categories or quantiles can be rejected.

First, the empirical survival and hazard functions are estimated for all the available observations without making distinctions based on variables. The estimated Kaplan-Meier and the Nelson-Aalen survival function are nearly identical as it can be seen in Figure 11. Furthermore, the hazard function has an increasing trend up to 100 months

after which the hazard rate is more stable. However, after 160 months, the patterns in the hazard rate is more volatile which can be explained by a smaller number of observations of older loans.

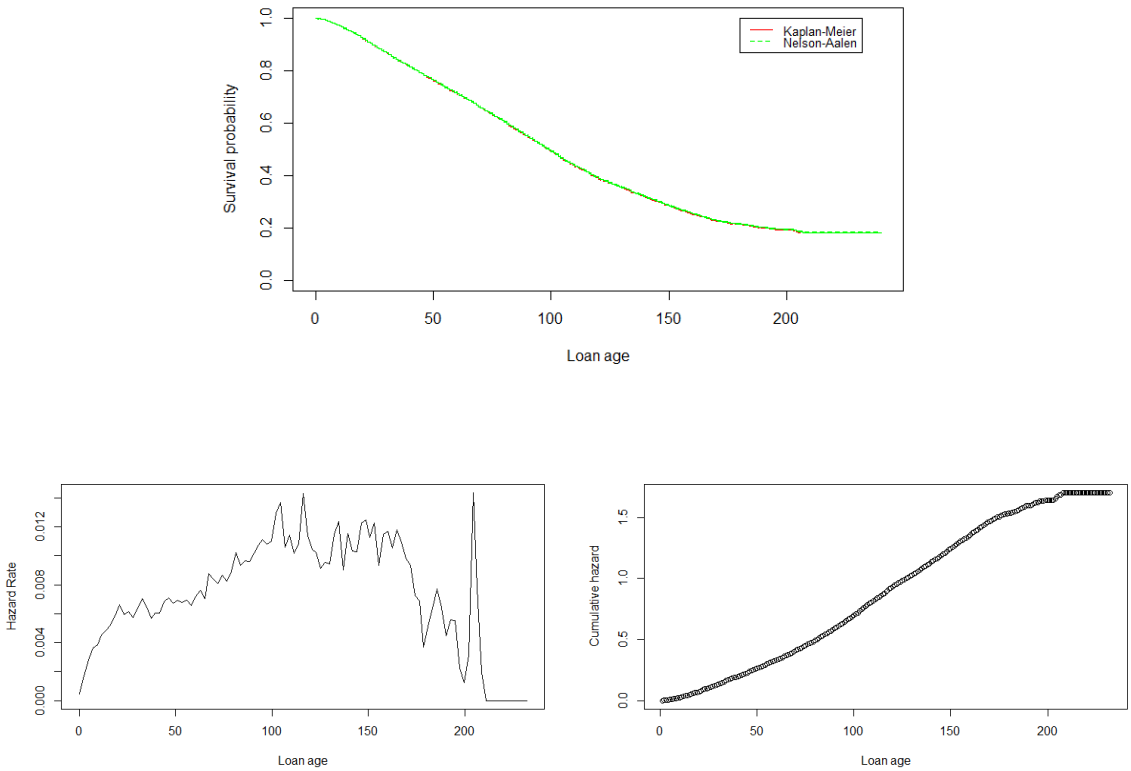


Figure 11: Survival function estimates (top), hazard (left bottom) and cumulative hazard functions (right bottom).

The relationship between interest rate incentive and prepayments is in line with the expectations: higher incentive buckets are associated with higher prepayments in Figure 12. However, one can also see that this relationship is not linear, there seem to be a lower and an upper asymptote. Hence, modelling this variable with a function that has similar properties can potentially increase model fit. Sigmoid functions such as the arc-tangent function appear to match these properties. This function is the inverse of the tangent function and has an inflection point at zero, with asymptotes $\frac{-\pi}{2}$ and $\frac{\pi}{2}$. The nonlinearity is also reflected in the estimated Kaplan-Meier survival curves based on the empirical quartiles: loans in the highest incentive quartile are the least likely to prepay, while the two middle quartiles are the most likely.

The relationship between the size of the loan and prepayment rates was clearly positive for values up to 200 thousand USD as it can be seen from Figure 13. Above this value, however, the relationship is less clear. From the survival plots it is clear that loans in the first quartile is significantly less likely to prepay, while loans in the second quartile also prepay more often than loans in the highest 2 quartile. From these relationships, it may be a viable approach to apply a logistic transformation to improve model fit.

In a similar manner, other variables were also examined. These additional plots are displayed in Appendix C. Loans with higher credit score appears to be less likely to prepay

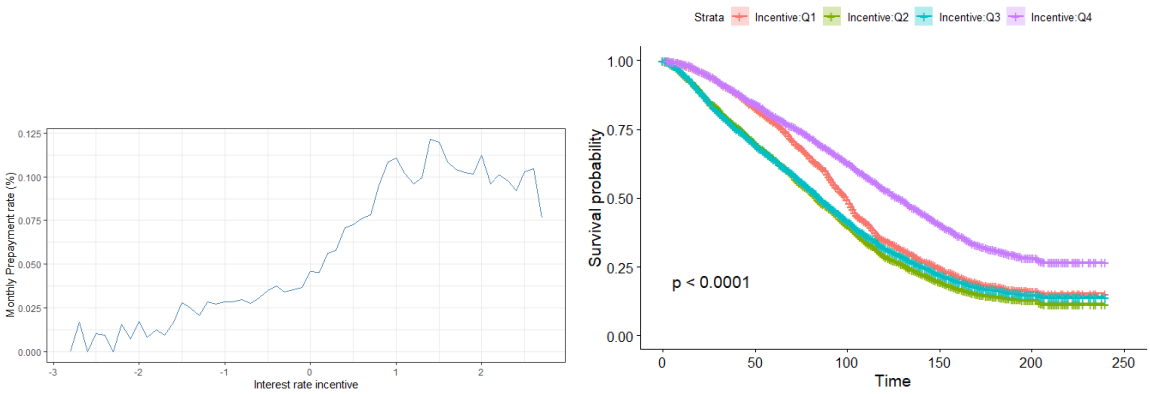


Figure 12: Relationship between prepayments and interest rate incentive.

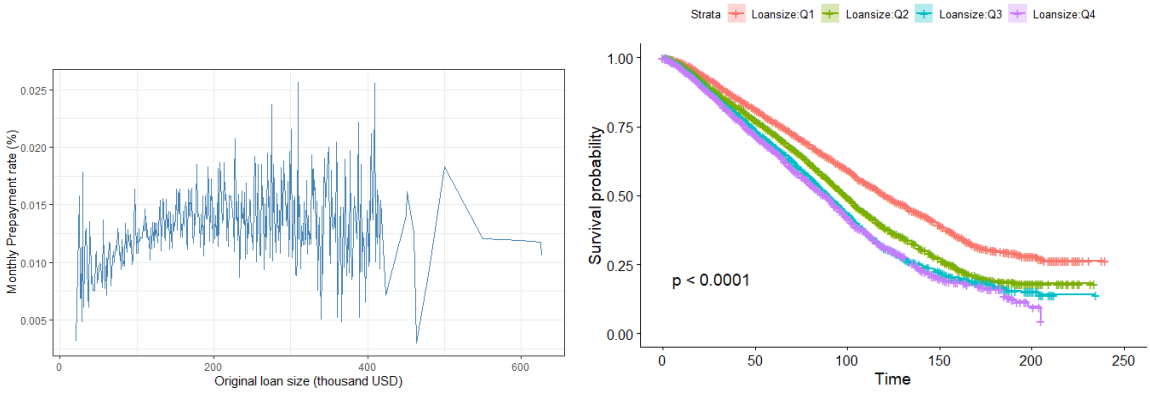


Figure 13: Relationship between prepayments and original loan amount.

(see Figure 30), while the relationship with dynamic Loan to Value is more complex (see Figure 31). There seem to be a positive association with prepayments for LTV values lower than 70%, however, for higher LTV values the relationship is somewhat reversed. Based on the survival curves, this reversal is not confirmed. This can be explained by the fact that the proportion of loans with LTV higher than 80% is only around 10%. The relationship between the original Debt to Income ratio does not seem to have an effect on prepayments while prepayment rates are higher in June and July and lower in December and January (see Figure 32).

Furthermore, prepayment rates seem to be higher in the Western region and for mortgages that were not taken by a first time home buyer as it can be seen in Figure 33. Prepayments also seem to grow with the number of borrowers in Figure 34, except when there are 4 borrowers. Moreover, as as Figure 35 displays, mortgages on homes that serve as primary residences have a higher propensity to prepay compared to investment-related homes. Secondary residences prepay more frequently that investment-related homes but less often than mortgages related to primary residences. Additionally, single-family homes and condominiums are also more likely to prepay than homes of manufactured or cooperative housing.

5. Model estimation

In this section, different prepayment models are estimated and the estimation processes are presented. First, a semiparametric extended Cox-model is discussed with time varying covariates, then a discrete time logistic model. This is followed by the discussion of two machine learning approaches: relative risk forest and conditional inference forest. Section 6 contains the more detailed evaluation of the models by different metrics.

5.1. Extended Cox model

5.1.1. Variable selection

Before estimating the final extended Cox model, variable selection has to be performed. This is because the inclusion of too many variables may result in an overly optimistic model performance that will not hold when the model is applied on other samples. In other words, the goal of variable selection is to avoid overfitting by penalizing the use of excessive number of variables.

There are different possibilities to perform variable selection. The method that is applied to obtain the final model is based on a stepwise selection procedure based on the Akaike Information Critierion (AIC), however the following alternatives were also considered:

- Best subset selection
- Selection based on likelihood ratio test
- Stepwise selection based on p-values
- K-fold cross-validation
- LASSO
- Bayesian Information Criterion (BIC)

Best subset selection would mean that different models are estimated for all of the combinations of the predictors and then the single best model is selected based on a metric (e.g. AIC or cross validated concordance index). However, this would require to fit all possible models which would mean at least 2^{20} models which is more than a million models. Since computation for one model is about 5 minutes, the computation time would be about 10 years. For this reason, this method is not considered to be feasible.

For comparing nested models it would be possible to use the likelihood ratio test for comparison:

$$LR = 2(L(\hat{\beta}_{full}) - L(\hat{\beta}_{reduced})) \tag{42}$$

This statistics can be compared to the χ^2 distribution with degrees of freedom corresponding to the difference in degrees of freedom between the comparable models. However, when models are not nested, this statistics cannot be used and for model selection alternative approaches are necessary.

Another alternative is to perform stepwise selection using p-values. The selection could start with the null model (forward stepwise selection) or with the full model (backward stepwise selection). With the forward selection method, in each step a model is fitted with one additional covariate and the variable that exhibits the smallest p-value is included. With backward stepwise seleciton, a model is fitted for all the reduced models where one variable is removed and the variable with the largest p-value is included in the next step. The process stops when all the p-values are smaller than some prespecified level. Although these selection procedures may seem logical and there were used in the past frequently, it turns out that the test performed at each step is dependent on the tests in the previous steps, so the uncertainty around p-values increases with each step. For this reason, approaches based on p-values are not recommended and hence they will not be used in this thesis either.

A commonly used alternative is called the K-fold cross-validation. This approach compares out-of sample errors by holding out a certain proportion of the sample multiple (K) times and then estimating the models based on the training samples on the test samples. Then, the model that corresponds to the lowest cross-validated error can be selected. However, this method would require to fit multiple models and calculate the cross-validated errors for each of them which would be very time-consuming, so this method is not used for model selection.

By using a regularized Cox method as explained in Section 3.2.3, one could combine estimation and feature selection in one step by modifying the objective function with a LASSO term. In the resulting model, the coefficients of the covariates that are not selected are shrunken toward zero. In may be attractive to follow this approach, however, it turns out to be infeasible with time-varying covariates due to memory allocation problems and computational time. In practice is only feasible to compute the LASSO model with a small subset of the loans with time-varying covariates. Furthermore, based on previous literature, it is not expected that too many coefficients would be shrunken toward zero, hence for these reasons, an alternative approach is preferred.

After reviewing these possibilities, it can be concluded that the use of a measure based on information criteria is preferred. Two possible such measures are considered: the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (P). Both of these methods include the values of the partial log likelihood corresponding to a given model and a penalty term. The difference between the two methods are in the penalty terms:

$$AIC(k) = -2(log(L(\hat{\beta})) + 2k \tag{43}$$

$$BIC(k) = -2(log(L(\hat{\beta})) + 2k \cdot (log(n)) \tag{44}$$

where k is the number of parameters and n is the number of observations. It can be seen that the penalty term for BIC is larger (if $n >= 3$). For this reason, BIC tends to select a more parsimonious model and it is commonly used for inference due to its asymptotic property that it chooses the correct model as $n \to \infty$. However, in a finite sample environment when the goal is to assess out-of sample performance, AIC is preferred, since it allows more flexibility by choosing a larger model (Hastie et al., 2009). For these reasons, the AIC method is chosen for variable selection.

The following algorithm is applied for variable selection:

1. Start with the full model with all the possibly relevant explanatory variables and obtain the AIC.
2. Estimate all models with removing or adding one variable and obtain the AIC for these models. (For the full model it is not possible to add variables.)
3. Compare the AICs obtained in step 2) to the AIC in the previous step and select the model with the smallest AIC.
4. Stop the process if the AIC cannot be further reduced by adding or removing a variable.

The steps of the variable selection procedures are included in Appendix D.

5.1.2. Estimated model after variable selection

After performing the variable selection, the following variables were excluded from estimation: Debt to Income, Mortgage Insurance %, Original Loan to Value, Unemployment rate, 6 month lagged unemployment rate and 6 and 9 months lagged interest rate incentive. The exclusion of Debt to Income and mortgage insurance is understandable based on Figure 32 and Figure 36, while original loan to value may not play a big role in prepayments since changes in loan to value due to house price changes were taken into account during the construction of the interest rate incentive variable. It appears that unemployment has a 1 month delayed effect on prepayments which is also understandable. The exclusion of some lagged variables from the final model are less intuitive; it seems that the inclusion of 3 months and 1 year lags can increase the model fit.

The selected model with the estimated coefficients and robust clustered standard errors are presented in Table 5. The relevant figures for month and state dummies are presented in Appendix E. Standard errors are clustered for each loan since one cannot assume independence between observations corresponding to the same loans over different periods.

Table 5: Estimated Extended Cox Model.

Variable	**Coef.**	**Hazard rate** exp(Coef.)	**Robust S.E.**	**P-value**

ATAN (Interest rate incentive)	1.3113441 ***	3.7111585	0.0423082	<0.000000000
LOG(Initial loan amount)	0.4717735 ***	1.6028343	0.0186121	<0.000000000
Number of borrowers	0.0262958	1.0266446	0.0177876	0.139322
Personal income/capita	0.0000129**	1.0000128	0.0000050	0.010615
CPI	- 0.0176791 ***	0.9824763	0.0016650	<0.000000000
Population/New residential sales	- 0.0000065 ***	0.9999934	0.0000022	0.003484
Yield curve steepness	0.0049172	1.0049293	0.0102192	0.630392
Delinquency indicator	- 0.6771383 ***	0.5080688	0.0597384	<0.000000000
Credit rating score	0.0009879 ***	1.0009884	0.0001781	0.0000000292
First homebuyer	- 0.1036568 ***	0.9015346	0.0351369	0.003177
Number of housing units	- 0.1429694 ***	0.8667806	0.0368609	0.000105
Loan modification	- 0.6357151 ***	0.5295567	0.1101792	0.0000000079
Occupancy-primary	0.2882692 ***	1.3341164	0.0363704	<0.000000000
Occupancy-secondary	0.1564257 ***	1.1693239	0.0563120	0.005472
Property - Central Planning	0.09655	1.101375	0.12026	0.422
Property - Manufactured housing	- 0.72599 ***	0.48384	0.11941	0.00000
Property - Single Family	-0.07285 *	0.92973	0.03326	0.0284
Purpose - Purchase	0.1082636 ***	1.1143414	0.0249983	0.000014854
Purpose - Non-Cash-out refinance	0.0257385	1.0260726	0.0217752	0.237200
Lag 1m Unemployment rate	- 0.0675696 ***	0.9346626	0.0154134	0.0000116612
Lag 3m Unemployment rate	- 0.0320347 *	0.9684729	0.0164779	0.051884
Lag 1y Unemployment rate	0.0451782 ***	1.0462143	0.0079567	0.0000000136
Lag 3m ATAN Incentive	- 0.2234936 ***	0.7997200	0.0432016	0.0000002300
Lag 1y ATAN Incentive	- 0.0958371 ***	0.9086121	0.0285277	0.000781
Months included	YES			
States Included	YES			
Time interactions	NO			
Concordance index: 0.694				
Wald statistic = 4066 (90DF), p=<0.00000000				
Robust score test statistic = 4163 (90DF), p=<0.00000000				

The results are mostly in line with the expectations based on the bivariate analysis. The hazard rates can be interpreted as multiplicative scaling parameters of the baseline hazard function. When the hazard rate is greater than 1, the baseline hazard increases proportionally by this factor and it decreases proportionally when it is smaller than one. The estimated hazard rates of the transformed coefficients can be interpreted similarly as for the untransformed ones, however, the hazard rate corresponds to one transformed unit of the variable. The transformed interest rate incentive and loan size coefficients are positive and the corresponding hazard rates are large and the p-values are small. All the

macroeconomic variables have the expected signs, however the coefficient of yield curve steepness is not significant. This can be possibly explained by assuming that most clients do not refinance their long-term mortgages frequently with shorter term loans as they prefer some protection against increasing interest rate movement. The coefficient of the 3-month lagged value of unemployment is less than half of the 1-month value and the 1 year lagged unemployment value is even positive. This may indicate that unemployment has a negative effect only on the short term on prepayments. Based on the negative lagged incentive values we can observe some burnout effects, it appears that most clients react quickly when their incentive increases suddenly.

The signs of the other variables are also in-line with the expectations: the number of borrowers seem to increase prepayments, however the standard error for this coefficient is high. Delinquent clients, loans with modifications in the terms, first time home owners and homes with more housing units are less likely to prepay. Loans with investment purpose (compared to loans related to first and secondary homes) are also less likely to prepay. Compared to loans with a purpose of refinancing, mortgages related to home purchase are 11% more likely to prepay compared to cash-out refinance loans.

Since the scale of the coefficients are different, visually it can be more useful to look at the standardized coefficients. These coefficients with the confidence intervals are displayed in Figure 14.

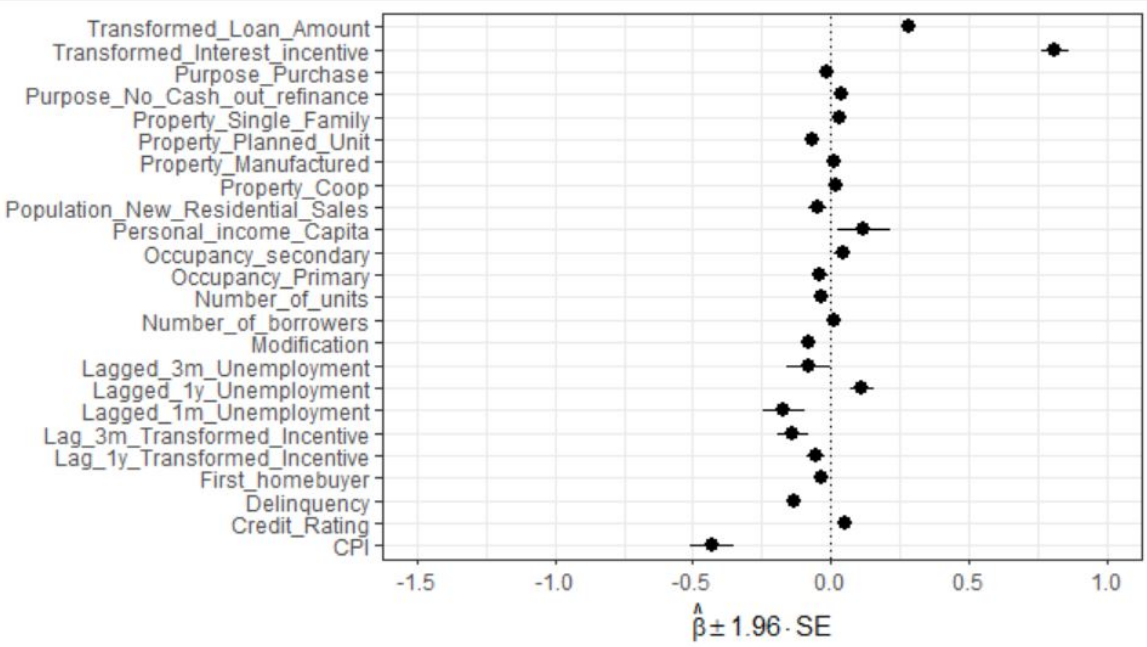


Figure 14: Standardized coefficients of the extended Cox model estimation.

The Extended Cox model includes time-varying and time-constant covariates. Although for time-varying covariates the proportional hazards assumption does not hold by construction (see (32)), for time-invariant covariates proportionality is still assumed. For this reason, the proportional hazard assumptions have to be tested for the time-constant covariates with Schoenfeld residuals (see (29)). Grambsch and Therneau (1994) developed a formal test which is based on these residuals. This is also called as the *zph* test. The null hypothesis of the test is that the covariate of interest meets the proportional

hazards assumption.

Table 6: Proportional hazards (zph) tests for time-constant covariates .

Variable	**Chi-squared value**	**P-value**
Number of borrowers	1.403	0.217
Credit rating	79.127	< 0.000
First homebuyer indicator	5.615	0.018
Number of units	5.298	0.0214
Property type	5.323	0.215
Loan purpose	4.658	0.199
Occupancy status	23.2	$<$0.000
Log Original Loan amount	0.438	0.508
States[5]	113	< 0.000

As one can see, credit score and occupancy status variables are the ones that exhibit the biggest non-proportionality. Furthermore, there are non-proportional effects in certain states, the number of housing units and the first home indicator. To account for this, these variables were interacted with grouped survival time variables. The residuals and the fit over time are plotted in Figure 15 and manual intervals were selected based on the shape of the observed fit over time. Intervals were created with a goal to contain relatively flat lines within them. For example, for credit scores, the time intervals are created based on the following times: 10, 15, 25 32, 45. In this way, time-dependent coefficients are also introduced in the model.

5.1.3. Final model after accounting for non-proportionality

The model is re-estimated after adding interaction terms to the time constant variables that exhibited non-proportional hazards. The final results of the Cox model is presented in Table 7.

Table 7: Estimated Extended Cox Model with some time-dependent coefficients.

Variable	**Coef.**	**Hazard rate** *exp*(Coef.)	**Robust S.E.**	**P-value**
ATAN (Interest rate incentive)	1.43804 ***	4.2124	0.0495292	$<$0.000000000
LOG(Initial loan amount)	0.4828 ***	1.6206	0.0223774	$<$0.000000000
Number of borrowers	-0.000996	0.9995	0.0213663	0.9628226
Personal income/capita	0.0000161**	1.000016	0.0000062	0.0089
CPI	- 0.0185791 ***	0.98160	0.0020236	$<$0.000000000
Population/New residential sales	- 0.0000063 **	0.9999937	0.0000028	0.022849
Yield curve steepness	0.00305	1.0030	0.0115983	0.7922421
Delinquency indicator	- 0.6160995 ***	0.540004	0.0702998	$<$0.000000000

[5]Non-proportional effects are present for Alabama, California, Delaware, Idaho, Illinois, Massachusetts, Michigan, Montana, Nevada, New York

Credit rating score	0.0019170 ***	1.0019	0.0004814	0.00000
First homebuyer	- 0.1409254 ***	0.9997	0.0398362	0.0004037
Number of housing units	- 0.1822161 ***	0.8334	0.0492226	0.000214
Loan modification	- 1.0891925 ***	0.3364	0.3117311	0.0004758
Occupancy-primary	0.4353036 ***	1.5454	0.0545738	<0.000000000
Occupancy-secondary	0.2596718 ***	1.2965	0.0820445	0.0015508
Property - Central Planning	0.107151	1.1131	0.151056	0.4781
Property - Manufactured housing	- 0.57685 ***	0.56166	0.13806	0.00003
Property - Single Family	-0.04388	0.957063	0.040036	0.273
Purpose - Purchase	0.1143808 ***	1.1211	0.0290393	0.000082
Purpose - Non-Cash-out refinance	0.0098677	1.0099	0.0264868	0.7094831
Lag 1m Unemployment rate	- 0.0644598 ***	0.9375	0.0180337	0.00035
Lag 3m Unemployment rate	- 0.0474818 **	0.9536	0.0191312	0.01306
Lag 1y Unemployment rate	0.0457419 ***	1.0468	0.0097028	0.0000024
Lag 3m ATAN Incentive	- 0.2129928 ***	0.80816	0.0514488	0.0000347
Lag 1y ATAN Incentive	- 0.1521739 ***	0.8588	0.0333040	0.000005
Credit score*time [10:15]	- 0.0023968***	0.9976	0.0007322	0.0010618
Credit score*time (15:20]	- 0.0018417 ***	0.9981	0.0006579	0.0051232
Credit score*time (20:25]	- 0.0022451***	0.997	0.0006613	0.0006866
Credit score*time (25:32]	- 0.0018137 ***	0.998	0.0006559	0.0056868
Credit score*time (32:45]	- 0.0008578 ***	0.9991	0.0006223	0.1680687
Credit score*time (45:60]	- 0.0018137 ***	0.998	0.0006559	0.0056868
Primary occupancy *time (40:85]	0.1813755 **	1.1988	0.0983361	0.0651183
Secondary occupancy *time (40:85]	0.1697449	1.185	0.1525731	0.2659026
First homebuyer *time (1:30]	- 0.249690 **	0.779	0.083536	0.0014
Number of units *time (85:240]	0.23834 ***	1.2691	0.0767070	.00097
Months included	YES			
States Included	YES			
Concordance index: 0.703				
Wald statistic = 3984 (114DF), p=<0.00000000				
Robust score test statistic = 3497 (114DF), p=<0.00000000				

The results are similar to those in Table 5, however interest rate incentive and CPI has an even bigger impact on prepayments. Furthermore, the time interactions address non-proportionality sufficiently that were observed in Figure 15. Early time interactions with credit scores have a bigger negative impact, while the peak in for occupancy type between 40 and 85 months is also reflected in the signs of the interaction terms. After

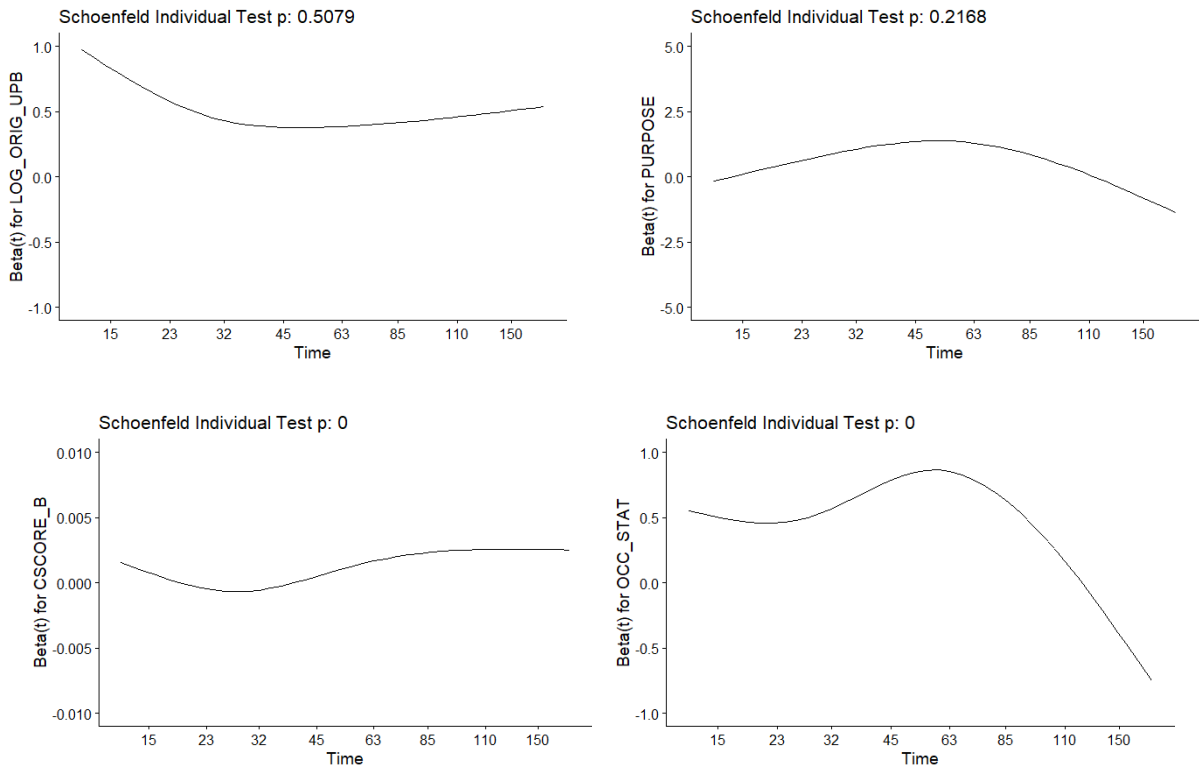


Figure 15: Plotted estimated coefficients over survival times for certain time-constant variables (loan amount, loan purpose, credit score, occupancy status from left to right).

adding time-varying covariates, the Grambsch and Therneau (1994) test does not indicate non-proportional effects (see Table 8).

Table 8: Proportional hazards (zph) tests for time-constant covariates after adding interaction terms.

Variable	**Chi-squared value**	**P-value**
Number of borrowers	1.01	0.31403
Credit rating	0.05142	0.82062
First homebuyer indicator	2.93861	0.0864
Number of units	1.18602	0.27613
Property type	6.31346	0.177
Loan purpose	0.74697	0.862
Occupancy status	2.73888	0.254
Log Original Loan amount	1.587	0.2077
States	64.50	0.166

5.2. Discrete time logistic model

In this section the discrete time logistic model is estimated as defined in (22). Discrete time survival analysis have potential advantages over continuous time models (such as the semiparametric Cox model) when data is intrinsically discrete, i.e. when data can be measured in specific cycles. One of these advantages is that discrete models can relax the proportional hazards assumption and can handle tied events without problems. Furthermore, it is computationally more feasible to include unobserved heterogeneity (frailty) compared to Cox models since it does not require multidimensional integrals (Hess & Persson, 2012). However, computation time with frailty can still be considerable for discrete models.

The estimated model (without unobserved heterogeneity) is presented in Table 9.

Table 9: Estimated Discrete Logistic Model.

Variable	**Coef.**	**Hazard rate** *exp*(Coef.)	**Robust S.E.**	**P-value**
Loan age	-0.0026998 ***	0.9973039	0.0003191	0.0000000
LOG(Initial loan amount)	0.4752798 ***	1.6084642	0.0182184	0.0000000
Occupancy-primary	0.292233 ***	1.3394160	0.038339	0.0000000
Occupancy-secondary	0.1612273 **	1.1749521	0.057158	0.0047917
Number of borrowers	0.0269008	1.0272659	0.0176548	0.1275815
Personal income per capita	0.0000122 **	1.0000122	0.0000050	0.0155775
CPI	-0.0175725 ***	0.9825810	0.0016402	0.0000000
Population/New residential sales	-0.0000059 ***	0.9999941	0.0000021	0.0059362
Yield steepness	-0.0006470	0.9993532	0.0105965	0.9513158
First homebuyer	-0.1045008 ***	0.9007740	0.0347175	0.0026122
Number of housing units	-0.1434754 ***	0.8663421	0.0399408	0.0003279
Delinquency	-0.6786460 ***	0.5073034	0.0531915	0.0000000
Modification	-0.6528274 ***	0.5205718	0.1117970	0.0000000
ATAN Incentive	1.3129348 ***	3.7170666	0.0422771	0.0000000
Property-central planning	0.0925827	1.0970038	0.1333720	0.4875765
Property-manufactured	-0.7377736 ***	0.4781774	0.1157838	0.0000000
Property-Single Family	-0.0744585 **	0.9282460	0.0333866	0.0257346
Purpose - Purchase	0.1062054 ***	1.1120503	0.0245223	0.0000148
Purpose - Non-Cash-out refinance	0.0270145	1.0273827	0.0217147	0.2134759
Credit score	0.0010208 ***	1.0010213	0.0001743	0.0000000
Lag 1m Unemployment rate	-0.0673180 ***	0.9348979	0.0146429	0.0000043
Lag 3m Unemployment rate	-0.0305010 **	0.9699594	0.0154621	0.0485377
Lag 1y Unemployment rate	0.0465516 ***	1.0476521	0.0079343	0.0000000
Lag 3m ATAN Incentive	-0.2239996 ***	0.7993154	0.0444546	0.0000005
Lag 1y ATAN Incentive	-0.0821808 ***	0.9211054	0.0289857	0.0045794

Months included	YES
States Included	YES
Time interactions	NO

The results are very much in line with those estimated with the Extended Cox model (without time interactions). All the signs are the same, except for yield curve steepness, but the coefficient is estimated with low precision in this case as well. One difference between the Cox model and the discrete logistic model is that in case of the latter loan age is one of the explanatory variables.

5.3. Relative risk forest

5.3.1. Hyperparameter tuning

The first machine learning approach that is examined is the relative risk forest method. The relative risk forest includes a number of optional parameters.The relevant hyperparameters for the relative risk forest are the following: number of trees, the number of candidate covariates randomly selected at each node (*mtry*) and the minimum number of subjects in a terminal node (*nodesize*). The final model is selected after hyperparameter-tuning i.e. selecting these parameters optimally. The final model is selected based on the combination of hyperparameters that results in the smallest out-of-bag error rate or equivalently to the highest concordance index since error rate is defined as $1 - Concordance\ index$ [6]. Unlike in the case of parametric or semiparametric methods, the interpretation of the outcome is less intuitive since it is an ensemble forest. However, the contribution of covariates can still be assessed by means of variable importance or minimal depth of a predictor.

The bigger number of trees used usually leads to a lower prediction error. However, the additional benefit of increasing the number of trees is not linear. Since more trees can increase computational power significantly, the OOB error rates are plotted as a function of the number of trees in Figure 16. Based on this plot, it appears that adding more than 200 trees does not reduce the OOB error significantly. For this reason, 200 trees are selected for the final model.

Next, the other two relevant tuning parameters (*mtry* and *nodesize*) are selected together. A two-step grid search algorithm is conducted in which first a larger number of grid points are selected with bigger steps between the grid points and in the second step the grid points are reduced and all integer values between the reduced range are examined. After this, the optimal hyperparameters for *mtry* and *nodesize* turned out to be 24 and 1, respectively with an error rate of 13.98%. The heatmap in Figure 17 shows the corresponding out-of-bag error rates combinations of the two parameters.

[6] Out-of-bag error refers the the average prediction error based on all trees where an OOB error of a given tree is the prediction error of that tree when fitted on the left-out sample.

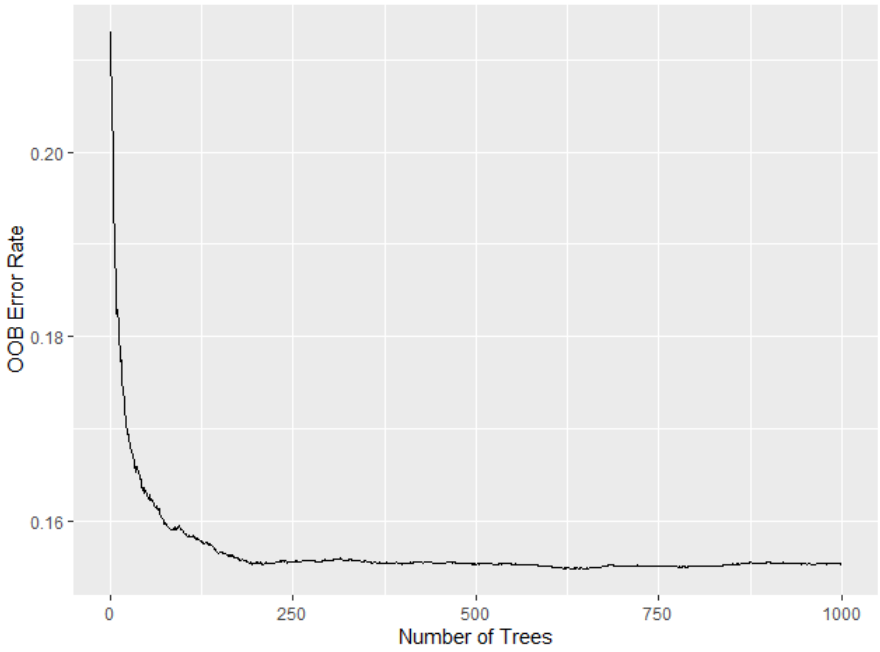


Figure 16: Out-of-bag error rate (1-concordance index) as a function of number of trees for relative risk forest.

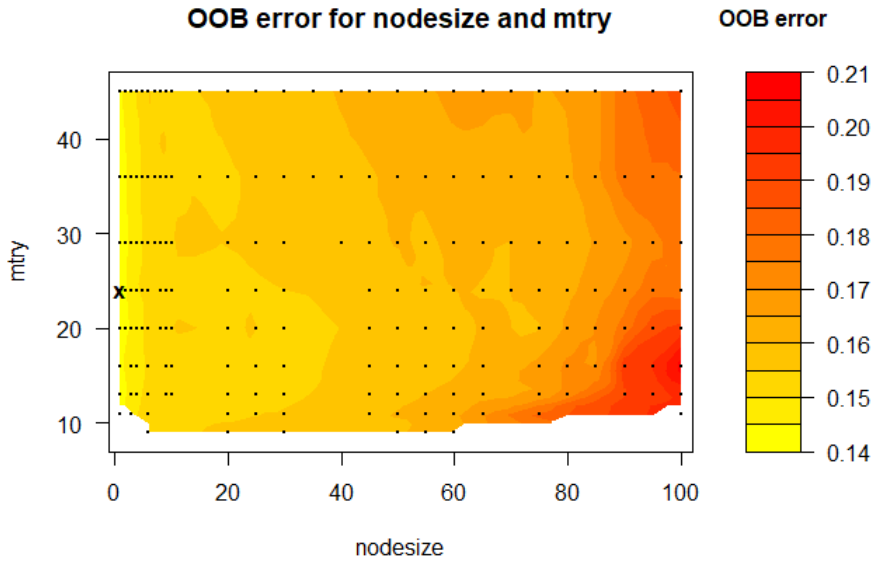


Figure 17: Out-of-bag errors heatmap for relative risk forest as a function of *mtry* and *nodesize* parameters.

5.3.2. Variable selection and interactions

Since it is not possible to perform variable selection based on an Information Criterion in this case, an alternative measure of variable selection is necessary. Two possible ways of assessing the importance of variables include the variable importance measure (VIMP) and the minimal depth measure. The former is defined as the difference between the out-of-sample error rate when the variable in question is randomly permuted and the observed out-of-sample error rate. If this difference is large, the variable is important

whereas if it is negative, it indicates that there is no predictive power in the variable.

On the other hand, the minimal depth measure relates to the tree construction splitting process. It is calculated by averaging over the depths of the trees (from the root node denoted by 0, where the first splitting happens based on the given variable. A smaller value indicates that more early splitting happens based on that variable so a lower value indicates that the variable is more important (Ehrlinger, 2016).

It is possible to visualize variable importance and minimal depth in one graph by ordering the covariates based on these two measures. In Figure 18 it appears that the ranking of most of the variables are mostly in line based on the two measures. While CPI ranks first for minimal depth, delinquency indicator is the most important variable based on VIMP. Some variables (e.g. occupancy status and certain states) are not marked as important by VIMP since they have a negative value. However, the conclusion regarding the most important variables are mostly in line: CPI, Delinquency status, Loan amount, Interest rate incentive and Personal Income per Capita appears to be the most important variables. Variables with both negative VIMP values and high minimal depth ranking are removed from the final model and the relative risk forest is recalibrated. Since all of the covariates that have negative VIMP values also have a high minimal depth, but not the other way around, this means in practice that only variables with negative VIMP values were removed from the final model.

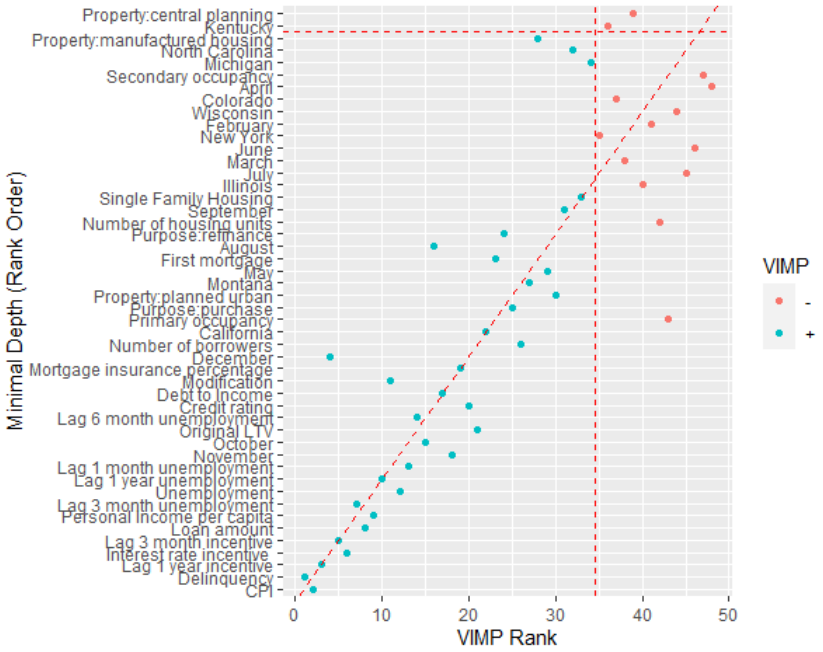


Figure 18: Variable importance based on VIMP and minimal depth measures for relative risk forest.

Variable interactions can be examined by the minimal depth of a variable with respect to a maximal subtree of another variable (Ehrlinger, 2016). The concept was formalized by Ishwaran et al. (2007) and he defined maximal subtree as the largest subtree that uses a certain variable for splitting. Based on this metric, interactions can be studied. In Figure 19 the normalized minimal depth of the maximal subtree of interest rate incentive and loan amounts are plotted. A larger value indicates a stronger interaction with the variable. One can see that both interest rate incentive and loan age are most strongly

interacted with CPI, credit score, LTV, November and lagged values of incentive while interactions with other variables are somewhat less common.

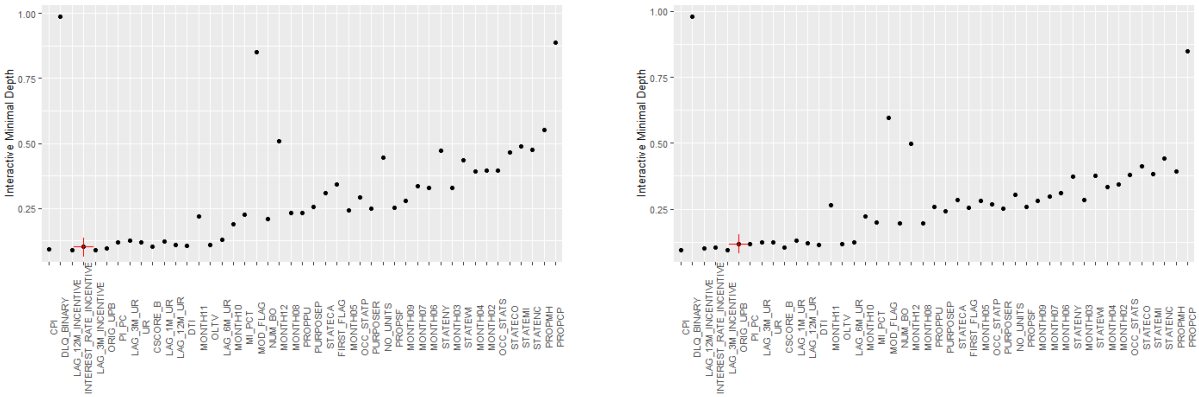


Figure 19: Interactions of interest rate incentive and loan amount with other variables.

5.4. Conditional inference forest

The conditional inference survival forest separates covariate selection and finding the best splitting rule, unlike the relative risk forest which does it in one step. As Strobl et al. (2007) pointed out, the relative risk forest algorithm is biased towards continuous variables, since the number of potential splits is higher compared to discrete variables and by random chance they splitting by a continuous variable can increase node purity compared to categorical variables. For this reason, hypotheses tests are performed for variable selection and in the second step the optimal splitting rule is determined. The variable selection where the desired α value is prespecified; 95% is considered as a typical value and this also used in the thesis. Hence, in the case of conditional inference forest, only the number of trees and the number of potential variables (*mtry*) are the parameters that have to be tuned.

Regarding the number of trees, a similar conclusion is reached as in the case of relative risk forests: there is no significant improvement after adding more than 200 trees to the forest. The *mtry* value turned out to be 22 in this case as it can be seen in Figure 20.

The variable importance is evaluated similarly as in case of the relative risk forest. In Figure 21 the variable importance metric (VIMP) of the two methods are compared. As one can see, in the conditional risk forest, the three most important variable are the delinquency indicator (whether the client has arrears), CPI, 1 year lagged interest rate incentive. Furthermore, it appears that the delinquency indicator is relatively more important for the conditional inference forest than for the relative risk forest. The importance ranking of the other variables are mostly similar, however, there are differences, for example the state California is more important for the relative risk forest while the number of borrowers is more important for the conditional inference forest.

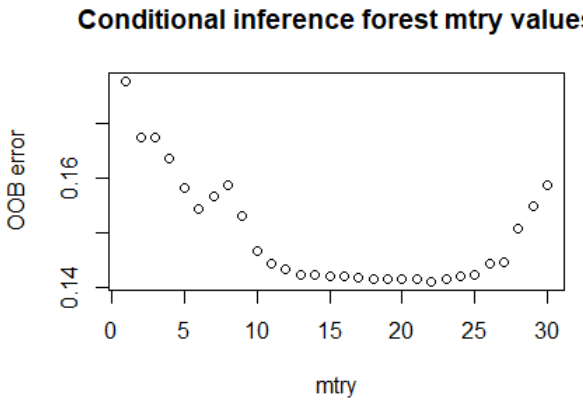


Figure 20: Tuning of the number of candidate variables at each split for the conditional inference forest.

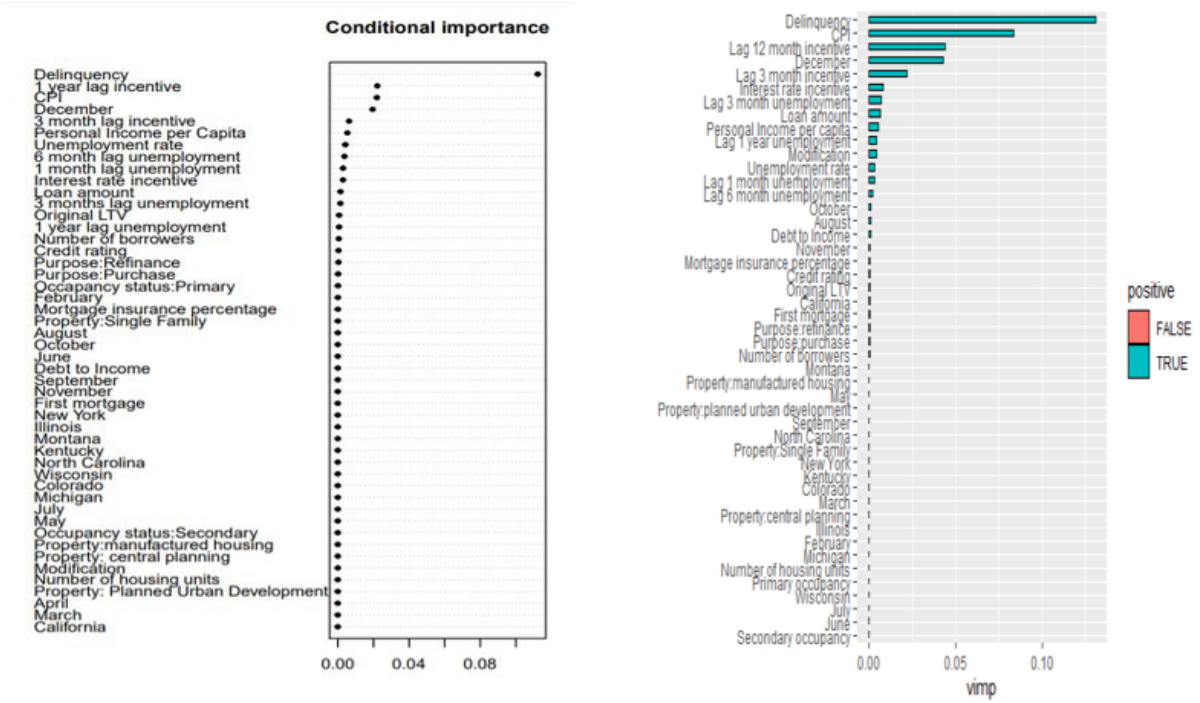


Figure 21: Variable importance plots for conditional inference forest and relative risk forest.

6. Model evaluation

The goal of this chapter is to evaluate the 4 models presented in Section 5 in different ways. In Section 6.1 the models are evaluated by means of metrics presented in Section 3.4. In Section 6.2 the fit is assessed over time. Lastly, in Section 6.3 sensitivity testing is performed with a focus on macroeconomic variables.

6.1. Out-of sample prediction accuracy

The selected models are evaluated on a different sample (that does not contain mortgages from the first one). This is necessary since if the models are evaluated on the training data, it can lead to overfitting. Three performance metrics are used for the four models: Concordance index, Brier score and Dynamic AUC. Since both Brier score and Dynamic AUC requires the estimation of individual survival curves at multiple time points (due to time-varying covariates) 5000 mortgages were selected randomly for the test sample since computation time with significantly more loans would be very long.

Table 10: Concordance index, Integrated Brier Score and Integrated AUC measures.

Model	**Concordance index**	**Integrated Brier score**	**Integrated Dynamic AUC**
Extended Cox model	0.695	0.0773	0.646
Discrete time logistic model	0.691	0.0784	0.641
Relative risk forest	0.873	0.0287	0.845
Conditional inference forest	0.879	0.0285	0.852

In Table 10 the main evaluation metrics are presented. From these metrics the two machine learning methods clearly outperform the Extended Cox model and the Discrete time logistic models. Conditional Inference Forest turns out to exhibit the best performance based on all the three metrics which is closely followed by Relative Risk forest. The Extended Cox model is the third best model while the Discrete time logistic model has only a slightly worse performance than the Extended Cox model.

The Dynamic Brier scores and the dynamic AUC values for the four models are presented in Figure 22 and Figure 23. From the Dynamic Brier score it appears that there is a decreasing trend in prediction accuracy over time. This is more prevalent in the case of the two traditional models where accuracy decreases at the highest rate between 100 and 160 months. For the machine learning models there the prediction accuracy decreases at a higher rate after 200 months. One can see that the dynamic Brier scores for the 2 machine learning and 2 traditional models move together and the difference is small. However, between the machine learning and traditional models, there is a considerable performance difference. Until 100 months the accuracy of the traditional and machine learning models are similar but the gap widens in the next 100 months.

Regarding dynamic AUC scores, one can observe a relatively stable performance over

time, performance only decreases significantly in the last few months. This may seem to contradict the Brier score performance over time, but this is not the case, since the Brier score compares the predicted probabilities with the realized outcomes while the AUC measures classification accuracy by comparing the true positive and false positive rates at each time point. The performance of the two class of models is very similar in this case as well and the best performing model turns out to be the conditional inference forest for most of the survival times, while the Extended Cox model also seem to outperform the logistic model.

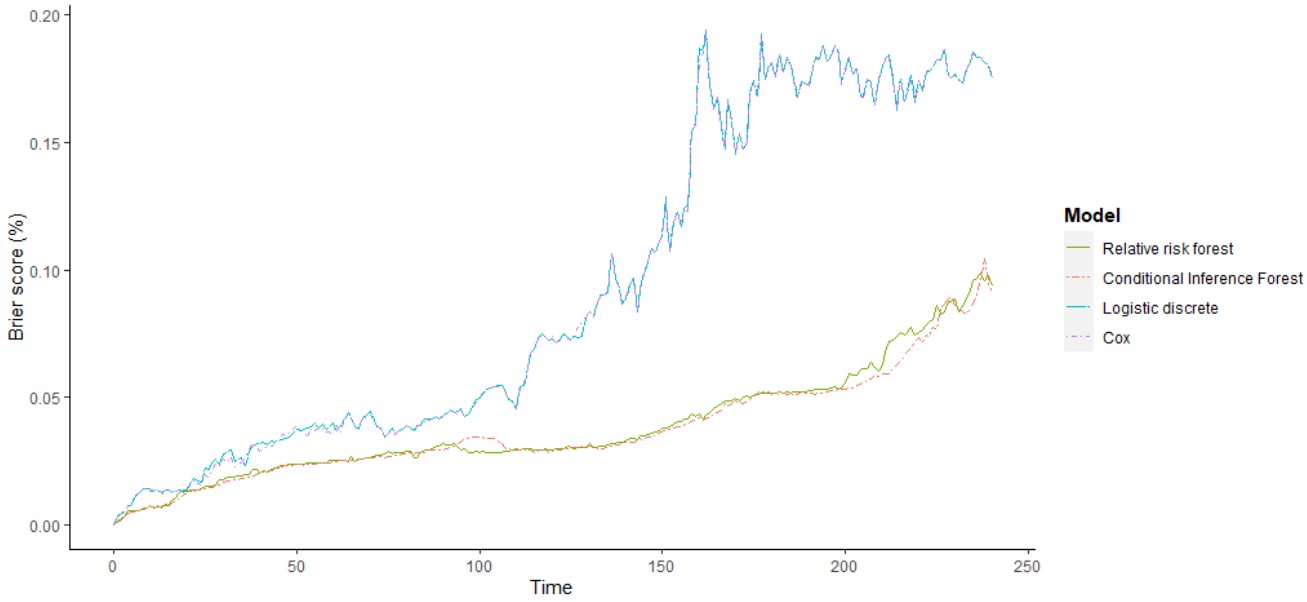


Figure 22: Dynamic Brier scores over time for the 4 models.

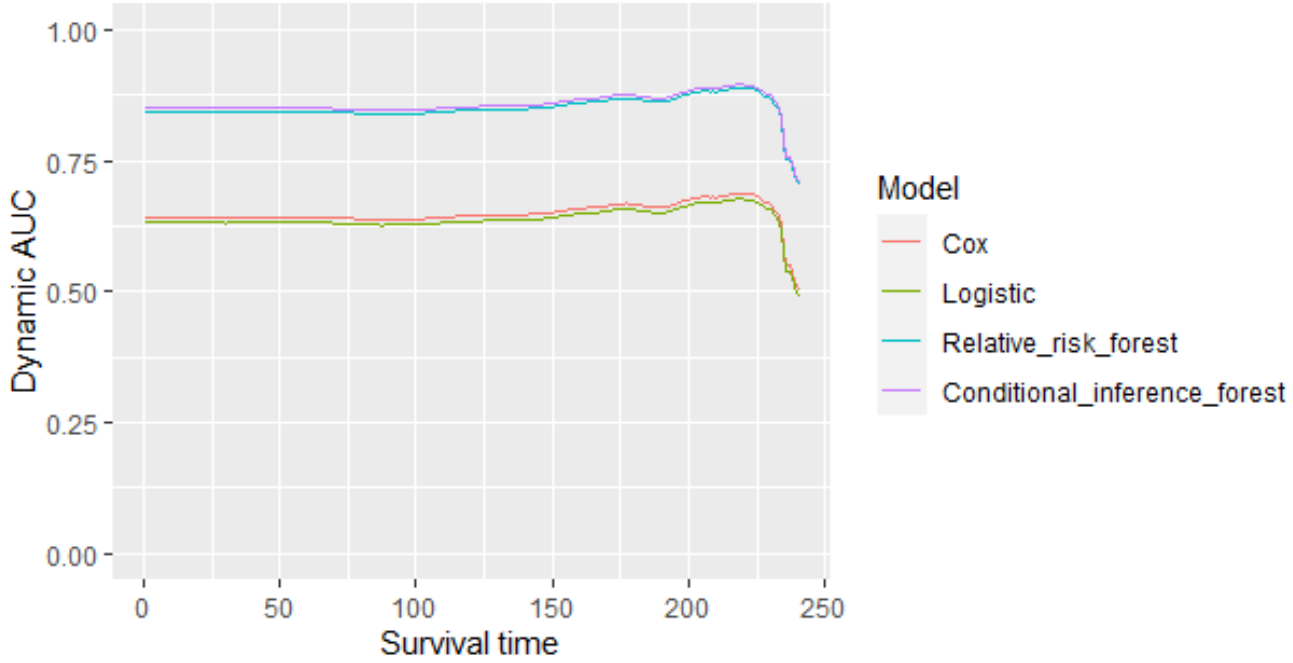


Figure 23: Dynamic AUC scores over time for the 4 models.

6.2. Prediction accuracy over time

In this section the realized prepayment percentage is compared to the predicted prepayment rates over time on the test sample. This measure is important from a practical point since consistent performance over time is desirable for a model that is supposed to be used for pricing or hedging prepayment risk frequently. The predicted prepayment probability in a given period is defined as the sum of the individual predicted prepayment probabilities divided by the number of observations in that month. The fit of the 4 different models are presented in Figure 24. Furthermore, the observation-weighted and unweighted root mean squared errors (RMSE) are also presented in Table 11.

As one can see from the figures and table, the average predicted monthly prepayment probability for the random survival forest and the conditional inference forest are following more closely the realized prepayment percentages than for the Extended Cox and the discrete time logistic model. Conditional Inference Forest exhibits the closest fit, However, a relatively big fluctuation can be still observed in the predicted prepayment rate fit close to the realized prepayment rate values for the machine learning models. Nonetheless, forest-based models generally seem to capture very well the underlying trend in prepayment rates over time. The fit for the extended Cox model and the discrete time logistic models are similar. For these models a bigger misalignment can be seen between the two lines in 2018-2019 and some prepayment spikes between 2008 and 2014 are somewhat underestimated by both traditional models. The RMSE values in Table 11 also support these conclusions.

Table 11: Weighted and unweighted RMSE difference between the fitted and realized prepayment rates.

Model	**Weighted RMSE**	**Unweighted RMSE**
Extended Cox model	0.006167141	0.007057281
Discrete time logistic model	0.006258633	0.00711237
Relative risk forest	0.004929258	0.004951041
Conditional inference forest	0.004563837	0.004582637

6.3. Sensitivity tests

To implement one of the models in practice, it is necessary to assess how sensitive the model outcome is with respect to time-varying variables used in the model. This is because while time-invariant covariates for a given loan remain the same in the future, for time-varying covariates this is definitely not the case and an expectation is necessary for future values to make predictions for prepayment rates. Hence, for successful prediction ahead in time one can think of possible scenarios for the future, especially for macroeconomic variables. Sensitivity is captured by the estimated hazard rates for parameteric and semiparmetric methods, hence this question is even more relevant for machine learning models. In this section macroeconomic sensitivities are tested in a way that in Section 6.3.1 interest rate sensitivities are presented while in Section 6.3.2 other macroeconomic scenarios are studied. A separate section is dedicated to interest

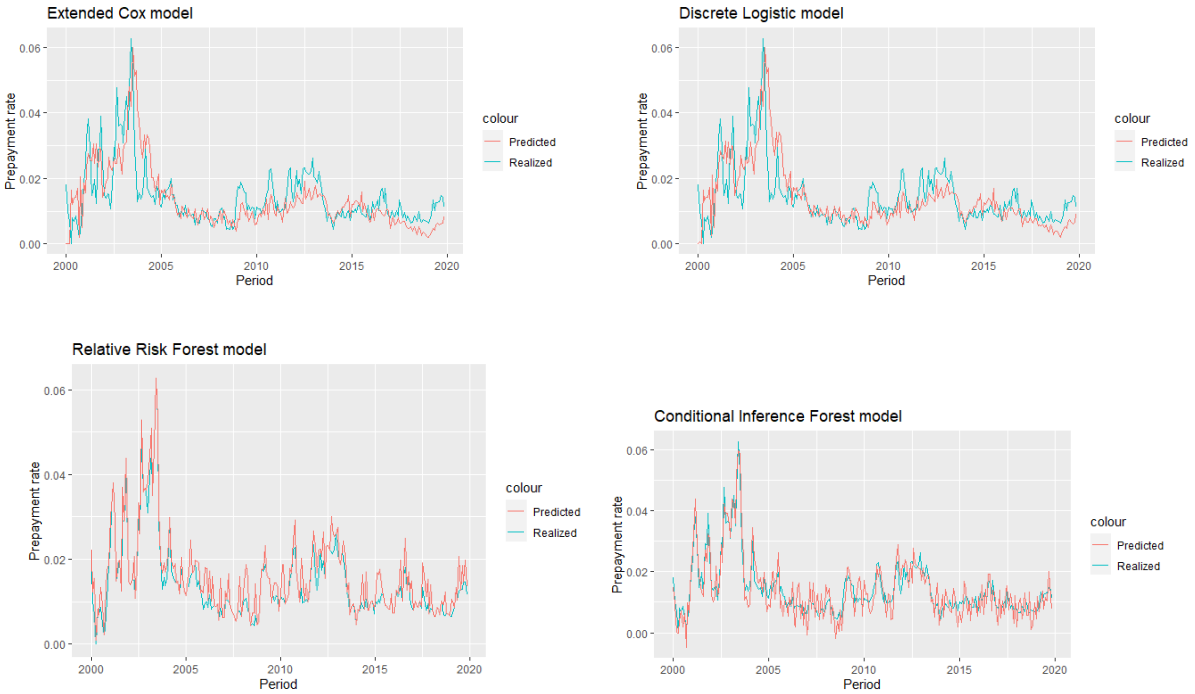


Figure 24: Predicted and realized monthly prepayment rate over time for different models.

rate sensitivity because interest rate incentive is one of the most important drivers and regulators also emphasize to assess the sensitivity of prepayment models to pre-specified interest rate scenarios. For example, in EBA (2018) Guidelines on the management of interest rate risk one can read:

> 106. In assessing the implications of optionality, institutions should take into account:
>
> - The potential impact on current and future loan prepayment speeds arising from the interest rate scenario, underlying economic environment and contractual features. Institutions should take into account the various dimensions influencing the embedded behavioural options.

6.3.1. Interest rate sensitivity

To test the interest rate sensitivity of prepayments, the out-of time performance is assessed on the last available year, 2019. For this reason, the models are re-estimated by leaving out this year. Then, predictions are carried out for the baseline scenario with the original interest rate incentives for 2019 and with modified interest rate incentives based on the Guidelines of EBA (2018). The applied shocks can be viewed as instantaneous happening in 2019 January. EBA defines the following scenarios:

- Parallel upward shift by 200 basis points
- Parallel downward shift by 200 basis points

- Upward shift in short rates
- Downward shift in short rates
- Flatter curve
- Steeper curve

EBA (2018) also provides formulas for different scenarios. For parallel shifts the interest rate incentive can be changed directly (in the opposite direction as the interest rate shock), while for the remaining 4 scenarios, the following formulas are applied:

$$\Delta R_{short} = R_{short,USD} \cdot \exp(\frac{-t_k}{4}) \tag{45}$$

$$\Delta R_{long} = R_{long,USD} \cdot (1 - \exp(\frac{-t_k}{4})) \tag{46}$$

$$\Delta R_{flattener} = 0.8 \cdot \Delta R_{short} - 0.6 \cdot \Delta R_{long} \tag{47}$$

$$\Delta R_{steepener} = -0.65 \cdot \Delta R_{short} + 0.9 \cdot \Delta R_{long} \tag{48}$$

where 4 is a specified constant by EBA for decay, $R_{short,USD}$ is 300 basis points, $R_{long,USD}$ is 150 basis points and t_k is the midpoint in a given tenor bucket. Five tenor buckets are used for the Fannie Mae data, namely the [0,120), [120,180), [180,240), [240,300), [360,Inf) month buckets. After changing the interest rate incentive using the above equations (but with opposite signs) and taking into account the 0% interest floor, the different interest rate scenarios are calculated for the 4 models. These scenarios are plotted in Figure 25.

As observed in Section 5, the traditional models underestimate prepayments for 2019 which remains true after re-estimating the models by omitting this year. The patterns are in line with the previous expectations: the 200 basis point parallel downward shift in the interest rates is associated with the biggest increase in prepayments while an upward interest rate has a significantly smaller effect in absolute terms. This seem to indicate that when interest rates are already low, an upward interest rate shift may only reduce prepayments slightly. Furthermore, short rates shock appear to have only a moderate effect on prepayments. This is also understandable since most of the loans have a 20 or 30 years maturity. It is also visible that in the machine learning models prepayment rates are less sensitive to interest rate movements.

6.3.2. Macroeconomic scenarios

In this section the potential effects of different macroeconomic scenarios on prepayment rates are studied. The procedure is similar as in case of interest rates: the re-estimated

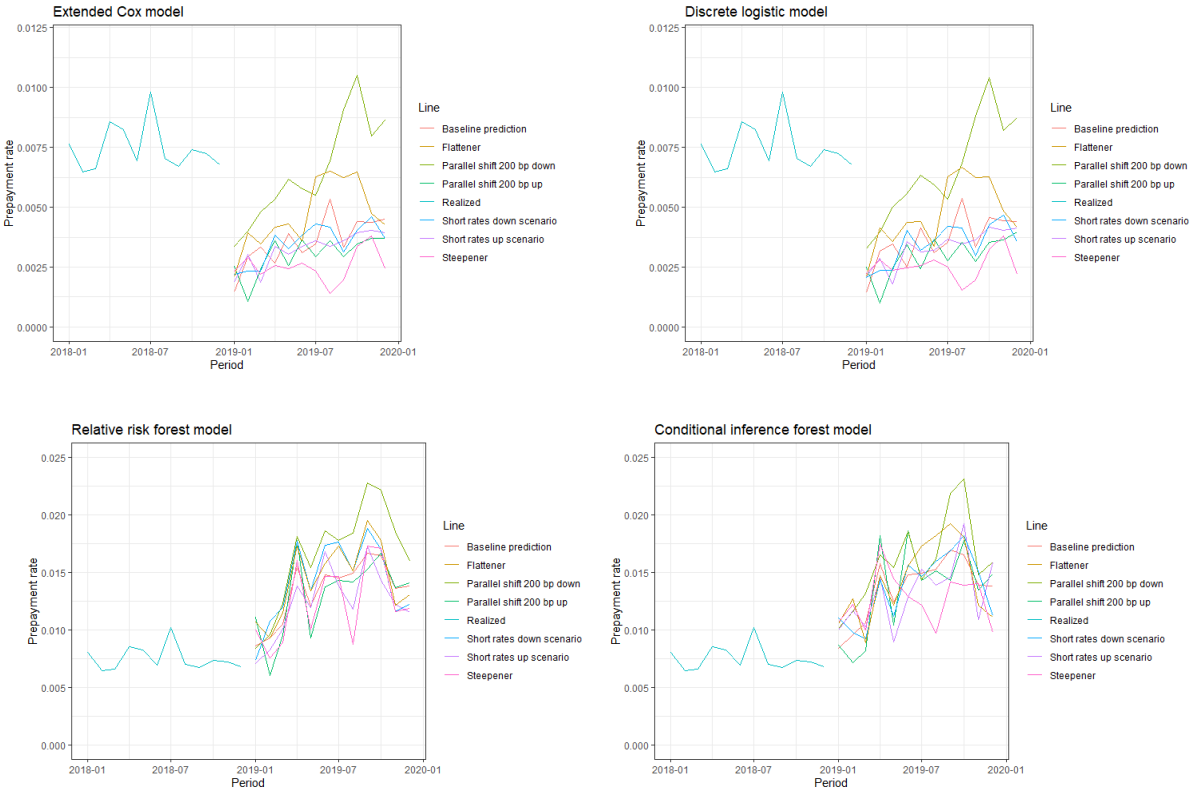


Figure 25: Interest rate scenarios forecast on 2019 for the 4 different models.

models are used and then the modified out-of-time fits for 2019 are compared. In this case, the macroeconomic scenarios by Federal Reserve (2019) are used as benchmark for 2019. In line with the requirements of the Dodd-Frank Act Stress Testing Rules and the Capital Plan Rule, $3 + 1$ scenarios are studied: baseline, adverse, severely adverse and an additional positive macroeconomic scenario which is not required. The positive macroeconomic scenario is defined as the inverse of the adverse scenario, except for the CPI forecast where reasonable values are used. The 4 scenarios are summarized in Table 12.

The four available macroeconomic variables that can be shocked are Consumer Price Index, House Price Index, Unemployment Rate and Nominal Disposable Income (which is expected to be highly correlated with Personal Income per Capita). House Price Index plays a role through interest rate incentive as described in Section 4.3, while other variables have a direct impact on prepayment rates in the models. The CPI and Nominal Disposable Income are defined in annual growth rates (with quarterly frequency) by Federal Reserve (2019) for 2019, hence for projections the values are converted to monthly frequency. Since there is great geographical heterogeneity in most of the macroeconomic variables, growth rates were calculated for Nominal Disposable Income, Unemployment Rate and HPI and these were applied to the out-of-time horizon for each geography. For example, the observed national unemployment rate was 3.8% in the fourth quarter of 2018, and 3.7% is used for the baseline scenario in the first quarter of 2019, then the equivalent unemployment growth for the months in the first quarter of 2019 is compared to the last month in 2018. To use this method successfully, loans originated in 2019 were excluded.

The different macroeconomic scenarios for the 4 models are presented in Figure 26. The

perfect forecast prediction corresponds to the realized macroeconomic values in 2019. In these figures it appears that in the positive scenario, especially in the case of the Extended Cox and Discrete Logistic models, prepayment rates are higher. This is also consistent with the previous expectations since prepayment rates tend to be pro-cyclical. However, it is more difficult to evaluate the impact of an adverse and severely adverse scenarios. Using traditional methods, a higher prepayment rate can be observed for the adverse scenarios in comparison to the baseline macro scenario. This can be attributed to the high sensitivity of CPI to prepayments in these models; namely that lower CPI for the adverse scenarios can counteract the negative effect of increased unemployment and decreased personal income and HPI (through incentive) on prepayments. For the two machine learning models, however, it is possible to observe a decreased prepayment rate when the severely adverse macroeconomic scenario is in place.

Table 12: Scenario projections for 2019 by Federal Reserve (2019).

Scenario	**Date**	**Nominal Disposable Income Growth**	**Unemp. Growth**	**CPI**	**HPI Growth**[7]
Base	Q1	4.2	-2.63	1.5	0.98
Base	Q2	4.5	-5.26	2.3	1.46
Base	Q3	4.4	-5.26	2.3	1.95
Base	Q4	4.3	-5.26	2.3	2.44
Adverse	Q1	-2.6	13.16	1.3	-1.95
Adverse	Q2	-2.4	34.21	2.0	-3.41
Adverse	Q3	-0.6	50.0	1.9	-5.37
Adverse	Q4	0.7	63.16	2.0	-7.32
Severely Adverse	Q1	-4.2	23.28	1.2	-2.93
Severely Adverse	Q2	-5.8	65.79	1.6	-5.85
Severely Adverse	Q3	-3.4	97.37	1.7	-9.27
Severely Adverse	Q4	-1.6	128.95	1.8	-13.17
Positive (extra)	Q1	2.6	-13.16	2	1.95
Positive (extra)	Q2	2.4	-34.21	2.5	3.41
Positive (extra)	Q3	0.6	-50.0	2.5	5.37
Positive (extra)	Q4	-0.7	-63.16	2.5	7.32

[7]Unemployment Growth and HPI Growth are defined in comparison to December 2018 values.

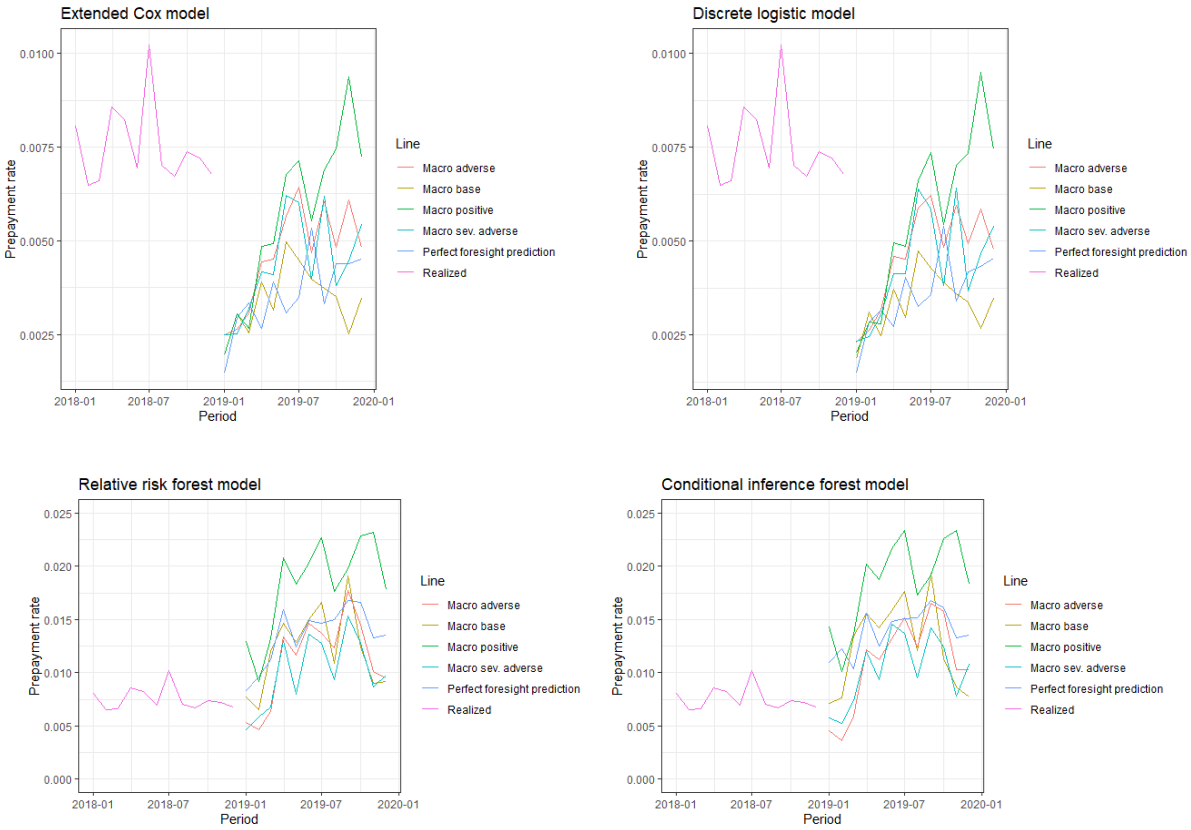


Figure 26: Macroeconomic scenarios forecast on 2019 for the 4 different models.

7. Conclusion

This thesis compared four methods to study the mortgage market in the US. All of these methods can be associated with survival analysis since they take into account censoring and they make it possible to predict individual survival curves by using the latest available information for the loan and a macroeconomic forecast. To capture the effect of macroeconomic variables effectively, highly granular and (relatively) high frequency macroeconomic data was collected from different sources and this data was matched with a randomly selected sample from the Single-Family Fixed Rate Mortgage dataset provided by Fannie Mae.

Four different models were estimated after data cleaning and transforming some variables (for the traditional models) based on explanatory driver analysis. Since there are time-dependent covariates, estimation was done by using all the observations for all the loans by using a *start-stop* data format. Since observed partial prepayments (or curtailments) are proportionally very small compared to full prepayment amounts, the modelling scope was limited to full prepayment.

Two of the applied models can be classified as traditonal methods: the semiparametric Extended Cox model and the Discrete logistic model. The other two methods - Relative Risk Forest and Conditional Inference Forest - are machine learning forest approaches applied to survival data. Machine learning models can uncover relationships that may remain hidden by traditional methods but on the other hand they are more difficult to interpret since the output does not include coefficients for different drivers, however variable importance can be quantified. Before estimating the traditional models, variable selection has been performed based on the Akaike Information Critierion (AIC). Since the Extended Cox model requires the proportional hazards assumption and some variable exhibited nonproportionality, time interactions were added to this model. For the machine learning models, relevant hyperparameters have been tuned before performing the model estimation.

One of the research questions posed in the beginning of the thesis focused on identifying the drivers of prepayment. To answer this question, based on the 4 models it appears that macroeconomic variables, especially incentive and CPI have the biggest impact. Furthermore, loan amount, and December appears to have a big explanatory power. For the machine learning models loan delinquency indicator is the most important variable and the traditional models also confirm that delinquency status is in a relatively strong negative relationship with prepayments.

Next, model performance was evaluated on a held out test sample based on three different metrics: the concordance index, the Brier score and dynamic AUC. The performance of the semiparametric Extended Cox model and the Discrete Logistic model is similar to each other which is also reflected in the similarity in the magnitudes and signs of the coefficients. However, in comparison to the performance metrics calculated on the machine learning models, it turns out that the traditional models underperform significantly. The difference between the relative risk forest and the conditional inference forest performance is relatively small, but the latter slightly performed better based on all the evaluation metrics. The conclusion of the model fit evaluation over time is similar: the two machine

learning models produced a better fit than the traditional models, however the fit of the traditional models is also relatively good with the exception of 2019 and some spikes in prepayments in the 2008-2013 period. This is also reflected in the calculated weighted Root Mean Squared Error which is close to 0.6 percentage points.

Lastly, interest rate and macroeocnomic sensitivity tests were performed for all the 4 models. The interest rate scenarios were based on EBA (2018) Guidelines while for the macroeconomic scenarios the supervisory scenarios of Federal Reserve (2019) and an additional positive scenario for 2019 were used by applying the growth rates for different geographies. For the interest rate scenarios, the 200 basis point downward interest rate shift appears to have the biggest (positive) impact on prepayment rates. Furthermore, it appears that short rate shocks have a limited effect on prepayments while flatter and steeper yield curves result in higher and lower prepayment rates, respectively. Regarding macroeconomic scenarios, the positive scenario is unambiguously related to higher prepayments based on all the 4 models. However, the impact of other scenarios is less conclusive, although machine learning models indicate that severely adverse scenarios reduce the propensity to prepay.

There are a number of limitations and possible improvement for future research in this thesis. Firstly, the thesis solely focuses on the US mortgage market, more specifically to Fannie Mae penalty-free fully amortizing mortgages. Extending this scope to other countries or different products could increases the external validity of the models. Secondly, the number of risk drivers is limited. This is partially due to the scarcity of the loan level data and to the fact that some variables are only available at loan origination. In practice, banks may know more about their customers and may also include more macroeconomic variables. Thirdly, the models do not distinguish between different terminations, it treats defaults for example as censoring events. Although prepayment are by far the most common terminating events, applying a competing risk approach could be investigated. Fourth, as mentioned before, partial prepayments are out of scope of the analysis. Lastly, a number of different methods such as neural networks and Bayesian approaches were not investigated in this thesis in order to limit the scope. Studying these methods, however, could provide additional insights into prepayment patterns.

Appendices

A. Additional summary statistics about the Single-Family Fixed Rate Mortgage dataset

Table 13: Summary statistics on the pooled 1% sample of the Single-Family Fixed Rate Mortgage dataset.

Variable name	Category	Mean	Med.	S.D.	Min	Max	25% Prc.	75% Prc.	Miss.%
Actual Period	-	2012-01-19	2012-06-01	1914 days	2000-01-01	2019-12-01	2007-10-01	2016-01-01	0
Original int. rate	-	5.139	5.125	1.19	2.125	11.375	4.250	6.000	0
Current int. rate	-	5.10	5.12	1.2	2.00	11.38	4.12	5.88	1.37%
Original Principal	-	178K	153K	105K	7K	1403K	100K	233K	0
Current Principal	-	136K	116K	108K	0	1396K	59K	193K	0
Original term	-	302.6	360	84.25	60	360	180	360	0
Origination date	-	2008-05-12	2008-02-01	1906 days	1999-01-01	2019-12-01	2003-06-01	2012-10-01	0
Loan age	-	43.18	32	38.15	-2	254	14	63	1.37%
Remaining months	-	261.3	303	95.38	0	483	171	340	1.37%
Adjusted Rem. months	-	252.8	294	97.6	0	480	166	337	2.94%
Maturity date	-	2032-02-09	2033-12-01	21717 days	2005-06-01	2060-06-01	2027-11-01	2041-05-01	0
Original LTV	-	69.49	75	17.58	1	97	59	80	0
Number of borrowers	-	1.577	2	0.51	1	2	1	10	0.02%
Debt to Income	-	33.2	33	11.7	0	64	24	42	1.98%
Credit Score (B)	-	742	755	54.5	300	850	706	786	0.43%
Number of units	-	1.04	1	0.25	1	4	1	1	0
Insurance percentage	-	24	25	7.3	1	50	17	30	84.3%
First home	Yes	0.1	0	0.3	0	1	0	0	0.5%
Modification	Yes	0.02	0	0.125	0	1	0	0	1.37%

Loan purpose	Cash out refinance	0.3	0	0.46	0	1	0	1	0
	Purchase	0.33	0	0.47	0	1	0	1	0
	No-Cash Out Refinance	0.36	0	0.48	0	1	0	1	0
Prop type	Condominium	0.077	0	0.27	0	1	0	0	0
	Cooperative	0.05	0	0.07	0	1	0	0	0
	Planned Urban Development	0.096	0	0.097	0	1	0	0	0
	Manufactured	0.16	0	0.37	0	1	0	0	0
	Single Family Home	0.75	1	0.43	0	1	0	1	0
Occupancy status	First	0.885	1	0.32	0	1	1	1	0
	Second	0.04	0	0.19	0	1	0	0	0
	Investment	0.075	0	0.26	0	1	0	0	0
Region	Midwest	0.23	0	0.42	0	1	0	0	0
	Northeast	0.16	0	0.37	0	1	0	0	0
	South	0.33	0	0.47	0	1	0	1	0
	West	0.28	0	0.45	0	1	0	0	0
	Outside	0.005	0	0.07	0	1	0	0	0
Zero Balance Code	Prepayment	0.95	1	0.21	0	1	0	1	98.6%
	Third Party Sale	0.002	0	0.04	0	1	0	0	98.6%
	Short Sale	0.003	0	0.06	0	1	0	0	98.6%
	REO Disposition	0.015	0	0.11	0	1	0	0	98.6%
	Matured	0.025	0	0.16	0	1	0	0	98.6%
	Notes Sale	0.001	0	0.03	0	1	0	0	98.6%
	Reperforming Notes Sale	0.002	0	0.05	0	1	0	0	98.6%

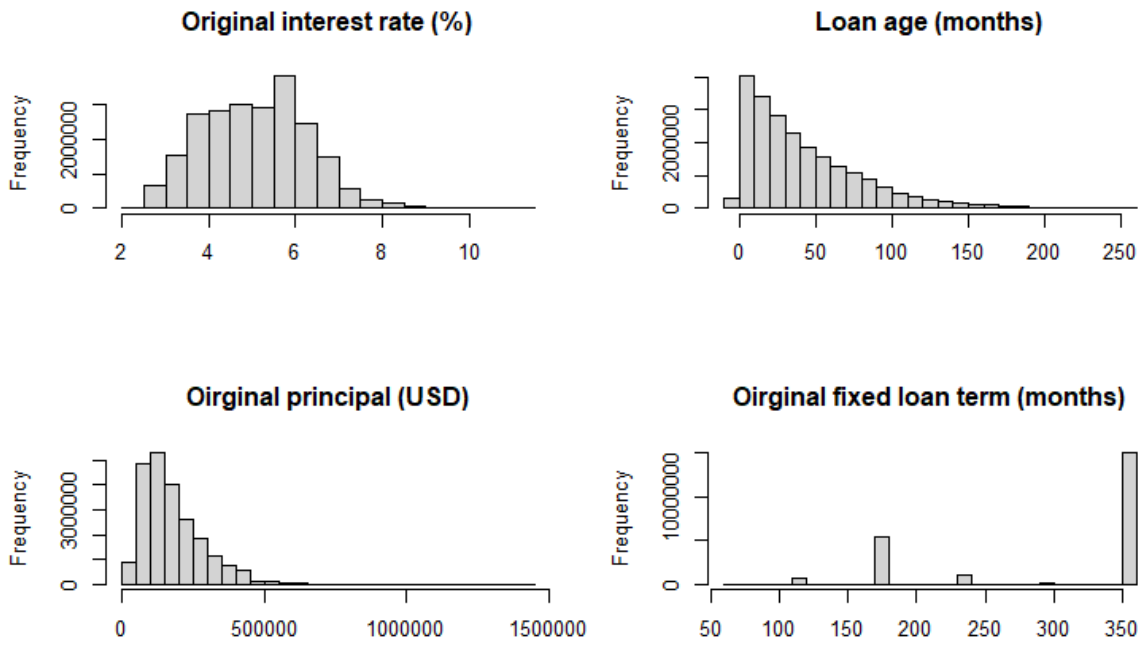


Figure 27: Histograms of some selected variables (1).

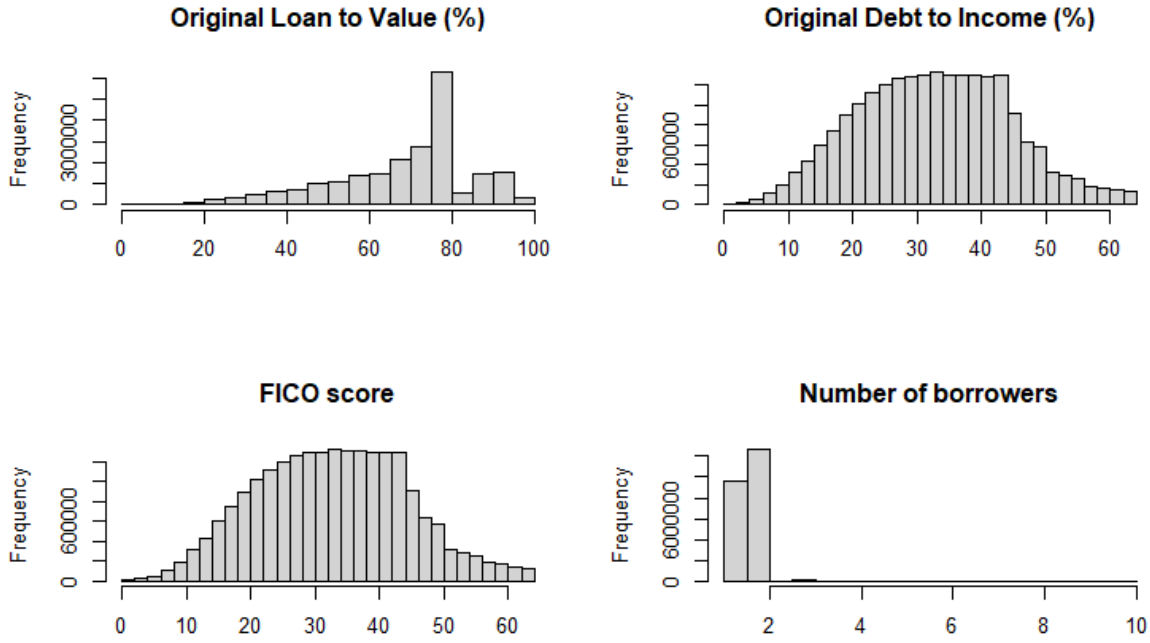


Figure 28: Histograms of some selected variables (2).

B. Regional differences in the relevant macroeconomic variables

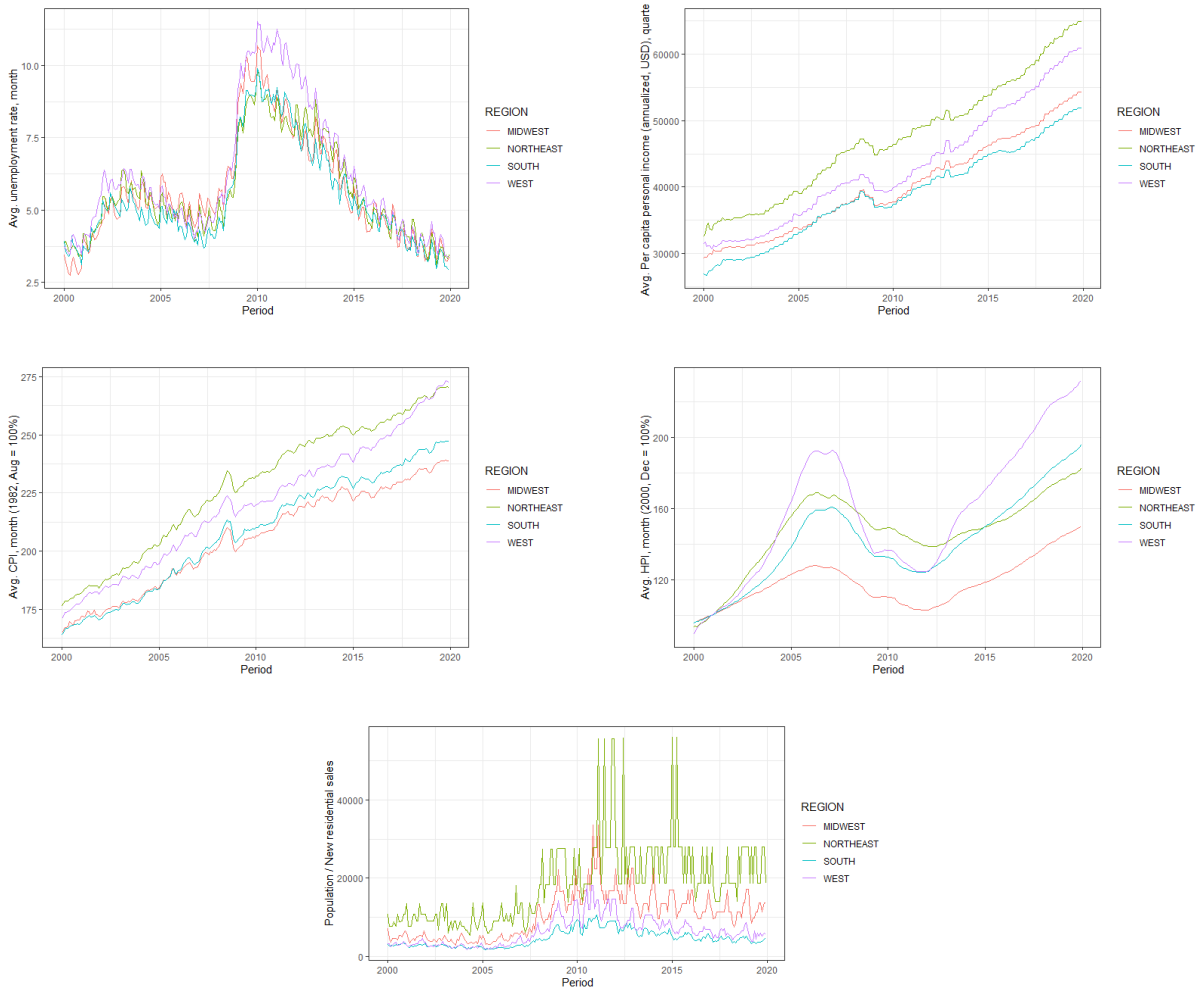


Figure 29: Average values of macroeconomic variables over time per region.

C. Additional bivariate and survival plots.

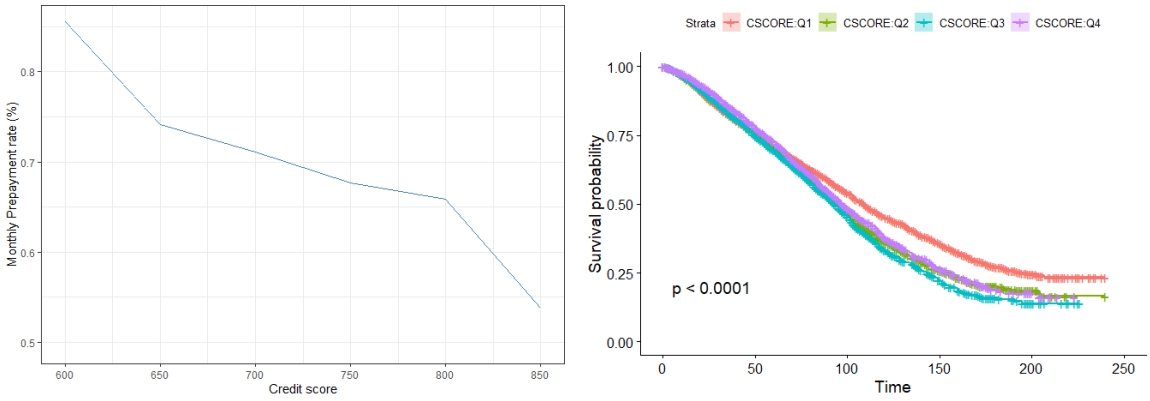


Figure 30: Relationship between prepayments and credit scores.

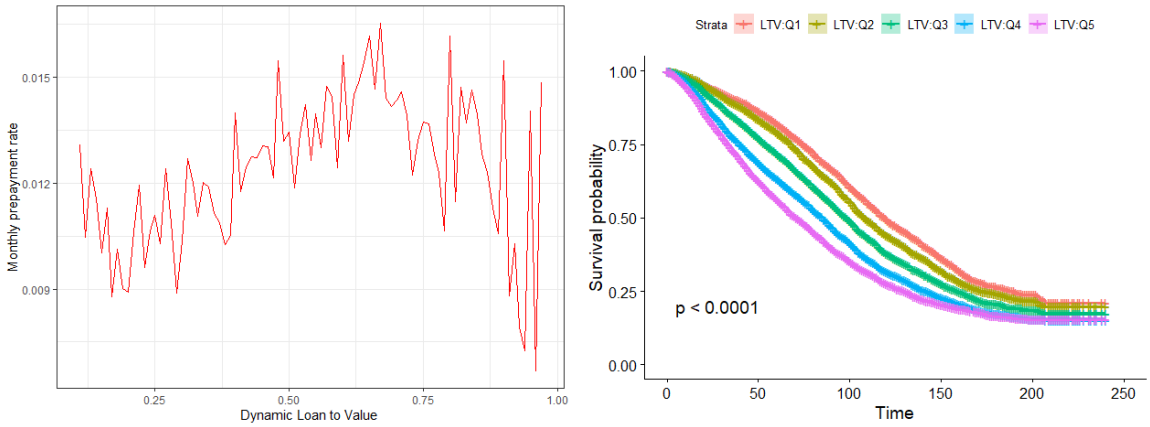


Figure 31: Relationship between prepayments and Loan to Value.

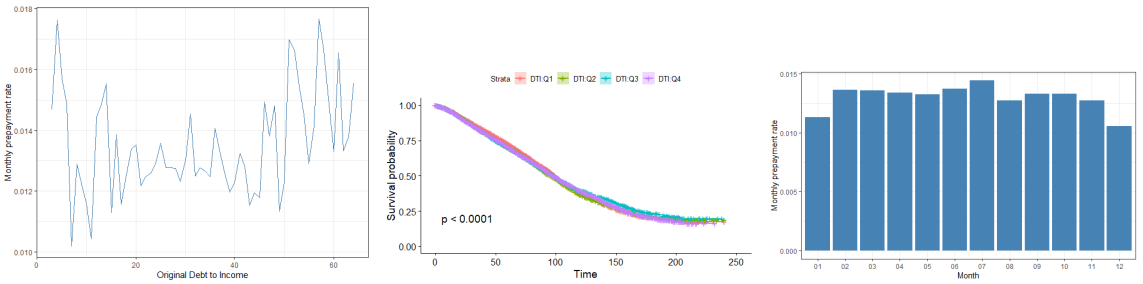


Figure 32: Relationship between Debt to Income and Months and Prepayments.

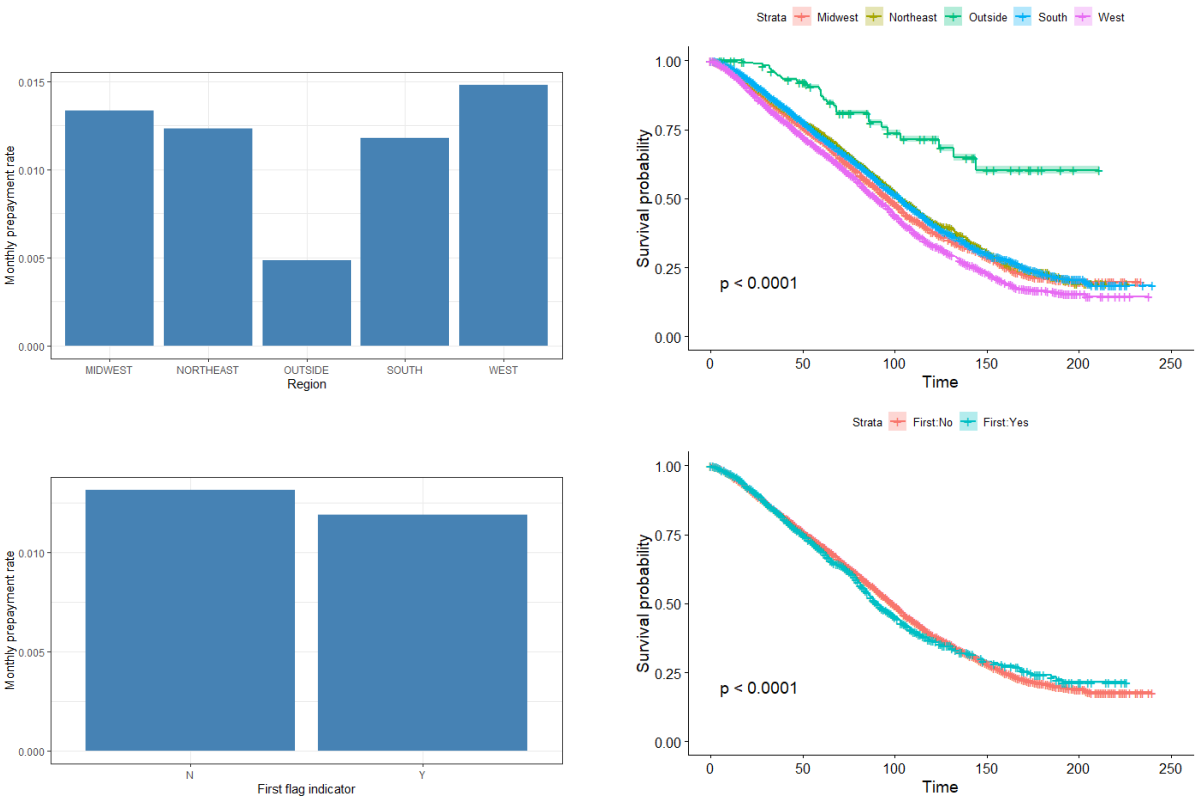


Figure 33: Relationship between Debt to Income, First homebuyer flag and Prepayments.

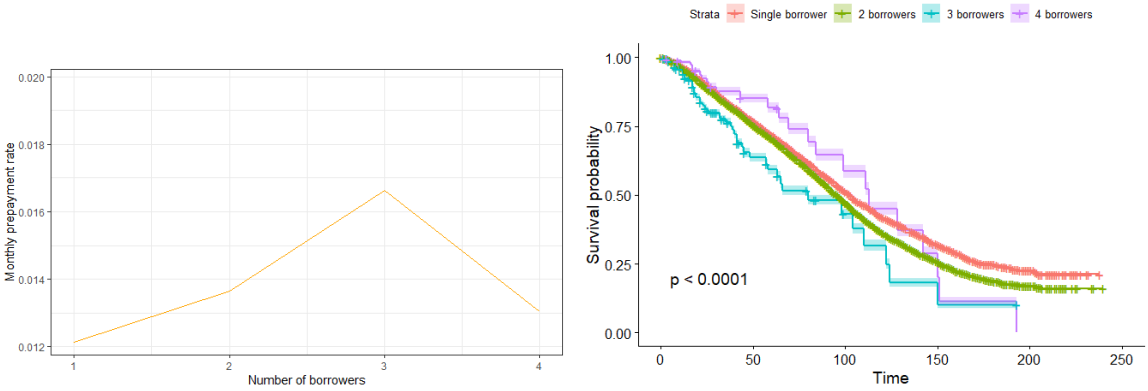


Figure 34: Relationship between prepayments and number of borrowers.

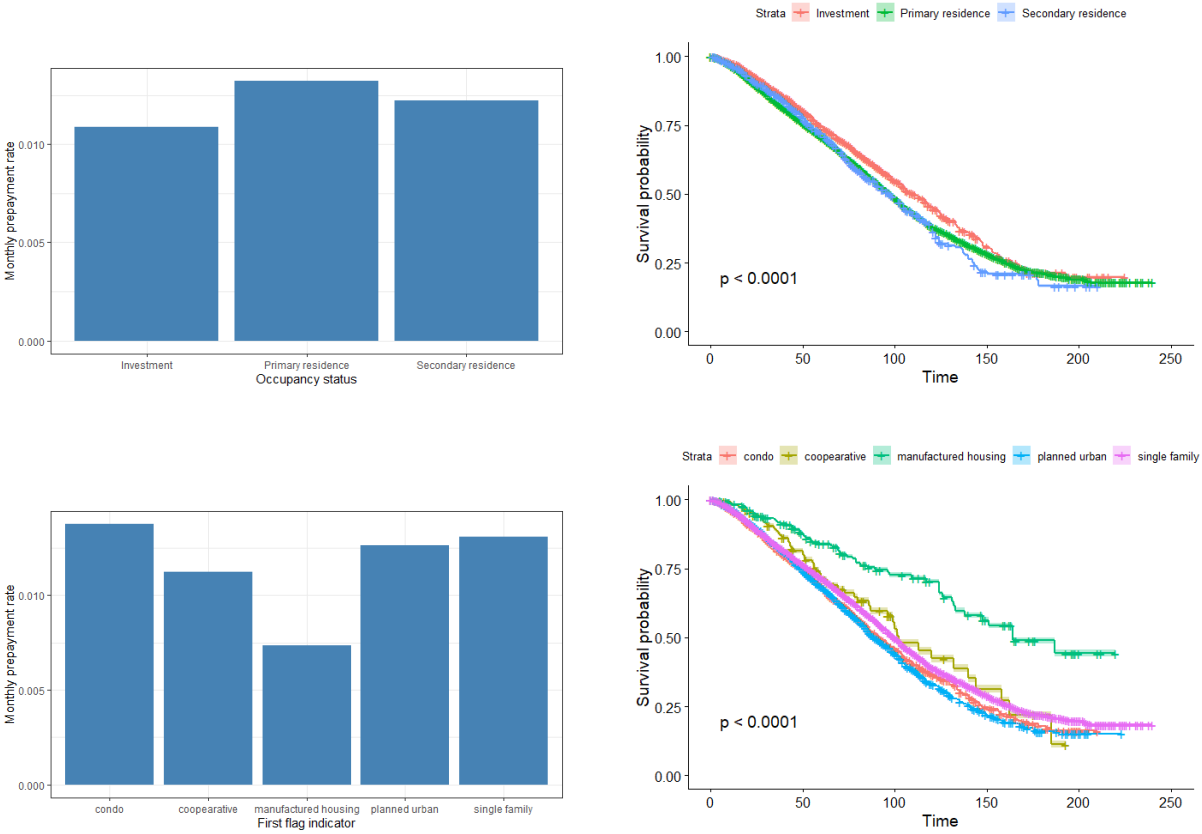


Figure 35: Relationship between prepayments, property types and occupancy status.

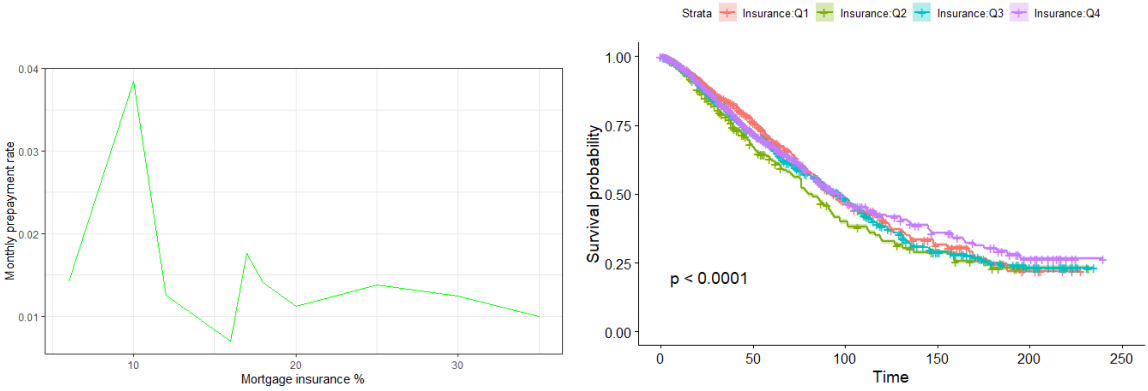


Figure 36: Relationship between prepayments and insurance percentage.

D. Variable selection steps with AIC for the Extended Cox Model

Table 14: AIC-based selection procedure example steps.

Step 1		**Step 2**		**Step 3**		**Steps...**	**Step 7**	
Var.	AIC	Var.	AIC	Var.	AIC	...	Var.	AIC
-L9MI	245325	-L6MI	245326	-L6MUR	245323	...	**Mod. 6**	245320
-L6MI	245326	-L6MUR	245326	-DTI	245323	...	+UR	245320
-L6MUR	245327	-DTI	245326	-Ins%	245323	...	+OLTV	245320
-DTI	245327	-Ins%	245327	-L3MUR	245323	...	+ DTI	245321
-Ins%	245328	-L3MUR	245327	-UR	245323	...	+L6MUR	245321
-L3MUR	245328	-UR	245327	**Mod. 2**	245323	...	+Ins%	245321
-UR	245328	**Mod. 1**	245327	-OLTV	245325	...	+L6MI	245321
Mod. 0	245329	-OLTV	245325	+L6MI	245325	...	-L3MUR	245321
-OLTV	245329	-L1MUR	245326	+L9MI	245326	...	-#Bor.	245321
-L1MUR	245329	+L9MI	245327	-L1MUR	245326	...	+L9MI	245322
-#Bor.	245330	-#Bor.	245327	-#Bor	245327	...	-Inc/cap	245322
-Inc/cap	245332	-Inc/cap	245327	-Inc/cap	245328	...	-Prop. t.	245322
-Prop. t.	245333	-Prop. t.	245327	-Prop. t.	245331	...	-First	245322
-First	245336	-First	245329	-P/NRS	245332	...	-P/NRS	245323
+P/NRS	245337	-P/NRS	245329	-Purpose	245332	...	-Purpose	245324
-Purpose	245338	-Purpose	245332	-First	245331	...	-#Units	245326
-#Units	245343	-#Units	245333	-#Units	245338	...	-L1MUR	245327
-Yield c.	245350	-Yield c.	245334	-Yield c.	245345	...	-Yield c.	245329
-L3MI	245354	-L3MI	245338	-L3MI	245337	...	-L3MI	245345
-Mod.	245366	-Mod.	245339	-CSC.	245368	...	-CSC	245346
-CSC	245374	-CSC	245339	-L1YUR	245370	...	-L1YUR	245348
-L12MI	245374	-L12MI	245340	-Delin	245345	...	-Delin	245352
-L1YUR	245375	-L1YUR	245346	-Mod.	245355	...	-L12MI	245354
-Delin	245379	-Delin	245349	-L12MI	245378	...	-Mod.	245356
-Occ. s.	245384	-Occ. s.	245362	-Occ. s.	245398	...	-Occ. s.	245357
-Month	245404	-Month	245370	-Month	245415	...	-Month	245434
-CPI	245447	-CPI	245444	-CPI	245441	...	-CPI	245441
-State	245744	-State	245741	-State	245739	...	-State	245732
-lnLoan	245852	-lnLoan	245850	-lnLoan	245848	...	-lnLoan	245887
-AtanInc	245943	-AtanInc	245945	-AtanInc	245944	...	-AtanInc	245948

Abbreviations a) Lagged variables: L+Number of months + year + Variable., b) I: incentive c) UR: Unemployment rate, d) OLTV: Original Loan to Value, e) Ins%: Insurance %, f) DTI: Debt to Income, g) #Bor.: Number of borrowers, h) Prop: Property, h) #Units: Number of units, i) CSC: Credit Score, j) First: First loan indicator k) Mod: Modification indicator l) Occ.s.: Occupancy status, m) Inc/Cap: Personal income per capita, n) P/NRS: Population/New Residential Sales.

E. Estimation results for month and state dummies

Table 15: States and Months result, Extended Cox Model.

Variable	**Coef.**	**Hazard rate** *exp*(Coef.)	**Robust S.E.**	**P-value**
Alabama	- 0.0671	0.935102	0.2051	0.743
Arkansas	0.0279	1.028293	0.2193	0.898
Arizona	0.2706	1.310751	0.2037	0.184
California	0.1821	1.199734	0.1858	0.327
Colorado	0.3061	1.358118	0.1931	0.113
Connecticut	0.0258	1.026136	0.2049	0.899
D.C.	- 0.3853	0.680247	0.2623	0.141
Delaware	- 0.3050	0.737123	0.2277	0.180
Florida	- 0.2173	0.804689	0.1904	0.253
Georgia	- 0.0099	0.990149	0.1972	0.960
Guam	- 0.6868	0.503184	1.0176	0.499
Hawaii	0.0112	1.011263	0.2181	0.958
Iowa	0.1934	1.213368	0.2037	0.342
Idaho	0.6519 ***	1.919184	0.2192	0.003
Illinois	0.0315	1.032001	0.1897	0.867
Indiana	- 0.1273	0.880469	0.2007	0.526
Kansas	- 0.0925	0.911649	0.2084	0.657
Kentucky	- 0.0206	0.979611	0.2135	0.923
Louisiana	- 0.0656	0.911649	0.2040	0.748
Massachusetts	0.4251 **	1.529743	0.2050	0.038
Maryland	- 0.3133	0.731031	0.1958	0.109
Maine	0.1519	1.164044	0.2237	0.497
Michigan	0.0774	1.080474	0.1941	0.690
Minnesota	- 0.1931	0.8244	0.1940	0.319
Missouri	0.1028	1.10827	0.1985	0.604
Mississippi	- 0.0354	0.965219	0.2229	0.873
Montana	0.2747	1.316136	0.2347	0.241
North Carolina	- 0.0903	0.913657	0.1989	0.649
North Dakota	0.0432	1.044147	0.2359	0.854
Nebraska	- 0.2253	0.798277	0.2177	0.300
New Hampshire	0.5609 ***	1.752249	0.2137	0.008
New Jersey	0.1001	1.105281	0.1930	0.604
New Mexico	0.1388	1.148894	0.2295	0.545
Nevada	0.1740	1.190056	0.2084	0.403
New York	- 0.0671	0.935102	0.1919	0.726
Ohio	- 0.1763	0.838366	0.1942	0.363
Oklahoma	- 0.2141	0.807268	0.2155	0.320
Oregon	0.4129 **	1.511194	0.2031	0.042
Pennsylvania	0.2328	1.262129	0.1965	0.236
Puerto Rico	- 0.5292 **	0.589076	0.2731	0.053
Rhode Island	0.6456 ***	1.907131	0.2194	0.003
South Carolina	- 0.1612	0.851122	0.2112	0.445

South Dakota	- 0.1322	0.876166	0.2541	0.602
Tennessee	- 0.0444	0.956571	0.1996	0.823
Texas	- 0.1628	0.849761	0.1903	0.392
Utah	0.3468	1.414534	0.2146	0.106
Virginia	- 0.3350 *	0.715338	0.1910	0.079
U.S. Virgin Islands	- 1.039 *	0.353808	0.5549	0.061
Vermont	0.0347	1.035309	0.2425	0.886
Washington	0.1262	1.134509	0.1918	0.510
Wisconsin	0.1576	1.170698	0.1927	0.413
West Virginia	- 0.1458	0.864331	0.2670	0.584
Wyoming	0.2087	1.232075	0.2750	0.447
February	0.2304 ***	1.259104	0.0450	<0.000000
March	0.2522 ***	1.286853	0.0449	<0.000000
April	0.3024 ***	1.353102	0.0456	<0.000000
May	0.2525 ***	1.287239	0.0462	<0.000000
June	0.2473 ***	1.280563	0.0448	<0.000000
July	0.3386 ***	1.402982	0.0441	<0.000000
August	0.2810 ***	1.324454	0.0451	<0.000000
September	0.3299 ***	1.390829	0.0449	<0.000000
October	0.2779 ***	1.320354	0.0450	<0.000000
November	0.2226 ***	1.249321	0.0447	<0.000000
December	- 0.0030	0.997004	0.0464	0.947460

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 701–726.

Abrahams, S. W. (1997). The new view in mortgage prepayments: Insight from analysis at the loan-by-loan level. *The Journal of Fixed Income*, *7*(1), 8.

Alink, B. J. (2002). Mortgage prepayments in the netherlands. explanatory research on prepayment variables and the effects on asset & liability management and securitisation.

Amar, S. (2020). Modeling of mortgage loan prepayment risk with machine learning.

An, X., Clapp, J. M., & Deng, Y. (2010). Omitted mobility characteristics and property market dynamics: Application to mortgage termination. *The Journal of Real Estate Finance and Economics*, *41*(3), 245–271.

Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The annals of statistics*, 1100–1120.

Banasik, J., Crook, J. N., & Thomas, L. C. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society*, *50*(12), 1185–1190.

Berger, M., & Schmid, M. (2018). Semiparametric regression for discrete time-to-event data. *Statistical Modelling*, *18*(3-4), 322–345.

Bhattacharya, A., Wilson, S. P., & Soyer, R. (2019). A bayesian approach to modeling mortgage default and prepayment. *European Journal of Operational Research*, *274*(3), 1112–1124.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

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

Bureau of Economic Analysis. (2021). *Regional data, gdp and personal income, personal income by major compenent and earnings by industry, sqinc5n*. Retrieved April 26, 2021, from https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index-Datasets.aspx

Census Bureau. (2021). *New residential sales, houses sold*. Retrieved April 26, 2021, from https://www.census.gov/construction/nrs/historical_data/index.html

Charlier, E., & Van Bussel, A. (2003). Prepayment behavior of dutch mortgagors: An empirical analysis. *Real estate economics*, *31*(2), 165–204.

Chau, K., Pretorius, F., & Yu, C. (2000). Factors affecting mortgage prepayment in hong kong.

Collett, D. (2015). *Modelling survival data in medical research*. CRC press.

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

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

Das, A., Abdel-Aty, M., & Pande, A. (2009). Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. *Journal of safety research*, *40*(4), 317–327.

Deng, Y., Zheng, D., & Ling, C. (2005). An early assessment of residential mortgage performance in china. *The Journal of Real Estate Finance and Economics*, *31*(2), 117–136.

Dunn, K. B., & McConnell, J. J. (1981). Valuation of gnma mortgage-backed securities. *The Journal of Finance*, *36*(3), 599–616.

EBA. (2018). *Eba/gl/2018/02, guidelines on the management of interest rate risk arising from non-trading book activities*. Retrieved May 24, 2021, from https://www.eba.europa.eu/sites/default/documents/files/documents/10180/2282655/169993e1-ad7a-4d78-8a27-1975b4860da0/Guidelines%5C%20on%5C%20the%5C%20management%5C%20of%5C%20interest%5C%20rate%5C%20risk%5C%20arising%5C%20from%5C%20non-trading%5C%20activities%5C%20%5C%28EBA-GL-2018-02%5C%29.pdf?retry=1

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

Ehrlinger, J. (2016). Ggrandomforests: Exploring random forest survival. *arXiv preprint arXiv:1612.08974*.

Fannie Mae. (2021a). *Fannie mae single-family loan performance data frequently asked questions (faqs)*. Retrieved April 26, 2021, from https://capitalmarkets.fanniemae.com/media/8921/display

Fannie Mae. (2021b). *Single-family loan performance dataset and credit risk transfer - glossary and file layout*. Retrieved April 26, 2021, from https://capitalmarkets.fanniemae.com/media/6931/display

Federal Housing Finance Agency. (2020). *Prepayment monitoring report third quarter 2020*. Retrieved April 24, 2021, from https://www.fhfa.gov/AboutUs/Reports/Pages/Prepayment-Monitoring-Report-Third-Quarter-2020.aspx

Federal Housing Finance Agency. (2021a). *Metropolitan statistical areas and divisions (not seasonally adjusted)*. Retrieved April 26, 2021, from https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index-Datasets.aspx

Federal Housing Finance Agency. (2021b). *Technical note. transitioning to the new omb 2013 metropolitan area definitions*. Retrieved April 26, 2021, from https://www.fhfa.gov/DataTools/Downloads/Documents/HPI_Focus_Pieces/2013Q2_HPIFocus_N508.pdf

Federal Reserve. (2019). *2019 supervisory scenarios for annual stress tests required under the dodd-frank act stress testing rules and the capital plan rule*. Retrieved May 24, 2021, from https://www.federalreserve.gov/newsevents/pressreleases/files/bcreg20190213a1.pdf

Foote, C., Gerardi, K., Goette, L., & Willen, P. (2010). Reducing foreclosures: No easy answers. *NBER Macroeconomics Annual*, *24*(1), 89–138.

FRED. (2018). *Who holds mortgages?* Retrieved April 24, 2021, from https://fredblog.stlouisfed.org/2018/05/who-holds-mortgages/?utm_source=series_page&utm_medium=related_content&utm_term=related_resources&utm_campaign=fredblog

Geanakoplos, J., Axtell, R., Farmer, J. D., Howitt, P., Conlee, B., Goldstein, J., Hendrey, M., Palmer, N. M., & Yang, C.-Y. (2012). Getting at systemic risk via an agent-based model of the housing market. *American Economic Review*, *102*(3), 53–58.

George, E. I., Popova, I., & Popova, E. (2008). Bayesian forecasting of prepayment rates for individual pools of mortgages. *Bayesian Analysis*, *3*(2), 393–426.

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

Green, J. R., & Shoven, J. B. (1983). The effects of interest rates on mortgage prepayments.

Green, R. K., & Wachter, S. M. (2005). The american mortgage in historical and international context. *Journal of Economic Perspectives*, *19*(4), 93–114.

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

Hall, A., & Maingi, R. Q. (2019). The mortgage prepayment decision: Are there other motivations beyond refinance and move?

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Hayre, L. S. (2003). Prepayment modeling and valuation of dutch mortgages. *The Journal of Fixed Income*, *12*(4), 25–47.

Hayre, L. S., Chaudhary, S., & Young, R. A. (2000). Anatomy of prepayments. *The Journal of Fixed Income*, *10*(1), 19–49.

Hess, W., & Persson, M. (2012). The duration of trade revisited. *Empirical Economics*, *43*(3), 1083–1107.

Ho, K. H. D., & Su, H. (2006). Structural prepayment risk behavior of the underlying mortgages for residential mortgage life insurance in a developing market. *Journal of housing economics*, *15*(3), 257–278.

Hosmer, D. W., & Lemeshow, S. (1999). *Applied survival analysis: Time-to-event* (Vol. 317). Wiley-Interscience.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651–674.

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

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

Ishwaran, H., Blackstone, E. H., Pothier, C. E., & Lauer, M. S. (2004). Relative risk forests for exercise heart rate recovery as a predictor of mortality. *Journal of the American Statistical Association*, *99*(467), 591–600.

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

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

Jenkins, S. P. (2005). Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, *42*, 54–56.

Kalbfleisch, J. D., & Prentice, R. L. (2011). *The statistical analysis of failure time data* (Vol. 360). John Wiley & Sons.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457–481.

Kara, A., Marques-Ibanez, D., & Ongena, S. (2016). Securitization and lending standards: Evidence from the european wholesale loan market. *Journal of Financial Stability*, *26*, 107–127.

Kau, J. B., & Keenan, D. C. (1995). An overview of the option-theoretic pricing of mortgages. *Journal of housing research*, 217–244.

Kau, J. B., Keenan, D. C., Muller, W. J., & Epperson, J. F. (1992). A generalized valuation model for fixed-rate residential mortgages. *Journal of money, credit and banking*, *24*(3), 279–299.

Kleinbaum, D. G., & Klein, M. (2010). *Survival analysis*. Springer.

LeBlanc, M., & Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics*, 411–425.

Park, M. Y., & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(4), 659–677.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Saito, T. (2018). *Mortgage prepayment rate estimation with machine learning* (Doctoral dissertation). Master's thesis, Delft University of Technology.

Schultz, G. M. (2016). *Investing in mortgage-backed and asset-backed securities,+ website: Financial modeling with r and open source analytics*. John Wiley & Sons.

Schwartz, E. S., & Torous, W. N. (1989). Prepayment and the valuation of mortgage-backed securities. *The Journal of Finance*, *44*(2), 375–392.

SIFMA. (2021). *Fixed income outstanding*. Retrieved April 24, 2021, from https://www.sifma.org/resources/research/fixed-income-chart/

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, *39*(5), 1.

Sirignano, J., Sadhwani, A., & Giesecke, K. (2016). Deep learning for mortgage risk. *arXiv preprint arXiv:1607.02470*.

Statista. (2021a). *Value of mortgage debt outstanding in the united states from 2001 to 2020*. Retrieved April 24, 2021, from https://www.statista.com/statistics/274636/combined-sum-of-all-holders-of-mortgage-debt-outstanding-in-the-us/

Statista. (2021b). *Value of mortgage debt outstanding on family residences in the united states from 2001 to 2020*. Retrieved April 24, 2021, from https://www.statista.com/statistics/274638/mortgage-debt-outstanding-on-us-family-residences/

Statista. (2021c). *Value of mortgage debt outstanding on farm property in the united states from 2001 to 2020*. Retrieved April 24, 2021, from https://www.statista.com/statistics/274649/mortgage-debt-outstanding-on-us-farm-property/

Statista. (2021d). *Value of mortgage debt outstanding on nonfarm and nonresidential property in the united*. Retrieved April 24, 2021, from https://www.statista.com/statistics/274644/mortgage-debt-outstanding-on-nonfarm-nonresidential-us-property/

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, *8*(1), 1–21.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, *16*(4), 385–395.

Tsiatis, A. A. et al. (1981). A large sample study of cox's regression model. *The Annals of Statistics*, *9*(1), 93–108.

U.S. Bureau of Labor Statistics. (2021a). *Consumer price index, regional resources*. Retrieved April 26, 2021, from https://www.bls.gov/cpi/regional-resources.htm

U.S. Bureau of Labor Statistics. (2021b). *Metropolitan area employment and unemployment archived news releases*. Retrieved April 26, 2021, from https://www.bls.gov/bls/news-release/metro.htm

U.S. Department of Treasury. (2021). *Daily treasury yield curve rates*. Retrieved April 26, 2021, from https://www.treasury.gov/resource-center/data-chart-center/interest-rates/pages/textview.aspx?data=yield

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

Yao, W., Frydman, H., Larocque, D., & Simonoff, J. S. (2020). Ensemble methods for survival data with time-varying covariates. *arXiv preprint arXiv:2006.00567*.

Yiwen, C. M. F. (2007). The risk and the management of the prepayment behavior in commercial bank's consumer credit loans: Theory and empirical analysis [j]. *Journal of Financial Research*, *7*.

Zenios, S. A., & Kang, P. (1993). Mean-absolute deviation portfolio optimization for mortgage-backed securities. *Annals of Operations Research*, *45*(1), 433–450.