PressAcademia
www.pressacademia.org/journals/jefa

JEFA
ISSN 2148-6697

Journal of Economics, Finance and Accounting YEAR 2024 VOLUME 11 ISSUE 1

PREPAYMENT AND DEFAULT RISK: A REVIEW

DOI: 10.17261/Pressacademia.2024.1896
JEFA- V.11-ISS.1-2024(6)-p.43-59

Sukriye Tuysuz
Yeditepe University, International Finance, Istanbul, Turkiye.
sukriye.tuysuz@yeditepe.edu.tr , ORCID: 0000-0001-8391-6521

Date Received: February 12, 2024 **Date Accepted:** June 1, 2024

To cite this document
Tuysuz, S., (2024). Prepayment and default risk: a review. Journal of Economics, Finance and Accounting (JEFA), 11(6), 43-59.
Permanent link to this document: http://doi.org/10.17261/Pressacademia.2024.1896

ABSTRACT

Purpose- The main purpose of this article is to make a comprehensive review of existing studies on prepayment and default (competing risk). This review enables to shed light on the main determinant of prepayment and default as well as on methods used to model competing risk.
Methodology- A comprehensive review of existing studies/articles.
Findings- More recently proposed machine learning methods (Random Survival Forest and Random Competing Risks Forests, as well as the DeepHit model and Dynamic DeepHit model) enable to take into account the complex/no-linear response of prepayment and default to their determinant more efficiently.
Conclusion- To model properly/correctly the prepayment and default risks it is important to consider the fact that the exercise of the prepayment option brings an end to the default option, and vice versa. These both risks should be modelled together: competing risk. Furthermore, models/methods accounting the complex/no-linear impact of explanatory variables on prepayment and default risks should be used; such as the Random Survival Forest and Random Competing Risks Forests, as well as the DeepHit model and Dynamic DeepHit model.

Keywords: Prepayment risk, default risk, multinomial logit models, multi-state model, random survival forest, Dynamic DeepHit model
JEL Codes: G10, G12, C58

1. INTRODUCTION

A mortgage loan is an agreement between a lender, known as the mortgagee, and a borrower, known as the mortgagor. Within this contract, the mortgagee and mortgagor establish various aspects of the loan, such as its size, repayment schedule, attached life insurance policy, interest rate, collateral, and other relevant details. The mortgagors may unexpectedly default on or prepay the loan. They possess loans that contain American straddles, which are combined call and put options, sold by the mortgagee.1[1] The prepayment or default of loans exposes the mortgagee to several risks, including credit risk, interest rate risk, liquidity risk, mispricing towards customers, and mispricing towards special-purpose vehicles (securitization). To assess and mitigate these risks, as well as for asset-liability management (ALM) purposes and to comply with accounting rules by marking their banking books to market, banks need to evaluate the prepayment and default risks associated with mortgage loans.

Modelling prepayment and default in financial analysis and risk management is of significant importance for financial agents, portfolio managers as well as for policymaker several reasons. Prepayment and default events can have a substantial impact on the risk profile of loan portfolios. By accurately modelling these events, financial institutions can better assess and manage the associated credit and interest rate risks. Furthermore, proper modelling of prepayment and default allows investors and market participants to accurately value cash flows and expected returns of mortgage-backed securities (MBS) and other asset-backed securities (ABS), aiding in investment decisions and pricing. Similarly, lenders and investors need reliable models to price mortgage loans, estimate the expected cash flows, and assess the risk associated with prepayment and default. These models help in determining appropriate interest rates, loan terms, and risk premiums. In term of regulatory frameworks, such as Basel III, require financial institutions to assess and adequately reserve capital against potential credit losses. Robust prepayment and default models assist in estimating expected credit losses, ensuring compliance with regulatory capital

[1] The borrower receives two options: prepayment as a call option and default as a put option.

requirements. Prepayment and default models provide also valuable insights for strategic decision-making. Lenders can evaluate the impact of different loan products, underwriting standards, and risk management strategies on prepayment and default rates, enabling them to optimize their loan origination and servicing processes. Finally, prepayment and default behavior can reflect broader economic trends and market conditions. By incorporating these models into macroeconomic analysis, policymakers and researchers can gain insights into the health of the housing market, consumer behavior, monetary policy implications, and potential systemic risks.

A mortgage loan can end due to two factors: prepayment and default, which represent competing risks. In the context of mortgage loans, competing risks arise because exercising the prepayment option results in the termination of the default option, and vice versa (Ambrose and Sanders, 2003). Competing risks modelling has gained popularity in analyzing duration data where an event can have multiple causes of termination across various fields, including economics, finance, and risk management.

Prepayment and default risks can be assessed using either financial frameworks or econometric methods rooted in behavioral analysis. Financial methods rely on option pricing models and are commonly employed in pricing callable securities or corporate loans (Chen, 1996; Cheyette, 1996; Deng et al. 2000; Longstaff, 2002; Levin and Davidson, 2005). According to this approach, the rational exercise of the call option by the mortgagor for prepayment or default occurs only when the option is "in the money." In other words, the actual value of the asset should exceed the remaining mortgage balance plus transaction costs. By employing option pricing theory and models, this approach enables the calculation of prepayment risk, default risk, and the value of the mortgage.

However, mortgagors often do not exercise their options rationally. In fact, some mortgagors choose to prepay or default even when it is not financially advantageous, while others fail to exercise the option when it would be beneficial to do so. This irrational behavior among mortgagors can primarily be attributed to behavioral factors. Empirical evidence has shown that prepayment risk is mainly influenced by factors such as refinancing considerations (coupon effects, interest rates, yield curve), personal motives (job changes, relocation, divorce, increased household income, emigration, etc.), and market conditions (business cycle, economic situation, housing market conditions, etc.). As for default risk, its primary determinants are house equity, loan-to-value (LTV) ratio, loan-to-income (LTI) ratio, and debt-coverage ratio (DCR). Additionally, the default rate is influenced by other borrower and loan-specific characteristics, as well as macroeconomic variables.

To address these components, statistical approaches have been proposed and utilized to model and quantify default and/or prepayment risks. Previous studies mostly examined these risks separately, disregarding the fact that they are interdependent. A defaulted loan cannot be prepaid, and a prepaid loan cannot default. Therefore, it is essential to investigate them as competing risks. Several authors have employed competing risk methods to analyse prepayment and default risks in residential mortgages (Ambrose and Capone, 2000; Ambrose and LaCour-Little, 2001; Clapp et al., 2001; Deng et al., 2000). Ciochetti et al (2002) and Ambrose and Sanders (2003) investigated prepayment and default risks (competing risks) in commercial mortgages. Traditional approaches, such as multinomial logistic models, multi-state models, and those based on the Cumulative Incidence Function (CIF), have been commonly used in existing studies to analyse these competing risks. These models rely on functional forms and predominantly consider the linear impact of covariates. However, mis specifying the model can have adverse effects on the results. Additionally, some authors have demonstrated the nonlinear impact of covariates on prepayment and default risks (Sirignano et al., 2018). To overcome these limitations, machine learning and deep learning methods have been introduced. Ishwaran et al. (2014) introduced the Competing Risks Forest method, which is based on the random forest technique and Random Survival Forest (RSF). Lee et al. (2018) developed the DeepHit model, and Sirignano et al. (2018) proposed a deep learning multi-state method to investigate competing prepayment and default risks. These methods directly estimate the CIF, model the nonlinear effects and interactions of covariates, and are free from model assumptions. By not assuming a specific underlying stochastic process, these models allow for a flexible relationship (both linear and nonlinear) between covariates and risks over time.

2. LITERATURE REVIEW

Prepayment rate and default rate have their own determinants as well as some common explanatory variables such macroeconomic and financial variables. These specific and common explanatory variables are presented in what follow.

2.1. The Main Determinants of Prepayment Risk

Determinants of prepayment are often classified into four classes (Hayre, 2003; Jacobs et al., 2005): Refinancing components, Housing turnover, Defaults, and Curtailments.

Refinancing components: The incentive to refinance is closely tied to the movement of interest rates. When market rates significantly drop below the coupon rate, it may trigger the option to prepay for fixed-rate loans. Conversely, for floating-rate loans, an increase in interest rates can lead to prepayment. However, it has been observed that this incentive is not always rational in practice (Hayre, 2003; Jacobs et al., 2005). In the study conducted by Sirignano et al. (2018), they discovered a complex relationship between the actual prepayment rate and the prepayment incentive, which is measured by the difference between the initial mortgage rate and the prevailing market rate. Moreover, they observed that the sensitivity of prepayment rates to changes in the incentive varied significantly in magnitude and direction. According to these authors, the refinancing component exhibits a nonlinear impact on the prepayment rate.

Mortgagors consider not only the current observed interest rates but also the anticipated future movement of interest rates when deciding whether to prepay or not. The expected future interest rate dynamics are often assessed by examining the shape of the yield curve (Goodarzi et al., 1998; Ambrose and Sanders, 2003). In their analysis of prepayment and default risk, Ambrose and Sanders (2003) incorporated the spread between the current contract interest rate and the 10-year Treasury rate as well as a measure of the term structure. The term structure, which serves as an indicator of market expectations, is calculated as the difference between the 10-year Treasury bond rate and the 1-year Treasury bond rate. Moreover, building on the findings of Kau et al. (1993), Ambrose and Sanders (2003) also took into account the volatility of the 10-year Treasury rate. Kau et al. (1993) discovered that interest rate volatility negatively influences the prepayment rate. The results obtained by Ambrose and Sanders (2003) revealed that the yield curve exerts a negative and statistically significant impact on both the prepayment and default rates.

In relation to coupon effects, a lower coupon rate will result in a slower pace of prepayment. The coupon rate is influenced by borrower and loan-specific factors, as well as the year in which the loan was originated. Borrowers with higher credit scores are typically offered lower coupon rates compared to those with lower scores. Additionally, during periods of high interest rates, such as inflationary periods, the proposed coupon rates are higher than during periods of lower interest rates. This situation predominantly arises during economic recessions and times of crisis.

Housing turnover: The rate of prepayment can also be influenced by personal motivations and seasoning effects. Personal reasons may include factors such as job changes or relocations, divorce, increased household income, emigration, and credit score. Several authors, including Agarwal and Taffler (2008), Consalvi and di Freca (2010) and Sirignano et al., (2018), have found that higher credit scores have a positive impact on the prepayment rate.

Seasoning effects are closely associated with the business cycle, the overall economic situation, and prevailing conditions in the housing market. During a booming housing market, the rate of prepayment tends to increase. Hoff (1996) found that the likelihood of prepayment for 30-year Fixed-Rate Mortgages (FRMs) is influenced by upward trends in housing prices and personal income growth. This probability also rises during expansionary periods, as more job changes occur and unemployment decreases.[2] While most existing studies have found limited or no evidence for the influence of unemployment rates on prepayment (Deng, 1997; Elul et al., 2010; Foote et al., 2010). Sirignano et al. (2018) discovered that unemployment rates are the most influential factor in explaining borrower behavior, particularly in terms of prepayment. According to their findings, the relationship between prepayment and unemployment is highly nonlinear and heavily dependent on the borrower's credit score and the prevailing level of unemployment. An increase in the unemployment rate affects the gap between the prepayment rates of borrowers with high and low FICO scores. Similarly, Agarwal and Taffler (2008) put forth the argument that a decrease in borrower credit quality, as measured by the Fair Isaac Company (FICO) score, would diminish the likelihood of prepayment. They also asserted that similar effects can be observed when there is an increase in interest rates and the unemployment rate. In other words, lower borrower credit quality, higher interest rates, and a rise in the unemployment rate all contribute to reducing the probability of prepayment.

Loan-specific characteristics play a crucial role in determining the prepayment rate. Factors such as the type of mortgage product, loan size, and remaining time until maturity have been identified as important determinants (Jacobs et al., 2005; Smith et al. 2007). Additionally, Smith et al. (2007) discovered that the locations of real estate properties, interest rates, and loan amounts borrowed also exert a significant influence on prepayments.

2.2. Determinants of Default Rates

The main determinants of default are the House equity, the Loan-to-value (LTV), the Loan-to-income (LTI), and the Debt-Coverage-Ratio (DCR). Default rate is also impacted by other borrower, loan specific features, and macroeconomic variables.

The significance of decreasing and potentially negative home equity in relation to the default rate has been highlighted in several studies (Vandell, 1978; Campbell and Dietrich, 1983; Bajari et al., 2008). Home equity refers to the difference between the value of the property and the remaining loan balance. According to the option pricing framework, borrowers strategically default when they enter a negative equity position.[3] Empirical evidence from authors such as Mayer et al. (2009) and

[2] In contrast, in downturn situation prepayments are triggers by defaults.

[3] See Vandell (1995) for a review of empirical literature.

Goodman et al. (2010) have supported the primary role of negative equity in explaining mortgage defaults. However, these findings have been challenged by other authors who argue that homeowners do not default immediately upon reaching negative equity but rather experience a delay (Bhutta et al., 2010; Deng et al., 2000; An and Qi, 2012; Guiso et al., 2013; Campbell and Cocco, 2015; Sirignano et al., 2018). Bhutta et al. (2010) and Guiso et al. (2013) have demonstrated through different methods that equity shortfalls must exceed 50% before strategic default becomes prevalent. In the case of Sirignano et al. (2018), they discovered a nonlinear relationship between home equity and the default rate. Specifically, their findings indicated that the sensitivity of the delinquency rate to changes in house prices strongly depends on the realized appreciation and the state of unemployment. Moreover, according to some authors, defaults can be explained by a combination of negative equity and affordability shocks, known as the "double-trigger hypothesis" (Bajari et al., 2008; Bhutta et al., 2010; Elul et al., 2010; Gerardi et al., 2013; Lydon and McCarthy, 2013; Gyourko and Tracy, 2014; McCarthy, 2014; Sirignano et al., 2018). Bhutta et al. (2010) found that defaults are influenced by a combination of negative equity and income or employment shocks.

The standard contingent claims approach to mortgage pricing suggests that default is primarily explained by the loan-to-value (LTV) ratio. Several authors have demonstrated the positive impact of the LTV ratio on the likelihood of negative home equity and mortgage default (Schwartz and Torous, 2003; Mayer et al., 2009; Campbell and Cocco, 2015). For example, Campbell and Cocco (2015) observed that the unconditional default probabilities significantly increase for LTV ratios exceeding ninety percent. However, in contrast to these findings, a few authors have not found a significant relationship between default/prepayment and the LTV ratio (Ambrose and Sanders, 2003).

The default rate of loans associated with income-generating properties, such as hotels, commercial properties, or traditional multifamily loans, is also influenced by the debt-servicing coverage ratio (DCR) (Ambrose and Sanders, 2003). The DCR is a measure that assesses the property's income relative to its debt obligations. Properties with higher DCR are considered more profitable, while those with low DCR indicate potential financial challenges. Generally, properties with an exceptionally low DCR face difficulties in repaying loans on time. Lenders prefer to provide loans to borrowers with higher DCR as it signifies a greater ability to meet debt obligations. The impact of DCR on the default rate depends on the level of property risk. Riskier property types, such as hotels or motels, may require a higher DCR to secure a loan compared to traditional multifamily or commercial properties (e.g., apartment buildings or shopping centers with anchor tenants).

Ambrose and Sanders (2003) found no significant influence of Loan-to-Value (LTV) and Debt Coverage Ratio (DCR) on the rates of prepayment and default. Their findings align with the explanation provided by Archer et al. (2001). According to Archer et al., LTV and DCR are factors that are determined by the loan origination process itself. Additionally, lenders take into consideration the risk of default when offering mortgage terms. For example, applicants with higher risk loans may be required to provide a larger down payment, resulting in a lower LTV ratio. Conversely, lower-risk applicants may be allowed a smaller down payment, leading to a higher LTV ratio. Similarly, higher-quality borrowers may have a lower Debt Coverage Ratio (DCR) requirement, while lower-quality borrowers may face a higher DCR requirement.

The loan-to-income (LTI) and mortgage-payments-to-income (MTI) ratios are also crucial factors in determining default probabilities. Campbell and Cocco (2015) explain that while the loan-to-value (LTV) ratio measures the initial equity stake of the household, the LTI and MTI ratios reflect the initial affordability of the mortgage. According to Campbell and Cocco (2015), the LTI ratio impacts default rates through a distinct mechanism. They argue that a higher initial LTI ratio does not directly increase the likelihood of negative equity; instead, it reduces mortgage affordability, making it more likely for borrowing constraints to become binding. This, in turn, lowers the threshold of negative home equity that triggers default, leading to an increase in default probabilities. Our model suggests that mortgage providers and regulators should consider the combined effects of LTV and LTI ratios and avoid controlling these parameters in isolation.

Apart these aforementioned factors, certain loan-specific characteristics also play a significant role in determining default rates. These include the age of the mortgage, borrowed amounts, origination year, and time until maturity. Some authors, such as Lawrence et al. (1992) and Schwartz and Torous (2003), have identified the impact of mortgage age on default rates. Smith et al. (2007) discovered that interest rates, loan amounts, and the proportion of mortgages also have a positive influence on defaults. Lawrence et al. (1992) conducted an analysis of mortgage loan defaults, taking into account borrower characteristics, mortgage durations, terms and conditions, economic factors, and default costs. While many of these variables have a significant impact on default behavior, mortgage leverage and repayment pressure emerge as the most influential factors.

Borrower-specific variables also play a crucial role in determining default events. In fact, defaults can be triggered by various stressful events experienced by borrowers, such as job layoffs, divorces, and so on (An and Qi, 2012).

Several economic and financial variables also serve as important determinants of default rates (Smith et al., 2007; LaCour-Little, 2008; Campbell and Cocco, 2015). An increase in interest rates has a positive impact on default rates for adjustable-rate mortgages (ARMs) due to the resulting higher required payments, while the opposite is true for fixed-rate mortgages

(FRMs) (Campbell and Cocco, 2015). Additionally, Sirignano et al. (2018) discovered that macroeconomic variables, particularly the unemployment rate, interact in complex ways with various other factors such as loan-to-value ratios, mortgage rates, and house price appreciation. They also found that the impact on prepayment and default rates is nonlinear.

2.3. Additional Information on Prepayment and Default Determinants

As presented in the previous section, main determinants of prepayment and default are:

The risk of prepayment is influenced by various refinancing factors, including financial incentives, the term structure of interest rates, the yield curve, and spreads. Additionally, prepayment is affected by variables associated with housing turnover. These variables encompass economic indicators such as GDP and the unemployment rate, housing market conditions indicated by the house price index (HPI), loan-specific characteristics such as loan age, loan type, loan amount, origination date, and property location, as well as borrower attributes like age, income, and credit score (FICO).

The default rate is primarily influenced by various factors, including house equity, loan-to-value (LTV) ratio, loan-to-income (LTI) ratio, debt-servicing coverage ratio (DCR), macroeconomic variables such as the unemployment rate, GDP, and policy rate, loan-specific variables such as loan age, loan type, time to maturity, origination date, and property location, as well as borrower characteristics like age, income, and credit score (FICO).

2.3.1 Lagged Values and Time-Varying Values

According to several authors (Goodarzi et al., 1998; Elie et al., 2002; Li et al., 2019), it has been observed that mortgagors tend to react to market conditions with a delay of several weeks. In light of this, these authors incorporated lagged values of explanatory variables in their analyses. For example, Goodarzi et al. (1998) considered lagged values ranging from 0 to 12 weeks for variables such as Burnout, Yield Curve Slope, and Present Value Ratio. Li et al. (2019) examined default and prepayment rates as competing risks using a multinomial logit model, and they included lagged values of explanatory variables (loan-specific, personal, and macroeconomic variables) from three periods (months) prior. Elie et al. (2002) stated that borrowers' reactions to changes in market rates occur between 4 to 7 months later. In comparison to these mentioned studies, most existing research on prepayment and default risks did not take into account the lagged values of explanatory variables.

In addition, the majority of existing studies on prepayment and default risks have focused on time-invariant covariates, using only the initial values of the variables. However, certain covariates, such as macroeconomic indicators, loan-to-value ratio (LTV), financing incentives, and credit score, are subject to changes over time. It is more appropriate to consider the current values of these variables, taking into account their time-varying nature. Only a few authors have incorporated time-varying variables into their analyses, such as (Visible Equity, Ambrose and Sanders (2003), Ding et al. (2012), and Sirignano et al. (2018).

2.3.2. Refinancing Incentive Modelling

Determining the refinance incentive is crucial in understanding both the prepayment and default rates. Various methods have been employed to calculate this incentive. The simplest approach involves calculating the refinance incentive as the difference between the loan rate and the market rate, as suggested by Ambrose and Sanders (2003) and Sirignano et al. (2018). Elie et al. (2002) observed that borrowers typically react to changes in market rates with a delay of 4 to 7 months. Taking this delayed reaction into account, they defined the refinance incentive as an average of the spread between the loan coupon rate and the market rate over this time period.

Other methods have also been considered. For example, some authors defined the refinance incentive as the Present Value Ratio (PVR) of the existing mortgage's payments compared to the annuity value of a new mortgage, as proposed by Richard and Roll (1989) and Goodarzi et al. (1998). Jacobs et al. (2005) computed the refinance incentive (RFI) as the net present value of the interest payments saved (until the next interest reset date) if the mortgage could be refinanced at the prevailing interest rate, relative to the size of the loan.

These methods are presented in the Appendix.

2.3.3. Initial and Current LTV and House Equity

Most of existing empirical studies have considered the initial loan-to-value (LTV) ratio, which is calculated as the initial loan value divided by the initial property value. However, it is worth noting that the realized LTV at the event time or a few periods before (referred to as the current LTV) may be a more significant explanatory variable for prepayment and default rates than the initial LTV. The current LTV can be determined by multiplying the initial property value (computed as the initial loan value divided by the initial LTV) and the monthly cumulative return of the house price index, as suggested by Ambrose and Sanders (2003) and Visible Equity. By dividing the current outstanding loan amount by the determined current property value, the current LTV can be calculated. Additionally, the computed current property value (obtained by multiplying the initial property value by the monthly cumulative return of the house price index) allows for the determination of the current house equity. The current house equity is defined as the difference between the property value and the outstanding loan balance.

3. METHODOLOGY

Prepayment risk and default risk have often been modelled separately using survival analysis techniques (Dunn and McConnell; 1981; Brennan and Schwartz, 1985; Green and Shoven, 1986; Deng et al., 2000; Jacobs et al., 2005; Consalvi and di Freca, 2010; Stanton and Wallace, 2011; Nijescu, 2012).

The analysis of prepayment and default as competing risks in existing studies has predominantly relied on hazard models or multinomial logistic models. Multinomial Logit (MNL) models have been commonly employed in the mortgage termination literature to model competing risk events (Clapp et al, 2000, 2001, 2006; Dunsky and Ho, 2007; VisibleEquity; Li et al. 2019). These models are based on a specific functional form and primarily consider the linear impact of covariates. However, the limitations of these models can be overcome by utilizing methods based on machine learning and deep learning.

Ishwaran et al. (2014) introduced the Competing Risks Forest method, which is based on the random forest approach and incorporates elements of the Random Survival Forest (RSF). Lee et al. (2018) developed the DeepHit model, which leverages deep learning techniques. Similarly, Sirignano et al. (2018) proposed a deep learning multi-state method to investigate the competing risks of prepayment and default. These methods directly estimate the Cumulative Incidence Function (CIF), enabling the modelling of non-linear effects and interactions among covariates, and they are free from assumptions imposed by traditional models. By not assuming any specific underlying stochastic process, these models allow for a flexible relationship, both linear and non-linear, between covariates and risks over time. The following sections provide a detailed presentation of these methods.

3.1. Classical Approaches

3.1.1 Survival Analysis

Prepayment risk and default risk have often been modelled separately using survival analysis techniques (Dunn and McConnell; 1981; Brennan and Schwartz, 1985; Green and Shoven, 1986; Deng et al., 2000; Jacobs et al., 2005; Consalvi and di Freca, 2010; Stanton and Wallace, 2011; Nijescu, 2012). Survival models are commonly employed to analyse the time elapsed before a specific random event takes place. In our context, the event occurrence time (T) represents either the prepayment time, the default time, or censoring. The survival function, in this case, denotes the probability that the event will not occur until time t: S(t) = P(T > t). The complement of the survival function is the cumulative distribution function, which describes the cumulative incidence of the event of interest up to time t: F(t) = 1 − S(t) = P(T ≤ t). The survival function is related to the hazard function h(t), which represents the instantaneous failure at time t, conditional on no occurrence of the event until time t. This hazard function, the cumulative hazard function (H_t) and the survival function (S(t)) are expressed as:

$$
\begin{aligned}
h(t) &= lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{S'(t)}{S(t)} = \frac{f(t)}{S(t)}, && (1)\\
H(t) &= \int_0^t h(u)du = -lnS(t), && (2)\\
S(t) &= exp(-H(t)), && (3)
\end{aligned}
$$

where f(t) represents the density function of the occurrence of the analyzed event.

Understanding the relationship between covariates and survival times (times-to-event) is crucial. However, in many existing studies, this relationship has been investigated by assuming a specific form for the underlying stochastic process, such as the Cox proportional hazards model (Cox, 1972) or the Accelerated Failure Time (AFT) model. These approaches rely on strong assumptions about the underlying stochastic process and the relationship between covariates and the parameters of that process. Additionally, these methods do not account for competing risks, which is an important consideration in many survival analysis scenarios.

When dealing with competing risks, it is important to consider the joint distribution of the event occurrence time and the competing events. Some researchers have utilized classical survival models to study the default and prepayment behavior of loans as competing risks. Proportional hazard models have been commonly employed in these studies (Deng et al., 2000; Ambrose and LaCour-Little, 2001; Pavlov, 2001; Ciochetti et al., 2002; Ambrose and Sanders, 2003; An and Qi, 2012). In the classical model, right-censoring is treated as uninformed or providing no information about the eventual terminal event of the subject.

However, it is important to note that the occurrence of other events leading to termination, besides the main event being analysed, is not uninformed. For example, when modeling default with the Cox Proportional Hazards model, it assumes that default is the terminal event and prepayments are treated as right censored. However, after the point of prepayment, prepaid loans will never default, and thus censoring at that point is not uninformed.

3.1.2. Cumulative Incidence Function (CIF)

To account for competing risks, two formulations of the Cumulative Incidence Function (CIF), representing the probability of failing from cause k before time t, were proposed: 1) CIF based on cause-specific hazard function and 2) CIF based on the sub-distribution.

Cause-specific CIF

The probability that an event occurs in a specific time period [0, t] depends on the cause-specific hazards of the other events (Gray, 1988). The probability of experiencing an event k by time t, determined using the cumulative incidence function (CIF: $F_k(t|x)$), is related to cause-specific hazard function (λ_k) as follows:

$$\begin{aligned} F_k(t|x) &= P[T \leq t, D = k|x], && (4) \\ &= \int_0^t S(s-|x)h_k(s|x)ds = \int_0^t exp[-\int_0^s \sum^K h_l(u|x)du]h_k(s|x)ds, && (5) \end{aligned}$$

where $S(t|x) = P[T \geq t|x]$ is the event-free survival probability function given covariates x. Event k can only occur for those surviving other risks. Regarding the cause-specific hazard function ($h_k(t|x) = f_k(t|x)/S(t|x)$), it describes the instantaneous risk of event k for subjects that currently are event-free. This hazard function is expressed as:

$$h_j(t|x) = lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, k = j|T \geq t)}{\Delta t} = \frac{f_j(t)}{S(t)}. \qquad (6)$$

The covariates (x), haven an impact on the CIF of event k, are those that change the cause-specific hazard function of event k as well as the cause-specific hazard functions of the competing risks (Ishwaran et al., 2014). Several authors determined the CIF by modelling the cause-specific hazards with the Cox' regression model (Lunn and McNeil, 1995; Cheng et al., 1998; Shen and Cheng, 1999; Scheike and Zhang, 2002, 2003). Other survival method can also be used to model the cause-specific hazard function.

CIF based on the sub-distribution - Fine and Gray (1999)

Fine and Gray (1999) proposed to model the Cumulative Incidence Function (CIF) as:

$$\begin{aligned} F_k(t) &= P(T \leq t, D = k), && (7) \\ F_k(t|X) &= 1 - exp(-\Lambda_k(t)exp(x^T\beta), && (8) \end{aligned}$$

where $\Lambda_k(t)$ is an unknown increasing function and β is a vector of coefficients to be estimated. Fine and Gray proposed to use an inverse of censoring weighting technique to estimate $\Lambda_k(t)$ and β.

Fine and Gray proposed a proportional hazards model for the subdistribution hazard function, specifically for the k_{th} type of event.

$$h_k^{st}(t) = lim_{\Delta t \to 0} \frac{Prob(t \leq T \leq t + \Delta t, D = k|T \geq t \cup (T \leq t \cap D \neq k))}{\Delta t}, \qquad (9)$$

where D represents the type of event that occurred. Each of the K different types of events has its own subdistribution hazard function, which is defined as the instantaneous rate of occurrence of the given type of event in subjects who have not yet experienced an event of that type. The risk set consists of those subjects who are either currently event-free or who have previously experienced a competing event.

This model is also based on strong assumptions related to the hazard rates and on the impact of covariates on model's parameters.

Scheike and Zhang (2011) introduced a class of flexible models for CIF. According to these authors, previously proposed model for CIF are special sub-models of their models (Fine and Gray, 1999; Shen and Cheng, 1999; Scheike and Zhang, 2002, 2003) . The general formulation of their models is given as:

$$h(F_k(t|X, z)) = X^T\alpha(t) + g(z, \lambda, t), \qquad (10)$$

where h and g are link functions and α(t) and λ are unknown regression coefficients.

Covariates z are assumed to have proportional effect on CIF whereas covariates X are allowed to change their effects on the cumulative incidence function over time. Any link function can be used. In their study, Scheike and Zhang (2011) retained two classes of flexible models: proportional models (P) and additive models (A); which are expressed as: [4]

$$\begin{aligned} cloglog[1 - F_k(t|X, z)] &= X^T\alpha(t) + z^T\lambda, \qquad (P) \qquad (11)\\ -log[1 - F_k(t|X, z)] &= X^T\alpha(t) + (z^T\lambda)t, \qquad (A). \qquad (12)\end{aligned}$$

3.1.3. Multi-State Models

Lando and Skodeberg (2002) proposed a Markov multi-state model of transitions between multiple states in continuous time.[5] This model captures movements between n states where the probability of moving away from the current state depends on the previous state. Survival analysis is a special case with two states, "alive" and "dead". Competing risks can also be considered as a special case, where there are multiple causes of termination (death). The transition intensities from the initial state X(t) at time t to the next state are defined as:

$$q_{rs}(t, X(t), \Im_t) = lim_{\delta t \to 0} P(X(t + \delta t) = s|X(t) = r, X(t), \Im_t)/\delta t, \qquad (13)$$

where r, s = 1, ...,K represent the states. The transition intensities can depend on covariates (X(t)), the time (t), and the "history" of the process up to that time ($\Im_t$).[6] [7] The KxK matrix Q is composed with the transition intensities q_{rs}. The rows of this matrix sum to zero, so that the diagonal entries are defined by $q_{rr} = -\sum_{r \neq s} q_{rs}$.

An individual may change states several times ($t_1,...,t_n$ are event times).

In semi-Markov (clock-reset) models, the transition intensity ($q_{rs}(t)$) from state r to s at time t depends only on the time t since entry into the current state, and the time since the beginning of the process is not taken into account. Whereas, in an inhomogeneous Markov (clock-forward) model, the time t represents the time since the beginning of the process, but the intensity $q_{rs}(t)$ does not depend further on $\Im_t$.

Multi-state models can be represented with of survival models. In the multi-state model, each transition intensity can have a corresponding survival model (time-to-event model), with hazard rates representing the transition intensities q_{rs}. Precisely, an individual who entered in state r at time t can transit to n_r competing risks ($s_1,...,s_{nr}$) at time t + δ_t. They are then n_r survival models related to the state r and $\sum_r n_r$ models over all states r. If the individual transits to state s_k at time (t+δ_t) then the transitions to all other states are censored at this time (t + δ_t).

Retained survival models can be parametric models as well as semi-parametric. Parametric models can include different probability distributions such as Weibull, Gompertz, gamma, log-logistic, lognormal, and generalized gamma. Parametric models are useful for extrapolating beyond the time horizon in the existing data (Jackson, 2018).

[4] Other link functions are also possible.

[5] Lando and Skodeberg (2002) pointed out significant advantages of this continuous multi-state model over the discrete time, "cohort method" approach.

[6] This history represents the states previously visited and the time spent in them.

[7] For time-homogeneous models, the transition intensities are determined as:

$$\begin{aligned} q_{rs}(t, X(t)) &= q_{rs},\\ q_{rs}(z(t)) &= q^0_{rs} exp(\sum_{k=1}^{K} \beta_{rs,k} x_k),\\ ln(q_{rs}(X(t))) &= ln(q^0_{rs}) + \sum_{k=1}^{K} \beta_{rs,k} x_k,\end{aligned}$$

where X_k, k = 1, 2, . . . , k is a set of exogenous variables affecting the instantaneous risk of going from state r to state s.

3.1.4. Multinomial Logit (MNL) Models

Multinomial Logit (MNL) models were used to model competing risk events in mortgage termination literature (Clapp et al, 2000, 2001, 2006; Dunsky and Ho, 2007; VisibleE-quity; Li et al. 2019). Multinomial logistic regression assumes mutual independence of choices for a given record during an observation period. For example, each periodic (monthly for example) observation is treated as though it were independent from the prior observation. Furthermore, multinomial logistic regression directly estimates the probabilities of each outcome which sum up to one. In this study, at each period t, the possible outcomes are active ($D_i = 0$), defaulted ($D_i = 1$) and prepaid ($D_i = 2$) for each loan. The probability of each outcome category conditioning on the covariates are given as:

$$P(D_i = 0|x_i) = \frac{1}{1 + \sum_{k=1}^{2} exp(\beta_k^T x_i)}, \qquad (14)$$

$$P(D_i = k|x_i) = \frac{exp(\beta_k^T x_i)}{1 + \sum_{l=1}^{2} exp(\beta_l^T x_i)} \qquad k = 1, 2, \qquad (15)$$

where x_i is a vector of observed independent variables including borrower characteristics w_i, loan features λ_i and macroeconomic factors z_i ($x_i = (w_i; \lambda_i; z_i)$). β represents the vector of parameters to estimate.

The MNL method consists of fitting the ratio of the expected proportion for each response category over the expected proportion of the reference category. The logit functions are defined as

$$ln(\frac{P(D_i = k|x_i)}{P(D_i = 0|x_i)}) = \beta_k^T x_i, \qquad k = 1, 2, \qquad (16)$$

The parameters are estimated by using maximum likelihood function defined as

$$lnL = \sum_i ln(P(D_i)|x_i)). \qquad (17)$$

The probabilities, defined in eq. 14 and 15, represent the probabilities of staying active, prepaying, or defaulting during a particular period given that the loan was active at the beginning of that period. The probability of these events after few periods can be deduced.[8] For example, the probability that the event k realizes for a certain loan in the future at time T knowing that it was active at time t (t < T) is simply the product of probabilities obtained from eq. 14 and 15, as

$$P_k(T, t) = P(D_i = k|x_i)(T)P(D_i = 0|x_i)(T-1)...P(D_i = 0|x_i)(t+1). \qquad (18)$$

Clapp, Deng, and An (2006) presented evidence that the multinomial logistic regression model is an attractive alternative to proportional hazard models in a case of mortgage termination, by either refinancing, moving or defaulting. However, compared to proportional hazard models' multinomial logistic regression enables to predict the occurrence of the events over the retained period rather than when they will take place.

3.2. Methods based on Machine Learning and Deep Learning

Several researchers have highlighted the non-linear effects of explanatory variables on prepayment and default rates (Elie et al., 2002; Dunsky et Ho, 2007; Elul et al., 2010; Agarwal et al., 2012; Sirignano et al., 2018). In particular, Sirignano et al. (2018)

discovered that many explanatory variables have a highly non-linear impact on borrower behavior, specifically on prepayment and default events. These authors emphasized that interactions between variables play a significant role in generating the non-linear effects. They also noted that the influence of a covariate on the dependent variable (prepayment, default, etc.) can be influenced by one or more other covariates.

Furthermore, the non-linear relationship may arise due to changes in the sensitivity of borrower behavior with respect to an explanatory variable, depending on the level of that variable (Sirignano et al., 2018). For example, Sirignano et al. (2018) observed that the sensitivity of borrowers to fluctuations in unemployment rates is not constant (linear), but varies significantly with the level of unemployment itself as well as the level of other variables.

There are methods available that allow for capturing the non-linear impact of covariates on dependent variables. In the Cox model, the nonparametric baseline hazard function can capture the non-linear influence of covariates on prepayment and

[8] Multinomial logistic regression assumes mutual independence of choices for a given record during an observation period. For example, each periodic (monthly for example) observation is treated as though it were independent from the prior observation.

default (Sirignano et al., 2018). Additionally, both the Cox and logistic regression models can incorporate non-linearity by including quadratic or other nonlinear transformations of specific variables (Agarwal et al., 2012; Sirignano et al., 2018). For example, Agarwal et al. (2012) introduced the squared loan age as a risk factor in addition to the loan age itself.

Another approach, used by Elul et al. (2010), involves discretizing continuous variables such as the loan-to-value ratio. However, this method requires identifying and analyzing the variables that need to be transformed, which can be time-consuming, especially in studies with a large number of explanatory variables.

The non-linear impact of covariates and the consideration of competing risks can also be addressed using methods based on machine learning and deep learning. There have been approaches proposed that combine survival analysis with machine learning techniques. For example, Ishwaran et al. (2008) introduced random survival forests for competing risks, while Sirignano et al. (2018) utilized a deep learning model to capture all non-linear effects, including variable interactions of any order, in the data. The deep learning model was employed to model different states of mortgages, including prepayment and default. Similarly, Lee et al. (2018) developed the DeepHit model specifically for investigating competing risks. These methods leverage the power of machine learning and deep learning to effectively handle non-linearity and competing risks in survival analysis.

3.2.1. Random Survival Forest - Random Competing Risks Forest

The Random Survival Forest (RSF), initially proposed by Ishwaran et al. (2008), is an ensemble method comprised of randomly grown survival trees. Each tree in the forest is built using an independent bootstrapped sample from the learning data, and random feature selection is applied at each node during tree growth. Similarly, a competing risk forest can be constructed using the same methodology as RSF, with the key difference lying in the splitting rules.

There are two main approaches to building a competing risk forest. The first approach involves growing separate competing risk trees for each event in each bootstrap sample, utilizing event-specific splitting rules to guide the tree growth. The second approach entails growing a single tree in each bootstrap sample, where the splitting rules can either be event-specific or a combination of event-specific rules across the events. The latter approach is more commonly used as it is more efficient and generally sufficient for most tasks (Ishwaran et al., 2014; Hamidi et al., 2017). Precisely, a competing risks forest is grown as follows:

. B bootstrap samples from the original data are drawn. Each competing risk tree is grown by using 63% of the data and the remaining 37% of data (out-of-bag OOB) are used to determine the out-of-bag cross-validated survival as well as variable importance (VIMP) and minimum depth measures for each independent variable.

. Using each bootstrap sample, a competing risk tree is grown based on randomly selected M ≤ p covariates at each node of the tree. The retained variable for splitting each node is the one that maximizes a splitting rule. This rule can be based on generalized log-rank test, Gray's test, or composite splitting rule. The generalized log-rank test is most suitable for selecting variables that affect cause specific hazards whereas Gray's test is most suitable for identifying variables directly affecting cumulative incidence function. These splitting rules and tests are presented in detail in Ishwaran et al. (2014).

. Grow each tree to full size and calculate the cumulative cause-specific hazards for all events (k) for each tree (b).

. Take the average of each estimator over the B trees to obtain its ensemble.

The CIF of the b_{th} tree is defined as:

$$\hat{F}_{k,b}(t|x) = \int_0^t \hat{S}_b(u-|x) Y_b(u|x)^{-1} N_{k,b}(du|x), \tag{19}$$

$$with,$$

$$\hat{S}_b(t|x) = \prod_{u \le t}(1 - \frac{\sum_k N_{k,b}(du|x)}{Y_b(u|x)}), \tag{20}$$

where $\hat{S}_b(t|x)$ is the Kaplan-Meier estimate of event-free survival. t_j represents ordered event times. $d_k(t_j)$ and $N_k(t)$ correspond to the number of type k events at t_j and in the interval [0, t], respectively. $d(t_j)$ and N(t) represent the number of all events at t_j and in the interval [0, t], respectively. Y(t) is the number of individuals at risk (event-free and uncensored) just prior to t.

The ensemble estimates of the CIF and the cause-k mortality are given as:

$$\bar{F}_k(t|x) = \frac{1}{B}\sum_{b=1}^{B} \bar{F}_{k,b}(t|x), \qquad \bar{M}_k(\tau|x) = \int_0^\tau \bar{F}_k(t|x)dt := \frac{1}{B}\sum_{b=1}^{B} \bar{M}_{k,b}(\tau|x).$$

To determine and compare the performance of models, the ensemble estimates of the CIF and the cause-k mortality are also determined by using out-of-bag (OOB) data as:

$$\bar{F}_k^{OOB}(t|x_i) = \frac{1}{|O_i|}\sum_{b\in O_i} \bar{F}_{k,b}(t|x_i), \qquad \bar{M}_k^{OOB}(\tau|x_i) = \int_0^{\tau} \bar{F}_k^{OOB}(t|x_i)dt := \frac{1}{|O_i|}\sum_{b\in O_i} \bar{M}_{k,b}(\tau|x_i),$$

where $O_i \subset 1, ...,B$ is the index set of trees. The OOB predicted value for a case is a cross-validation based estimator. It can be used for estimation of the prediction error.

The prediction performance can be measured with the concordance index (C–index); which is related to the area under the receiver operating characteristic curve (Ishwaran et al., 2014). The classical concordance index represents the probability that, in a randomly selected pair of cases, the case failing first had a worse predicted outcome. In the case of competing risk forest, the ensemble prediction of the CIF is concordant with the outcome if either the case with the higher cause-k mortality has event k before the other case haven an event of cause k or a competing event. Case (individual) i has a higher risk of event k then case i' if $\bar{M}_k(\tau|x_i) > \bar{M}_k(\tau|x_{i'})$. The time-truncated concordance index for competing risks, proposed by Wolbers et al. (2013), is given as:

$$C_k(\tau) = P[\bar{M}_k(\tau|x_i) > \bar{M}_k(\tau|x_{i'})|T_i^0 \leq \tau, \delta_i^0 = j \quad and \quad (T_i^0 < T_{i'}^0, \quad or, \delta_{i'}^0 \neq k)]. \quad (21)$$

Regarding the prediction error, it can be measured with the integrated Brier score (BS), which is the squared difference between actual and predicted outcome (Ishwaran et al., 2014). A static and time dependent BS were proposed (Graf et al., 1999; Gerds and Schumacher, 2006). To assess the performance of the ensemble CIF, the integral of the time-dependent BS is given as:

$$IBS_k(\tau) = \int_0^{\tau} BS_k(t)dt = \int_0^{\tau} E[I(T_i^0 \leq t, D_i = k) - \bar{F}_k(t|X)]^2 dt. \qquad (22)$$

The importance of covariates is determined with the variable importance (VIMP) and minimum depth measures. The VIMP measures the increase (or decrease) in prediction error for the forest ensemble when a covariate is randomly "noised-up" (Breiman, 2001). The larger VIMP of a covariate is the higher predictiveness of the variables is. The value of VIMP that is greater than 0.002 is assumed as an effective variable whereas the smaller values of the minimal depth reveal higher predictiveness of the variables.

Minimal depth assesses the predictiveness of a variable by the depth of the first split of a variable relative to the root node of a tree (Ishwaran et al. 2010).

All these measures (concordance index, prediction error, the variable importance (VIMP) and minimum depth and the minimal depth) are determined in-sample as well as out-sample (Out-of-Bag - OOB).

3.2.2. Deep Learning

DeepHit model

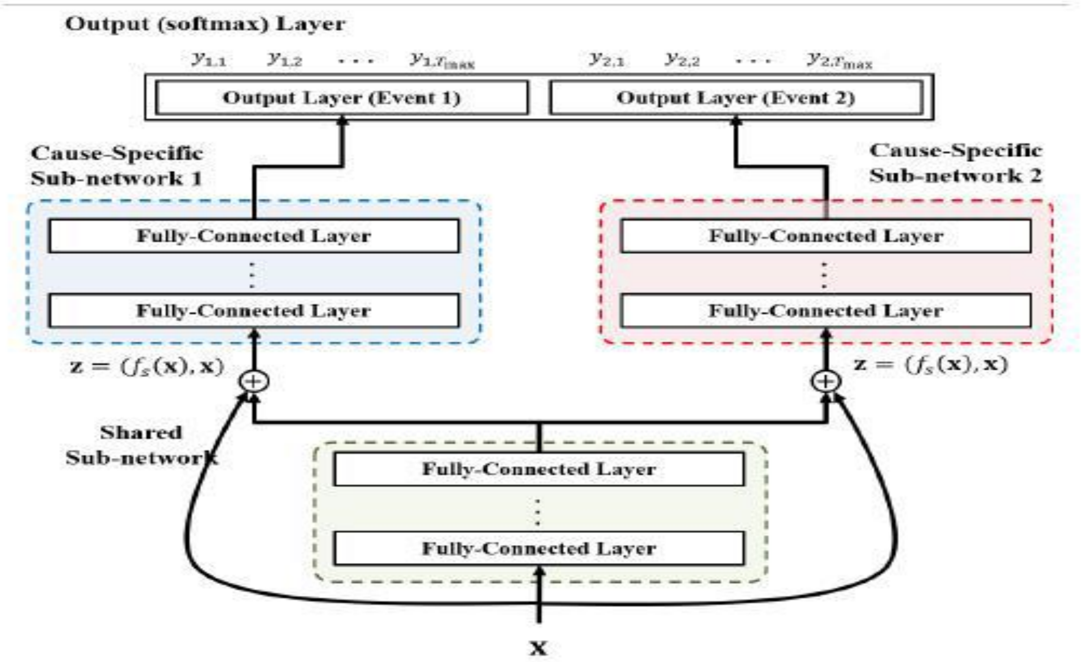
Faraggi and Simon (1995) were the pioneers in applying neural networks to the Cox Proportional Hazard model, which is commonly used in survival analysis. Subsequent authors, such as Katzman et al. (2016) and Luck et al. (2017), expanded on this work by introducing more advanced network architectures and loss functions. These authors relaxed the specific functional relationship between covariates and the hazard function while maintaining other key assumptions. However, their models still assumed a constant hazard rate over time, did not capture the time-dependent influence of covariates on survival, and did not account for competing risks.

In comparison, Lee et al. (2018) developed the DeepHit model, which utilizes a deep neural network to directly learn the distribution of survival times without making assumptions about the underlying stochastic process. Unlike the previous models, the DeepHit model acknowledges that the relationship between covariates and risks may change over time. Additionally, the DeepHit model is capable of handling competing risks, providing a more comprehensive framework for survival analysis.

The network architecture of this DeepHit model consists of a single shared subnetwork and a family of cause-specific sub-networks. In our case, there are two cause-specific sub-networks corresponding to prepayment and to default. The shared sub-network and the k−th cause-specific sub-network (k = 1, ...,K, K = 2 in our case) are formed with L_S ans $L_{c,k}$ fully-connected layers, respectively. The graph 1 represents a DeepHit model with 2 competing risks (K = 2).

By using the covariates x as inputs, the shared sub-network produces a vector of output fs(x) that is common to retained competing events. This vector and the original covariates (X) are used by each cause-specific sub-network as inputs (z = (f_s(x), x)). Each cause-specific sub-network produces a vector f_{ck} (z) as output, representing the probability of the first hitting time of a specific cause k. Precisely, the cause-specific sub-networks are learning the distribution for the first hitting time for each cause in parallel. All outputs represent the joint probability distribution on the first hitting time and event. A single softmax layer is used as the output layer of DeepHit in order to ensure that the network learns the joint distribution of K competing events not the marginal distributions of each event. This output layer produces a probability distribution y = [$y_{1,1}$, ..., $y_{1,T}$, $y_{2,1}$, ..., $y_{2,T}$] representing the probability (P(s, k|x)) that a borrower characterized with features (x) will experience the event k at time s.

Figure 1: Architecture of DeepHit in case of 2 Competing Risk (Lee et al. (2018))


Output (softmax) Layer
$y_{1,1}$ $y_{1,2}$ ... $y_{1,T_{\max}}$
$y_{2,1}$ $y_{2,2}$ ... $y_{2,T_{\max}}$
Output Layer (Event 1)
Output Layer (Event 2)
Cause-Specific Sub-network 1
Cause-Specific Sub-network 2
Fully-Connected Layer
Fully-Connected Layer
Fully-Connected Layer
Fully-Connected Layer
$\mathbf{z} = (f_s(\mathbf{x}), \mathbf{x})$
$\mathbf{z} = (f_s(\mathbf{x}), \mathbf{x})$
Shared Sub-network
Fully-Connected Layer
Fully-Connected Layer
x

The network is trained by using a loss function based on both survival times and relative risks. The loss function is defined as the sum of two terms: L_{Total} = L_1 +L_2, where L_1 is the log-likelihood of the joint distribution of the first hitting time and corresponding event, modified to take account of the right-censoring. As for L_2, based on estimated CIFs calculated at different times, it enables to fine-tune the network to each cause-specific estimated CIF. As Fine and Gray (1999), Lee et al. (2018) defined the CIF function as the probability that a particular event k^* occurs on or before time t^* conditional on covariates x^*. As the true CIF is not known, Lee et al. (2018) used the estimated CIF. For the event k^* the estimate CIF is defined as: $\hat{F}_k^*(t^*|x^*) = \sum_{m=0}^{s^*} y_{k,m}^*$.

Lee et al. (2018) compared the performance of DeepHit model with the performance of other retained model by using the time-dependent concordance index. Compared to the classical concordance index, the time-dependent concordance index enables to take into account the possible change in risk into account.

Dynamic DeepHit Model

The initial DeepHit model allows for predicting competing risks by utilizing the most recent available measurements of retained covariates. In contrast, the Dynamic-DeepHit model goes a step further by incorporating longitudinal data that consists of infrequently repeated measurements. This updated model can dynamically update survival predictions for one or multiple competing risks. Importantly, the Dynamic-DeepHit model learns the distributions of time-to-event without assuming any specific underlying stochastic models for the longitudinal and time-to-event processes.

The Dynamic-DeepHit model is constructed as a multi-task network, comprising of a shared subnetwork and a family of cause-specific subnetworks. The shared subnetwork consists of two components: i) an RNN structure that handles the longitudinal data, and ii) an attention mechanism that identifies the importance of the measurement history in making risk predictions. This shared subnetwork predicts the next measurements of time-varying covariates. Using the context vector (the output of the shared subnetwork model) and the last measurements as input, the cause-specific subnetworks estimate the joint distribution of the first hitting time and competing events. Each cause-specific subnetwork employs a feed-forward network with fully connected layers to capture the relationship between the cause-specific risk and the measurements (context vector and last measurements). The output layer utilizes a softmax layer to generate an estimated joint distribution of the first hitting time and competing events.

Training the Dynamic-DeepHit model involves minimizing a total loss function that incorporates longitudinal measurements and accounts for right-censoring. The model utilizes both time-invariant and time-varying covariates, treating the survival time as discrete. Discretization is performed by transforming continuous-valued times into a set of contiguous time intervals. In the study by Lee et al. (2019), who introduced the Dynamic-DeepHit model, monthly data was retained, and the time was discretized with a resolution of one month.

Lee et al. (2019) developed a cause-specific time-dependent concordance index ($C_k(t;\Delta t)$) to assess the discriminative performance of various methods. This index is an extension of the time-dependent concordance index, proposed by Gerds et al. (2013. Lee et al, (2019) adapted this time-dependent concordance to the competing risks setting with longitudinal measurements. The ($C_k(t;\Delta t)$) index uses prediction and evaluation times to reflect possible changes in risk over time. The ($C_k(t;\Delta t)$) for event k is defined as;

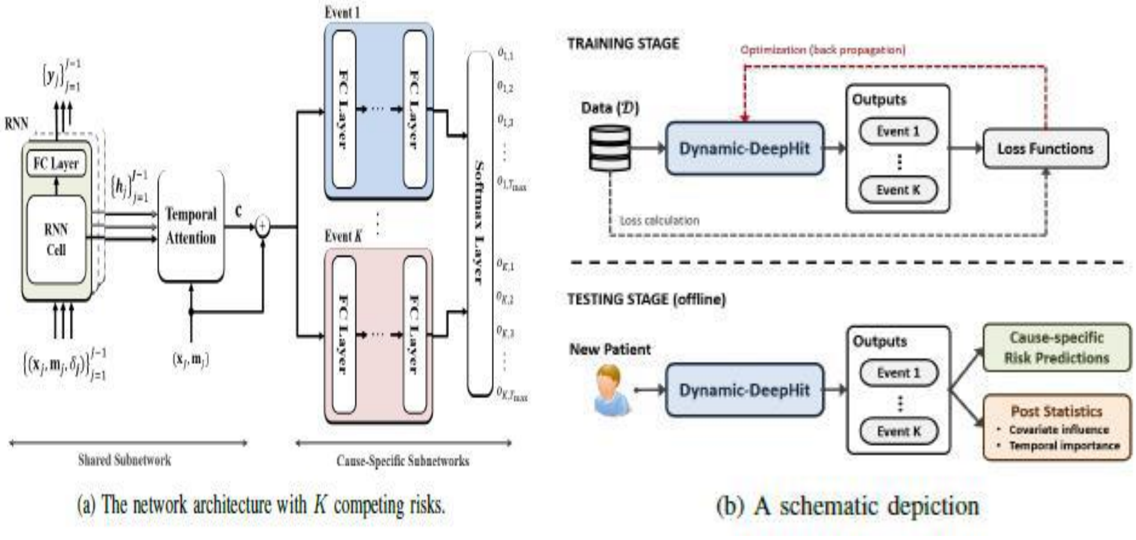
$$C_k(t;\Delta t) = P(\hat{F}_k(t+\Delta t|\chi^i(t)) > \hat{F}_k(t+\Delta t|\chi^j(t))|\tau^i < \tau^j, k^i = k, \tau^i < t+\Delta t), \quad (23)$$

where $\hat{F}_k^{\square}(.)$ is the estimated CIF for cause k. t and Δt represent the prediction time and the evaluation time (time elapsed since the prediction is made), respectively. $\hat{F}_k^{\square}(t + \Delta t|\chi^i(t))$ represents the predicted risk of event k occurring in Δt years by using the longitudinal measurements until t.

3.2.3. Deep Learning and Multi-State Model

Sirignano et al. (2018) investigated the transition of mortgages between 7 states (current, 30 days delinquent, 60 days delinquent, 90+ days delinquent, foreclosed, REO, and paid off) over its lifetime. A mortgage will transit between retained stated over its lifetime. These authors proposed to model the dynamic of the state process with a deep learning model; which is a nonlinear extension of the familiar logistic regression model. They used the empirical frequency of the different types of transitions between states and assumed that the dynamics of the state process be influenced by a vector of explanatory variables $X^n_t \in R^{dX}$. They modeled the marginal conditional probability for the transition of the n-th mortgage from its state U^n_{t-1} at time t − 1 to state u at time t given the explanatory variables X^n_{t-1} as:

Figure 2: Architecture of Dynamic DeepHit in case of 2 Competing Risk (Lee et al. (2019))



(a) The network architecture with K competing risks.

(b) A schematic depiction

$$h_{\theta,l}(x) = g_l(W_l^T h_{\theta,l-1}(x) + b_l), \quad (24)$$

where $W \in R^K x R^{dK}$ and $b \in R^K$. l = 1, ...,L−1 where L−1 is the number of layer. The nonlinear transformation $g_l(z)$ is given by the softmax function defined as follows:

$$g(z) = (\frac{e^{z_1}}{\sum_{k=1}^{K} e^{z_k}}, ..., \frac{e^{z_K}}{\sum_{k=1}^{K} e^{z_k}}), \qquad z = (z_1, ..., z_K) \in R^K. \tag{25}$$

The hidden nodes enable the nonlinear transformations of input variables (explanatory variables) and connect them to the output nodes (the conditional probabilities of the different mortgage states). The hidden nodes represent the nonlinear transformations of input variables.

At each layer l, the output $h_{\theta,l}(x)$ is a simple nonlinear link function g_l of a linear combination of the nonlinear basis functions $h_{\theta,l-1}(x)$, where the nonlinear basis function $h_{\theta,l-1}(x)$ must be learned from data via the parameter θ. The output $h_{\theta,l}(x)$ from the l-th layer of the neural network becomes the basis function for the (l + 1)-th layer.

Sirignano et al. (2018) investigated the explanatory power of explanatory variables by considering the behavior of the out-of-sample negative average loglikelihood $(\frac{1}{N} L_{T,N}(\hat{\theta}))$ with respect to changes of the set explanatory variables. This negative average loglikelihood is a standard measure of fit, called the cross-entropy error or simply the loss. The explanatory power of variables is measured by how much the loss increases when the variable is removed as an explanatory variable. A large increase in the loss means large explanatory power of the variable.

Regarding the economic significance of a particular variable on borrower behavior, it is measured by the magnitude of the derivative of a fitted transition probability with respect to the variable (averaged over the data). The sensitivity of the fitted probability with respect to j-th variable for a transition from state u to v is determined as:

$$E[|\frac{\partial}{\partial x_j} h_{\hat{\theta}}(V, X)|V = v, U = u].| \tag{26}$$

A sensitivity of value a for a given variable means that the probability for a transition from state u to v will approximately change by zΔ if that variable is changed by a small amount Δ.

4. CONCLUSION

Effectively modelling prepayment and default risks is very important for assessing risk, valuing assets, making informed decisions, and ensuring the stability and profitability of financial institutions and markets. In the specific context of mortgage loans, the exercise of the prepayment option brings an end to the default option, and vice versa: competing risks. These competing risks are influenced by various factors. The impacts of these factors on these risks may not be simple, but complex.

Classical methods, such as classical survival models, model each risk separately without considering them as competing risks. Classical models like multi-state models and multinomial logit models have traditionally been used to model these competing risks. However, these models do not take into account possible complex responses of prepayment and default risk to their determinants. These complex influences of independent variables on considered risks are taken into account by recently proposed methods/models, such as the Random Survival Forest and Random Competing Risks Forests, as well as the DeepHit model and Dynamic DeepHit model.

REFERENCES

S. Agarwal, Y. Chang, and A. Yavaz (2012). Adverse selection in mortgage securitization. Journal of Financial Economics, 105, 640–660.

V. Agarwal and R. Taffler (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. Journal of Banking and Finance, 32(8), 1541–1551.

B. Ambrose and C.A. Capone (2000). The hazard rates of first and second default. Journal of Real Estate Finance and Economics, 20(3), 275-293.

B. Ambrose and A.B. Sanders (2003). Commercial mortgage-backed securities: prepayment and default. Journal of Real Estate Finance and Economics, 26(2-3), 179–196.

M.Y. An and Z. Qi (2012). Competing Risks models using mortgage duration data under proportional hazards assumption. Journal of Real Estate Research 34(1). 15-29.

P. Bajari, S. Chu, and M. Park (1997). An Empirical Model of Subprime Mortgage Default from 2000 to 2007. Working Paper (2008).

N. Bhutta, J. Dokko, and H. Shan (2010). The Depth of Negative Equity and Mortgage Default Decisions. Finance and Economics Discussion Series, Board of Governors of the Federal Reserve System 2010-35.

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

M. Brennan and E.S. Schwartz, (1985). Determinants of GNMA mortgage prices. AREUEA Journal, 13, 291-302.

JY Campbell and JF. Cocco (2015). A model of mortgage default. Journal of Finance 70(4), 1495–1554.

T.S. Campbell and J.K. Dietrich (1983). The determinants of default on insured conventional residential mortgage loans. Journal of Finance 38(5), 1569–1581.

R.-R. Chen (1996). Understanding and managing ınterest rate risks. World Scientific- Singapore.

S.C. Cheng, J.P. Fine, and L.J. Wei (1998). Prediction of cumulative ıncidence function under the proportional hazards model. Biometrics, 54, 219–228.

B.A. Ciochetti et al. (2002). The termination of mortgage contracts through prepayment and default in the commercial mortgage markets: a proportional hazard approach with competing risks. Real Estate Economics, 30(4), 304-321.

J.M. Clapp, Y. Deng, and X. An. (2006). Unobserved heterogeneity in models of competing mortgage termination risks. Real Estate Economics 34(2), 243-273.

J.M. Clapp, J.P. Harding, and M. LaCour-Little (2000). Expected mobility: Part of the prepayment puzzle. The Journal of Fixed Income, 10(1), 68–78.

J.M. Clapp et al. (2001). Movers and shuckers: ınterdependent prepayment decisions. Real Estate Economics, 29(3), 411–450.

M. Consalvi and G.S. di Freca (2010). Measuring prepayment risk: an application to UniCredit Family Financing. Tech. Rep. Working Paper Series, UniCredit Universities.

Y. Deng, J.M. Quigley, and R. van Order (2000). Mortgage terminations, heterogeneity and the exercise of mortgage options. Econometrica, 68(2), 275-307.

K. Dunn and J. McConnell (1981). Valuation of GNMA mortgage-backed securities. Journal of Finance, 36, 599–616.

R.M. Dunsky and T.S.Y. Ho (2007). Valuing fixed rate mortgage loans with default and prepayment options. Journal of Fixed Income, 16(4), 7–31.

R. Elie et al (2002). A Model of Prepayment for the French Residential Loan Market. Working Paper Groupe de Recherche Operationnelle, Credit Lyonnais http://gro.creditlyonnais.fr.

R. Elul et al. (2010). What ’triggers’ mortgage default? The American Economic Review, Papers and Proceedings 200(2), 490–494.

D. Faraggi and R. Simon (1995). A neural network model for survival data. Statistics in Medicine, 14, 73–82.

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

T.A. Gerds et al (2013). Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. In: Statistics in Medicine, 32(13), 2173–2184.

L.S Goodman et al. (2010). Negative equity trumps unemployment in predicting defaults. Journal of Fixed Income, 19(4), 67–72.

E. Graf et al. (1999). Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine, 18(17-18), 2529–2545.

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

J. Green and J. Shoven (1986). The effects of interest rates on mortgage prepayments. Journal of Money, Credit and Banking, 18 (1986), 41–59.

L. Guiso, P. Sapienza, and L. Zingales (2013). The determinants of attitudes toward strategic default on mortgages. Journal of Finance, 68(4), 1473– 1515.

J. Gyourko and J. Tracy (2014). Reconciling theory and empirics on the role of unemployment in mortgage default. Journal of Urban Economics, 80, 87–96.

O. Hamidi et al (2017). Application of random survival forest for competing risks in prediction of cumulative incidence function for progression to AIDS. Epidemiology Biostatistics and Public Health 14(4), 323-241.

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

J.V. Hoff (1996). Adjustable and fixed rate mortgage termination, option values and local market conditions: An empirical analysis. Real Estate Economics, 24(3), 379–406.

H. Ishwaran et al. (2014). Random survival forests for competing risks. Biostatistics, 15(4). 757–773.

H. Ishwaran et al. (2010). Random survival forests for high-dimensional data. Statistical Analysis and Data Mining, 4, 115–132.

J.B. Kau et al. (1993). Option theory and floating-rate securities with a comparison of adjustable and fixed-rate mortgages. Journal of Business, 66(4), 595–618.

M. LaCour-Little (2008). Mortgage termination risk: a review of the recent literature. Journal of Real Estate Literature, 16(3), 295–326.

D. Lando and T.M. Skodeberg (2002). Analyzing rating transitions and rating drift with continuous observations. Journal of Banking and Finance, 26(2-3), 423–444.

E.C. Lawrence, L.D. Smith, and M. Rhoades (1992). An analysis of default risk in mobile home credit. Journal of Banking and Finance 16(2), 299–312.

C. Lee, J. Yoon, and M. van der Schaar (2019). Dynamic-deephit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. IEEE Ttransactions on Biomedical Engineering 20(5), 1–12.

C. Lee et al. (2018). Deephit: A deep learning approach to survival analysis with competing risk. Procedia, 32th AAAI Conf. Artif. Intell, 2314–2321.

A. Levin and A. Davidson (2005). Prepayment and option adjusted valuation of MBS. The Journal of Portfolio Management 31(4), 286-299.

Z. Li et al. (2019). Predicting prepayment and default risks of unsecured consumer loans in online lending. Emerging Markets Finance and Trade 55(1), 118–132.

F. Longstaff (2002). Optimal recursive refinancing and the valuation of mortgage-backed securities. Working paper, University of California, Los Angeles.

M. Lunn and D. McNeil. (1995). Applying Cox regression to competing risks. Biometrics, 51, 524–532.

R. Lydon and Y. McCarthy (2013). What lies beneath? Understanding recent trends in Irish mortgage arrears. The Economic and Social Review, Economic and Social Studies, 44(1), 117–150.

C. Mayer, K. Pence, and S. Sherlund (2009). The rise in mortgage defaults. Journal of Economic Perspectives, 23(1), 27–50.

F. McCann (2014). Modelling default transitions in the UK mortgage market. Central Bank of Ireland- Research Technical Paper 17/RT/14.

D.C. Nijescu (2012). Prepayment risk, impact on credit products. Theoretical and Applied Economics,19(8), 53–62.

A.D. Pavlov (2001). Competing risks of mortgage termination: Who refinances, who moves, and who defaults? The Journal of Real Estate Finance and Economics, 23(2), 185–211.

S. Richard and R. Roll (1989). Prepayments on fixed-rate mortgage-backed securities. Journal of Portfolio Management, 15, 73–82.

T.H. Scheike and M-J. Zhang (2011). Analyzing competing risk data using the R timereg package. Journal of Statistical Software 38(2), 1–15.

T.H. Scheike and M.J. Zhang (2002). An additive-multiplicative Cox-Aalen model. Scandinavian Journal of Statistics in Medicine, 28, 75–88.

T.H. Scheike and M.J. Zhang (2003). Extensions and applications of the Cox-Aalen survival models. Biometrics 59, 1033–1045.

E.S. Schwartz and W.N. Torous (2003). Commercial office space: tests of a real options model with competitive ınteractions. Working Paper, University of California, Los Angeles.

Y. Shen and S.C. Cheng (1999). Confidence bands for cumulative ıncidence curves under the additive risk model. Biometrika, 55, 1093–1100.

L.D. Smith, E.C. Lawrence, and S.M. Sanchez (2007). A comprehensive model for managing credit risk on home mortgage portfolios. Decision Sciences, 27(2), 291–317.

R. Stanton and N. Wallace (2011). The bear lair: Index credit default swaps and the subprime mortgage crisis. Review of Financial Studies 24(10), 3250–3280.

K.D. Vandell (1978). Default risk under alternativemortgage ınstruments. The Journal of Finance, 33(5), 1279–1296.

K.D. Vandell (1995). How ruthless is mortgage default? A review and synthesis of the evidence” Journal of Housing Research, 6(2), 245–264.

M. Wolbers, M.T. Koller, and J.C. Witteman (2013). Concordance for prognostic models with competing risks. Research Report, University of Copenhagen, Department of Biostatistics 3.

Appendix 1: Refinance Incentive

Ambrose and Sanders (2003) formalized the refinance incentive as follows:

$$PPOPTION(t) = \frac{r_c(t) - r_G(t)}{r_G(t)}$$

where r_c and r_G denote the contract interest rate at time t and the current 10-year Treasury rate at time t, respectively. The choice of this latter rate is due to the fact that new contracts are indexed to the 10-year Treasury rate.

Goodarzi et al., 1998): The PVR is expressed as:

$$RFI_{it} = \frac{\sum_{j=0}^{n} \frac{(cr_i - mr_{n,t})PP_i}{(1+d)^j}}{PP_i} 100,$$

where I and R denote the note rate monthly and the current mortgage refinance rates on a monthly basis, respectively. M represents the remaining life of the loan in months.

Jacobs et al. (2005) determined the refinance incentive as follows:

$$RFI_{it} = \frac{\sum_{j=0}^{n} \frac{(cr_i - mr_{n,t})PP_i}{(1+d)^j}}{PP_i} 100,$$

where "n" represents the number of months remaining until the next interest reset date, which is specific to the contract and changes over time. cr_i represents the contract interest rate, while $mr_{n;t}$ represents the market rate at time "t" for a loan with a duration of "n" months. Additionally, "d" represents the discount rate, assumed to be 3% annually. The loan amount, denoted as PPi" appears in both the numerator and denominator of the equation but is ultimately irrelevant to our measure of refinance incentive.

Elie et al. (2002) considered the average of the spread between the loan coupon rate and the market rate (δ(t)) from month t − 4 to month t − 7 in prepayment modelling. As Bennett et al. (1997), Elie et al. (2002) modelled the refinancing incentive as the exponential of the piecewise linear function $l(\bar{\delta})$, where l() is defined as:

$$\begin{aligned} l(\bar{\delta}(t)) &= exp[\alpha_1(\bar{\delta}(t) - u_1)^+ + \alpha_2(\bar{\delta}(t) - u_2)^+], \\ \bar{\delta}(t) &= \frac{1}{4}[\delta(t-4) + \delta(t-5)\delta(t-6)\delta(t-7), \end{aligned}$$

where the market interest rate is represented by the 10-year swap rate at time t.