Erasmus University
Erasmus School of Economics
Master Econometrics & Management Science
Specialisation: Quantitative Finance

Mixture Cure Model For Credit Scoring Of Mortgage Loans

Author:
Gertrude Yaka Tine - 543774

Supervisor:
Dr. Mikhail Zhelonkin

Second Assessor:
Prof Chen Zhou

26th November 2020

Abstract

Contrary to most survival analysis models used for credit risk scoring, the Logistic-CoxPH Mixture Cure accounts for the existence of a proportion of borrowers that does not experience default. It can yield higher performance in predicting the probability of default if used properly. It is known from the medical science area that the model performs best on large samples with long observation period and few censoring. However, in the literature of credit scoring its advantage over standard models have been studied with data specifications that do not fully match these requirements. In this paper, the benefits of using the Logistic-CoxPH Mixture Cure instead of the Logistic Regression or the Cox Proportional Hazards for predicting the probability of default of mortgage loans are investigated when accounting for the competing risks of default and prepayment. The data set is a large sample from the Fannie Mae Mortgage Loans Data and allows for a long observation period of 10 years. The discrimination and calibration performances of the different models are compared. The study reveals that the survival models are better suited than the Logistic Regression to identify defaulters on intermediate time intervals. Similarly, the calibration test favors the survival models. However, it appears that the Cox PH is preferable to the Mixture Cure in terms of both discrimination and calibration performances.

Contents

List of Figures

List of Tables

1 Introduction

This paper investigates evidences to recommend a Logistic-Cox Proportional Hazards (Logistic-CoxPH) Mixture Cure model over the Cox Proportional Hazards(Cox PH) model and the standard Logistic Regression(LR) model for mortgage loan credit scoring in presence of early repayments. In the two survival analysis models, the study considers prepayments as a competing risk to the event of default. The three models are compared through measures of their discrimination and calibration performance. The research provides two main results. First, the Mixture Cure model and the Cox PH better differentiate between defaulters and non-defaulters than the Logistic Regression and show more evidences of good fit. In fact, all the cross-validated discrimination measures reveal higher performance for the two survival models. The cross-validated calibration test shows a poor fit of the Logistic Regression for all observation years with p-values equal to zero. On the contrary, for Cox PH and Mixture Cure, the null hypothesis of poor fit is rejected at both 1% and 5% significance levels in most validation samples during the first three years. Thus, the study favors the survival models over the Logistic Regression for timely prediction of the yearly Probability of Default(PD) of mortgage loans in presence of prepayments at least for a three years ahead prediction. The second result outlines that accounting for a fraction of borrowers that do not default does not always yield higher predictive performances even when early repayments are taken into account and longer observation time is considered. In fact, in terms of discriminatory performance, the Mixture Cure outperforms the Cox PH only in three years out of ten. And, in terms of calibration performance it is outperformed by the Cox PH for which the null hypothesis of poor fit is rejected for all validation samples in the first three years. Thus, based on these results, the standard Cox PH is in most cases preferable to the Mixture Cure for timely prediction of the PD of mortgage loans.

Banks occupy a huge part of the financial sector which counts for a sensible proportion of the economy. They mainly make profits by granting loans to their clients using available deposits. This activity carries a risk that the borrower fails to fulfill its engagement of repaying in time called credit risk. A good management of such risk is indispensable to ensure good performance for banks (Jiang et al., 2019) and stability for the whole financial system. For that purpose, credit risk scoring models have been used in order to distinguish which loans applicants are less likely to default and creditworthy. The Logistic Regression is the standard PD model used in the bank industry. For mortgage loans, the longer terms require a good timely risk management. Therefore survival analysis models are suitable as they provide a dynamic estimation of probability of default contrary to binary models such as Logistic Regression. A dynamic PD provides lenders with a timely appreciation of the repayment ability of a borrower and allows to take informed decisions about a loan application. The Basel 2 accord already gives an important place to long-run probability of default and encourages the use of credit risk assessment methods that consider

evaluating the distribution of probability over time (Wycinka & Jurkiewicz, 2017).

On the other hand, PD models with high predictive performances are also important for the mortgage market given its substantial size. In the USA, it clocks in at $10.5 trillion in 2019. Therefore, due to the pertaining credit risk, even an improvement of 1% can make a big difference because it can yield a great decrease of losses (Hand & Henley, 1997). However, the quest for higher predictive power usually results in sacrificing explanatory understanding of the effect of features on default probability (Jiang et al., 2019). It is the case for machine learning techniques that have been used in the context credit scoring (Lessmann et al., 2015). Despite their high predictive performances, they are considered as 'black boxes' for which the conclusions about a loan application cannot be fully explained. Thus, they cannot be used for regulatory purposes. Consequently, with their acceptable trade-off between explainability and predictive power, survival analysis models have been studied in various (Section 2) researches as an alternative to Logistic Regression. The success of survival analysis models in credit risk scoring is mainly due to the fact that they allow for a timely risk management, yields competitive predictive performance and provides more information about loans risk profiles compared to Logistic regression. In fact, for a chosen horizon, they inform about whether the loan will default. Therefore, timely PD are produced and can be used for further analysis such as studying seasoning patterns of default (Tong et al., 2012). Subsequently, contrary to Logistic Regression, survival analysis models can handle censored observations. Thus, less information about loans is thrown away during the data sampling process preceding the modelling. This latter feature may explain why in some cases survival models perform better than LR (Section 2.1). Cox PH is the most used survival analysis model in credit risk scoring and one with usually the highest predictive performance (Dirick et al. (2017), Jiang et al. (2019)).

One drawback of most standard survival models such as Cox PH is their implicit assumption that all borrowers will default in the long run which does not stand in practice (J. Zhang & Thomas, 2012). In fact, examples in credit risk show that a large proportion of borrowers does not default during the lifetime of their loan. To alleviate this inconvenient, Tong et al. (2012) introduce in the area of credit scoring the Logistic-CoxPH Mixture Cure model proposed by Kuk and CHen (1992). Indeed, the Mixture Cure model is a survival analysis model that relaxes this assumption. It considers two sub-populations in set of the borrowers. The first population corresponds to that fraction of borrowers that never default during the lifetime of their loan and the second one designates those who experience it at some point in time during the loan term. Thus, the model has two components. The first component is a binary classification model which tells if the default event will occur, e.g Logistic Regression. The second component is a survival analysis model which indicates when will the default occur,e.g Cox PH . With this setting, the Logistic-CoxPH Mixture Cure combines both advantages of LR and Cox PH. In terms of interpretation, it has a particularly

attractive feature for risk managers as it identifies variables that determine default from those that define the time to default.

Furthermore, the predictive power of the Logistic-CoxPH Mixture Cure can be increased in some cases without losing its interesting interpretation feature.Wycinka and Jurkiewicz (2017) suggest for that purpose to account for early repayments if there exits any. They argue that in presence of early repayments it is suitable to modify the standard Logistic-CoxPH mixture Cure model by including the idea of competing risks. Thus, instead of considering the occurrence of default as the only event of interest in the survival model, the prepayment of the loan is added as another event to consider. This procedure reduces the bias on parameters estimates in the survival part of the mixture cure model due to heavy censoring. Thus, it can be advantageous to apply this approach of the Logistic-CoxPH Mixture Cure model on mortgage loans data where early repayments are frequent. However, reviewing the literature of Mixture Cure models in credit risk scoring (Section 2.2) shows that their advantages over standard PD models have not been studied with mortgage loans data and the extension including competing risks is hardly applied in the comparison process.

In fact, most of the related studies concern personal or consumer loans data and do not account for the existence of early repayments. As data specifications can affect survival models, the real advantage of using the Mixture Cure model for mortgage loans credit scoring can not be drawn from these previous results. Sy and Taylor (2000) highlight that Mixture Cure model performs better on large samples with long-term follow-up without too much censoring which is the features of mortgage loans data after accounting for early repayments as an event of interest. Additionally, as an possible extension of their study, Tong et al. (2012) suggest to apply the Logistic-CoxPH model on mortgage loans as their longer terms (20-30 years) allow for longer observation time. Thus, it is of interest to know whether the conclusions of the previous researches on the advantages of using a Mixture Cure model for credit scoring generalize to mortgage loans with prepayments. The research in this paper aims to contribute in the literature by answering to the following question : Does the Logistic-CoxPH Mixture Cure model performs better than the standard Logistic Regression or Cox Proportional Hazards models to predict the probability of default of mortgage loans in presence of early repayments? The quest to an answer directs to the two following questions:

- Are the dynamic PDs produced by survival models more accurate than the static ones provided by the Logistic Regression?
- Does accounting for a cured fraction of loans provides higher performance than with a Cox PH for mortgage loans credit scoring with competing risks?

In this thesis, the research pioneered by Tong et al. (2012) is extended by accounting for early repayments as a competing risk in the survival models. Moreover, the study uses mortgage loans

data, more observations and a longer observation time. In previous studies (Tong et al. (2012), Wycinka and Jurkiewicz (2017)) Cox PH and the Logistic-CoxPH are found to be competitive with the Logistic Regression in terms of discrimination performance when not considering early repayments as a competing risk. The same comparison criteria is used in this thesis on the three implemented models in order to highlight deviations from the literature due to the inclusion of competing risks in the survival models and the use of mortgage loans data. A calibration performance test is also used to assess the goodness of the fit which in banks has become equally important as the ability to identify creditworthy borrowers since Basel 2. The discrimination and calibration measures used in this paper to compare the PD predictions, combine a set of measures frequently used in the literature of credit scoring with some measures recommended by the regulator (*Studies on the Validation of Internal Rating Systems*, 2005) in order to better account for what is done practically by risk managers. Therefore, the AUC (Area Under the Curve), the Gini Coefficient, the H-measure, the Kolmogorov-Smirnov statistic (KS) and the Brier Score are computed using a 100-folds cross-validation to assess the discrimination power. The Hosmer-Lemeshow test is performed at each cross-validation round to evaluate the calibration performance. The p-values are afterwards combined using the Fisher's method to draw final conclusions about the calibration test. The cross-validation provides more accurate appreciation of the performances.

2 Literature Review

This section outlines some relevant academic papers related to the use of survival analysis, and more specifically Mixture Cure models, in credit risk scoring. The first part is a reminding of major researches on survival analysis models for credit scoring . The second part goes deeper into the literature of Mixture Cure models in that context.

2.1 Survival Analysis Literature in Credit Risk Scoring

First considered as a binomial problem (Rosenberg and Gleit (1994), Hand and Henley (1997)) that classifies borrowers in 'good' and 'bad' depending on their probability of default, credit scoring has started to be viewed in the perspective of survival analysis with Nairan (1992). His work is further developed in other researches as an alternative to Logistic Regression. Thus, Banasik et al. (1999), Hand and Kelly (2001), Stepanova and Thomas (2002) or Andreeva (2006) apply and compare various parametric and non parametric survival models to the standard Logistic Regression for credit risk scoring. Particularly, they find that the Cox PH model performs better than Logistic Regression in that context.

Stepanova and Thomas (2002) show that survival analysis models can provide credit risk scoring while predicting time to early repayment as it handles competing risks. They use cause-specific

regressions that model default times and early repayment times separately using Cox PH. Baesens et al. (2005) extend survival analysis to a non-linear framework with the introduction of neural networks methods. Some researches such as Bellotti and Crook (2009) or Im et al. (2012) contribute by introducing time dependency in basic survival models used in credit risk scoring. Bellotti and Crook (2009) suggest introducing macroeconomic variables in the set of features used by the Cox PH and Accelerated Failure Time (AFT) models. However, Im et al. (2012) propose to only modify the Cox PH into a Time Dependant Hazards (TDH) and argues that it is a better approach to account for the interaction between the failure time and covariates than adding to the model time-dependent covariates such as macroeconomic variables as done by Bellotti and Crook (2009).

The standard survival analysis models assume that as time extends, the probability to experience the event of interest increases. Therefore, in the context of credit risk scoring, each borrower will experience default. However, in practice a large proportion of loans reach maturity. (Amico & Keilegom, 2018) compares the Cox PH to the Mixture Cure and shows that not accounting for such a reality leads to overestimating the baseline hazard and to more biased parameters estimates in the Cox PH. To overcome this drawback, Mixture Cure models are introduced in the credit scoring domain by Tong et al. (2012).

Another important feature of credit performance data is the presence of early repayments. It is a major source of censoring if not considered as in event of interest in the model. As heavy censoring reduces the performance of survival analysis models, competing risks are used in credit risk scoring survival models for more accurate PD estimations.Wycinka (2019) compares predictive performances of four common methods used to evaluate probability of default over time with competing risks. It finds that probabilities of default are best predicted with the Cumulative Incidence Function (CIF) calculated on cause-specific Cox PH models for default and early repayment under the assumption of independence between these two events.

The benchmark study of Dirick et al. (2017) wrap almost two decades of research on applications of survival analysis in credit scoring. They compare standard survival models used in major papers of the literature of credit scoring. The models are applied to 10 different personal loans data sets of five financial institutions in UK and Belgium. The results show that Cox PH models particularly Cox PH with Penalised Splines are the best performing survival models along with Single-Event Mixture Cure model based on discrimination, calibration and economic criteria.

2.2 Mixture Cure Models for Credit Risk Scoring

This section first reviews the origins of the Mixture Cure models and presents two main papers which compare the Logistic-CoxPH Mixture Cure to the standard Logistic Regression and Cox Proportional Hazard models. Contrary to those two papers, this thesis assesses differently the performance of the Logistic-CoxPH Mixture Cure over the two standard models. In fact, here, the

survival analysis models are adapted to account for the existence of competing risks. Moreover, a larger data set with longer observation periods and less censoring is used. The second part of this section discusses studies that investigate some possible improvements of the standard Logistic-CoxPH Mixture Cure model for credit risk scoring. They have inspired the research question of this thesis.

2.2.1 Comparison to Standard Credit Scoring Models

Mixture Cure models are borrowed from medical statistics where they have been used to model survivors in cancer clinical trials in terms of two sub-populations (Sy & Taylor, 2000). Long term survivors form a sub-population called the cured fraction or insusceptible population. They are considered cancer free after the trial and therefore will not relapse. The remaining sub-population is the uncured fraction or susceptible population and may relapse during the follow-up or after. The theory behind this type of models is pioneered by Boag (1952) and Berkson and Gage (1952) and is further developed by Farewell (1982). Later, using Kaplan-Meier(KM) estimators, Farewell (1986) proves the existence of such an insusceptible fraction in a breast cancer survival research. These models has two components to predict if and when a patient will relapse.

There exist similarities between credit risk scoring and cancer survival studies. In fact, cancer patients can be replaced by loans and therefore, the event of interest becomes the default instead of the relapse. Thus, the two components of the model will enable to predict if and when a borrower will default. Based on these correspondences, Tong et al. (2012) introduce a Mixture Cure model in the area of credit risk scoring which is considered as a standard. It is a Single-Event Mixture Cure model developed by Kuk and CHen (1992) and usually called Logistic-CoxPH Mixture Cure. It considers the occurrence of default as the only event of interest. It has a latency part which is a Cox PH model that predicts the time to default conditional on the borrower being susceptible to default (Jiang et al., 2019). The incidence part is a Logistic Regression model which predicts loans are likely to default.

Tong et al. (2012) investigates the performances of the Logistic-CoxPH Mixture Cure model over LR and Cox PH models using consumer accounts of a major UK retail bank. Loans are observed up to their term which is 12, 24 or 32 months and yearly PDs are predicted. They find that in terms of discrimination performance Logistic-CoxPH Mixture Cure was competitive with LR. However the calibration performance measures show the Mixture Cure model outperforming LR for intermediate time intervals. In fact, as a survival analysis tool, the Logistic-Cox Mixture can adjust for early repayments and estimate the baseline hazard function across time which justify its better estimates of PD for intermediate time intervals. LR was found to provide good calibration performances only for end of loan term estimates.

Their study is later reproduced by Wycinka and Jurkiewicz (2017) with different data spe-

cifications and a focus on discrimination performance only. They use a sample from a 60-months consumer loans portfolio data of one of the Polish financial institutions. Loans are observed up to 24 months. Their research comes to the same conclusion of not recommending Logistic-CoxPH Mixture Cure model over LR and Cox PH based on discrimination power as they are competitive. These two previous papers both highlight the advantage of the Mixture Cure model for interpreting separately the effects of explanatory variables on PD and default time. Moreover, they underline some ways to improve its predictive power such as including time dependency by using macroeconomic variables, accounting for competing risks such as early repayments or using more sophisticated models in the latency and incidence parts .

2.2.2 Extending the Standard Logistic-CoxPH Mixture Cure Model

The use of Mixture Cure model for credit risk scoring is a topic studied in various research papers after Tong et al. (2012). Liu et al. (2015) extends the standard version of the model by using a hierarchical Bayesian approach to predict future defaults. Dirick et al. (2015) develops a proper version of the Akaike Information Criterion (AIC) for the Logistic-CoxPH Mixture Cure model. It also provides steps to estimate Multi-Event Logistic-CoxPH while accounting for a mature fraction using the Expectation-Maximization algorithm. Thus, the paper can be seen as a semi-parametric version of the work of Watkins et al. (2014) which aims to jointly model competing risks in the Multi-Event Mixture model. Other researches have explored three main extensions of the standard Logistic-CoxPH Mixture Cure model suggested by Tong et al. (2012) and Wycinka and Jurkiewicz (2017) to improve its predictive power in credit scoring.

For example, to capture the underlying macroeconomic cycle, Leonardis and Rocci (2014) replace the discrete time Cox PH baseline function in the Logistic-CoxPH Mixture Cure with a time-varying system-level covariate. Their model is estimated with proprietary data provided by Intesa Sanpaolo and collected from a sample of small and middle-sized Italian firms. For the same purpose of accounting for the macroeconomic situation and following the approach of Bellotti and Crook (2009) on Cox PH, Dirick et al. (2019) studied the use of macroeconomic factors in the latency part of the Single-Event Logistic-CoxPH Mixture Cure model for credit risk scoring. They extend their research to the Multi-Event Mixture Cure model by also considering early repayment as an event of interest. The two models are estimated in a simulation study but also applied on actual 36-months personal loans data provided by a Belgian financial institution. They outline that despite the advantage of providing a valuable economic interpretation of default or early repayment, the main inconvenient of such models is the fact that they may require information about macroeconomic factors and default or early repayment events on a daily or weekly basis. This inconvenient gives more credit to the approach of Leonardis and Rocci (2014) and is in line with the arguments of Im et al. (2012) for not using macroeconomic factors in the set of features

to accommodate for time dependency in Cox PH.

On the other side, N. Zhang et al. (2019) extend the Single-Event Logistic-CoxPH Mixture Cure to a Multi-Event Logistic-CoxPH Mixture Cure that considers early repayments along with default as events of interest. However, they use a parametric baseline function in the Cox PH component and consider four(4) assumptions about the susceptibility to risks of the borrowers. The new model with competing risks is first developed in a simulation study and then used to score online consumer loans from Lending Club,one of the largest Peer-to-Peer lending platform in the world.They find that the model under competing risks was competitive to LR in terms of predictive performance. And, regarding AIC values, it performs best under the assumption that a segment of the loans is immune to default and all loans are susceptible to prepayment.

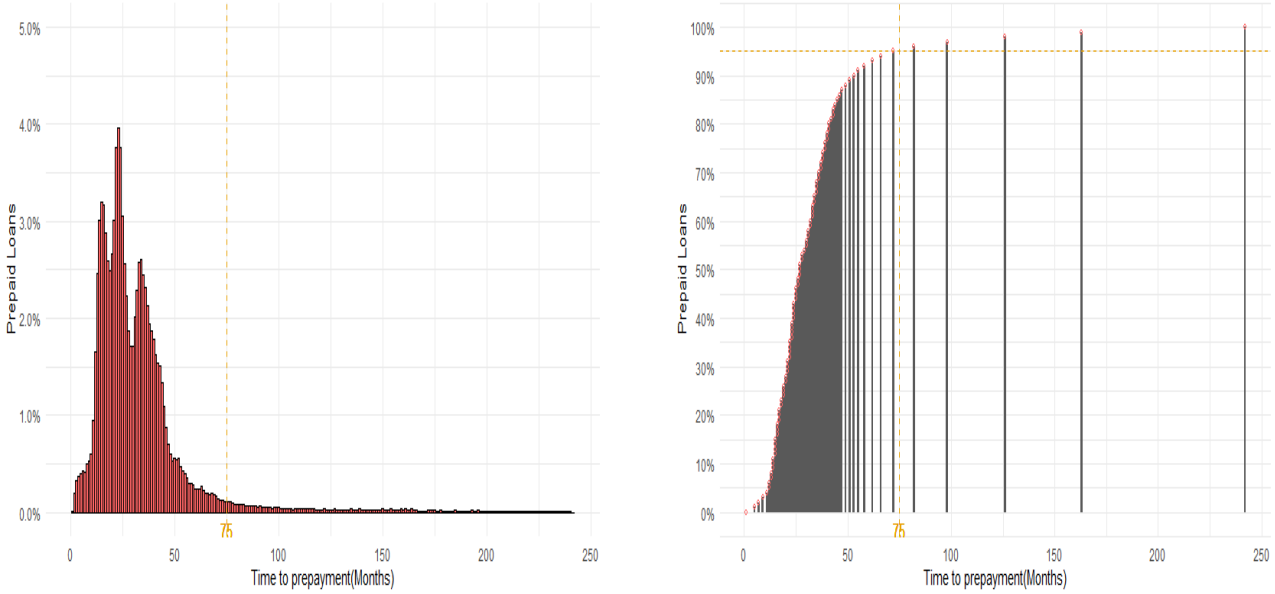
In the same line, Jiang et al. (2019) develop a prediction driven Mixture Cure model combining Time-Dependant Hazards (TDH) (Im et al., 2012) with Random Forest(RF) as an alternative to the standard Logistic-CoxPH Mixture Cure. Their study uses personal loans of a Peer-to-Peer lending institution in China. By using RF as incidence model and TDH as latency model, they follow the suggestion of Tong et al. (2012) to use tree based Machine Learning models in the incidence component and enhanced survival models that incorporate time dependency in the latency component of the Mixture Cure model. They justify the choice of TDH by the work of Im et al. (2012) which showed its higher predictive performance compared to standard Cox PH and its better ability to capture time dependency than including macroeconomic factors. They find that, even though the two extensions sacrifice the interpretation advantage that pertain in Logistic-CoxPH Mixture Cure, a significant improvement in both calibration and discrimination performances is obtained.

3 Fannie Mae's Single-Family Fixed Rate Morgage Data

In the US, Fanny Mae stands for the Federal National Mortgage Association (FNMA). After purchasing mortgages from different banks and financial institutions across the states, this association repackages them into mortgage-backed securities that are sold to investors. Thus, Fanny Mae participates to expand the secondary mortgage market and provides liquidity to financial institutions that allow them to provide more mortgages in return. The 'Fannie Mae Single-Family Loan Performance Data' contains fully amortizing, fully documented, single-family, conventional fixed-rate mortgage loans acquired by Fanny Mae. It is publicly available and promotes a better understanding of credit performance of mortgage loans portfolios. The data set is structured into two data files per quarter, Acquisition file and Performance file. The Acquisition file contains information on each mortgage loan at the moment it enters the database. The Performance file contains the monthly performance of each loan from its start date to the most recent date available, it is up-

dated quarterly. For example, the Acquisition file of 2000Q1 contains all mortgages added in the first quarter of 2000 with possible start dates in 1999 and the Performance file of 2000Q1 includes the performance of these mortgages in each month between the start date of the mortgage and December 2019.

A subset of this large database is used in this paper. It contains mortgage loans with terms of 20 years at least, bought from major financial institutions in the US (Table 5) and added in the Fannie Mae's database during the first quarter of 2000. The choice of this quarter allows to have 20 years of historical data for more informed modelling choices in the survival analysis. Some preparation steps are processed to make the data suitable for the intended study. First, none of the loans in the studied portfolio have been modified during their lifetime. They are fixed term mortgage loans and the only possible reasons a balance is reduced to zero is the occurrence of either prepayment or default or maturity. The sets of defaulted, prepaid and matured loans are mutually exclusive. There is no default-recovered loans in the portfolio. Subsequently, indicators of default and early repayment have been added to the data set as well as time duration variables indicating the number of months between a mortgage starting date and the first occurrence of those events. A loan is considered as defaulted if it has at least 90 days past-due at some point during its lifetime. Some basic statistics are computed in order to accurately define the follow-up scheme in survival models.

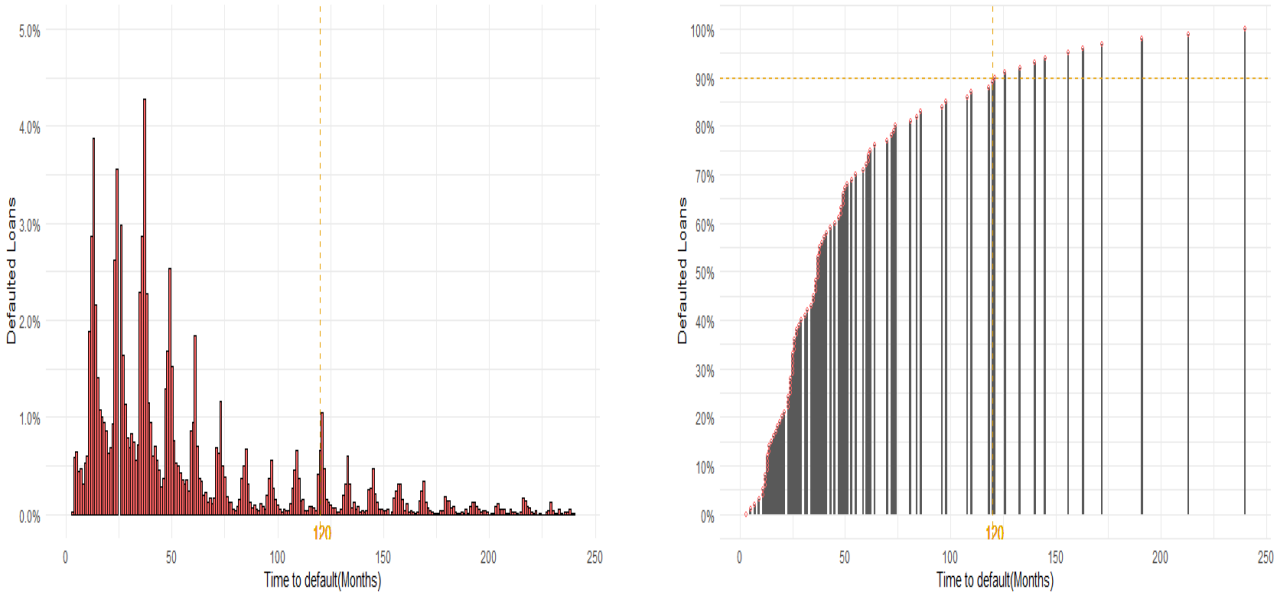


(a) Histogram of prepayment times

(b) Percentiles of prepayment time

Figure 1: Distribution of prepayment time

Figures 1 and 2 show respectively the distributions of prepayment and default times through



(a) Histogram of default times

(b) Percentiles of default time

Figure 2: Distribution of default time

percentiles plots and histograms. One can notice that 90% of defaults occur in the first 10 years (120 months) after origination date. For prepayments, 95% of them are realised before the 7^{th} year of the loan(75 months). Furthermore, both histograms show right skewed distributions which one characteristic of survival data. Concerning the follow-up scheme, Wycinka and Jurkiewicz (2017) suggests to fix the minimum observation time at the period in which most of the loans susceptible to default have been revealed in order to supply the incidence part of Mixture Cure model with enough data. This is also relevant for the other models. Thus, each loan is observed up to 10 years after its origination date. The enrolment period in the study starts from January 1999 to March 2000.

The final data set contains 206,326 mortgage loans with terms of 20 to 30 years. It has 21 features to which are added default and prepayment indicators as well a time variable indicating the number of months of follow-up for a loan. The 21 features are static and correspond to acquisition data. They cover relevant aspects for credit scoring such as financial information, borrower and loans' characteristics and behavioral information at origination. A full description of their contents is stated in Table 6. A summary of the the follow-up data reveals far more early repayments than defaults, 93.4% against 3.69%. Thus, when default is the only event of interest, 96.31% of the observations are censored which is quite heavy. This rate drops to 2.91% once early repayment is introduced as a competing event.

This data is suitable for the research conducted in this thesis as it meets the requirements for good performance of both the binary classification and survival analysis models that are compared.

In fact, binary classification models such as Logistic Regression are known to need enough amount of data to provide good predictions. However, there is still some class imbalance in the default that can be handled during the modelling process. Subsequently, according to Sy and Taylor (2000) Mixture Cure models performs better on large samples with long term follow-ups and no excessive censoring. The data set meets the first two conditions. Concerning the effect of heavy censoring it is alleviated by accounting for early repayment as a competing risk.

4 Models

This section presents Logistic Regression (LR), Cox Proportional Hazards(Cox PH) and Logistic-CoxPH Mixture Cure models which are implemented in this thesis to predict the probability of default of mortgage loans. Changes in the standard setting of the later two models to accommodate for competing risk are outlined.

4.1 Logistic Regression (LR)

Credit risk scoring has first been seen as a binomial problem consisting in classifying clients in 'good' or 'bad' depending on their probability to default. Therefore, appropriate statistical models have been used. Logistic Regression is the standard model for modelling PD in the bank industry (Lessmann et al., 2015). It is a statistical model used to explain a dichotomous dependant variable by a set of explanatory variables that can be numerical or categorical. Some extensions allow for a categorical dependant variable that have more than two classes. In PD modelling, its basic form with a binary dependant variable is used to estimate the probability that a loan defaults given some loan and borrower's characteristics. That is considering as response variable a binary variable Y that has values $Y = 1$ if a loan defaults and $Y = 0$ if it does not. Then, the probability that a loan i defaults given its set of explanatory variables X_i is as follow:

$$\Pr\left(Y_i = 1|X_i\right) = \frac{1}{1 + \exp\left(-(\alpha + \beta^T X_i)\right)},$$

where $g\left(\theta^T X\right) = \frac{1}{1+\exp\left(-\theta^T X\right)}$ is the logistic function evaluated on $\theta^T X$ for $\theta = \begin{bmatrix}\alpha \\ \beta\end{bmatrix}$ and $X = \begin{bmatrix}1 \\ X_i\end{bmatrix}$, α is the intercept and β is a vector of regression coefficients. Then, the probability that a loan i does not default given its characteristics X_i is $1 - \Pr\left(Y_i = 1|X_i\right)$. The parameters α and β are estimated by Maximum Likelihood. The log-likelihood function for a set of n loans is given by:

$$l(\theta) = \sum_{i=1}^{n} Y_i \log(p_i) + (1 - Y_i) \log(1 - p_i),$$

where $p_i = Pr(Y_i = 1|X_i;\theta)$ and $\theta = \begin{bmatrix}\alpha\\ \beta\end{bmatrix}$. Logistic Regression is linear in the log-odds which makes it convenient to interpret parameters using odd-ratios. The log-odd formulation of the model is given by:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta^T X. \tag{1}$$

4.2 Survival Analysis Models

This section describes two survival analysis models used in credit scoring , the standard Cox PH and the recently introduced Logistic-CoxPH Mixture Cure. The event of default and its consequences on a company's profits are related to the point in time it occurs(Jiang et al., 2019). Banasik et al. (1999) argue that not only if the borrower will default is important but also when it happens. In general, survival analysis models are used to analyze the expected time duration until a certain event occurs. In the context of credit risk scoring, if the occurrence of default is the only event of interest, survival analysis models the first time a loan defaults given its characteristics.

4.2.1 General Framework and PD Calculation

Considering T as the time until the occurrence of default, the survival, hazard and the cumulative hazard functions are different ways to describe the distribution of T. The survival function is given by $S(t) = P(T > t)$ and represents the probability of a loan not having defaulted by some time t. It can be approximated non-parametrically using the $Kaplan - Meier$ estimator given as:

$$\hat{S}(t) = \prod_{i:t_i \leqslant t}\left(1 - \frac{d_i}{n_i}\right),$$

where t_i is a time where at least one loan defaulted, d_i and n_i are respectively the number of defaulted and non-defaulted loans at t_i. The hazard function denotes the instantaneous default risk. It is the probability that a borrower defaults at time t conditional on not having defaulted before. It is stated as follow:

$$h(t) = \lim_{\delta_t \to 0}\left(\frac{P(t \leqslant T < t + \delta_t | T \geqslant t)}{\delta_t}\right) = \frac{f(t)}{S(t)}, \tag{2}$$

with $S(t)$ the survival function and $f(t) = -\frac{d}{du}S(u)$ the probability density function of T. The cumulative hazard function at time t is the accumulated default risk up to that time. It is expressed

as follow:

$$H(t) = \int_{-\infty}^{t} h(u)du,$$

where $h(.)$ is the hazard function. Similarly to the survival function, the cumulative hazard can also be estimated non-parametrically with the $Nelson - Aelen$ estimator given by:

$$\hat{H}(t) = \sum_{i:t_i \leqslant t} \left(\frac{d_i}{n_i}\right),$$

where t_i is a time where at least one loan defaulted, d_i is the number of defaults and n_i the number of non-defaulted loans.

As regression models, survival analysis models provide an estimation of the distribution of T through $h(t)$, $S(t)$ or $H(t)$ conditionally on a vector of covariates X. The covariates are measurements regarding the loan, the borrower or the underlying economic situation. The set corresponding set of parameters β can be estimated in different ways such as direct maximization of the likelihood or Expectation Maximization (EM) algorithm . In both cases, the estimation takes into account the existence of censored observations. An observation is censored when none of the events of interest have been observed at the moment of collecting the data (Dirick et al., 2017). Two different definitions of censoring are usually encountered in the literature of credit scoring depending on the inclusion of early repayment of the mortgage loan as a competing risk to the default event. Competing risks refer to the existence of two or more events of interest resulting in different reasons to stop observing a loan during the follow-up. For the case when the default of the loan is the only event of interest, censored observations are loans that did not experience default during the observation time. Thus, early repaid, matured and outstanding loans are censored. When early repayment enters as a competing risk, then censored cases are loans that have neither defaulted nor repaid early during the observation period.

Conditionally on a set of features X, yearly PDs are estimated from the Cumulative Incidence Function (CIF) of the default event using classical probabilities calculations. When the occurrence of default is the only event of interest, the CIF at any year t is the probability of defaulting at or before t written $P(T \leqslant t)$. It has an one-on-one relationship with the survival function because the probability of default during t is the probability of not surviving that year given as $1 - S(t)$. Thus, the CIF can be estimated non-parametrically with the $Kaplan - Meier$ estimator. And, as the survival function is an exponential function of the cumulative hazard given by $e^{-H(t)}$, then the CIF is $1 - e^{-H(t)}$. Consider a monthly credit risk performance data and a two years observation period

starting at t. The probability of default at the first year is simply the CIF at t, thus $1 - S(t)$. The probability that a loan defaults the second year provided that it did not default in the first one is the probability not to survive an additional year after t. It is derived using a simple Bayes rule. Consider D_t the default indicator at any year t, then it follows(Jung et al., 2018):

$$\begin{aligned} P\big(D_{t+1} = 1 | D_t = 0\big) &= \frac{P\big(D_{t+1} = 1, D_t = 0\big)}{P\big(D_t = 0\big)} \\ &= 1 - \frac{S\big(t+1\big)}{S\big(t\big)}. \end{aligned} \tag{3}$$

However, when considering early repayment as a competing risk these derivations do no longer hold (Wycinka, 2019). In that case the CIF of the default event cannot be directly linked to its survival function. It is the probability of defaulting at or before a given time t and before the occurrence of early repayment. Denote E_t the variable indicating the type of event that occurs at time t, e the prepayment event, d the default event and T a time duration variable indicating the moment of occurrence of the first event. The CIF of the default event is given by (Andersen et al., 2012):

$$CIF_d\big(t\big) = P\big(T \leqslant t | E = d\big) = \int_0^t S\big(s-\big) h_d\big(t\big) ds, \tag{4}$$

where $s-$ is the right sided limit, $S(.)$ which is equal to $e^{-H_e(s)-H_d(s)}$ at time s is the event-free survival function with $H_e(.)$ and $H_d(.)$ the cumulative hazard functions of respectively early repayment and default events, $h_d(.)$ is the default specific hazard function. From this expression of the CIF, one can notice that the incidence probability of default at t is simply the probability of surviving from both risks in the previous periods and defaulting at time t. The CIF estimated with the $Kaplan - Meier$ estimator in the previous case is always larger than the one obtained when accounting for competing risks(Z. Zhang, 2017). Thus, it is preferable to include competing risks if there are any in order to avoid overestimating the probability of default. Using the previous two years observation setting, the default probability at year t is equivalent to the CIF of the default event at that year as formulated in Equation (4). And, the probability of defaulting at year $t+1$ conditionally on surviving both risks at t is similarly given as:

$$
\begin{aligned}
P\big(D_{t+1}=1|D_t=0,E_t=0\big) &= \frac{P\big(D_{t+1}=1,D_t=0,E_t=0\big)}{P\big(D_t=0,E_t=0\big)} \\
&= \frac{P\big(t<T\leqslant t+1|D_{t+1}=1\big)}{P\big(T>t\big)} \\
&= \frac{CIF_d\big(t+1\big)-CIF_d\big(t\big)}{S_{e+d}\big(t\big)},
\end{aligned} \tag{5}
$$

where $S_{e+d}\big(.\big)$ is the event free survival function given by $e^{-H_e(s)-H_d(s)}$ with $H_e(.)$ and $H_d(.)$ the cumulative hazard functions of respectively early repayment and default events.

4.2.2 Cox Proportional Hazards (Cox PH)

Cox Proportional Hazards is a semi-parametric survival analysis model proposed by Cox (1972) and first used in credit scoring by Banasik et al. (1999). Since, it has been commonly applied for predicting PD (Dirick et al., 2017). Cox PH estimates the instantaneous default probability given a set X of borrower and loan's characteristics, $h(t|X)$, by a non-parametric baseline function along with a parametric component. The corresponding hazard function called Proportional Hazard (PH) is as follow in its continuous form:

$$
h\big(t|X\big) = h_0\big(t\big)exp\big(\beta^T X\big), \tag{6}
$$

where β is the vector of parameters to estimate and $h_0(.)$ is the baseline hazard function giving the hazard for $X=0$. The corresponding survival function is given by :

$$
\begin{aligned}
S\big(t|X\big) &= S_0\big(t\big)^{exp\big(\beta^T X\big)} \\
&= exp\big(-exp\big(\beta^T X\big)\int_0^t h_0\big(u\big)du\big) \\
&= exp\big(-exp\big(\beta^T X\big)H_0\big(t\big)\big),
\end{aligned} \tag{7}
$$

with $S_0(.)$ and $H_0(.)$ respectively the baseline survival and cumulative hazard functions.

With the approach proposed by Peng and Dear (2000) and Sy and Taylor (2000), the parameters β can be estimated without knowing the baseline hazard h_0. It consists of maximizing the following partial likelihood:

$$
L\big(\beta\big) = \prod_{i=1}^{k}\left(\frac{exp\big(x_{(i)}\beta\big)}{\sum_{l\in R(t_{(i)})}exp\big(x_{(l)}\beta\big)}\right),
$$

where $t_{(1)} < t_{(2)} ... < t_{(k)}$ are k ordered default times and $R(t_{(i)})$ the set of risky loans at time $t_{(i)}$. However, the assumption of continuous hazard function assumed by Proportional Hazard models does not hold for credit performance data (Stepanova & Thomas, 2002). In fact, for the data set used in this thesis, default times are tied because mortgage loans performances are reported on a monthly basis. Therefore, the exact set of risky loans at each default time appears unclear and the exact likelihood is difficult to compute. Cox (1972) suggests an approximation of the likelihood that uses a discrete hazard function. It replaces Equation (6) by a discrete logistic model. After some derivations, the corresponding approximated likelihood function in presence of ties is as follow:

$$L_{Cox}(\beta) = \prod_{i=1}^{k} \left(\frac{exp(s'_{D_i}\beta)}{\sum_{R \in R(t_{(i)};d_i)} exp(s'_R\beta)} \right),$$

where d_i is the number of defaults at time t_i, $R(t_{(i)}; d_i)$ contains all possible subsets of d_i loans from the set of risky loans at the ordered default time $t_{(i)}$, $R \in R(t_{(i)}; d_i)$ is a set of d_i risky loans at $t_{(i)}$, D_i is a set of d_i loans that default at $t_{(i)}$, $s'_R = \sum_{l \in R} x_l$ is the sum of covariates vectors over the loans in R and $s'_{D_i} = \sum_{l \in D_i} x_l$ is the sum of covariates vectors of loans in D_i.

Easier approximations of the denominator of the above likelihood have been developed (Stepanova & Thomas, 2002). In this thesis, the approximation proposed by Breslow (1974) is used. It is the most applied one in the literature of survival analysis for credit scoring and most implemented in softwares. It is given as follow:

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

4.2.3 Logistic-CoxPH Mixture Cure

The formulation of the survival function of standard survival analysis models used in credit scoring assumes that as time extends, the survival time goes to zero. Consequently, all borrowers will default at some point in time which in practice is not verified. To relax this assumption Tong et al. (2012) use the Logistic-CoxPH Mixture Cure of Kuk and CHen (1992) for credit risk scoring. It is since, the standard Mixture Cure model in this area (Dirick et al., 2017). In general, a Mixture Cure model considers two sub-populations in the set of borrowers. The first sub-population called 'cured' or 'insusceptible' corresponds to borrowers that will never default during the lifetime of their loan. The second sub-population called the 'uncured' or 'susceptible' will experience default at some point during the loan term. The model has therefore two components. The first component called the incidence part and is a binary classification model. The second component called the

latency is a survival analysis model. The Logistic-CoxPH Mixture Cure model is a Single-Event Mixture Cure that uses Logistic Regression(Section 4.1) as incidence model and Cox Proportional Hazards (Section 4.2.2) as latency model. It considers the occurrence of default as the only event of interest and thus prepaid and matured loans are censored (Section 4.2). The notations used in this section to describe the standard Mixture Cure are inspired from Tong et al. (2012).

Model Formulation

The incidence and latency components of Mixture Cure models are modelled separately and can have different explanatory variables. Consider X as the set of p observed features used by the latency model to estimate probability of default over the observation period. Denote Z the vector of q observed covariates used in the incidence component to estimate survival probabilities for different time horizons. Define δ as a censoring indicator with $\delta = 1$ referring to non-censored loans and $\delta = 0$ indicating censored loans. Let Y be a binary variable that denotes loan's susceptibility to default. $Y = 0$ states that the loan will never experience default during its term. $Y = 1$ indicates that the loan is susceptible to default at some point in its lifetime. Susceptible loans may be right-censored as the default can occur after the observation period. The couple (δ, Y) defines three types of data observations. ($\delta = 1$,Y=1) indicates loans that will default during the observation period, ($\delta = 0$,Y=1) is for those that will default after the observation period and ($\delta = 0$,Y=0) defines long-term survivors that will never experience default in their lifetime. The Logistic-CoxPH Mixture Cure model is defined by its unconditional survival function given as follow:

$$S\big(t|X,Z\big) = \pi\big(Z\big)S\big(t|Y=1,X\big) + 1 - \pi\big(Z\big), \tag{8}$$

with S$\big($t|X,Z$\big)$ stating the probability that a loan with features X and Z does not default up to time t. This formulation shows the two parts of the Mixture Cure model. The incidence part $\pi\big(Z\big)$ which denotes the probability that a loan defaults given its features Z is modelled by a Logistic Regression. Using odd-ratios, the corresponding incidence model is given as in Equation (1) by:

$$log\left(\frac{\pi\big(Z\big)}{1-\pi\big(Z\big)}\right) = \beta_0 + \beta_1 Z_1 + ... + \beta_q Z_q = \beta^T Z, \tag{9}$$

with β the set of parameters. The latency part $S\big(t|Y=1,X\big)$ represents the probability that a susceptible loan defaults after time t given its features X, $P(T > t|Y = 1, X)$. It is modelled by a Cox PH. Thus, as in Equation (7), the distribution of default time conditionally of being

susceptible is represented by the following conditional survival function :

$$\begin{aligned} S\big(t|Y=1,X\big) &= S_0\big(t|Y=1\big)^{exp\big(X^T b\big)} \\ &= exp\big(-exp\big(X^T b\big)\int_0^t h_0\big(u|Y=1\big)du\big) \\ &= exp\big(-exp\big(X^T b\big)H_0\big(t|Y=1\big)\big), \end{aligned} \tag{10}$$

where b is a vector of parameters, $S_0(t|Y=1)$, $h_0(t|Y=1)$ and $H_0(t|Y=1)$ are respectively the conditional baseline survival, hazard and cumulative hazard functions. They are not functions of covariates X.

Model Estimation: EM Algorithm

Consider a set of n loans and $O = (t_i, \delta_i, Z_i, X_i)$ the observed data for the i^{th} loan. Denote $Y = (Y_1, Y_2, ..., Y_n)$ the indicators of susceptibility to default. Denote the unknown parameters of the Logistic-CoxPH Mixture Cure model as $\Theta = (b, \beta, S_0)$. The corresponding full likelihood function considers that the contribution of a loan i to the likelihood depends on its censoring status δ_i. Thus, it is given by:

$$L\big(b,\beta,O,Y\big) = \prod_{i=1}^{n}\{\pi\big(Z_i\big)f\big(t_i|Y_i=1,X_i\big)\}^{\delta_i} \times \{\big(1-\pi\big(Z_i\big)\big) + \pi\big(Z_i\big)S\big(t_i|Y_i,X_i\big)\}^{1-\delta_i}, \tag{11}$$

where $f(.)$ is the probability density function of the default time T. Using the relation between $f(.)$ and $h(.)$ given in Equation (2), the log likelihood is $l\big(b,\beta,O,Y\big) = l_1\big(b,O,Y\big) + l_2\big(\beta,O,Y\big)$ with:

$$l_1\big(b,O,Y\big) = \sum_{i=1}^{n} Y_i \log[\pi\big(Z_i\big)] + \big(1-Y_i\big)\log[\big(1-\pi\big(Z_i\big)\big)] \tag{12}$$

and

$$l_2\big(\beta,O,Y\big) = \sum_{i=1}^{n} \delta_i Y_i \log[h\big(t_i|Y_i=1,X_i\big)] + Y_i \log[S\big(t_i|Y_i,X_i\big)]. \tag{13}$$

This expression of the log likelihood shows clearly that when all borrowers are susceptible to default the model reduces to a standard Cox PH. Furthermore, supposing the susceptibility status Y known for all borrowers, $l_1(b,O)$ is the log likelihood of the Logistic Regression and $l_2(b,O)$ is the log likelihood of Cox Proportional Hazards model including an offset variable $\log(Y_i)$ (Tong et al., 2012). Thus, if the susceptibility status Y is observed completely, the maximization problem of

the log likelihood is straightforward. It can be treated as two separated problems using the same approaches described in sections 4.1 and 4.2.2 to estimate respectively Cox PH and LR.

However the reality is different. In fact, the susceptibility to default is not known for all borrowers. It is only observed to be $Y = 1$ for uncensored observations ($\delta = 1$) which have for sure experienced the default event. Otherwise, Y is considered missing. Many approaches have been proposed in the literature (Amico & Keilegom, 2018) to estimate the parameters of the Logistic-CoxPH Mixture Cure with this missing data issue. In this thesis, the approach proposed by both Peng and Dear (2000) and Sy and Taylor (2000) is adopted. It is based on the Expectation-Maximisation (EM) algorithm developed by Dempster et al. (1977). The motivation behind this approach is to stick as much as possible to an easy maximization problem where the default status is completely observed. The algorithm first estimates the expected value of the complete data log likelihood given by Equations (12) and (13). The estimated values are obtained by replacing each susceptibility status Y_i with its expectation given observed data and parameters. Thus, the EM algorithm has two steps, expectation (E-step) and maximization (M-step). As the expected value of Y_i depends on the parameters and vice-versa, the algorithm starts with initial parameters values Θ^0 and iterates the two steps until it converges. Convergence occurs when the likelihood does not significantly change between two iterations. At the m^{th} iteration of the EM algorithm, the parameters $\Theta^{(m-1)} = (b^{(m-1)}, \beta^{(m-1)}, S_0^{(m-1)})$ derived from the previous iteration are used to compute for each loan the expectation of its susceptibility indicator. It corresponds for the i^{th} loan to $E(Y_i|\Theta^{(m-1)}, O)$ and is given by:

$$w_i^m = \delta_i + \big(1 - \delta_i\big) \frac{\pi\big(Z_i\big) S\big(t_i|Y_i = 1, X_i\big)}{1 - \pi\big(Z_i\big) + \pi\big(Z_i\big) S\big(t_i|Y_i = 1, X_i\big)} \Big|_{\left(\Theta^{(m-1)}, O\right)} .$$

Then, the E-step produces the following expectations (Cai et al., 2012) for the expressions in Equations (12) and (13) :

$$E\big(l_1\big) = \sum_{i=1}^{n} w_i^m \log[\pi\big(Z_i\big)] + \big(1 - w_i^m\big) \log[\big(1 - \pi\big(Z_i\big)\big)] \tag{14}$$

and

$$E\big(l_2\big) = \sum_{i=1}^{n} \delta_i \log[w_i^m h\big(t_i|Y_i = 1, X_i\big)] + w_i^m \log[S\big(t_i|Y_i, X_i\big)]. \tag{15}$$

The corresponding M-step maximizes above Equation (14) to produces parameters estimates of the Logistic Regression in the incidence part. For Equation (15), it can be rewritten as the log likelihood of the Cox PH model with an additional offset variable $log(w_i^m)$. Thus, the partial

likelihood approach presented in Section 4.2.2 is used to estimate the Cox PH model in the latency part. Similarly, the estimation accounts for the existence of ties and uses a Breslow estimator of the likelihood. For the following E-step, the survival function in the latency part is updated with an estimator of the baseline survival function derived from this step. Cai et al. (2012) proposes a Breslow-type estimator of the baseline survival function as:

$$\hat{S}_0\big(t|Y=1\big) = exp\left(-\sum_{j:t_{(j)}\leqslant t} \frac{d_{t_{(j)}}}{\sum_{i\in R(t_{(j)})} w_i^m e^{\hat{\beta}X_i}}\right),$$

where $t_{(1)} < t_{(2)} ... < t_{(k)}$ are k distinct ordered default times, $R(t_{(j)})$ and $d_{t_{(j)}}$ are respectively the set of risky loans and the number of defaults at the ordered default time $t_{(j)}$. To ensure that this estimator tends to zero as time extends, they set $\hat{S}_0\big(t|Y=1\big) = 0$ for $t > t_{(k)}$. The corresponding estimate of the baseline hazard function of the latency model can be obtained as:

$$\hat{h}_0\big(t|Y=1\big) = \frac{d_{t_{(j)}}}{\sum_{i\in R(t_{(j)})} w_i^m e^{\hat{\beta}x_i}}.$$

This approach of EM algorithm to estimate the Logistic-CoxPH Mixture Cure provides the estimated incidence model and estimations of the baseline survival and hazard functions of the Cox PH model as well as its parameters. Thus, the estimated survival function of the Logistic-CoxPH is easily derived using Equation (8).

4.2.4 Including Competing Risks: Default And Early Repayment

The survival models presented in the previous sections do not account for early repayment as a competing risk to default. However, early repayments reduce the size of the credit portfolio over time which in return changes the estimates of timely PDs. In some portfolios as the one studied in this thesis, there are far more early repayments than defaults during the observation time. Thus, not accounting for them in the survival models leads to more censoring. Yet, heavy censoring tends to bias parameter estimates in Cox PH and Logistic-CoxPH Mixture Cure models (Wycinka (2019) ; Amico and Keilegom (2018)).

There are two approaches to handle competing risks in survival models. The first one considers a bivariate variable (T, C) where T is the occurrence time of the first event and C the type of the corresponding event. The second analyses a bivariate latent variable $T = (T_d, T_e)$ of unobserved event times (Wycinka, 2019). In this later approach, the distribution of T can be seen as the joint

distribution of the default time T_d and prepayment time T_e. Thus, if default and prepayment times are independent, their marginal distributions define entirely the distribution of T. Yet, this condition of independence of the two failure times cannot be verified as only the time of occurrence of the first event given by $min(T_d, T_e)$ can be observed.

Additionally, in presence of competing events, when a specific event is observed in isolation from all the others, the distribution of the corresponding failure time is known as the sub-distribution. It is not always equivalent to its marginal distribution which is derived from the join distribution of failure times. In fact, it has been shown that for a given event, the sub-distribution is inferior to the marginal distribution (Wycinka, 2019). However, if the failure times are independent there is no longer difference between the sub-distribution and the marginal distribution. Thus, the hazard of the marginal distribution called the cause-specific hazard is equal to the hazard of the sub-distribution called sub-hazard. Consequently, using the second aforementioned approach of handling competing risk and supposing that prepayment times are independent from default times, the survival analysis model under the two competing risks can be derived from two isolated survival models. In the literature of survival analysis in credit scoring, this approach is frequently used.

For example, Banasik et al. (1999), Stepanova and Thomas (2002) or Bellotti and Crook (2009) use two Cox PH models separately to model default and prepayment times and derive afterwards the default probabilities using the appropriate CIF formula (Equation (4)). Though the underlying assumption of independence of default and prepayment times is unverifiable, this approach is considered by Wycinka (2019) as the best performing to predict default probabilities among four other possible methods reviewed in medical science literature. The advantage of this method lies in its simplicity and ability to provide different sets of features for modelling default and prepayment time. Thus, it is used in this paper to include competing risks in the two survival models used to predict the probability of default.

Cause-Specific Cox PH

Cause specific regressions are used to extend the standard Cox PH to competing risks. For each event of interest, default and prepayment, the corresponding Cox PH model is modelled by supposing that observations of the competing risk as censored. The same estimation procedure of Cox PH described in Section 4.2.2 is used. Subsequently, conditional on a set of features X, the CIF of default can be derived by estimating the event-free survival function as suggested by Benichou

and Gail (1990). They propose the following product integral estimator:

$$S\big(t|X\big) = \prod_{s \leqslant t} \big(1 - dH_d(t|X) - dH_e(t|X)\big),$$

where $H_d(.)$ and $H_e(.)$ denote the cause-specific cumulative hazard rates for respectively default and early repayment. This product integral is the asymptotic equivalent of the exponential estimator of the event-free survival function given by $\hat{S}(t|X) = e^{-\hat{H}(t_e|X)-\hat{H}(t_d|X)}$. It ensures that at time t, a loan with characteristics x will either default or prepay or remain in the portfolio as $S(t|X) + CIF_d(t|X) + CIF_e(t|X) = 1$. This feature is in accordance with the specifications of the data used in this study where there is no matured loan during the follow-up time. Moreover, Ozenne et al. (2017) recently developed a R-package which allows to compute CIF for cause-specific Cox PH regressions using this above estimation of the event-free survival function. Thus it is convenient to use this approach as parametric CIF calculations are not available in most softwares.

Cause Specific Logistic-CoxPH Mixture Cure

To account for competing risks in the standard Logistic-CoxPH, the assumption of existence of a 'cured' fraction is applied only to the default event. Therefore, it is assumed that all loans are susceptible to early repayment whereas some of them are immune to default. This assumption is relevant because in practice the rate of early repayment is high. In fact, in the credit risk data used in this thesis 93.4% of loans are prepaid. Not accounting for the small fraction that do not prepay is not expected to have a negative impact in the model. Additionally, N. Zhang et al. (2019) uses a Logistic-CoxPH Mixture model under competing risks to score online consumer loans and concludes that in terms of AIC, the model performs better under that later assumption. Thus, considering that the two risks are independent, a cause-specific approach can be used as previously. The choice of this procedure is also to ensure that survival models used in this paper are compared with the same assumptions under competing risks.

The prepayments times are modelled with a Cox PH model assuming that default observations are censored. And, a standard Logistic-CoxPH Mixture Cure model is used for the defaults times supposing prepayment observations as censored. The same set of covariates are used in both models. The CIF of the default event at time t_d given a covariates vector X can be approximated from the data as :

$$CIF_d(t_d|X) = \sum_{i=1}^{d} \hat{S}\big(t_{i-1}|X\big) \times \hat{h}_d\big(t_i|X\big),$$

where $\hat{h}_d\big(.\big)$ is the estimated hazard function from the default specific Logistic-CoxPH model and

$\hat{S}(.)$ is the estimated event-free survival function. The estimation of the event-free survival function is derived from its exponential approximation given as $\hat{S}(t|X) = e^{-\hat{H}(t_e|X)-\hat{H}(t_d|X)}$ where $\hat{H}(t_e|X)$ and $\hat{H}(t_d|X)$ are the estimated cumulative hazard functions in respectively default and prepayment survival models. Conditional on a set of features X, the hazard function of the Logistic-CoxPH Mixture Cure model is estimated by exploiting its relation with its survival function as given in Equation (2). In fact, this equation suggests that the hazard function at t is equal to $-\frac{d}{dt}log(S(t))$ which is in return equivalent to $-\frac{\frac{d}{dt}S(t)}{S(t)}$). Thus, using the survival functions given in Equations (8) and (7), conditional on X the later expression can be developed as follow:

$$\begin{aligned} h_d\big(t|X\big) &= -\frac{\frac{d}{dt}S(t|X)}{S(t|X)} \\ &= \frac{\pi\big(X\big)h_{c_0}(t)exp(\beta^T X)S_c(t|X)}{S(t|X)} \\ &= \frac{\pi\big(Z\big)h_c(t|X)S_c(t|X)}{S(t|X)}, \end{aligned}$$

where $S(t|X)$ is the survival function of the Logistic-CoxPH Mixture Cure, $\pi\big(X\big)$ its incidence part, $h_c(t|X)$ the hazard of the Cox PH model used in the latency part and $S_c(t|X)$ its corresponding survival function. All components of this expression are estimated by the outputs the EM algorithm.

5 Model Building and Validation

This section describes the main steps in building and validating the models compared in this thesis. It outlines major transformations performed on the data before modelling, summarizes how yearly PDs are obtained from the models and presents the measures of discrimination performance.

5.1 Weight of Evidence (WoE) and Feature Selection

The models implemented in this paper use the same set of features to predict the probability of default as no feature selection is performed. All the acquisition data of mortgage loans as recorded by Fannie Mae are included in the set of covariates (Table 6). They are relevant features for credit scoring as they contain most characteristics used by lenders to decide on a loan application. In this thesis, the interpretation of the effect of features on the probability of default is not of interest. Rather, the main interest is the differences in the discrimination and calibration performances and not their values per se. Thus, only necessary data management is performed such as handling high dimensional categorical variables or balancing data sets for Logistic Regression models.

Unlike Logistic Regression, survival analysis models such as Cox PH and Logistic-Cox Mixture Cure are regression models that cannot handle categorical variables. Thus, usually they are transformed into continuous variables by encoding them into dummies. This procedure can be beneficial as a categorical feature can be insignificant for predicting PD while one or some of its levels are. However, this approach may considerably increase the number of features. For example, the data set used in this study contains categorical variables with hundreds of levels such as Zip Codes. Thus, the regular dummy encoding procedure is not feasible. An alternative is to use the Weight Of Evidence (WoE). Moeyersoms and Martens (2015) demonstrate that transforming a feature with high-cardinality attributes into a continuous feature using WoE is preferable for the model predictive performance.

The WoE defines new values for a categorical variable relatively on how the observations in each category are classified comparatively to the rest of the population based on a binary target variable. In the context of this thesis as in credit scoring in general, the indicator of default during the follow-up is used as target variable (Jiang et al., 2019). Thus, the value of WoE generated depends on whether the category contains more or less defaults than the overall population. This method is applied on all categorical features present in the data set. The transformation of a feature Y is performed as follow:

$$WOE_i^Y = \ln\left(\frac{C_i^Y/TC}{N_i^Y/TN}\right),$$

where C_i^Y and N_i^Y denote respectively the number of defaults and non defaults for observations with the i^{th} attribute of variable Y, TC is the total number of defaulted loans in the data set and TN is the total number of non-defaulted ones.

5.2 Yearly Probability of Default

Similarly to previous studies (Tong et al. (2012), Stepanova and Thomas (2002)), the models used in this paper estimate yearly probabilities of default. It is in accordance with what banks use in practice in their risk management process. Furthermore, a yearly forecast cycle is appropriate for mortgage loans data where terms are typically long. For the two survival models implemented, yearly PD can be derived using estimated CIF and survival functions as described in Sections 4.2.1 and 4.2.4. Note that contrary to Tong et al. (2012), survival models are not stratified by loan terms. In fact, in this case, terms do not affect the observation time of different loans. They are sufficiently long to cover the 10-years of follow-up which have been shown to be optimal to observe most of defaults and prepayments (Section 3). Furthermore, the loan term variable does not come as the most significant in predicting default. The Cox PH models are implemented using

the R-packages *Survival* (Therneau et al., 2020) and *riskRegression* (Ozenne et al., 2017). For the Logistic-CoxPH Mixture Cure, the estimation procedure has been coded in R using the *smcure* package (Cai et al., 2012) as an inspiration.

Logistic Regression is a static model and does not produce dynamic PD. A single run of the model is therefore not enough to obtain the yearly PD during the 10 years follow-up. Thus, to produce a dynamic estimation of the probability of default from the Logistic Regression, the model is run 10 times with 10 different data sets derived from a one year rolling window on follow-up data. In fact, all loans are supposed to have entered the study at the same time t_0 and years are defined in reference to that date and not as calendar years. Thus, the different Logistic Regression models provide the default probabilities of mortgage loans at respectively t_1, t_2,...,t_{10} years after their origination date. The model at t_i uses data of mortgage loans that have not defaulted at t_{i-1} years after origination. It gives the probability that corresponding mortgages default within the upcoming year which is equivalent to defaulting between the $t_{i-1} - th$ and $t_i - th$ year after origination. Loans that have prepaid between the $t_{i-1} - th$ and $t_i - th$ year after origination are not included in the modelling data. Table 1 describes the data sets used by the 10 Logistic Regression models given the remaining loans in the portfolio at the beginning of each year. One can notice the imbalance of the default class and that as time extends, less data is used by LR models. The models are estimated with R using the *glm* function.

Table 1: Summary of the data used by Logistic Regression models

	Total	Excluded (%)	Included (%)	Defaulted (%)
year 1	206326	12761 (6.18)	193565 (93.82)	724 (0.37)
year 2	192841	73682 (38.21)	119159 (61.79)	1687 (1.42)
year 3	117472	54342 (46.26)	63130 (53.74)	1771 (2.81)
year 4	61359	33401 (54.44)	27958 (45.56)	1248 (4.46)
year 5	26710	9359 (35.04)	17351 (64.96)	797 (4.59)
year 6	16554	4287 (25.90)	12267 (74.10)	475 (3.87)
year 7	11792	1966 (16.67)	9826 (83.33)	316 (3.22)
year 8	9510	1258 (13.23)	8252 (86.77)	184 (2.23)
year 9	8068	890 (11.03)	7178 (88.97)	179 (2.49)
year 10	6999	755 (10.79)	6244 (89.21)	234 (3.75)

5.3 Class Imbalance: SMOTE Resampling

The data set contains only 3.69% of defaulted loans during the follow-up. For the Logistic Regression model this imbalance affects performances. In fact, with class imbalance, the model has fewer examples to learn from in order to model defaults accurately. Thus, it will considers the few defaulted observations as outliers and focus mainly on the majority class, which can lead to poor predictive performance for defaults. There are many different methods to handle this class imbalance in the Logistic Regression setting among which a cost-sensitive approach or resampling techniques. A cost-sensitive Logistic Regression could be used if the true proportion of defaults in the Fannie Mae's portfolio was known. However, this information is not available. Thus, at each round of cross-validation, the training set for the Logistic Regression is balanced by a Synthetic Minority Over-sampling Technique (SMOTE) together with under-sampling from the majority class.

Chawla et al. (2012) argues that this method is particularly suitable when classifying a defaulted loan as non-default is much more costly than the other way around. Similarly, Marques et al. (2013) examines resampling techniques for credit scoring and finds SMOTE method to ensure good discrimination performance based on AUC. In this method, the over-sampling of the minority class is done by creating synthetic observations. They are generated by supposing the feature space of loans in the minority class to be similar. For each vector of features in the minority class, its k nearest neighbors are identified in the feature space, and one of them is randomly selected. Then, a new synthetic observation is generated as a combination of both observations. The SMOTE method is implemented in R within the package $DMwR$(Torgo, 2013).

5.4 Cross Validated Discrimination Performance Measures

Similarly to Tong et al. (2012), models implemented in this study are estimated using a 100-folds cross-validation for more accurate measure of the discrimination performance. For each Logistic Regression model, the splitting of the data set into 100-folds is done such that the proportion of defaults in the data set remains unchanged in the folds. This procedure prevents from having validation sets empty of defaulted loans. For the survival analysis models, the split ensures that in the 100 folds, the proportion of defaulted loans at any year after origination is identical to what is observed in the whole data set. Thus, the performance of the survival models in predicting the probability of default at any year within the 10-years horizon can be estimated on all the 100 validation sets.

In both Logistic Regression and survival analysis models, the cross-validation procedure provides

at each predicted yearly probability of default, 100 values for each measure of discrimination performance. Thus, their means and standard deviations are derived and used to compare the models. In the literature of survival analysis for credit risk scoring, three discrimination measures are mainly used, the ROC measure(AUC), the H-measure and the Kolmogorov-Smirnov(KS) statistic. In this paper, the Gini Coefficient and Brier Score are added as discrimination measures to cover most of the methods suggested by the Basel Committee for validating credit scoring models in Banks(*Studies on the Validation of Internal Rating Systems*, 2005).

5.4.1 AUC and H-Measure

The AUC is the area under the curve of the Receiver Operating Characteristic (ROC) which plots the true positive rate or sensitivity against the true negative rate or specificity. In terms of credit scoring, the plot represents the percentage of defaulted loans that have been classified as such against the percentage of non-defaulted loans that have been classified as such. The AUC is a number between 0.5 and 1 with 0.5 corresponding to a random classification of loans. The higher is the AUC the better is the model at determining creditworthiness. However, the AUC uses different misclassification cost-distributions depending on the model. As a result, it compares different classification models with different metrics. In many researches, the AUC has shown its inability to accurately discriminate between algorithms (Dirick et al. (2017),Tong et al. (2012)). Thus, the H-measure is suggested by Hand (2009) as an alternative. In fact,The H-measure operates more coherently by using the same cost-distributions independently of the model. Therefore, in this thesis, if the two measures lead to different results in comparing two models, the conclusion of the H-measure will be considered.

5.4.2 Gini Coefficient

The Gini Coefficient or Accuracy Ratio is a summary statistic of the Cumulative Accuracy Profile (CAP) which plots the cumulative population in the x-axis and the corresponding cumulative defaults in y-axis. The Gini Coefficient is the ratio of the area between the CAP of the model and the CAP of the random model with the area between the CAP of the model and the CAP of the perfect model. It is also a simplified representation of the AUC and equals $AUC * 2 - 1$. Like the AUC, it measures the models ability to correctly classify a loan as defaulted or not. It values are between 0 and 1. The value 1 is an asymptotic value which is not reached in practice and the value 0 correspond to a model classifying loans randomly. The closer the Gini is to 1 the better is the model in identifying creditworthy loans. This measure is generally used on imbalanced data, hence its application in credit scoring where default rate are usually low.

5.4.3 Kolmogorov-Smirnov (KS) Statistic and the Brier Score

The Kolmogorov-Smirnov test is used to compare two samples under a null hypothesis stating that the two samples are drawn from the same distribution. Thus, for large values of the Kolmogorov-Smirnov statistic the null hypothesis is rejected. In credit scoring, this statistic is used to quantify the maximum difference of the cumulative distributions of good and bad borrowers. It is measures the maximum vertical distance between their curves. It values range from 0 to 1. A KS statistic equal to zero signifies that the credit scoring model is unable to differentiate defaulters from non-defaulters. The value 1 for the KS designates a perfect ability to identify creditworthy borrowers. Thus the higher is the KS the better is the discriminatory performance of the model.

In the context of credit scoring the Brier score is not a calibration measure. In fact, it does not compare the true conditional probabilities of default with their estimates. Instead, it estimates the mean squared difference between the default indicators of l the loans and their predicted defaulted probabilities. It can be thought as the residual sum of squares resulting from the regression of the default indicators on the rating. In that sense, the smaller is the Brier score the larger is the variance of the default probability forecasts (*Studies on the Validation of Internal Rating Systems*, 2005) and therefore the ability of the model to distinguish between defaulters from non-defaulters.

5.5 Hosmer-Lemeshow Calibration Test and the Fisher's Method

In practice there is a difference between the predicted default rates and the observed ones. In banks, depending on the magnitude of this difference the computed capital requirement to cover the risk can be inappropriate. Thus, the quality of a credit risk model can not only be based on how well it identify defaulters and non-defaulters. Instead, its calibration performance which focuses on the appropriateness of the PD estimates (*Studies on the Validation of Internal Rating Systems*, 2005) should be investigated. In fact, it provides a measurement of how likely a borrower classified as defaulter will actually default. The Hosmer-Lemeshow or Chi-Square test is one of the most used method to assess the goodness of fit or calibration of credit risk models. It tests simultaneously for different rating categories whether or not the observed default rates match the expected ones. It is based on the assumption of independence of default times and normal approximation. However, these assumptions can be unverified in practice. In that case, the test can erroneously reject the null hypothesis of 'correct PD forecast'. In this thesis, the test is performed at each round of cross validation by using the R-Package *ResourceSelection* and identifying the deciles of PD forecasts as categories. Thus, for each model and at each year, 100 results of the test are obtained.

For every model, a cross-validated p-value at each year is obtained by applying the Fisher's combined probability test or Fisher's method (Hosmer et al., 1925) under the assumption that the 100 tests are independent. This assumption is relevant because, in the cross-validation procedure

the data sets on which are based the Hosmer-Lemeshow tests are independent. The Fisher's method combines the 100 p-values into one single statistic given by:

$$X^2 = -2\sum_{i=1}^{k} \log\left(p_i\right),$$

where k is the number of p-values form the independent tests and p_i the p-value derived from the i^{th} test. Under the null hypothesis stating that all the k tests do not reject their null hypothesis, X^2 follows a Chi squared distribution with $2k$ degrees of freedom. Thus, rejecting the null hypothesis of the Fisher's method means that at least one of the independent tests rejects its null hypothesis. In the case of this study it would mean that the null hypothesis of good fit is not commonly accepted across the validation samples. The Fisher's combined probability test is implemented in R with the package *metaseqR* inspired by Moreau et al. (2003).

6 Results

In this section, the Logistic-CoxPH Mixture Cure model is compared to the Logistic Regression and Cox PH models in terms of discrimination and calibration performances for predicting the yearly probability of default of mortgage loans in presence of early repayments. The two survival models are implemented by accounting for competing risks using a cause-specific approach. The discrimination performances are measured by 100-folds cross-validated AUC, Gini Coefficient, H-measure, Brier Score and Kolomogorov-Smirnov (KS) Statistic. The calibration performance is evaluated through the 100-folds cross validated Hosmer-Lemeshow test. For each model, the cross-validation procedure provides, at each year of follow-up, 100 values for each discrimination measure and 100 p-values for the Hosmer-Lemeshow test. The yearly means and standard deviations of the discrimination measures are computed to compare the models. For each year, the overall p-value of the Hosmer-Lemeshow tests is obtained by combining the 100 cross-validated p-values through the Fisher's method.

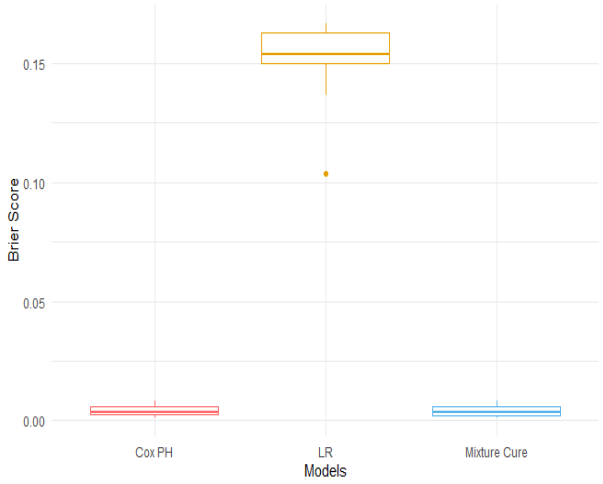
6.1 Survival Models Outperforming the Logistic Regression

In Figure 3, the distribution of the average discrimination measures over the 10 years of prediction are compared for the different models. First, it reveals that across the years, survival analysis models perform better than the Logistic Regression on average. In fact, for all discrimination measures, the Cox PH and Logistic-CoxPH Mixture Cure yield higher and more stable average values. Secondly, the boxplots reveal that the two survival models are competitive across the years. Table 2 provides more details about these results. It shows that in terms of discrimination

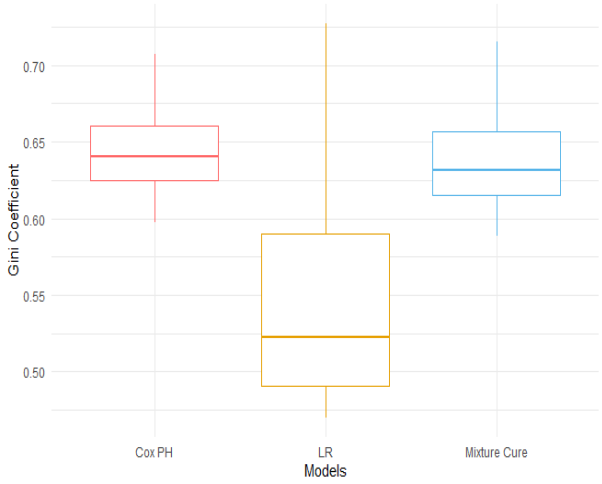
performances, the three models are better than a random classification. For a given year, the coefficients of variation of the different measures across the 100-folds are all smaller than one which ensures accurate comparison of the means. In overall, this variation increases with time and is slightly larger for the Logistic Regression. Thus, the survival models tend to better generalise on independent data for predicting PD in intermediate time intervals.

Figure 4 plots for each year the 100 p-values of the Hosmer-Lemeshow tests derived from the cross-validation. The tests are evaluated at 1% and 5% significance levels. These thresholds are represented by dotted lines in Figure 4. The plots show that for the Logistic Regression, at each year, the null hypothesis of good fit is always rejected for both significance levels. For the two survival models, there are evidences of a good fit of the model only for the first three years. However, for these years, p-values obtained for the Cox PH model are on overall larger and more stable than for the Mixture Cure model. Table 3 summarises theses observations with the p-values of Fisher's combined probability test. Considering the same significance levels as previously, this test results in the the rejection of the null hypothesis at all years for the Logistic Regression and its acceptation at the first three year in the case of the Cox PH. For the Mixture Cure, the null hypothesis of the Fisher test is accepted only for the second year.

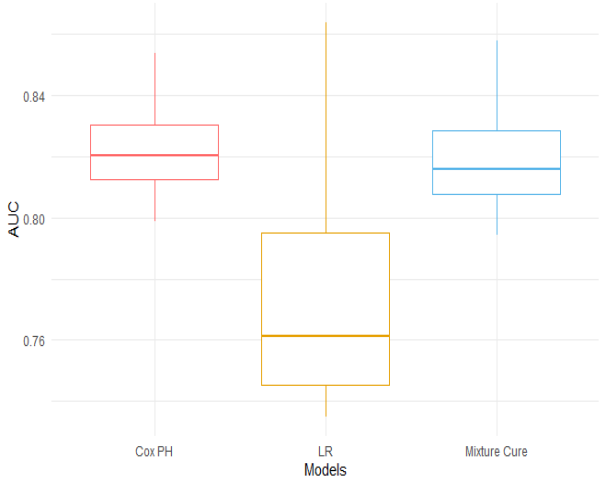
The calibration tests favor the survival models for the first three years. For the remaining years, all models have poor calibration performance. These conclusions about the calibration performance can related to one of the results of Tong et al. (2012) where the Logistic Regression was found to calibrate well only at the end of the loan estimates. The results prove that in terms of discrimination performance, survival models can be better than the Logistic Regression for long term PD prediction.This conclusion deviates from the previous studies of Tong et al. (2012) and Wycinka and Jurkiewicz (2017) who did not find any evidence in terms of discriminatory power to recommend the Cox PH and the Logistic-CoxPH Mixture Cure over the Logistic Regression. Their study differ from the one conducted in this thesis in two points. First, they did not account for the presence of early repayments in their data and have therefore higher rates of censoring which affect the performances of the survival models. Secondly, the mortgage loan data used in this thesis match all the requirements for good performance of the Mixture Cure Model outlined by Sy and Taylor (2000). On another side, the results obtained in terms of discrimination power are in accordance with the conclusions of Banasik et al. (1999), Hand and Kelly (2001), Stepanova and Thomas (2002) and Andreeva (2006) who outline the same advantages of using survival models as alternatives to the Logistic Regression for credit risk scoring. The over-performance of survival analysis models in this thesis can be explained by their ability to handle censored data contrary to Logistic Regression. The higher discrimination performance of the Logistic Regression in the first year can be explained by the fact that there are much less censored observations during the first year to exclude from its modelling data (Table 1). For the remaining years, the cumulative



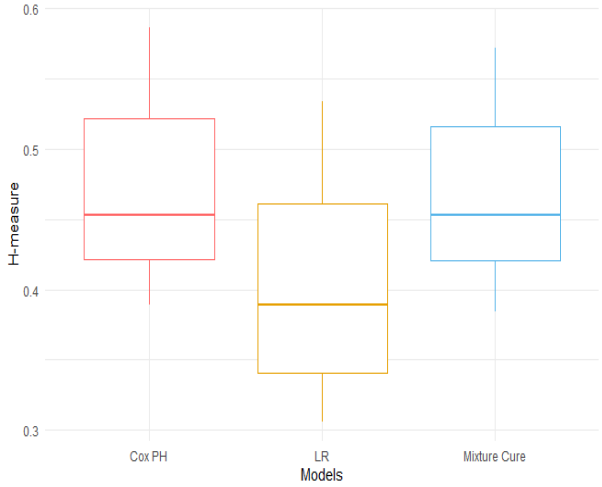
(a) Brier Score



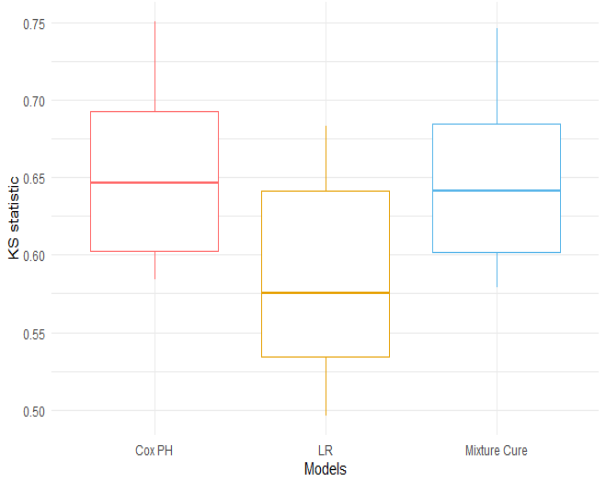
(b) Gini Coefficient



(c) AUC



(d) H-measure



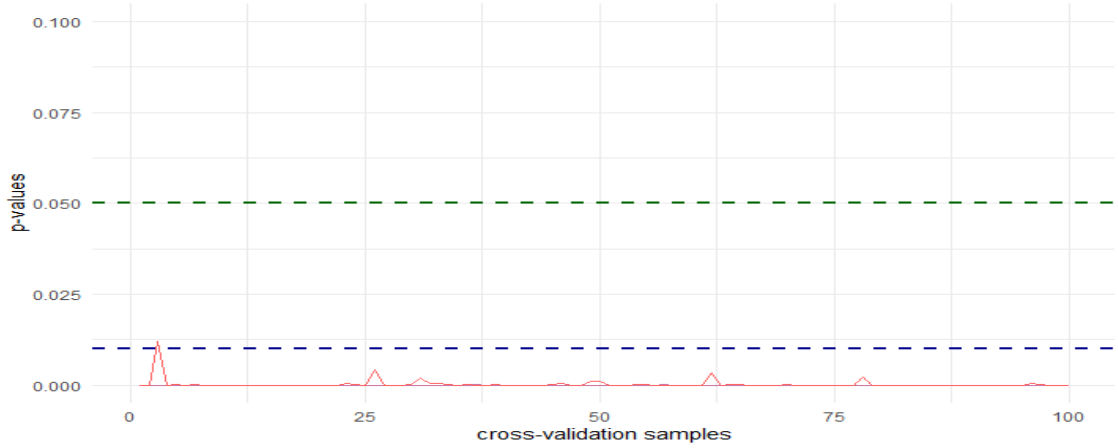
(e) KS Statistic

Figure 3: Distribution of average discrimination measures over the 10 years

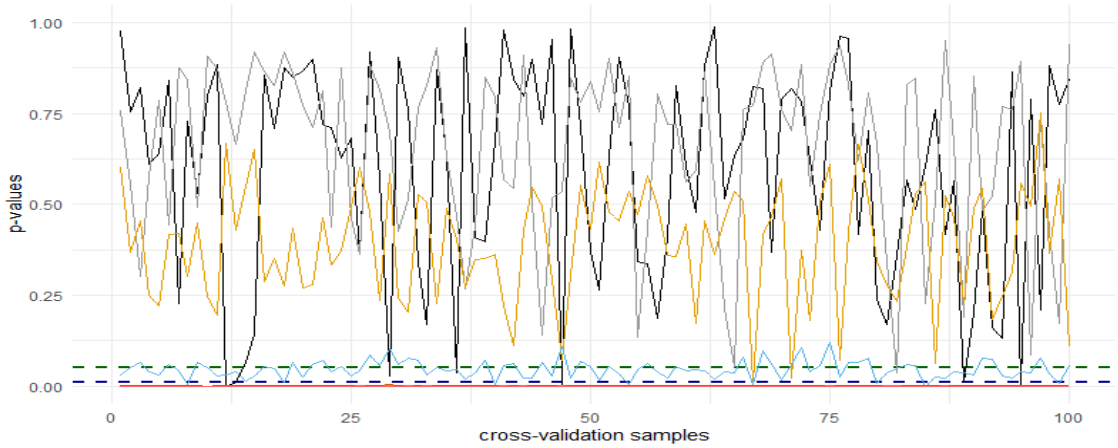
Table 2: Comparison of cross-validated discrimination performances

	Model	H-measure	AUC	Gini	KS	Brier
year 1	LR	0.534 (0.143)	0.863 (0.068)	0.727 (0.136)	0.683 (0.126)	0.104 (0.005)
	Cox PH	0.515 (0.126)	0.854 (0.064)	0.707 (0.129)	0.674 (0.111)	0.003 (0.000)
	Mixture Cure	0.523 (0.123)	0.858 (0.064)	0.715 (0.128)	0.679 (0.109)	0.003 (0.000)
year 2	LR	0.393 (0.085)	0.829 (0.042)	0.659 (0.084)	0.576 (0.079)	0.136 (0.006)
	Cox PH	0.417 (0.085)	0.842 (0.041)	0.683 (0.081)	0.596 (0.075)	0.008 (0.000)
	Mixture Cure	0.419 (0.084)	0.842 (0.041)	0.684 (0.081)	0.599 (0.075)	0.008 (0.000)
year 3	LR	0.349 (0.085)	0.806 (0.044)	0.612 (0.088)	0.546 (0.081)	0.151 (0.009)
	Cox PH	0.392 (0.082)	0.832 (0.041)	0.664 (0.083)	0.588 (0.078)	0.008 (0.000)
	Mixture Cure	0.391 (0.080)	0.832 (0.041)	0.663 (0.082)	0.589(0.075)	0.008 (0.000)
year 4	LR	0.306 (0.111)	0.762 (0.065)	0.525 (0.131)	0.496 (0.113)	0.164 (0.010)
	Cox PH	0.389 (0.113)	0.819 (0.060)	0.637 (0.120)	0.584 (0.111)	0.006 (0.000)
	Mixture Cure	0.385 (0.111)	0.815 (0.061)	0.631 (0.122)	0.579 (0.111)	0.006 (0.000)
year 5	LR	0.337 (0.141)	0.760 (0.088)	0.521 (0.176)	0.521 (0.135)	0.165 (0.013)
	Cox PH	0.434 (0.131)	0.824 (0.068)	0.648 (0.137)	0.620 (0.125)	0.005 (0.000)
	Mixture Cure	0.426 (0.131)	0.819 (0.071)	0.638 (0.142)	0.610 (0.125)	0.005 (0.000)
year 6	LR	0.333 (0.153)	0.736 (0.104)	0.471 (0.208)	0.530 (0.144)	0.167 (0.019)
	Cox PH	0.438 (0.146)	0.810 (0.086)	0.621 (0.172)	0.640 (0.133)	0.003 (0.000)
	Mixture Cure	0.433 (0.148)	0.805 (0.089)	0.610 (0.178)	0.628 (0.141)	0.003 (0.000)
year 7	LR	0.385 (0.179)	0.735 (0.115)	0.470 (0.229)	0.575 (0.168)	0.160 (0.021)
	Cox PH	0.469 (0.195)	0.799 (0.115)	0.597 (0.231)	0.653 (0.169)	0.002 (0.000)
	Mixture Cure	0.473 (0.198)	0.799 (0.115)	0.599 (0.231)	0.655 (0.168)	0.002 (0.000)
year 8	LR	0.488 (0.236)	0.763 (0.142)	0.525 (0.283)	0.676 (0.191)	0.150 (0.022)
	Cox PH	0.572 (0.248)	0.819 (0.135)	0.638 (0.269)	0.750 (0.175)	0.001 (0.000)
	Mixture Cure	0.567 (0.246)	0.816 (0.133)	0.632 (0.267)	0.746 (0.175)	0.001 (0.000)
year 9	LR	0.465 (0.216)	0.745 (0.132)	0.489 (0.263)	0.649 (0.188)	0.152 (0.021)
	Cox PH	0.586 (0.247)	0.821 (0.133)	0.642 (0.266)	0.746 (0.186)	0.001 (0.000)
	Mixture Cure	0.572 (0.245)	0.814 (0.131)	0.629 (0.263)	0.737 (0.189)	0.001 (0.000)
year 10	LR	0.449 (0.201)	0.748 (0.135)	0.426 (0.270)	0.616 (0.185)	0.155 (0.025)
	Cox PH	0.523 (0.213)	0.809 (0.124)	0.618 (0.247)	0.699 (0.180)	0.003 (0.000)
	Mixture Cure	0.495 (0.234)	0.794 (0.130)	0.588 (0.259)	0.686 (0.189)	0.002 (0.000)

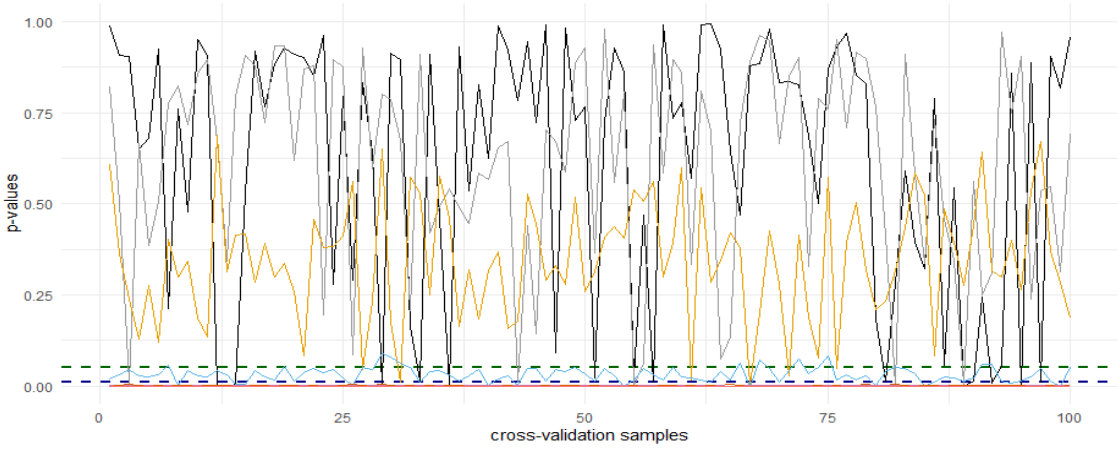
*The means and standard deviations are calculated using 100 folds-cross validation



(a) Logistic Regression



(b) Cox Proportional Hazard



(c) Mixture Cure

Figure 4: P-values of Hosmer-Lemeshow tests evaluated at 0.05 and 0.01 significance level

prepayments and defaults rates increase as time extends and thus less and less data are included in Logistic Regression models (Table 1).

Table 3: P-values of Fisher's method applied to Hosmer-Lemeshow tests

Year	LR	Cox PH	Mixture Cure
1	0	0.8153868283	0.0000002416
2	0	0.9999999899	0.9872599298
3	0	0.2278343725	0.0042889939
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	0	0	0

6.2 Cox PH Performing Slightly Better than the Mixture Cure

An unanswered part of the research question of this thesis remains which one between the Cox PH and Mixture Cure model perform better in predicting yearly PDs in presence of early repayments. This question is of interest because the main reason to introducing Mixture Cure models to credit scoring is to circumvent the false assumption of standard survival models that all borrowers will experience default at some point in time. Thus, a higher performance of the Mixture Cure model would signify a possible negative impact of this assumption. Except for the Gini Coefficient and the AUC, a graphical comparison in Figure 3 shows that the two models have on average close discrimination performances across the years and can hardly be differentiated. Closely examining the results in Table 2 reveals that except for the Brier Score for which the two models are highly competitive, the Cox PH tends to discriminate between defaulters and non-defaulters slightly better than the Mixture Cure model according to most of the discrimination measures. Additionally, Cox PH yields higher calibration performances in the first three years (Figure 4 and Table 3). Even though the discrimination performances of the two models are numerically close, the small differences should not been ignore because in credit risk modelling even an increase of 1% of the predictive power is important (Hand & Henley, 1997). The H-measure, KS statistic and Gini Coefficient are used to compare the two survival models because the conclusions of the AUC disagree with the H-measure in some years and the Brier Scores are similar for all years. Table 4 shows that across the 10 years of comparison, the Cox PH is more often identified as the model that better differentiate defaulters form non-defaulters. However the Mixture Cure Model performs

best in the first two years and in the seventh year.

Table 4: Comparison of cross-validated discrimination performances

	Model	H-measure	Gini	KS
year 1	Cox PH			
	Mixture Cure	✓	✓	✓
year 2	Cox PH			
	Mixture Cure	✓	✓	✓
year 3	Cox PH	✓	✓	
	Mixture Cure			✓
year 4	Cox PH	✓	✓	✓
	Mixture Cure			
year 5	Cox PH	✓	✓	✓
	Mixture Cure			
year 6	Cox PH	✓	✓	✓
	Mixture Cure			
year 7	Cox PH			
	Mixture Cure	✓	✓	✓
year 8	Cox PH	✓	✓	✓
	Mixture Cure			
year 9	Cox PH	✓	✓	✓
	Mixture Cure			
year 10	Cox PH	✓	✓	✓
	Mixture Cure			
Most recommended model		Cox PH	Cox PH	Cox PH

In terms of discrimination and calibration performance this study concludes that the Cox PH model should be used over the Mixture Cure to predict the probability of default of mortgage loans at intermediate time intervals. The conclusions are consistent with the results in Dirick et al. (2017) whose comparison of survival models for credit scoring which show that the Cox PH is the best performing followed closely by the Mixture Cure model. There may be two possible causes of this conclusion. The first one is that the data set is robust to the assumption of existence of a cure fraction and therefore a standard Cox-PH is more efficient. The second one concerns the method use in this thesis to incorporate competing risk in the standard Mixture Cure model. In fact, one of the strengths of the Cox PH model is that it can be estimated without knowing the baseline hazard. Thus, an additional source of error have been added in the Mixture Cure

model by approximating the baseline hazard function of the latency part with a Breslow type estimator in the EM algorithm. Modelling this baseline hazard is an important issue in Mixture Cure models. In the literature, some researchers propose a more flexible modelling of the hazard function of susceptible loans using a penalised likelihood approach with M-splines. In the context of this thesis, this approach was not applied because the pre-processing steps needed to use splines is hardly compatible with a 100-folds cross validation.

7 Conclusion

This thesis studied the performance of a Mixture Cure model compared to Cox PH or Logistic Regression for predicting the probability of default have been studied for mortgage loans in presence of early repayments. The survival models are implemented using a cause-specific approach to handle the competing risks. Different discrimination and calibration measures have been chosen as comparison criteria. The discrimination measures are computed using 100-folds cross validation as well as the calibration tests. Contrary to similar studies that found the Mixture Cure model, the Cox PH and Logistic regression competitive, the results of this thesis show that the two survival analysis models can be good alternatives to Regression in terms of discrimination and calibration performance. The survival models show good calibration performances in the first three years whereas the Logistic Regression reveals a poor fit during the entire follow-up period. The discrimination measures also have higher values in the survival models. However, it seems there is not always significant advantage in accounting for a cured fraction of borrowers even when competing risks are considered along with a long period of follow-up . In fact, there are only three years out of ten where the Mixture Cure model outperforms the Cox PH in terms of discrimination performance. Moreover, the Cox PH has shown better and more stable calibration performances in the first three years. Thus, the Cox PH model outperforms the Mixture Cure in most cases and is therefore be a first choice recommendation. It can be concluded that accounting for the competing risks of early repayments and using mortgages loan data as recommended by Tong et al. (2012) can leads to different conclusions regarding the performance of Mixture Cure models in credit scoring over the Logistic Regression. These conclusions are not meant to be generalised to all type of data but to contribute to the literature of survival analysis in credit scoring and more specifically to the studies regarding the use of the Mixture Cure model.

The study conducted in the thesis has some limitations. In fact, the comparison of the performance measures could be more precise by adding confidence intervals. Moreover, the study focuses on one calibration measure, the Hosmer-Lemeshow test. It could be interesting to add more calibration measures as for the discrimination performance. Regarding the Mixture Cure model, the baseline hazard of the latency part used to compute the CIF is based on a Breslow approximation

which is not flexible and can be unstable. A further research including a more sophisticated estimation of the baseline hazard such as the penalised likelihood approach may yield higher predictive performance. The data used in this analysis are static acquisition features whereas it is common to have time dependant variables in credit risk data. These static feature may predict the default accurately in the early years but as time extends, different states of the economy and changes in the borrower characteristics may no longer be captured. Introducing time dependency provides higher predictive performance for the Mixture Cure (Dirick et al., 2019) and Cox PH (Im et al., 2012) models. For example, adding macro-economic variables or monthly performance data to the set of features used in this study may lead to different results. Finally, the survival models have been estimated using a cause-specific approach with an unverifiable assumption of independence of the default and prepayment times and supposing that all borrowers are susceptible to prepayment. If these suppositions do not hold in practice, the results of the comparison of models can be misleading. An extended study may use other approaches of handling competing risks in survival models such as in Dirick et al. (2017) which jointly estimate the parameters of the Mixture Cure model under competing risks.

References

Amico M. & Keilegom I. V. (2018). Cure models in survival analysis. *Annual Review of Statistics and Its Application*, *5*, 311–342.

Andersen P., Geskus R., de Witte T. & Putter H. (2012). Competing risks in epidemiology: Possibilities and pitfalls. *International Journal of Epidemiology*, *41*, 861–870.

Andreeva G. (2006). European generic scoring models using survival analysis. *Journal of the Operational Research Society*, *57*, 1180–1187.

Baesens B., Gestel T., Stepanova M., Poel D. & Vanthienen J. (2005). Neural network survival analysis for personal loan data. *Journal of the Operational Research Society*, *56*(9), 1089–1098.

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

Bellotti T. & Crook J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, *60*, 1699–1707.

Benichou J. & Gail M. H. (1990). Estimates of absolute cause-specific risk in cohort. *Biometrics*, *46*(3), 813–826.

Berkson J. & Gage R. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, *47*(259), 501–515.

Boag J. (1952). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society*, *11*(2), 15–53.

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

Cai C., Zou Y., Peng Y. & Zhang J. (2012). Smcure: An r-package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, *108*(3), 1255–1260.

Chawla N. V., Bowyer K. W., Hall L. O. & Kegelmeyer W. P. (2012). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

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

Dempster A. P., Laird N. M. & Rubin D. B. (1977). Maximum likelihood from incomplete data via em algorithm. *Journal of the Royal Statistical Society*, *39*(1), 1–38.

Dirick L., Claeskens G. & Baesens B. (2015). An akaike information criterion for multiple event mixture cure models. *European Journal of Operational Research*, *241*(2), 449–457.

Dirick L., Bellotti T., Claeskens G. & Baesens B. (2019). Macro-economic factors in credit risk calculations: Including time-varying covariates in mixture cure models. *Journal of Business & Economic Statistics*, *37*(1), 40–53.

Dirick L., Claeskens G. & Baesens B. (2017). Time to default in credit scoring using survival analysis: A benchmark study. *Journal of The Operational Research Society*, *68*, 652–665.

Farewell V. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, *38*(4), 1041–1046.

Farewell V. (1986). Mixture models in survival analysis: Are they worth the risk? *The Canadian Journal of Statistics*, *14*(3), 257–262.

Hand D. (2009). Measuring classifier performance: A coherent alternative to the area under the roc curve. *Machine Learning*, *77*, 103–123.

Hand D. & Henley W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of The Royal Statistical Society*, *160*(3), 523–541.

Hand D. & Kelly M. (2001). Lookahead scorecards for new fixed term credit. *Journal of the Operational Research Society*, *56*(9), 989–996.

Hosmer D., Lemeshow S. & May S. (1925). Statistical methods for research workers.

Im J.-K., Apley D. & Shan X. (2012). A time-dependent proportional hazards survival model for credit risk analysis. *Journal of the Operational Research Society*, *63*, 306–321.

Jiang C., Wang Z. & Zhao H. (2019). A prediction-driven mixture cure model and its application in credit scoring. *European Journal of Operational Research*, *277*, 20–31.

Jung S.-H., Lee H. Y. & Chow S.-C. (2018). Survival analysis in the presence of competing risks. *Statistical Methods for Conditional Survival Analysis*, *28*(5), 927–938.

Kuk A. Y. C. & CHen C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, *79*, 531–541.

Leonardis D. D. & Rocci R. (2014). Default risk analysis via a discrete time cure rate model. *Applied Stochastic Models in Business and Industry*, *30*, 529–543.

Lessmann S., Baesen B., Seow H.-V. & Thomas L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring:an update of research. *European Journal of Operational Research*, *247*, 124–136.

Liu F., Hua Z. & Lim A. (2015). Identifying future defaulters: A hierarchical bayesian method. *European Journal of Operational Research*, *241*, 202–211.

Marques A., Gracia V. & Sanchez J. (2013). On the suitability of resampling techniques for the class imbalance problem in credit scoring. *The Journal of the Operational Research Society*, *64*(7), 1060–1070.

Moeyersoms J. & Martens D. (2015). Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems*, *72*(8), 72–81.

Moreau Y., Aerts S., Moor B. D., Strooper B. D. & Dabrowski M. (2003). Comparison and meta-analysis of microarray data: From the bench to the computer desk. *Trends Genetics*, *19*(10), 570–577.

Nairan B. (1992). Survival analysis and the credit granting decision. In L. Thomas, J. Crook & D. Edelman (Eds.), *Credit scoring and credit control* (pp. 109–121). Clarendon Press:Oxford.

Ozenne B., Sørensen A. L., Scheike T., Torp-Pedersen C. & Gerds T. A. (2017). Riskregression: Predicting the risk of an event using cox regression models. *R Journal*, *9*(2), 440–460.

Peng Y. & Dear K. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, *56*(1), 237–243.

Rosenberg E. & Gleit A. (1994). Quantitative methods in credit management: A survey. *Operations Research*, *42*(4), 589–613.

Stepanova M. & Thomas L. (2002). Survival analysis methods for personal loan data. *Operations Research*, *50*(2), 277–289.

Studies on the validation of internal rating systems. (2005). Version No. 14. Basel Committee On Banking Supervision.

Sy J. & Taylor J. (2000). Estimation in a cox proportional hazards cure model. *Biometrics*, *56*(1), 227–236.

Therneau T. M., Lumley T., Atkinson E. & Crowson C. (2020). *Survival: Survival analysis*. Version 3.1-12. https://cran.r-project.org/web/packages/survival/

Tong E. N., Mues C. & Thomas L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research*, *218*, 132–139.

Torgo L. (2013). *Dmwr: Functions and data for "data mining with r"*. Version 0.4.1. https://CRAN.R-project.org/package=DMwR

Watkins J., Vasnev A. & Gerlach R. (2014). Multiple event incidence and duration analysis for credit data incorporating nonstochastic loan maturity. *Journal of Applied Econometrics*, *29*(4), 627–648.

Wycinka E. (2019). Competing risk models of default in the presence of early repayments. *Econometrics*, *23*(2), 99–120.

Wycinka E. & Jurkiewicz T. (2017). Mixture cure models in prediction of time to default: Comparison with logit and cox models (K. Jajuga, L. T. Orlowski & K. Staehr, Eds.). In K. Jajuga, L. T. Orlowski & K. Staehr (Eds.), *Contemporary trends and challenges in finance, proceedings from the 2nd wroclaw international conference in finance*, Springer.

Zhang J. & Thomas L. C. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling lgd. *International Journal of Forecasting*, *28*(1), 204–215.

Zhang N., Yang Q., Kelleher A. & Si W. (2019). A new mixture cure model under competing risks to score online consumer loan. *Quantitative Finance*, *19*(7), 1243–1253.

Zhang Z. (2017). Survival analysis in the presence of competing risks. *Annals of Translational Medecine*, *5*(3), 47.

8 Appendix

8.1 Repartition of Loans by Seller

Table 5: Fannie Mae Mortgage Loan Sellers

Financial Institution	Loans	Defaulted (%)	Prepaid (%)
AMTRUST BANK	6810	2.57	95.95
BANK OF AMERICA, N.A.	21002	4.13	92.54
BISHOPS GATE RESIDENTIAL MORTGAGE TRUST	8286	3.45	93.25
CITIMORTGAGE, INC.	9563	3.94	93.05
FIRST TENNESSEE BANK NATIONAL ASSOCIATION	7195	3.35	94.73
FLAGSTAR BANK, FSB	6386	4.95	92.11
GE MORTGAGE SERVICES, LLC	147	2.72	90.48
GMAC MORTGAGE, LLC	10284	4.10	93.10
HARWOOD STREET FUNDING I, LLC	1998	3.05	95.25
JP MORGAN CHASE BANK, NA	3358	5.69	90.41
JPMORGAN CHASE BANK, NA	38321	3.41	94.25
JPMORGAN CHASE BANK, NATIONAL ASSOCIATION	11924	4.01	93.31
NETBANK FUNDING SERVICES	4801	2.83	95.06
OLD KENT MORTGAGE COMPANY	109	7.34	86.24
PNC BANK, N.A.	252	3.57	92.06
RBC MORTGAGE COMPANY	1990	3.32	95.78
REGIONS BANK	4401	4.43	91.05
SUNTRUST MORTGAGE INC.	9480	2.33	94.88
USAA FEDERAL SAVINGS BANK	2183	2.02	94.46
WASHTENAW MORTGAGE COMPANY	197	8.63	85.79
WELLS FARGO BANK, N.A.	5872	3.39	93.12
OTHER	51767	3.85	92.77

8.2 Acquisition Data in Fannie Mae's Database

Table 6: Fannie Mae's Mortgage Loan Acquisition data

Attribute	Type	Values
First home buyer indicator	Categorical	Y=Yes, N=No, U=Unknown
Minimum of borrower and co-borrower credit scores(Fico)	Categorical	300-850, Blank (if score is <300 or >850 or unknown)
Mortgage Insurance Type	Categorical	1=Borrower Paid, 2= Lender Paid, 3=Investor Paid, Blank=None
Number of units	Categorical	1-4
Number of borrowers	Categorical	1-10
Origination Channel	Categorical	R=Retail,C=Correspondent,B=Broker
Original debt to income ratio	Categorical	1% - 64%, Blank (if DTI is = 0, or $\geqslant$ 65, unknown, or if the mortgage loan is a HARP refinance)
Original Interest Rate	Continuous	Blank = Unknown
Original UPB	Continuous	Also known as original loan amount, original loan size or original principal balance
Original loan term	Continuous	60 - 419 months
Original Loan-To-Value (LTV)	Categorical	0% - 97% (or up to 200% for a mortgage loan acquired through a HARP refinance), Blank (if LTV is >97% or is >200% for a mortgage loan acquired through a HARP refinance, or unknown)
Original Combined Loan-To-Value (CLTV)	Categorical	0% 200%, Blank (if CLTV is > 200 or unknown)
Original Value at Origination	Categorical	Original House Price at loan origanation
Occupancy type	Categorical	P = Principal, S = Second, I = Investor, U = Unknown
Loan Purpose	Categorical	P = Purchase, C = Cash-out Refinance, R = No Cash-out Refinance, U = Refinance – Not Specified
Property type	Categorical	SF = Single-Family,CO = Condo, CP = Co-Op, MH = Manufactured Housing, PU = PUD
Primary mortgage insurance percent	Categorical	1% - 50%, Blank (if not applicable or is < 1% or > 50%)
Relocation Mortgage Indicator	Categorical	Y = Yes, N = No
Seller Name	Categorical	Name of the mortgage loan seller to Fannie Mae
Property State	Categorical	two-letter abbreviation indicating the state within which the property securing the mortgage loan is located
ZIP code short	Categorical	XXX=first three digits of the property's zip code

Source: Acquisition File Layout of Fannie Mae Loan Performance Data