Distressed Mortgages: A Machine Learning Assessment*

Elio Bolliger[a], Paul Lyons[b], Sameer Shaikh[b], Fang Yao[b]

[a]University of Lausanne
[b]Central Bank of Ireland

December 16, 2023

Abstract

This paper presents a machine learning based approach for assessing distressed mortgages. We combine bank-level loan data with financial information obtained from the Standard Financial Statement (SFS) in Ireland to identify loans in arrears that engage with their lenders to start alternative repayment arrangement renegotiations. We compare standard regression techniques to a machine learning method in an exercise to best predict those loans that are likely to engage with their lender and explore the most important drivers. We find that machine learning outperforms standard regression tools in terms of overall performance and has advantages in handling complex features of our dataset. This finding underscores the potential for machine learning models to offer new or complimentary insights to important policy questions. Finally, we highlight that early identification of distressed loans is crucial for engagement, as early engagement with distressed borrowers leads to more favourable outcomes for both lenders and borrowers.

Keywords: Mortgage arrears, Financial Stability, Macroprudential Regulation, Machine Learning

JEL Codes: C55, D14, G18

*The views, opinions, findings, and conclusions or recommendations expressed in this paper are strictly those of the authors. They do not necessarily reflect the views of the Central Bank of Ireland. The Central Bank of Ireland takes no responsibility for any errors or omissions in, or for the correctness of, the information contained in this paper. We thank David Byrne, Fergal McCann, Edward Gaffney, and all seminar participants at the Central Bank of Ireland as well as the Conference of Applied Theory, Macro and Empirical Finance (AMEF 2023) for helpful feedback, comments and suggestions.

1 Introduction

The financial crisis of 2008 - 2013 in Ireland resulted in a record number of mortgage loans becoming delinquent, which, under normal circumstances, would have resulted in foreclosure and repossession. However, in recognition of the large numbers of arrears cases that needed to be dealt with in an appropriate and sustainable manner and of the adverse effects of repossession, the Central Bank introduced a series of policy measures, which pivoted towards forbearance and mortgage modification as a response. This approach, while allowing defaulted borrowers remain in their homes, also resulted in a longer duration of default for those with delinquent mortgage loans (Kelly and McCann, 2016). Current prevailing inflationary pressures, lingering effects from the pandemic and a slowdown of major economies may all pose renewed challenges to borrowers' repayment capacities and may lead to new mortgage arrears cases. Ireland's experience of working through a large stock of long-term mortgage arrears cases can provide important lessons for policy makers around the world in the new era of hightened financial risks.

The first important step in the resolution of mortgage arrears process is getting distressed borrowers to engage with their lenders. In this paper, we apply Machine Learning (ML) techniques to study the behavior of the engagement between mortgage borrowers in arrears and their lenders. In doing so, our aim is threefold, (1) to introduce machine learning techniques which are increasingly used at Central Banks and among financial institutions more generally; (2) to compare and contrast ML models and more traditional econometric models and (3) to provide new insights on resolving distressed debt, particularly relevant given the current cost-of-living pressures borrowers are facing.

ML, in broad terms, focusses on developing models that can be applied to data for detecting patterns, making predictions and deriving policy recommendations. In particular, where data (especially "big data") are available, ML can be applied for the aforementioned outcomes. A key use case for ML in central banks is to support analysis, business intelligence, and data mining applications, by providing economists or other analysts with modern and robust tools beyond the traditional economic and statistical methods.

In this paper, we develop a ML model to predict the likelihood of borrowers engaging with their lender as set out in the Central Bank's Consumer Protection Code (CCMA). The Central Bank first introduced the CCMA in 2009 in the midst of an economic and employment crisis to provide statutory safeguards for vulnerable, financially-distressed borrowers in arrears or at risk of falling into arrears. Further strengthened in subsequent years, the CCMA, and within it, the Mortgage Arrears Resolution Process (MARP), is just one part of the national policy framework of supports and protections to assist people with mortgage arrears difficulties.[1] A key step in the MARP is the gathering of specific financial information from the borrower, to enable the regulated entity to undertake an assessment of the borrower's case and to consider whether an alternative repayment arrangement (ARA) is appropriate and sustainable for the borrower's individual circumstances. This information must be gathered via the Standard Financial Statement (SFS). Using both SFS and loan level data collected from the five main financial institutions in Ireland since 2012, we can apply ML techniques to assess characteristics of engaged versus non-engaged borrowers.

As a key finding of this paper in terms of identifying characteristics that predict whether a borrower will engage or otherwise, the number of days in arrears (days-past-due) is identified as the most important variable, along with general macroeconomic conditions. In light of the current macroeconomic uncertainty, getting borrowers to engage early in their arrears cycle is an important consideration and evidence from Ireland's financial crisis years (2008-2013) shows that this will lead to more favourable outcomes for both borrowers and banks. In addition, we also find that higher indebtedness appears to be associated with an increase in the probability of an engagement via SFS, while being a first-time-buyer (FTB) reduces it. Finally, we find evidence that macroeconomic conditions (e.g., higher interest rates or higher unemployment rate) are positively associated with the probability of SFS engagement.

[1] These supports include the national Mortgage Arrears Resolution Service (Abhaile), the Mortgage to Rent Scheme and Personal Insolvency Arrangements under the Personal Insolvency Act 2012, as amended, all of which can potentially assist borrowers to stay in their homes.

Specifically, we estimate survival models to predict the probability of borrower engagement. We first estimate a more traditional survival model, a Cox regression model. We then compare it with a Random Survival Forest (RSF) model. While, both models identify important characteristics that predict borrower propensity to engage via the SFS, we find that the ML model outperforms the Cox model in terms of overall predictive power. This has important implications for Central Bank policy and research, most notably being, that ML models should form an increasingly important complementary role in its analytical toolkit for policy making. Secondly, ML models can offer unique or supplementary insights to important policy questions. In particular, the key finding of both approaches are that engagement with borrowers in distress at an early stage in their arrears is known to lead to the most favourable outcomes for both borrowers and banks.

The paper is organized as follows. Section 2 describes the relevant literature our paper relates to. Section 3 provides a background about the history of distressed mortgages in Ireland. Section 4 outlines the data used. Section 5 describes the different models used to predict SFS engagement and discusses the effects of certain variables on the probability of SFS engagement. Section 6 compares the accuracy of all models. Section 7 analyses the most important variables for predicting SFS engagement. Section 8 provides two main robustness checks. Finally, section 9 concludes.

2 Literature Review

Our paper broadly relates to several strands of literature. First, we contribute new insights to the literature on borrower distress and debt relief programs (e.g., Mian and Sufi, 2012; Berger et al., 2020; Hsu et al., 2018; Agarwal et al., 2017). Within this literature, our paper is closely related to work that examines the importance of institutional frictions in effective implementation of stabilization programs. In particular, focusing on the Home Affordable Modification Program, Agarwal et al. (2023) provide evidence that competition among intermediaries could impact effective implementation of such policies. Focusing on the Irish context, Kelly et al. (2021) analyse the financial situation of those households that engage via SFS. In contrast to this literature, we focus on the characteristics of borrowers in predicting their engagement with lenders and we examine the extensive margin, i.e. predicting those loans that are most likely to engage via SFS. We find that an early detection of those loans that are in arrears is the most important predictor of engagement.

Second, we are related to the literature that studies loans and their determinants with respect to mortgage modifications. Ganong and Noel (2020) find liquidity relief measures are more effective than principal reductions that increase wealth without affecting liquidity in distressed debt restructuring. In Danne et al. (2016), they find that current loan characteristics are the most important variables that increase the chances of receiving a permanent loan modification and making full payment after modification. In contrast, a high mortgage repayment to income ratio, higher household leverage, and higher household expenditure reduce the probability. We add to this literature a novel survival analysis of distressed-borrower-lender engagement, which is a necessary first step to receive a mortgage modification.

Lastly, we contribute to fast and growing literature of ML applications in topics relevant to financial stability. In particular, ML has become particularly useful in prediction exercises where large datasets are available. Closely related to our work, Barbaglia et al. (2020) use a dataset of 12 million residential mortgages to investigate the loan default behavior for different countries in Europe. They compare the predictive power across different ML models where the benchmark model is the logistic regression. They find that the ML methods perform significantly better and that the most important variables for predicting loan default are the interest rate, loan to value, and the local economic variables.[2] Fouliard et al. (2021) use an online machine learning to forecast financial crises out-of-sample. We contribute to

[2]See also Hoang and Wiegratz (2021) that compare ML with OLS in predicting real estate prices. They find that ML outperforms OLS where the main reasons are that they can better handle non-linearities and interactions.

this literature by applying Random Survival Forest (RSF) model to the question of SFS engagement. Furthermore, we compare the standard Cox model in survival analysis to a RFS ML model. We find that our ML model cannot only handle more variables, but also produces predictions that are more accurate than the standard Cox model.

3 Distressed Mortgages: Background

Ireland experienced a deep financial crisis between 2008 and 2013 and a record number of mortgage borrowers faced difficulties to pay their monthly loan payments. As shown in Figure 1, the share of loan accounts in arrears for more than 90 days reached its peak in September 2013 with 12.9 percent, which amounted to 18.810 billion Euro in outstanding mortgage balance. Even though this amount decreased over time, there is still a substantial share of 4.1 percent of loans in arrears for more than 90 days leading to 5.776 billion of oustanding mortgage balance. As of December 2022, there were 46,743 in mortgage arrears, and of this, 22,626 are in long-term mortgage arrears (greater than one year (Central Bank of Ireland, 2022b)). These loans remain a material issue for lenders, borrowers, and policymakers.

As a reaction to the severe economic and financial crisis in Ireland, the CCMA was first introduced in 2009 to provide statutory safeguards for vulnerable borrowers who were in arrears or at risk of falling into arrears. Over the years, the CCMA has been further strengthened, and the MARP has been designed to help people with mortgage arrears difficulties. The SFS is a crucial component of the MARP. The MARP involves gathering specific financial information from the borrower via the SFS to enable the regulated entity to assess the borrower's case and consider whether an alternative repayment arrangement (ARA) is suitable and sustainable for the borrower's individual circumstances. The SFS collects data such as the borrower's income, expenses, assets, and liabilities. Since 2012, the five main financial institutions in Ireland have been collecting this data. The goal of the SFS is that lenders can better understand the financial situations of their borrowers and develop more effective strategies to assist them in staying in their homes (Central Bank of Ireland, 2022a). 87% of those loans with modified loan contracts meet the terms of their current restructure arrangement (Central Bank of Ireland, 2022b). Therefore, the SFS engagement is an important first step for resolving non-performing loans.

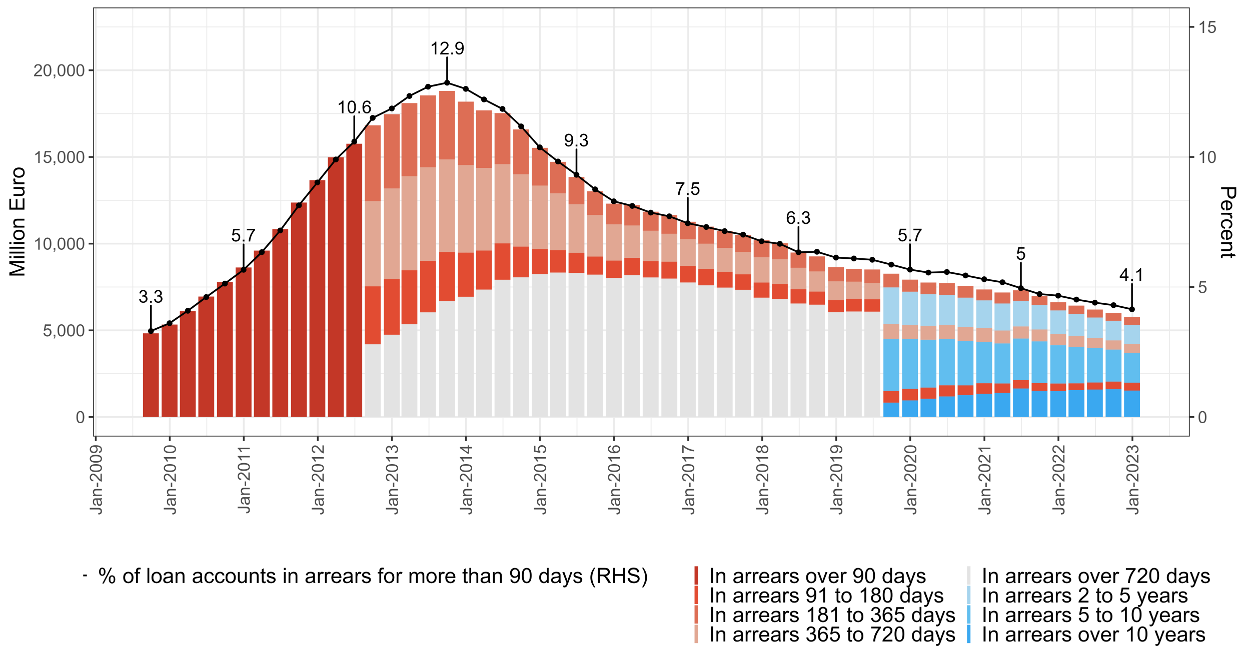


Figure 1: PDH Mortgage Accounts in Arrears over 90 Days

Note: The figure shows the share of loan accounts in arrears greater than 90 days (solid black line, right-hand-scale) together with the balance of mortgages (in million Euro). The breakdown of arrears greater than 90 days is not available pre-September 2012. The breakdown of arrears greater than 720 days is not available pre-September 2019.

4 Data

In the empirical analysis, we use two sources of data. First, we use loan level data from five Irish banks and focus on those loans that are in default (we define this as days-past-due > 90). Second, we join the loan level data with SFS fillings. This allows us to identify loans that engaged via SFS and also those that did not. As SFS fillings only started in 2012, we retrict our sample periods from 2012 H2 to 2021 H1.[3] Consistently, we only include loans

[3]We observe an exact date of the SFS filling which must not necessarily correspond to the loan level drop date. For this reason, we the date of the SFS filling to the nearest data drop date of the loan level data.

that have a maturity date after 2012. We remove outliers for numerical variables and focus on 25 variables that are described in Table 1.[4]

To build the data set suitable for survival analysis, we drop all those loans after an engagement has taken place. We acknowledge the fact that subsequent SFS engagement might be an interesting subset of observations to analyse as for those loan modifications have presumably not succeeded. However, for the purpose of this study, we limit observations on "first-engagers" only. As a consequence of our definition of engagement and the focus on the extensive margin, every loan in our dataset is observed at least twice.

Our sample contains a panel of 70,688 unique loans with a total number of observations of 170,933 over 9 years. All loans are in default with average length of time in arrears of 27 months and accumulated impairment of around 39 thousands Euros. Importantly, about 12% of defaulted loans in our sample has engaged with lenders via SFS. As majority of those loans are backed by properties that have experienced a significant fall in values since the origination, the loan-to-value (LTV) has increased from 0.68 at origination to 0.91 contemporaneously. The average interest rate that borrowers pay on their loan is 3% and two third of those loans are on variable rates.

Note that the lag change of installment corresponds to the difference in current installment compared to the first installment observed in our panel. Variations in installments usually stem from either a change in the interest rate or a change in the debt burden. A change in the debt burden, in turn, could follow from a loan modification where arrears are capitalized, and hence, the debt-repayment increases. In our model specifications, we will use the lag of this variable to ensure that such a loan modification would not occur at the same time as an SFS engagement to avoid reverse causality.

The loan data contains information on lender's name and geographic information of the collateral. These variables allow us to control fixed effects of lenders and regions. We also include information about the current payment type. Payment type can either be a

[4]Note that due to data confidentiality reasons, we do not display any bank-specific summary statistics or results. However, we use this information in all of our empirical specifications.

Table 1: Summary Statistics

Variable	Type	Unit	Mean/Mode	St.d.	Min.	Max.
Loan characteristics:						
SFS Engagement	Numeric	Share (%)	12.29	32.84	0	100
Term Months	Numeric	Months	322	83.97	12	653
Days-Past-Due (current)	Numeric	Days	820	560	90.01	2,910
Accumulated Impairment	Numeric	Euros (€)	38,858	44,695	0	182,981
LTV	Numeric	Ratio	0.91	0.51	0	2.57
LTV at origination	Numeric	Ratio	0.68	0.24	0.04	1.42
Current Interest Rate	Numeric	Rate (%)	0.03	0.01	0	0.09
Interest Rate Type at Origination		Share (%)				
Fixed	Categorical		32.77			
Other: Floating Index Based	Categorical		3.29			
SVR	Categorical		44.6			
Tracker	Categorical		19.35			
Lag Change in Installment	Numeric	Euros (€)	-1.2	167	-2,098	2,093
Current Payment Type	Categorical	Mode	1			
Loan Credit Quality	Categorical	Mode	5			
County Region of Collateral	Categorical	Mode	6			
Bankname	Categorical					
Borrower characteristics:						
First Time Buyer	Numeric	Share (%)	35.45	47.84	0	100
Number of Borrowers	Numeric	Number	1.32	0.49	0	14
Borrowers' Year of Birth	Numeric	Year	1969	9.1	1919	1992
Income at Origination	Numeric	Euros (€)	59,991	26,260	0	144,280
Macro variables:						
Regional Unemployment Rate	Numeric	Rate (%)	10.94	3.95	3.7	20.2
GDP Growth	Numeric	Rate (%)	6.41	6.75	-0.11	23.41
Inflation	Numeric	Rate (%)	0.42	0.54	-0.98	1.57
Residential House Price Index	Numeric	Units	97.59	20.73	75.6	143

Note:The table shows summary statistics of the variables considered in this analysis. Note that for categorical variables with more than four categories, we only report the mode (most frequent observation). The categorical variables with more than four categories are *Data Drop Date, Current Payment Type, Loan Credit Quality*, and *County Region of Collateral*. Note that *First Time Buyer* and *SFS engagement* are dummy variables (0/1). *Lag change in Installments* indicate the lag of the change in installments relative to the first observation in our panel. Due to data confidentiality reasons, we don't report individual bank names.

moratorium, interest only, less than interest only, full capital and interest, interest and part capital, a combination of interest and moratorium, or a mixed facility amortizing and interest only. The most common category is full capital and interest as payment type. The loan credit quality takes up 7 categories that range from upper good quality, good lower, watch upper, watch lower, impaired, more than 90 days-past-due and past due not impaired. The most common category are loans that are impaired.

In addition to the loan characteristics, the data also contains borrower-level information.

We have 35% of borrowers who are first-time-buyers and the average age of borrowers is 54 years (as of 2023). In terms of number of borrowers on each loan, we have 63.4% of single borrowers and the average borrowers' household total income is 59,991 Euros at origination.

To explicitly control for the effects of macro-financial environment on distressed loan engagement, we supplement the data with a range of macro variables, such as the regional unemployment rate, GDP growth, inflation and residential property prices.[5]

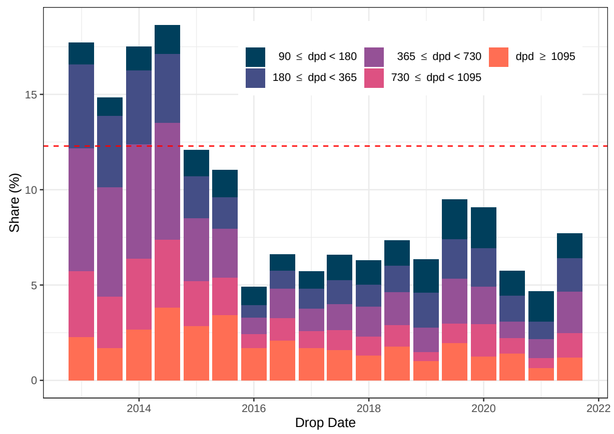
Figure 2 shows the share of loans that engaged via SFS over time and its breakdown into different buckets of days-past-due (DPD). As shown in the left panel, following the financial crisis in 2013, there was a big wave of SFS engagements, recording a high share of 17.71% SFS engagement at each data drop date. The share then falls to 7.72% in the latest drop in June 2021.[6] Overall, the average share of loans engaging via SFS amounts to 12.3%.

In the right hand panel of figure 2b, we plot the engagement share via SFS separately for each bucket of DPD. As an example, in the drop of June 2012, 19.8% of the loans that were less than 90 days-past-due engaged via SFS. Out of all loans at this drop date that were between 90 and 180 DPD, 28.5% engaged via SFS. The highest share of engagement per DPD-bucket is recorded for loans that are between 180 and 365 DPDs, followed by those loans that are in between 90 and 180 DPDs. On average, loans with days-past-due between 90 and 180 have the highest engagement share. Those loans in long-term arrears with DPD between 365 and 1095 have lower engagement rates. The lowest rates of engagement are recorded for loans that are above 1095 days-past-due.

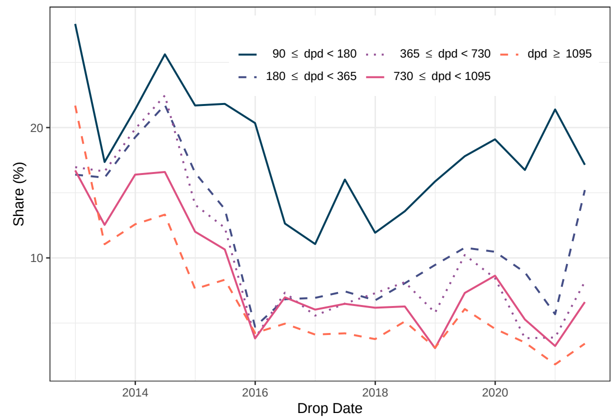
To motivate the survival analysis that we discuss in the next section, in Figure 3, we plot the number/share of SFS engagements conditional on a different time dimension. We define the "time in panel" as an increasing count variable that takes the value of 1 whenever we

[5] data source: Central Statistics Office (CSO) & Central Bank of Ireland

[6] Note that the first significant decline in the share of engagement in 2015 can be explained by quantitative targets in the MART framework introduced in March 2013. These targets were imposed on the six main mortgage lenders (which account for about 90 percent of the Irish mortgage market at that time). They focused on resolving accounts in default (arrears > 90 days) and were threefold, namely i) proposing sustainable solutions to borrowers, ii) concluding those sustainable solutions, and iii) subsequent performance rates on the concluded solutions. These targets were discontinued in 2014. Banks reported that they met all target requirements to end-2014 (Central Bank of Ireland 2016).



(a) Share of SFS Engagement by year and by DPD buckets



(b) Share of SFS engagement for each DPD bucket, by Year

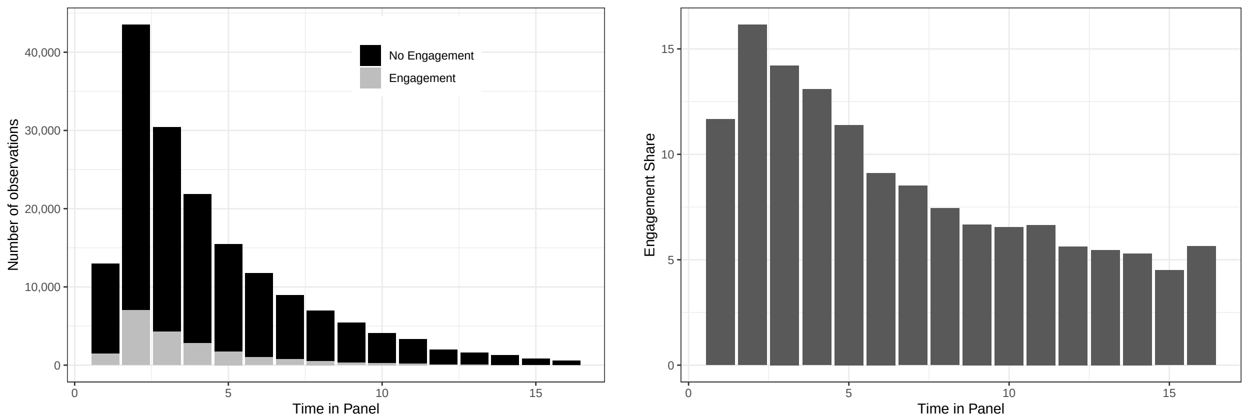
Figure 2: Share of SFS Engagement

Note: The figure shows the share of loans that engaged via a Standard Financial Statement (SFS) on calendar dates. In the left panel, the height of bars indicates the total share where the red line indicates the overall share of SFS engagement weighted by the number of observations in each drop. We also break down each bar conditional on the days-past-due (dpd) of the loans at the date. In the right panel, the figure calculates the mean share of those loans that engage via SFS for each DPD category individually.

first observe a loan in our sample with days-past-due $>$ 90 days. This time indicator informs about the time a loan spends in arrears.

In Panel 3a, the total number of observations by the time in the panel conditional on the SFS engagement is displayed. The number of observations decreases over time. The reasons for this decrease are mainly twofold. First, loans drop out of the panel because they engaged via SFS. Second, the loans drop out because they catch-up on their arrears. In this latter case, we do not observe an SFS engagement. In absolute terms, the most loans engage via SFS at the beginning of their time in the panel (with exception to the very first observation). The absolute number of engagements via SFS then decreases gradually over time.

In Panel 3b, we show the corresponding share of engagement via SFS over the time in the panel. The share is highest in the first five periods after the first time a loan was observed with days-past-due 90. The share of those loans that engage within five periods is 13.3%. After that, the mean share drops to 6.5%. In appendix A.1, we show the same figures but conditional on the drop date received.



(a) Obsevations by Engagement

(b) Engagement Share

Figure 3: Observations and Engagement over Time in Panel

Note: The Figure shows the total number of observations conditional on their time in the Panel. Panel 3a shows the total number of observations conditional on their time in the panel and their engagement. Panel 3b shows the share (in percent) of those loans that engage via SFS and those that do not, conditional on their time in the panel.

5 Predicting SFS engagement

The main purpose of this paper is to predict the SFS engagement of distressed loans comparing conventional empirical models and ML models. We also analyse what are the best predictors that predict loans that engage and those that do not. Learning about these key characteristics is policy relevant as getting borrowers to engage with lenders leads to more favourable outcomes for both borrowers and banks. In December 2022, 87% of those loans that engaged via SFS and received a loan modification have met the terms of their current restructure arrangement (Central Bank of Ireland, 2022b).

To predict SFS engagement, we use survival models. In an ideal case, our data should record each loan from the first day in default with $>$ 90 days-past-due days. After an SFS engagement, the loan drops out of our sample. Our models estimate the probability of SFS engagement conditional on the length of time past since the first default date. In reality, however, data can be subject to both "right" and "left" censorship problems. For the former case, the SFS engagements might not be observed, because our sample period is limited.

For those observations, we do not observe the actual outcome. In this respect, we have right-censored data. For the latter case, the first observation of a loan might pass the first dafault date. In such a case, we use the current DPD variable to inform us about the correct "time in panel" count.

In the following, we first apply a standard Cox Proportional-Hazards Model and then compare its performance to a random survival forest model.

5.1 Cox Model

5.1.1 Model Description

In this section, we present the Cox proportional hazards model (in short, Cox model) used to analyse those loans that engage via SFS. The Cox model is a standard statistic model in a survival analysis (see, for example, Blickle and Brown (2019) or Ampudia et al. (2021)) and is equipped to handle right-censored data. It allows us to model the baseline probability of engagement over the time a loan is observed in our sample, but also how this probability varies in response to explanatory covariates. Through this analysis, this model allows us to identify those variables that are the most important predictors of the SFS engagement.

While a complete technical description of the model is out of the scope of this paper, we provide a brief summary of the main idea of the model and let the reader refer to Cox (Cox, 1972) for more details. The Cox model is a semi-parametric model, meaning that it does not make assumptions about the shape of the underlying hazard function, instead it models the effect of the predictor variables on the hazard ratio. The hazard is defined as the likelihood of an event occurring during any given time point, given that it has not occured before that time. In our case, the risk of an event is the SFS engagement.

The hazard function in Cox hazard models is modeled as a function of the covariates X

(which are variables that might affect the hazard rate) and a set of coefficients, β:

$$h(t|X) = h_0(t)exp(X\beta) \tag{1}$$

where $h_0(t)$ is the baseline hazard function, which represents the hazard likelihood rate for a reference group. This function represents the likelihood of a failure in the absence of any predictor variables. This failure, in our case, refers to a non-engagement via SFS. $exp(XB)$ is the hazard ratio, which represents the effect of the predictor variables on the hazard.

This hazard ratio is a multiplicative factor that modifies the baseline hazard function. Therefore, if the hazard ratio is greater than 1, the hazard increases, and if the hazard ratio is less than 1, the hazard decreases. To estimate the parameters in the model, partial likelihood is used. The partial likelihood estimation allows to use information of all observations, including the right-censored ones.

Variable t in the model refers to the time stamp "time in panel". In the normal cases, we define it according to the drop date of a loan l in the sample, i.e. if the first time a loan is observed in the dataset and the days-past-due of the loan is in between 90 and 180 days, as we observe loans every 6 months, we set "time in panel" $t = 1$. In the "left censored" cases, however, those loans that we observe the first time in the panel have already had a days-past-due over 180 days. We then set $t = 2$ at first observation if $180 < dpd \leq 270$, $t = 3$ if $270 < dpd <= 360$, and so on.

An important assumption of the Cox model is the proportional hazard assumption - this means that the effect of a variable on the hazard is constant over time.[7]

To account model uncertainty, we consider two Cox models to analyse the question about the probability of SFS engagement. We include the following variables: *days-past-due, lag change in installments, age, LTV, LTV at origination, income at origination, first time buyer, current interest rate, current payment type, time since origination, residential real estate prices, regional unemployment rate, bank* and *date fixed effects.*

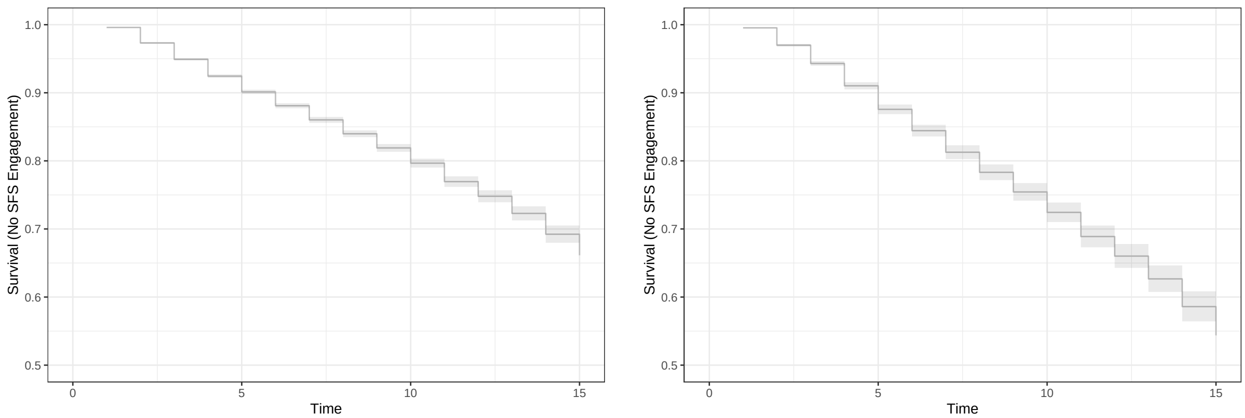
[7]This assumption, however, can be circumvented by interacting variables with a time indicator.

In the first specification, we exclude date fixed effects but include bank dummies and the current payment type information. The second specification is similar to the first but without the bank dummies and current payment type information, but with date fixed effects.The motivation behind two different model specifications stems from convergence challenges specific encountered with our dataset at hand. By employing two distinct Cox models, we can explore different variable selection techniques to mitigate convergence issues. This approach allows us to enhance the robustness of our findings, and provide a comprehensive analysis of SFS engagement.

5.1.2 Cox model results

In Figure 4, we plot the estimated survival rates for the two Cox Models together with the 95% confidence interval. The survival rate is mathematically related to the hazard function described in section 5.1.1 Equation 1. In fact, the hazard function describes how fast the survival function decreases over time.

We observe that the probability to engage via SFS varies slightly conditional on the model estimated as do the confidence intervals. While in model (1), after 15 periods in the panel 33.86% engaged via SFS, this share is 45.63% in model (2).



(a) Model (1)

(b) Model (2)

Figure 4: Survival Rate

Note: The Figure shows the estimated Survival Rate of two Cox models with the 95% confidence interval. The higher the survival, the lower the probability to engage via SFS. Mortality, in our model, refers to the probability that a loan engages via SFS.

Table 2 shows the results of the two estimated Cox models (standard errors clustered at the loan level). Most variables in the Cox model have a highly significant effect on the survival rate. The coefficients tell us how the change in the hazard rate, which is in our case the likelihood of SFS engagement, is associated with covariates. In more detail, the hazard rate describes the instantaneous rate of occurrence of the event at a particular point in time, given that the event has not occurred up to that point in time. An increase in hazard rates is therefore not necessarily the same as an increase in the probability, where the latter depends not only on the hazard rate, but also on other factors such as the length of the follow-up period and the baseline risk. It follows that a positive coefficient indicates an increase in the risk of engaging (i.e., an increase in the likelihood with respect to the hazard). A negative coefficient, in turn, shows that the likelihood of SFS engagement decreases.

Table 2: Cox Model - Results

	Dependent variable:	
	SFS Engagement	
	(1)	(2)
Days-Past-Due	-0.001^{***}	-0.001^{***}
	(0.00002)	(0.00002)
Lag Change in Installment	0.0003^{***}	0.0002^{***}
	(0.00005)	(0.00004)
First Time Buyer	-0.040^{**}	-0.130^{***}
	(0.017)	(0.016)
Age	-0.002	-0.007^{***}
	(0.001)	(0.001)
Loan-To-Value	-0.018	0.060^{**}
	(0.023)	(0.024)
Loan-To-Value at Origination	-0.301^{***}	-0.374^{***}
	(0.038)	(0.039)
Income at Origination	-0.00000^{***}	-0.00001^{***}
	(0.00000)	(0.00000)
PT: Interest and Part Capital	0.777^{***}	
	(0.090)	
PT: Interst Only	0.782^{***}	
	(0.028)	
PT: Less than Interest Only	0.310^{***}	
	(0.052)	
PT: Moratorium	-0.572^{***}	
	(0.046)	
PT: Other	0.183	
	(0.251)	
Time since Issuance (0/1)	-0.995^{***}	-1.022^{***}
	(0.051)	(0.051)
Residential Real Estate Prices	-0.022^{***}	-0.009^{***}
	(0.001)	(0.001)
Regional Unemployment	-0.005	0.009^{**}
	(0.003)	(0.003)
Current Interest Rate	2.285^{***}	
	(0.516)	
Time FE	No	Yes
Bank FE	Yes	No
Observations	170,933	170,933
R^2	0.101	0.102
Log Likelihood	$-230{,}687.900$	$-230{,}537.700$

Note: This table shows the partial effects estimated from two different Cox models. The dependant variable is the binary outcome of SFS engagement. PT stands for payment type. Time since issuance is a continous variable which takes the value 0 if the loan is observed at origination date and value 1 if it is observed at maturity date. *p<0.1; **p<0.05; ***p<0.01

First, on the micro level, we find that days-past-due (current) has a negative coefficient, implying that higher values of days-past-due reduce the likelihood of SFS engagement. To be precise, the coefficient indicates a 0.001 decrease in the expected log of the relative hazard of SFS engagement for each one unit increase in days-past-due (current), holding the other predictors constant. The hazard ratio equals $exp(-0.001) = 0.999$, which translates to an approximate decrease of 0.1% in the expected hazard of SFS engagement relative to a one day increase in days-past-due, holding all other predictors constant.

Furthermore, we find also a negative coefficient for the time passed since issuance. A loan that is closer to its origination is therefore associated with a higher likelihood of SFS engagement. There is an incentive for banks to make borrowers engage earlier when the loan is in its early stage. This allows them come up with a sustainable solution that is agreeable to both borrower and lender that increases the probability of meeting the new repayment arrangement over the rest of the loan term.

Next, we find that the lag of a change in the installment increases the likelihood of SFS engagement. A higher positive change in installment occurs either due to higher interest rates or higher debt burden. In our case, the change is likely to be due to an increase in the debt burden as we separately control for the interest rate change and more importantly interest rates have remained stable and at a low level during our observation period from 2012 to 2021. Thus, an increase in indebtedness appears to increase the chance of an engagement via SFS.

Lastly, being a first-time-buyer (FTB) decreases the likelihood of engagement. Compared to those households that have already owned more than 1 home, FTBs are less experienced with the option to engage with a bank and may also encounter some sort of stigma for engaging via SFS, potentially decreasing their likelihood of engagement.

Regarding the effects of macroeconomic variables, we find several interesting results. First, we find mixed signs for the two Cox models with respect to unemployment. Note that some of the effect of the macroeconomic variables is absorbed by the date fixed effects present in model (2), which might explain the opposite signs of the effect. We will further discuss these results when we look at the results of the machine learning model. Second, an increase in residential real estate prices leads to a decrease in the likelihood of SFS engagement. Indeed, a higher house price reduces the incentive of SFS engagement as the higher house price improves the home equity position of borrowers and might be also associated with more optimistic outlook of borrowers on their financial constraints. Third, we find that the current interest rate is positively associated with the likelihood of SFS engagement in Model 1. This finding is of particular interest in the current high inflation environment where interest rates have been increased subsequently.

Overall, the Cox model results point towards two aspects that potentially influence the probability of SFS engagement. First, the results for days-past-due and time since issuance underline that it is important for the lender to identify borrowers in financial difficulties and engage with them at an early stage. This increases the probability of finding a solution to the problem. Second, the results with respect to interest rate and the fact that macroeconomic variables seem to be important too might point toward the effect of current economic conditions on the probability of engagement. Potentially, when economic conditions are unfavorable, i.e. high interest rates and/or high unemployment, the probability that loans fall into arrears is higher, and therefore also SFS engagement.

5.2 Random Survival Forest Model

5.2.1 Model Description

In this section, we develop a ML model for the same survival analysis of the SFS engagement. In particular, we apply a Random Survival Forest (RSF) model as described in Ishwaran et al. (2021). The time indicator t is the same as for the Cox model and is set equal to 1 at first observation in our sample. δ is the censored variable that indicates whether a loan was engaging via SFS or not. The observed time t is defined as the minimum of the true survival event time t^0 and the true censoring time c^0. For some observations, we do not observe any engagement over the entire sample period. Either they did not want to engage or our time horizon is not long enough. Therefore, $t = min(t^0, c^0)$.

The right-censoring indicator is defined by $\delta = I\{t^0 \leq c^0\}$. Whenever $\delta = 1$, we observe an engagement via SFS. In this case, we observe the true event time $t = t^0$. If $\delta = 0$, the observation is right-censored and we only observe the censoring time $t = c^0$.

The RSF model takes this censoring into account when splitting the tree node into left and right subtrees. More specifically, a log-rank splitting rule is used to determine the split. The log-rank split-statistic is a measure of node separation with larger values indicating a greater survival difference between two nodes. Therefore, the higher the statistic, the better the split conditional on a specific characteristic X. The goal is to find the value of X for which the split-statistic is maximized.

The Log-rank test serves as the foundation for the Log-rank splitting criterion. This test is designed to compare the survival distributions of two or more groups. Within the context of a random survival forest, the Log-rank splitting criterion operates as follows.

To summarise the idea of log-rank splitting, let's consider the example of Ishwaran et al. (2021) and take a specific tree node for splitting, assuming it is the root node, for simplicity. We'll also assume non-bootstrapped data, so the root node's data is represented as $(T_1, X_1, \delta_1), \ldots, (T_n, X_n, \delta_n)$. Here, X denotes a specific variable (one of the feature vector coordinates, for example, days-past-due). A proposed split using X takes the form $X \leq c$ and $X > c$ (assuming nominal X), dividing the node into left and right daughter nodes, $L = \{X_i \leq c\}$ and $R = \{X_i > c\}$, respectively.

Consider the distinct times of SFS engagement as $t_1 < t_2 < \ldots < t_m$, and let $d_{j.L}$, $d_{j,R}$, and $Y_{j,L}$, $Y_{j,R}$ represent the number of SFS engagements and individuals at chance of SFS engagement at time t_j in the daughter nodes L and R. The "at chance" concept accounts for the number of individuals in a daughter node who have not engaged via SFS at time t_j or have an event (SFS engagement) at time t_j:

$$Y_{j,L} = \#\{T_i \geq t_j, X_i \leq c\}, \quad Y_{j,R} = \#\{T_i \geq t_j, X_i > c\}.$$

Define $Y_j = Y_{j,L} + Y_{j,R}$ and $d_j = d_{j,L} + d_{j,R}$. The log-rank split-statistic value for the split can be calculated as follows:

$$L(X, c) = \frac{\sum_{j=1}^{m}(d_{j,L} - Y_{j,L}d_j/Y_j)}{\sqrt{\sum_{j=1}^{m}(Y_{i,L}/Y_j)(1 - Y_{j,L}/Y_j)((Y_j - d_j)/(Y_j - 1))d_j}}.$$

The value $|L(X, c)|$ serves as a measure of node separation. A larger value indicates a more substantial difference in survival outcomes between L and R, signifying a more favorable split. The best split is determined by identifying the feature X^* and split-value c^* such that $|L(X^*, c^*)| \geq |L(X, c)|$ for all X and c.

Overall, the RSF allows for more flexibility with respect to the variables that we include. In contrast to Cox regression models, ML model is able to handle large amount of independent variable. As a result, we stay agnostic to the variables choice and add additional variables that might be of interest, such as bank-time fixed effects.

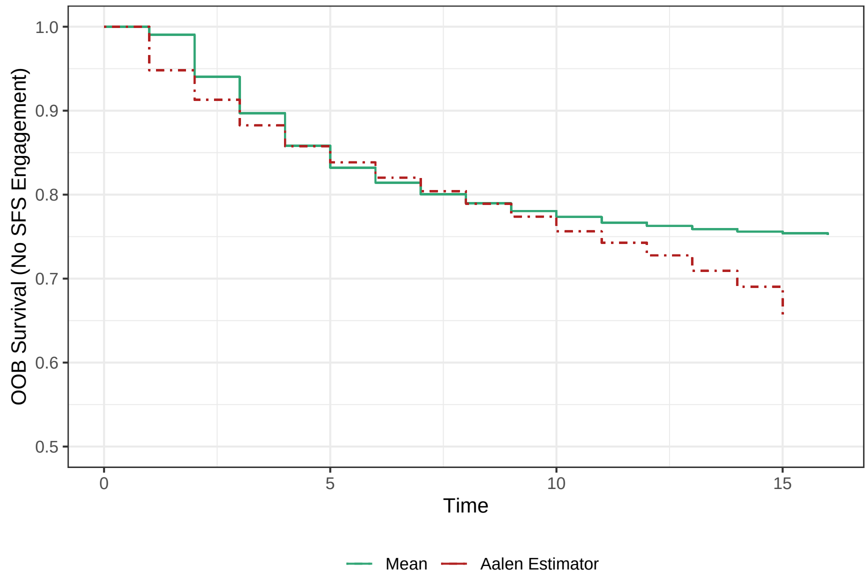


Figure 5: Out-Of-Bag Survival

Note: The figure plots the Out-Of-Bag (OOB) Survival Rate for the overall ensemble survival and the Nelson-Aalen estimator. At every step, the survival rate shows the probability of an SFS engagement where one step corresponds to the number of drop dates a loan is observed in our sample.

5.2.2 RSF results

Figure 5 plots the mean out-of-bag (OOB)[8] survival rate and an alternative estimator, the Nelson-Aalen estimator.[9] Compared to the standard Cox Model in section 5.1.2, the survival rate decreases faster at the beginning and shows a flatter survival curve towards the end of the time a loan is in the panel. Quantitatively, the results are closer to the second model estimated in 5.1.2. Potentially, the RSF model allows to better capture non-linearities and complex interactions present in the data. However, the final survival rate of approximately 70% (corresponding to 30% of the observations having engaged) at time period 15 is very similar when comparing the Nelson-Aalen estimator of the RSF to the standard Cox model.

[8]Note that OOB refers to the hold-out sets in the decision trees. Each tree out of the 601 trees in our random forest is estimated with a certain test sample and a hold-out sample. This hold-out sample was not used during training. OOB refers to the measures that were calculated using these hold-out sets. For example, the error rate was calculated predicting the SFS engagement for a loan in the hold-out set comparing it with the actual observed outcome in the hold-out set.

[9]Note that the Nelson-Aalen estimator is a statistical method commonly used in Survival Analysis with ML. Here, we use it as an alternative, non-parametric estimator in comparison to the overall ensemble survival.

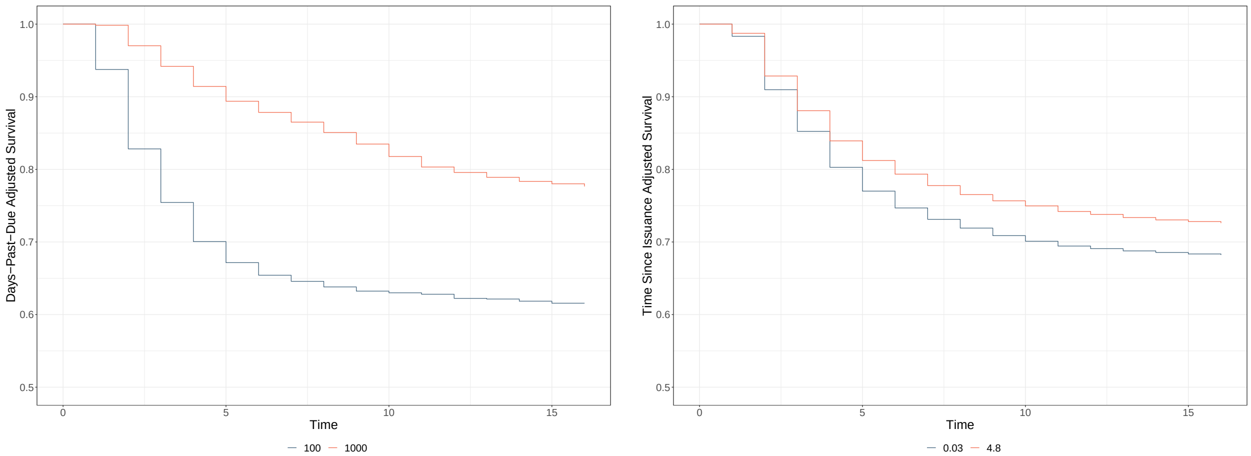
Partial effects: To compare the estimation results of RSF model to those of the Cox model, we construct so-called partial dependence plots (PDP). These plots were introduced by Friedman (2001) and explore the relationship between the right-hand variables and the outcome in supervised ML models. To create PDPs, the feature of interest is isolated by fixing its value while keeping other features unchanged. Predictions are generated based on this modified dataset. The selected feature is systematically varied, and the corresponding outcomes are predicted. These results can then be plotted, with the feature values on the x-axis and predicted outcomes on the y-axis.

In the main text, we show the partial effect plots for two exemplary variables in Figure 6. The figure shows the partial effects evaluated at the first and third quantile for each variable, respectively.

Panel 6a depicts the partial effect of having days-past-due at 100 days compared to days-past-due at 1000 days on the survival rate to engage via SFS. The faster a survival curve decreases, the more likely a loan with the corresponding characteristics is to engage via SFS. Therefore, from Panel 6a, consistent with the finding in Cox models, loans with lower days-past-due are more likely to engage via SFS.

Panel 6b shows that loans are more likely to engage via SFS when they are relatively closer to their origination date compared to their maturity date. The time since issuance variable is a ratio between 0 and 1 where 0 refers to the date of origination and 1 to the date of maturity. Values in between correspond to the number of days passed since their origination date when observed in the sample.

We summarise the rest of the results in Table 3 and compare them to the Cox model results. For additional partial effect plots of the random survival forest model we let the reader refer to appendix A.2.



(a) Days-Past-Due

(b) Time Since Issuance

Figure 6: Partial Effects of Selected Variables - I

Note: The Figure shows the partial effects of different variables on the probability to engage via SFS for different values of the respective variable. In Panel 6a, the partial effect of days-past-due is displayed. Panel 6b shows the partial effect of the time since issuance, which takes the value 0 if the loan is observed at origination date and value 1 if it is observed at maturity date. Every value in between corresponds to the number of days relative to total number of days of the entire loanspan time at observation date.

We find similar results for the RSF and Cox models with respect to the effect of the current interest rate. Namely, loans with a relatively higher interest rate tend to have a higher probability of engagement. Given the current high inflation rate environment and the subsequent rise in interest rates, it is likely that the number of loans engaging via SFS will increase in the near future.

All models point towards a similar positive effect of the lag of change in installments on the probability to engage via SFS. As previously mentioned, an increase in installment payments can be attributed to either elevated interest rates or an increased debt burden. In our specific situation, it is probable that the change is a result of a rising debt burden. This assertion is supported by our ability to isolate and account for changes in interest rates independently. Furthermore, it is noteworthy that interest rates have remained stable and at a low level throughout the duration of our observation period.

Consistent with the Cox models, residential real estate prices have a consistent negative effect on the probability of SFS engagement across all models estimated. It is likely that homeowners in distress in regions with higher real estate prices assume that potential arrears are covered by an increase in the value of their home, in case of a default. Therefore, they might have a lower incentive to engage via SFS compared to those homeowners where residential real estate prices have stayed constant or decreased.

For the unemployment rate, we find mixed results across the three models. However, in Cox model 1, the (negative) effect is non-significant whereas it is positive and significant in Cox model 2 and consistent with the result of the RSF. The interpretation of a higher probability of engagement with an increase in the unemployment rate is also consistent with our other macroeconomic indicators. If macroeconomic conditions worsen, households are more likely to engage via SFS.

Age, being a first time buyer, LTV and income at origination all have a consistent negative effect on the probability to engage via SFS. Potentially, elderly households are less aware of the possibility to engage via SFS, and so are first time buyers compared to buyers that have already purchased a home previously. While LTV and income at origination have a negative effect, its interpretation is more difficult as those variables might have changed considerably across our observation periods. The different partial effects of the current LTV (negative and positive in the two Cox models, positive in the RSF) underline that difficulty.

Overall, the results are similar to the findings of the Cox model (2) with respect to those variable where we find different signs of the coefficients between Cox model (1) and (2). In general, earlier identification of loans in arrears tends to increase the probability of SFS engagement. Moreover, loans in regions with more economic difficulties (or lender difficulties), such as lower house prices, higher LTV, and higher interest rates, or higher unemployment, increase the probability of an SFS engagement.

Table 3: Comparison Partial Effects between Models

	Cox 1	Cox 2	RSF
Days-Past-Due	$-^{***}$	$-^{***}$	$-$
Time since Issuance	$-^{***}$	$-^{***}$	$-$
Lag Change in Installments	$+^{***}$	$+^{***}$	$+$
Current Interest Rate	$+^{***}$		$+$
Residential Real Estate Prices	$-^{***}$	$-^{***}$	$-$
Regional Unemployment	$-$	$+^{**}$	$+$
Age	$-$	$-^{***}$	$-$
First Time Buyer	$-^{**}$	$-^{***}$	$-$
LTV	$-$	$+^{**}$	$+$
LTV at Origination	$-^{***}$	$-^{***}$	$-$
Income at Origination	$-^{***}$	$-^{***}$	$-$

Note: This table shows the partial effects estimated from the two different Cox models in section 5.1 and the Random Survival Forest (RSF) model in section 5.2. The dependant variable is the binary outcome of SFS engagement. Time since issuance is a continous variable which takes the value 0 if the loan is observed at origination date and value 1 if it is observed at maturity date. *p<0.1; **p<0.05; ***p<0.01

6 Prediction accuracy

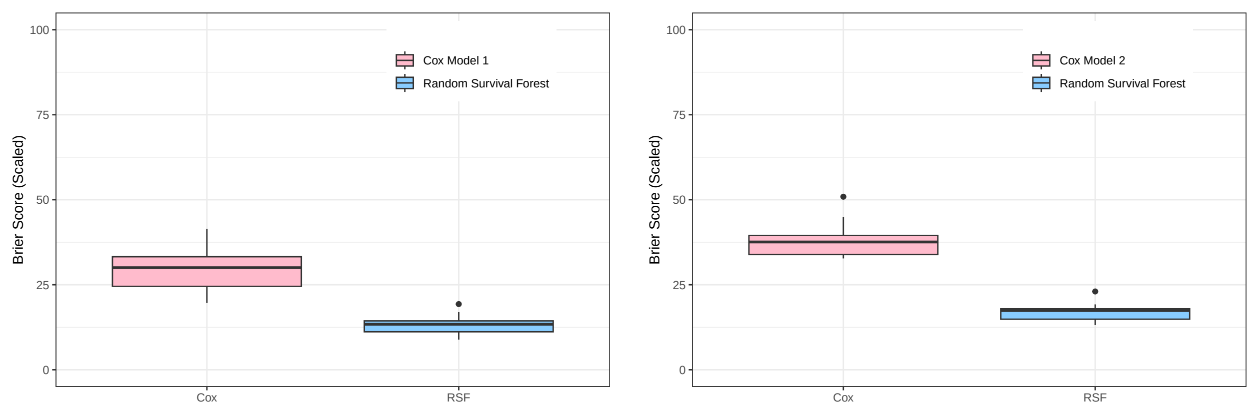
In this section, we compare the precision of the RSF model to the Cox models. To evaluate the prediction accuracy across these models we use the Brier Score, a commonly used metric in ML to evaluate the accuracy of probabilistic prediction. The Brier Score modifies the prediction error curve (PEC) to account for the inverse probability censoring weighting to evaluate the prediction performance of the survival models. Formally, we calculate

$$BS(t^*) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(\hat{S}(t^*|z_i))^2}{\hat{G}(X_i)} \cdot I(X_i < t^*, \delta_i = 1) + \frac{(1 - \hat{S}(t^*|z_i))^2}{\hat{G}(t^*)} \cdot I(X_i \geq t^*) \right],$$

where t^* is the time point at which BS is to be calculated, N is the sample size, z_i are the covariates corresponding to the sample i, X_i the observed survival time, $\hat{S}(\cdot)$ is the survival function predicted by the different models, $\hat{G}(\cdot)$ corresponds to the weight for the instance which is estimated by the KM (Kaplan–Meier) estimator of the censoring distribution (Zhou et al., 2022). The lower the Brier Score, the better is the models' performance. Note that the Brier Score can somewhat depend on the choice of t^*. We chose it to be the median. We scale the Brier score such that it is in between 0 and 100. A value of 100 corresponds to a model that is performing similar to a simple coin toss. A value of 0 means perfect prediction power.[10] In essential, the Brier score helps to assess the reliability of our probabilistic predictions in a quantitative manner.

The results are displayed in Figure 7.[11] We find that the RSF model outperforms both versions of the Cox model. It has a significantly lower Brier Score, and hence a higher precision in predicting the survival rates of loans engaging via SFS. Despite, we find that model (1) performs better than model (2). Also, all models perform significantly better than a coin toss.

Furthermore, we find that the variance of the Brier Score is lower for the RSF model. Over different cross-folds of the data, the RSF yields a similar performance whereas the Cox model results are slightly more dispersed.



(a) Cox Model (1) versus Random Forest (b) Cox Model (2) versus Random Forest

Figure 7: Precision using Brier Score

Note: The figure shows the Brier Score calculated for the median t^* for the two different Cox models as well as the Random Survival Model. The lower the Brier Score is, the more precise are the predictions of the survival rates. If a model scores a value of 100, it is no better than a coin toss. A value of 0 corresponds to perfect prediction power.

[10] In appendix B.4, we show another accuracy measure for the RSF, and show the results of the RSF finetuning on the out-of-bag error rate.

[11] Note that, due to computational limitations, we use a subsample of 50% of the data for the performance test.

7 Best Predictors

Cox Models. To identify the best predictors of loans that best predict the engagement via SFS, we conduct an Analysis of Variance (ANOVA). The ANOVA measures how much the explained variance of the dependent measure varies conditional on each variable. Table 4 shows the results for the Cox-regressions.

For Model (1), Days-Past-Due, bank, residential real estate prices, bank fixed effects, the current payment type and the time since issuance are among the 5 most important variables. For model (2) as well, Days-Past-Due is the most important variable and residential estate prices and time since issuance are among the top 5 predictors of SFS engagement. However, unemployment seems to play a lesser role in the Cox regression models, despite the significant effect in the regression results of the Cox models.

Table 4: Best Predictors Cox Models

Rank	Model (1)	χ^2	p-val.	Rank	Model (2)	χ^2	p-val.
1	Days-Past-Due	4998.77	0.00	1	Days-Past-Due	5295.22	0.00
2	Bank	2168.61	0.00	2	Income at Origination	524.19	0.00
3	Current Payment Type	1082.64	0.00	3	Residential Estate Prices	504.5	0.00
4	Residential Estate Prices	868.05	0.00	4	Time since Issuance	330	0.00
5	Time since Issuance	374.00	0.00	5	Date FE	193.72	0.00
6	Income at Origination	116.64	0.00	6	LTV at Origination	149.69	0.00
7	LTV at Origination	64.46	0.00	7	First Time Buyer	62.14	0.00
8	Lag Change in Installment	48.24	0.00	8	Lag Change in Installment	36.43	0.00
9	Current Interest Rate	19.58	0.00	9	Age	28.26	0.00
10	First Time Buyer	5.51	0.02	10	LTV	20.27	0.00
11	Age	2.64	0.10	11	Current Interest Rate	2.95	0.09
12	Unemployment	2.41	0.12	12	Unemployment	0.8	0.37
13	LTV	0.61	0.43				

Note: The table shows the most important variables for both Cox models estimated in section 5.1.2 using an Analysis of Variance. It reports the χ^2 test statistic and the p-value.

Random Survival Forest. The Random Survival Forest model allows to identify the most important variables for predicting the SFS engagement. In Figure 8, we plot the variable importance where larger values on the y-axis indicate a higher importance for predicting SFS engagement. We use 200 subsamples to calculate the variable importance.[12]

[12]The calculation of the most important variables is robust to different values of the subsample size.

To calculate the importance of a variable, the out-of-bag sample of the RSF model is used. Each tree is estimated only with a subsample of the data where usually 37% of the data is classified as hold-out data. With this hold-out data set, the prediction error can be estimated. To infer the variable importance from this hold-out dataset, a specific variable is permuted. Permutation with respect to a single variable means that the values of this variable are randomly shuffled. With this new feature, the prediction error is calculated. If a variable is important, then the prediction error of the model with original feature is lower than the prediction error with the randomly shuffled variable.

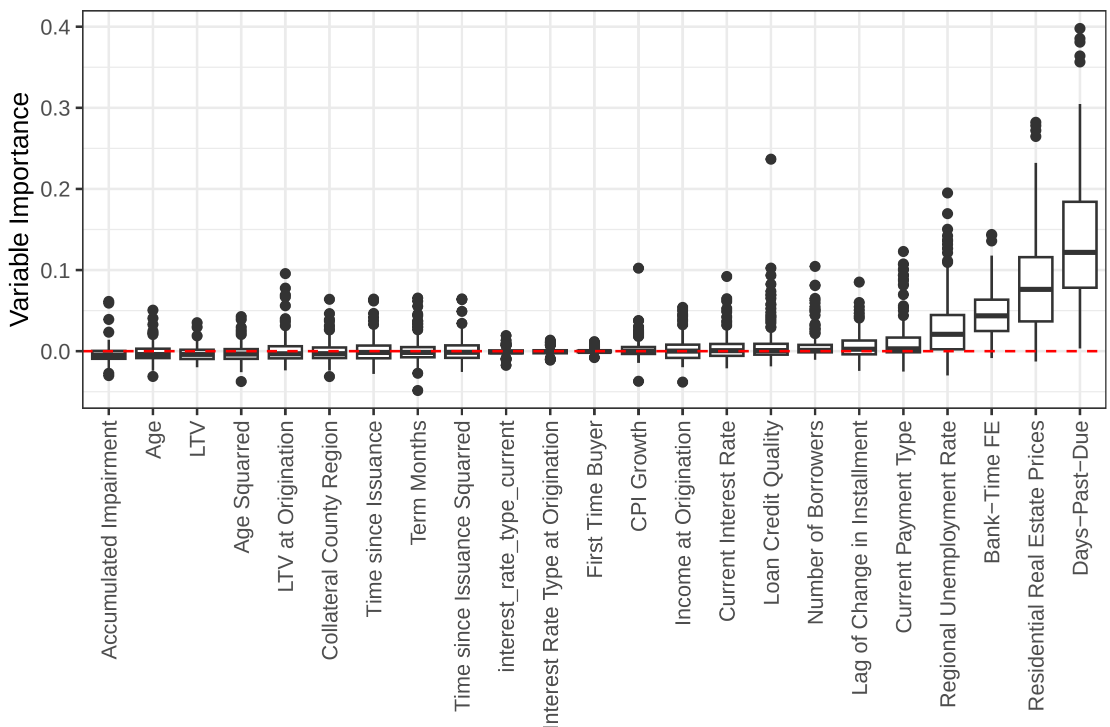


Figure 8: Most Important Variables identified by RF Survival Model

Note: The Figure shows the most important vairables identified by the Random Survival Forest Model. A value greater than zero indicates that the variable is important for predicting the SFS engagement. We used 200 subsamples of our data to estimate the importance of each variables. Larger values indicate a higher importance for predicting SFS engagement.

Similar to the Cox models, we also find that for the RSF model, Days-Past-Due is the most important variable. This variable is followed by the residential real estate prices, which is similar to the Cox model (1). Furthermore, bank time fixed effects are important predictors for the SFS engagement, which corresponds to the results of the two Cox model that find that the bank indicator but also date fixed effects are important variables. Furthermore, the unemployment rate plays an important role and so does the lag change in installment.

In contrast to both Cox models, the RSF model ranks the regional unemployment rate as an important indicator. Potentially, the relationship of regional unemployment and engagement is non-linear or features complex interactions that only the RSF model manages to identify.

In summary, the findings of the most important variables across models are generally consistent. An exception is the regional unemployment rate which is only identified as an important predictor in the RSF model. The consistent result that Days-Past-Due is the most important variable across all models suggests that an early detection of loans in arrears is crucial for a high probability of engagement, and in turn, is likely to lead to more favourable outcome for both borrowers and banks.

8 Robustness Checks

In this section, we conduct two robustness checks. First, we use a subsample restricting the data to observations after 2016 to check whether our results are driven by the MART targets. Second, we use an alternative precision measure to compare the accuracy of the RSF with the Cox models.

The share of engagement displayed in Figure 2a features a decrease in 2015. This decrease is likely to be be explained by the quantitative targets in the MART framework that were first introduced in March 2013 but discontinued in 2014. These targets focused on proposing and implementing sustainable solutions for borrowers, as well as maintaining good performance rates for those solutions. The targets were discontinued in 2014, and the banks reported that they had met all the requirements by the end of that year (Central Bank of Ireland, 2016).

We estimate all of our main results using data only after 2016 to analyse whether our results are solely driven by the MART targets. The results are displayed in appendix B.1. Overall, the results are very similar with respect to the most important predictors, as well as the accuracy. The RSF model outperforms the Cox model in terms of overall predictive power.

An exception is that the unemployment rate is less important, most likely due to the better macroeconomic environment after 2016. Also, the lag of change in installment seem to be more important in the Cox models for the sample after 2016. While this variable is more important in the RSF estimated over the entire sample, the lag of change in installment is an important variable across all models in the sample after 2016. This result is consistent with the observation of Byrne et al. (2022) who found a similar variable to be an important predictor of mortgage default.

As a second robustness check, we use an alternative precision measure to compare the accuracy of the RSF to the two Cox models. We use the precision measure called Concordance Index (CI) adapted to the Survival Framework as suggested in Ishwaran et al. (2021) . The CI index is similar to the commonly used receiver operating characteristic (ROC) curve and evaluates the models' performance at all classification thresholds. The higher the CI index is, the higher the accuracy of the model. We scale the index to be in between 0 and 100, where 100 corresponds to a perfect prediction accuracy. The results are displayed in appendix B.2. The results are similar, i.e. the RSF outperforms both Cox models.

9 Conclusion

This paper demonstrates the value of machine learning techniques in predicting borrower engagement in the mortgage arrears resolution process. By combining bank-level loan data with filings from the Standard Financial Statement in Ireland, we were able to identify loans in arrears that engage with their lenders to start debt renegotiation. Our comparison of standard regression techniques to machine learning methods shows that machine learning outperforms standard regression tools and has advantages in handling complex features of our dataset.

Our findings suggest that early identification of distressed loans is crucial for engagement. This is especially relevant given the current cost-of-living pressures borrowers are facing due to inflationary pressures, the lingering effects of the COVID-19 pandemic, and a slowdown of major economies. As such, our work provides valuable insights for policymakers and financial institutions looking to assist borrowers in financial difficulty.

We find that the number of days a borrower spends in arrears to be the most important variable in determining engagement, along with a borrower's installment changes and general macroeconomic conditions. Importantly, evidence from Ireland's financial crisis years (2008-2013) shows that getting borrowers to engage early in their arrears cycle leads to more favorable outcomes for both borrowers and banks.

The comparison between the Cox regression model and the Random Survival Forest (RSF) model highlights the importance of machine learning models for financial stability policy and research. While both models identify important characteristics that predict borrower propensity to engage via the Standard Financial Statement, we find that the RSF model outperforms the Cox model in terms of overall predictive power. This finding underscores the potential for machine learning models to offer new or complimentary insights to important policy questions.

The data shows that loans that went through modifications after SFS engagement meet their new requirements in 87% of the cases, leading to more favorable outcomes for both the lenders and borrowers. Our results ultimately suggest that early engagement is key to resolving distressed loans and avoiding the negative outcomes associated with non-engagement.

As a key finding of this paper in terms of identifying characteristics that predict whether a borrower will engage or otherwise, the number of days in arrears (day-past-due) is identified as the most important variable, along with general macroeconomic conditions. In light of the current macroeconomic uncertainty, getting borrowers to engage early in their arrears cycle is an important consideration and evidence from Ireland's financial crisis years (2008-2013) shows that this will lead to more favourable outcomes for both borrowers and banks. For engaging borrowers, there is an increased likelihood that an alternative solution to repossession (inside or outside the legal system) will be found. In contrast, for borrowers who do not engage the loss of ownership through repossession will remain. In the extreme cases of long-term non-cooperation, the functioning of the legal system to ensure the realisation of collateral for lenders will continue to be critical to the effective functioning of the mortgage market for all Irish citizens, not just those in long-term mortgage arrears.

References

Agarwal, S., G. Amromin, I. Ben-David, S. Chomsisengphet, T. Piskorski, and A. Seru (2017): "Policy intervention in debt renegotiation: Evidence from the home affordable modification program," *Journal of Political Economy*, 125, 654–712.

Agarwal, S., G. Amromin, S. Chomsisengphet, T. Landvoigt, T. Piskorski, A. Seru, and V. Yao (2023): "Mortgage refinancing, consumer spending, and competition: Evidence from the home affordable refinance program," *The Review of Economic Studies*, 90, 499–537.

Ampudia, M., M. Lo Duca, M. Farkas, G. Perez-Quiros, G. Rünstler, and E. Tereanu (2021): "On the effectiveness of macroprudential policy," .

Barbaglia, L., S. Manzan, and E. Tosetti (2020): "Forecasting loan default in Europe with machine learning," *Journal of Financial Econometrics*.

Berger, D., N. Turner, and E. Zwick (2020): "Stimulating housing markets," *The Journal of Finance*, 75, 277–321.

Blickle, K. and M. Brown (2019): "Borrowing Constraints, Home Ownership and Housing Choice: Evidence from Intra-Family Wealth Transfers," *Journal of Money, Credit and Banking*, 51, 539–580.

Byrne, D., R. Kelly, and C. O'Toole (2022): "How Does Monetary Policy Pass-Through Affect Mortgage Default? Evidence from the Irish Mortgage Market," *Journal of Money, Credit and Banking*, 54, 2081–2101.

Central Bank of Ireland (2016): "Report on Mortgage Arrears," *retrieved from* *https://www.centralbank.ie/docs/default-source/publications/correspondence/finance-reports/mortgage-arrears-report.pdf?sfvrsn=2*.

——— (2022a): "Mortgage Arrears: A Consumer Guide to Completing the Standard Financial Statement (SFS)," *retrieved from* *https://www.centralbank.ie/docs/default-source/Regulation/consumer-protection/other-codes-of-conduct/35-gns-4-2-7-mortgage-arrears---sfs.pdf*.

——— (2022b): "Residential Mortgage Arrears and Repossessions Statistics – Q4 2022," *retrieved from* *https://www.centralbank.ie/docs/default-source/statistics/data-and-analysis/credit-and-banking-statistics/mortgage-arrears/2022q4_ie_mortgage_arrears_statistics.pdf?sfvrsn=68e3991d_3*.

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

Danne, C., A. McGuinness, et al. (2016): "Mortgage modifications and loan performance," Tech. rep., Central Bank of Ireland.

Fouliard, J., M. Howell, and H. Rey (2021): "Answering the queen: Machine learning and financial crises," Tech. rep., National Bureau of Economic Research.

Friedman, J. H. (2001): "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, 1189–1232.

Ganong, P. and P. Noel (2020): "Liquidity versus wealth in household debt obligations: Evidence from housing policy in the great recession," *American Economic Review*, 110, 3100–3138.

Hoang, D. and K. Wiegratz (2021): "Machine learning methods in finance: Recent applications and prospects," *European Financial Management*.

Hsu, J. W., D. A. Matsa, and B. T. Melzer (2018): "Unemployment insurance as a housing market stabilizer," *American Economic Review*, 108, 49–81.

Ishwaran, H., M. S. Lauer, E. H. Blackstone, M. Lu, and U. B. Kogalur (2021): "randomForestSRC: random survival forests vignette," .

Kelly, J., P. Lyons, F. McCann, E. O'Brien, et al. (2021): "Long-term mortgage arrears: Analytical evidence for policy considerations," *Central Bank of Ireland Financial Stability Notes*.

Kelly, R. and F. McCann (2016): "Some defaults are deeper than others: Understanding long-term mortgage arrears," *Journal of Banking & Finance*, 72, 15–27.

Mian, A. R. and A. Sufi (2012): "What explains high unemployment? The aggregate demand channel," Tech. rep., National Bureau of Economic Research.

Zhou, H., H. Wang, S. Wang, and Y. Zou (2022): "SurvMetrics: An R package for Predictive Evaluation Metrics in Survival Analysis." *R Journal*, 14.

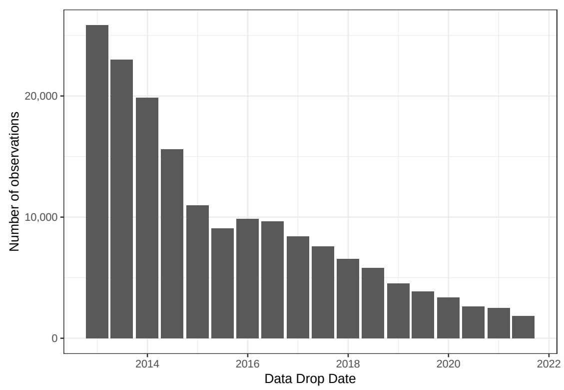
A Appendix

A.1 Additional Figures Background

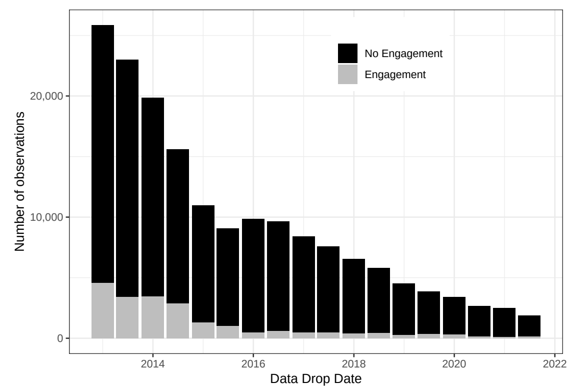
Figure A.1 shows similar information but instead of plotting the observations conditional on their time in panel we plot them conditional on their actual drop date. In Panel A.1a, the total number of observations by the time in the panel is displayed.

In Panel A.1b, we depict the total number of observations over the drop date and conditional on their engagement. Naturally, we only observe non-engagers in the first drop in 2012. To be able to switch from no engagement to engagement, we need to observe a loan for two consecutive drops.

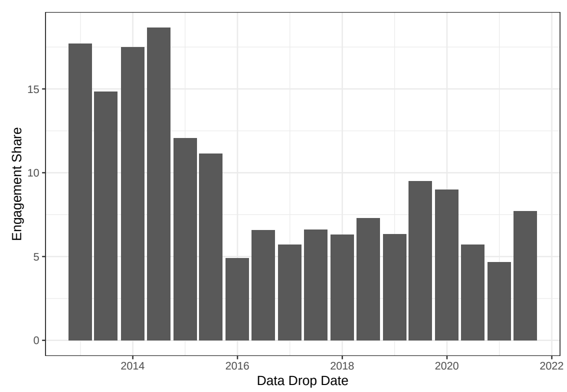
In Panel A.1c, we show the corresponding share of engagement via SFS conditional on the drop date. Consequently with the results from the previous figure, the engagement share is zero for the first drop date observed. It then stays above 10% until 2016 where it drops consistently below 10%.



(a) Number of Observations



(b) Observations by Engagement



(c) Engagement Share

Figure A.1: Observations and Engagement over Time in Panel

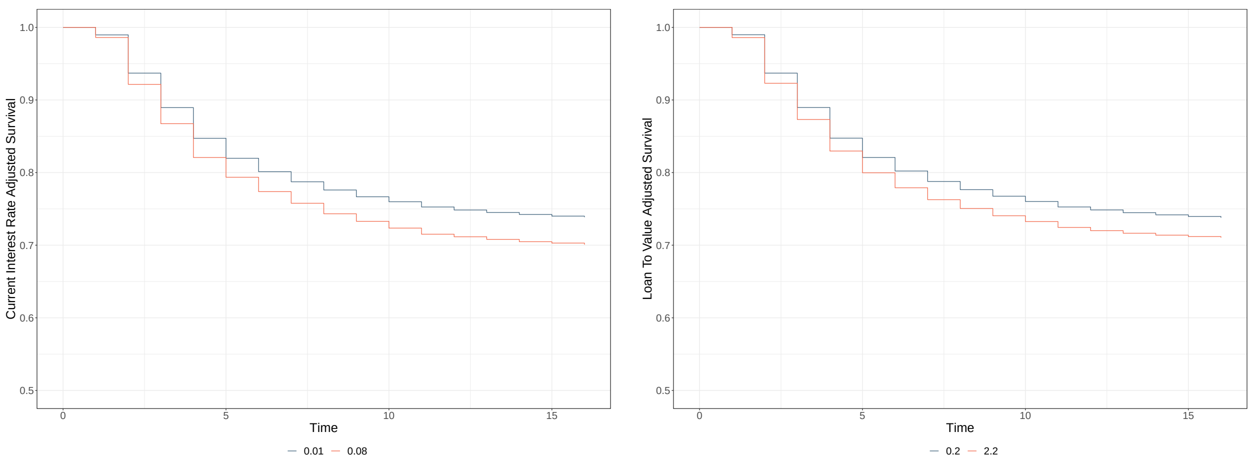
Note: The Figure shows the total number of observations conditional on their drop date. Panel 3 focuses on the total number of observations. Panel 3a shows the total number of observations conditional on their drop date and their engagement. Panel 3b shows the share (in percent) of those loans that engage via SFS and those that do not, conditional on their drop date.

A.2 Additional Partial Effects

Figure A.2 shows additional partial plots.

Panel A.2a plots the partial effect of the current interest rate. Loans that have a relatively higher interest rate tend to have a higher probability of engagement. Given the current high inflation rate environment and the following rise in interest rates, it is likely that the number of loans engaging via SFS will increase in the near future.

Panel A.2b shows the partial effect of different levels of LTV. The higher the LTV, the more likely a loan is to engage via SFS. However, the difference is not very pronounced despite the rather big difference in the LTV. This is consistent with the result from the Cox model, where we observed a change in sign (and significance) across the two regression models.



(a) Current Interest Rate

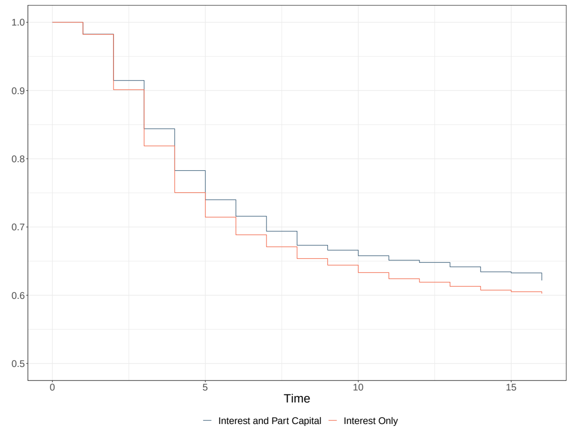
(b) Loan-To-Value

Figure A.2: Partial Effects of Selected Variables - II

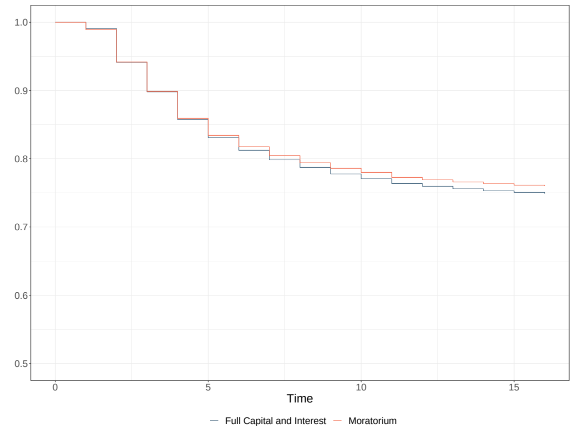
Note: The Figure shows the partial effects of different variables on the probability to engage via SFS for different values of the respective variable. In Panel A.2a, the partial effect of the current interest rate is displayed. Panel A.2b shows the partial effect of different levels of LTV.

Figure A.3 shows additional partial effects of categorical variables. Namely, in Panel A.3a and A.3b, the partial effects of different current payment type categories are displayed. Within each panel, the differences in the survival rate are marginal. The survival rate of "Interest and Part Capital" is in line with the survival rate of "Interest Only". Similarly, the survival rate of "Full Capital and Interest" and "Moratorium" is the same. However, across these two panels, the likelihood of an SFS engagement is much higher for the payment types displayed in Panel A.3a than Panel A.3b.

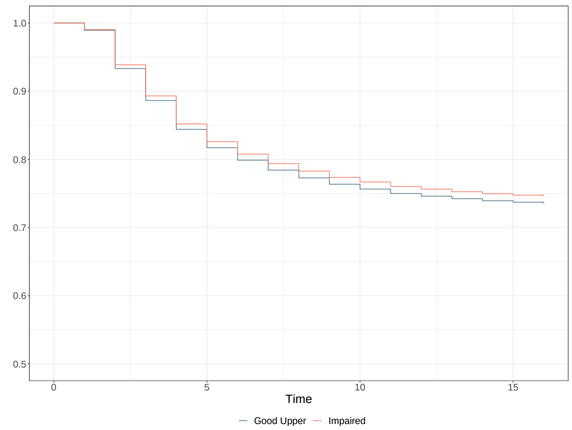
With respect to the loan credit quality in Panel A.3c, we see no strong differences across loans that are supposedly of good quality compared to those that are impaired.



(a) Current Payment Type - I



(b) Current Payment Type - III

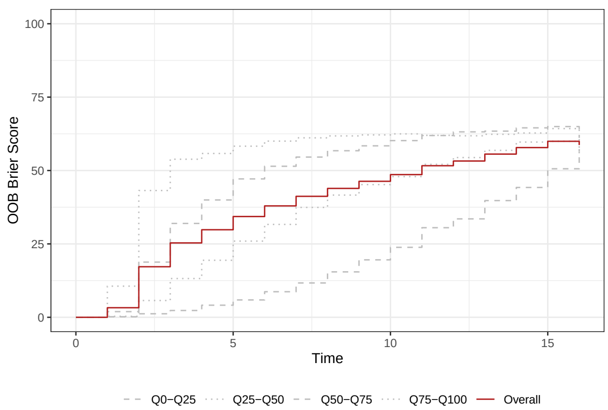


(c) Loan Credit Quality

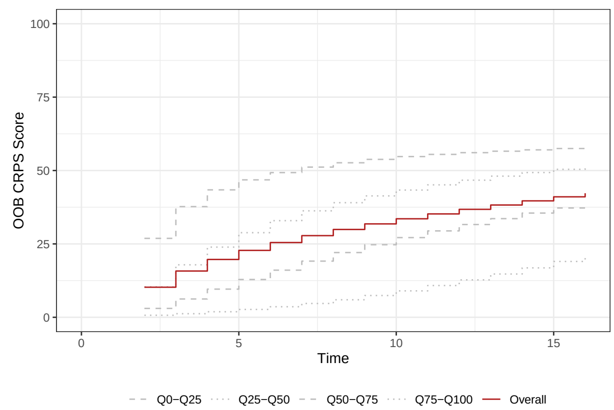
Figure A.3: Partial Effects of Selected Variables - III

Note: The Figure shows the partial effects of different variables on the probability to engage via SFS for different values of the respective variable. In Panel A.3a and A.3a, the partial effect of different payment types is displayed. Panel A.3c shows the partial effect of the loan credit quality.

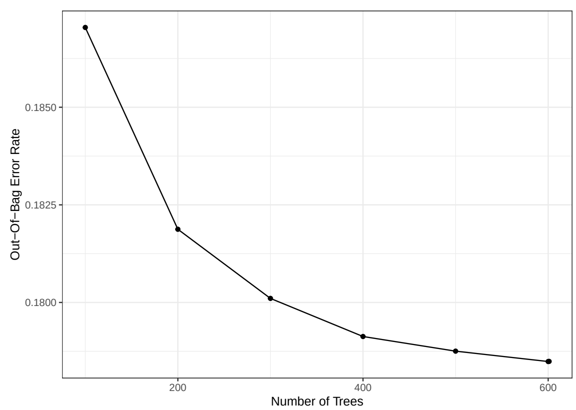
B Additional Prediction Accuracy Measure for RSF



(a) Out-Of-Bag Brier Score



(b) Out-Of-Bag CRPS Score



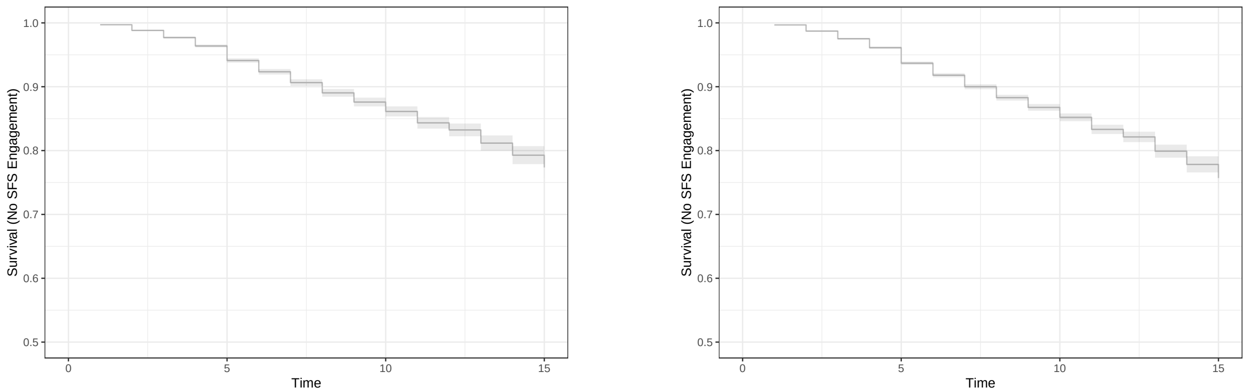
(c) Out-Of-Bag Error Rate

Figure B.4: Metrics from the RSF Model

Note:In Panel (a), we show the out-of-bag (OOB) Brier Score. The lower the brier score, the more accurate are predictions. The brier score ranges from 0 to 100, with 100 being a value that can be achieved if assignment was random. Panel (c) depicts the OOB Continous Rank Probability Score (CRPS). This score ranges from 0 to 100 with 0 being a perfect OOB prediction and 100 being wholly wrong. Panel (d) shows the results of the finetuning of the RSF model. It plots the cumulative OOB error rate conditional on the number of trees.

B.1 Subsample Estimation

B.1.1 Survival Functions

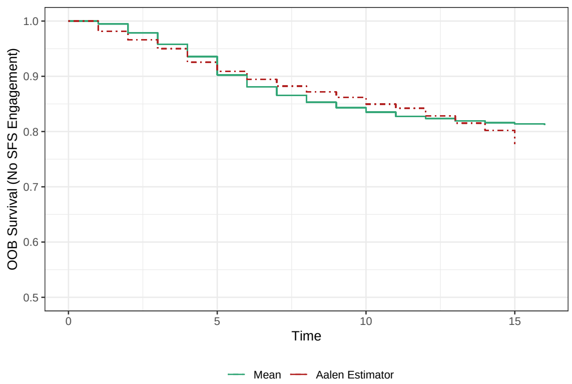


(a) Model (1)

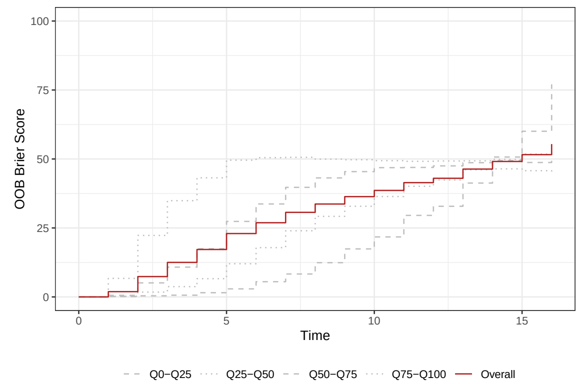
(b) Model (2)

Figure B.5: Survival Rate

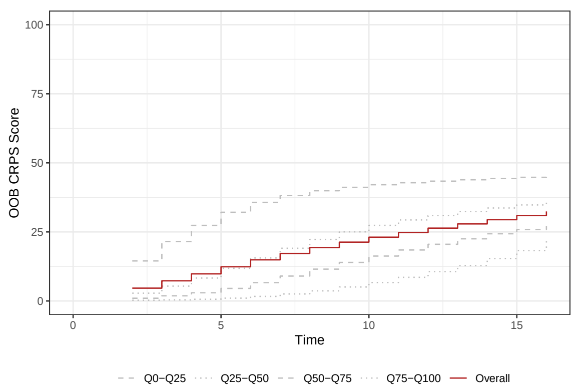
Note: The Figure shows the estimated Survival Rate of two Cox models with the 95% confidence interval. The higher the survival, the lower the probability to engage via SFS. Mortality, in our model, refers to the probability that a loan engages via SFS. The sample is restricted on data drops after 2016.



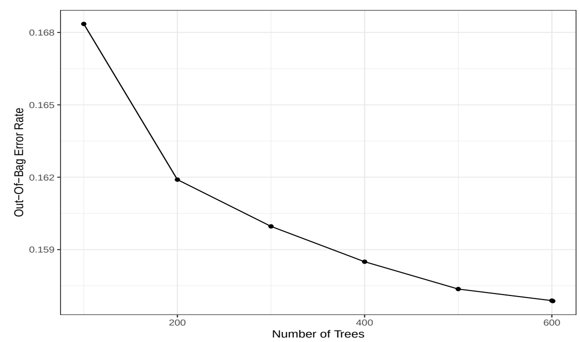
(a) Out-Of-Bag Survival



(b) Out-Of-Bag Brier Score



(c) Out-Of-Bag CRPS Score

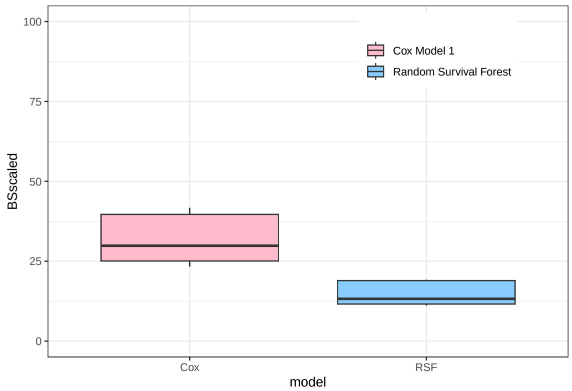


(d) Out-Of-Bag Error Rate

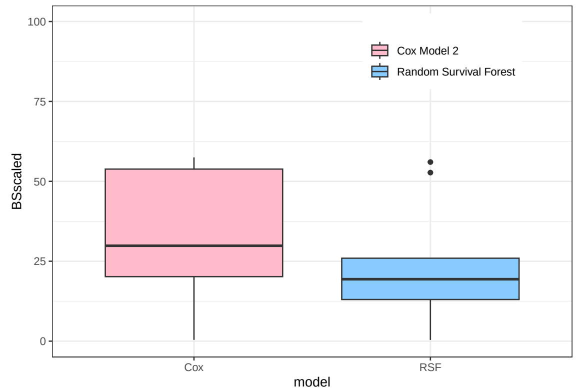
Figure B.6: Metrics from the RSF Model

Note: In Panel (a), the figure plots the Out-Of-Bag (OOB) Survival Rate for the overall ensemble survival and the Nelson-Aalen estimator. In Panel (b), the figure shows the OOB Brier Score. The lower the brier score, the more accurate are predictions. The brier score ranges from 0 to 100, with 100 being a value that can be achieved if assignment was random. Panel (c) depicts the OOB Continous Rank Probability Score (CRPS). This score ranges from 0 to 100 with 0 being a perfect OOB prediction and 100 being wholly wrong. Panel (d) shows the cumulative OOB error rate conditional on the number of trees. The sample is restricted on data drops after 2016.

B.1.2 Precision



(a) Cox Model (1) versus Random Forest



(b) Cox Model (2) versus Random Forest

Figure B.7: Precision using Brier Score

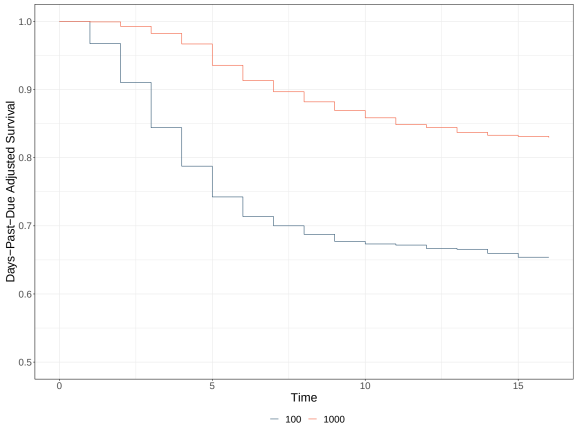
Note: The figure shows the Brier Score calculated for the median t^* for the two different Cox models as well as the Random Survival Model. The lower the Brier Score is, the more precise are the predictions of the survival rates. The sample is restricted on data drops after 2016.

B.1.3 Partial Effects

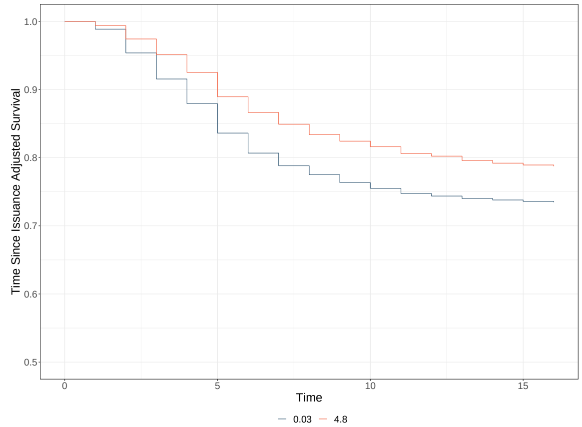
Table B.1: Cox Model - Results

	Dependent variable:	
	SFS Engagement	
	(1)	(2)
Days-Past-Due	-0.001^{***}	-0.001^{***}
	(0.00003)	(0.00003)
Age	-0.002	-0.003
	(0.002)	(0.002)
Loan-To-Value	-0.030	0.132^{***}
	(0.045)	(0.043)
Loan-To-Value at Origination	-0.647^{***}	-0.722^{***}
	(0.068)	(0.067)
Income at Origination	-0.00000^{***}	-0.00000^{***}
	(0.00000)	(0.00000)
First Time Buyer	0.118^{***}	0.080^{**}
	(0.029)	(0.028)
PT: Interest and Part Capital	0.986^{***}	
	(0.091)	
PT: Interst Only	0.802^{***}	
	(0.060)	
PT: Less than Interest Only	0.236^{**}	
	(0.092)	
PT: Moratorium	-0.523^{***}	
	(0.069)	
PT: Other	0.081	
	(0.252)	
Time since Issuance (0/1)	-1.356^{***}	-1.256^{***}
	(0.091)	(0.090)
Residential Real Estate Prices	-0.010^{***}	-0.001
	(0.001)	(0.003)
Regional Unemployment	0.052^{***}	0.047^{***}
	(0.008)	(0.008)
Lag Change in Installment	0.001^{***}	0.001^{***}
	(0.0001)	(0.0001)
Current Interest Rate	5.851^{***}	8.121^{***}
	(0.891)	(0.872)
Time FE	No	Yes
Bank FE	Yes	No
Observations	170,933	170,933
R^2	0.101	0.102
Log Likelihood	$-230{,}687.900$	$-230{,}537.700$

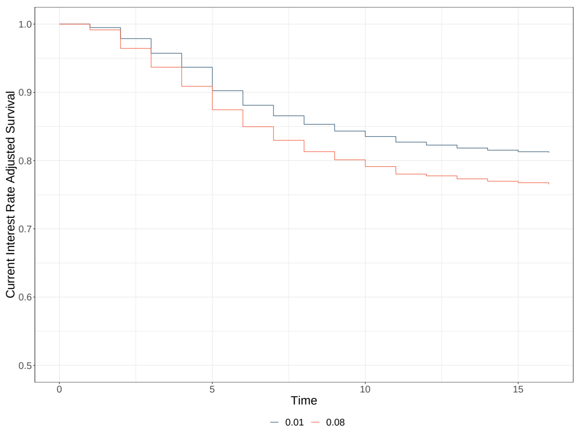
Note: This table shows the partial effects estimated from two different Cox models. The dependant variable is the binary outcome of SFS engagement. PT stands for payment type. Time since issuance is a continous variable which takes the value 0 if the loan is observed at origination date and value 1 if it is observed at maturity date. *p<0.1; **p<0.05; ***p<0.01



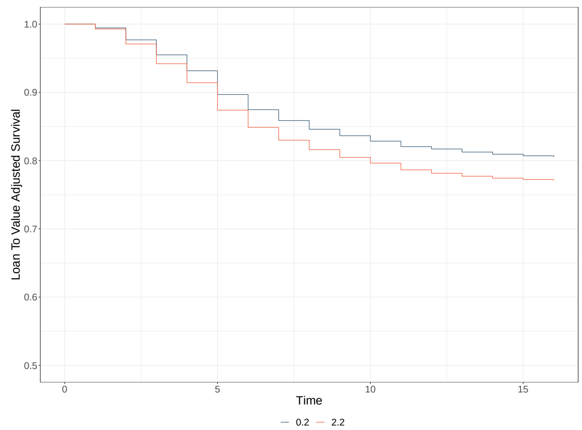
(a) Days-Past-Due



(b) Time Since Issuance



(c) Current Interest Rate



(d) Loan-To-Value

Figure B.8: Partial Effects of Selected Variables - I

Note: The Figure shows the partial effects of different variables on the probability to engage via SFS for different values of the respective variable. In Panel B.8a, the partial effect of days-past-due is displayed. Panel B.8b shows the partial effect of the time since issuance, which takes the value 0 if the loan is observed at origination date and value 1 if it is observed at maturity date. Every value in between corresponds to the number of days relative to total number of days of the entire loanspan time at observation date. Panel B.8c shows the partial effects of different levels of the current interest rate. Panel B.8d shows the partial effect of different levels of LTV. The sample is restricted on data drops after 2016.

Table B.2: Best Predictors Cox Models

Rank	Model (1)	χ^2	p-val.	Rank	Model (2)	χ^2	p-val.
1	Days-Past-Due	2413.50	0.00	1	Days-Past-Due	2704.54	0.00
2	Current Payment Type	367.23	0.00	2	Time since Issuance	194.55	0.00
3	Bank	222.91	0.00	3	Lag Change in Installment	119.4	0.00
4	Time since Issuance	220.26	0.00	4	LTV at Origination	114.5	0.00
5	Lag Change in Installment	121.22	0.00	5	Income at Origination	93.45	0.00
6	LTV at Origination	91.63	0.00	6	Current Interest Rate	86.78	0.00
7	Income at Origination	51.90	0.00	7	Unemployment	36.82	0.00
8	Residential Estate Prices	51.48	0.00	8	LTV	9.2	0.00
9	Unemployment	45.45	0.00	9	Date FE	9.14	0.00
10	Current Interest Rate	43.11	0.00	10	First Time Buyer	7.98	0.00
11	First Time Buyer	16.66	0.00	11	Age	3.09	0.08
12	Age	0.93	0.34	12	Residential Estate Prices	0.02	0.88
13	LTV	0.44	0.51				

Note: The table shows the most important variables for both Cox models estimated in section 5.1.2 using an Analysis of Variance for the subsample analysis. It reports the χ^2 test statistic and the p-value.

B.1.4 Best Predictors

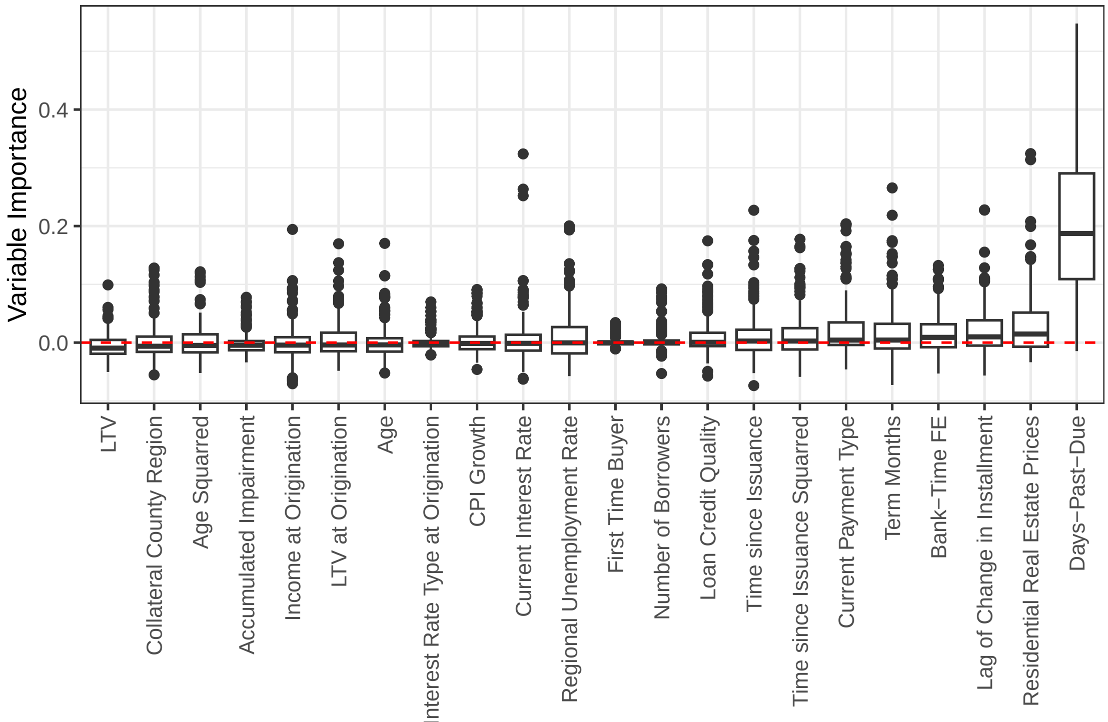
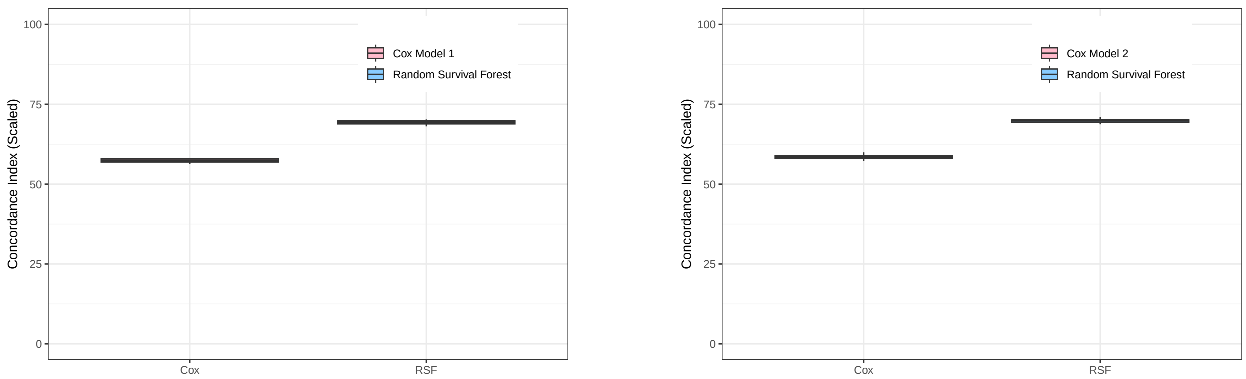


Figure B.9: Most Important Variables identified by RF Survival Model

Note: The Figure shows the most important vairables identified by the Random Survival Forest Model. A value greater than zero indicates that the variable is important for predicting the SFS engagement. We used 200 subsamples of our data to estimate the importance of each variables. Larger values indicate a higher importance for predicting SFS engagement. The sample is restricted on data drops after 2016.

B.2 Alternative Precision Measure



(a) Cox Model (1) versus Random Forest

(b) Cox Model (2) versus Random Forest

Figure B.10: Precision using Concordance Index

Note: The figure shows the Concordance Index calculated for the two different Cox models as well as the Random Survival Model. The index is scaled such that a value of 0 corresponds to a random outcome and a value of 100 to a perfect prediction.