Journal of the Operational Research Society

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjor20

Deep learning for survival and competing risk modelling

Gabriel Blumenstock, Stefan Lessmann & Hsin-Vonn Seow

To cite this article: Gabriel Blumenstock, Stefan Lessmann & Hsin-Vonn Seow (2022) Deep learning for survival and competing risk modelling, Journal of the Operational Research Society, 73:1, 26-38, DOI: 10.1080/01605682.2020.1838960

To link to this article: https://doi.org/10.1080/01605682.2020.1838960

JOURNAL OF THE OPERATIONAL RESEARCH SOCIETY
2022, VOL. 73, NO. 1, 26–38
https://doi.org/10.1080/01605682.2020.1838960

ORIGINAL ARTICLE

Deep learning for survival and competing risk modelling

Gabriel Blumenstock[a], Stefan Lessmann[a] and Hsin-Vonn Seow[b]

[a]School of Business and Economics, Humboldt-Universität zu Berlin, Berlin, Germany; [b]Nottingham University Business School, University of Nottingham Malaysia, Malaysia

ABSTRACT

The article examines novel machine learning techniques for survival analysis in a credit risk modelling context. Using a large dataset of US mortgages, we evaluate the adequacy of DeepHit, a deep learning-based competing risk model, and random survival forests. The observed results provide strong evidence that both models predict default and prepayment risk more accurately than statistical benchmarks in the form of the Cox proportional hazard model and the Fine and Gray model. The superiority of the machine learning models is robust across different periods including stressed periods. We also find machine learning models do not require larger amounts of training data than the statistical benchmarks. Finally, we extend methods for estimating feature importance scores to deep neural networks for survival analysis and clarify which covariates determine the estimated survival functions of DeepHit. An online companion with additional results is available in Supplementary Information.

ARTICLE HISTORY
Received 22 January 2020
Accepted 9 October 2020

KEYWORDS
survival analysis; competing risk model; deep learning; mortgage risk

1. Introduction

Statistical models for survival analysis are well-established in the credit scoring literature. Compared to static classification models, which predict the probability of default (PD) in a pre-defined time window, survival analysis models the time to a default event (e.g. Dirick et al., 2019). Regulatory frameworks such as the Basel regulation and IFRS9 highlight the importance of survival analysis, for example, to estimate loss-given-default (LGD) for defaulted exposures or loan loss provisioning (Cohen & Edwards, 2017; Eder, 2019). In this sense, the use cases of dynamic survival models versus static classification models differ and the choice of a type of model should depend on the requirements of the application.

Previous work has examined alternative survival modelling approaches or mixture cure models (Tong et al., 2012), which account for the fact that not every borrower is going to default. The potential of survival models to process time-varying covariates has received considerable attention, for example, to capitalise on behavioural characteristics (Stepanova & Thomas, 2002) or macroeconomic factors (Bellotti & Crook, 2009). We focus on survival models for competing risks, which model the event-time-distributions for mutually exclusive events. Drawing on the seminal work of Banasik et al. (1999), we consider the risk of default and prepayment.

Surveying the literature on survival analysis in credit scoring (see Table 1), we find that the field has paid limited attention to machine learning (ML) based model; with (Baesens et al., 2005) being a notable exception. This is surprising in that ML models have been considered for the estimation of various risk parameters including PD, LGD, and exposure at default (e.g. Lessmann et al., 2015; Loterman et al., 2012; Yao et al., 2017). Recent work also studies deep learning (DL) methods (Addo et al., 2018).

The appealing results observed in prior work on static ML/DL models for credit risk prediction motivates us to explore the potential of ML/DL-based dynamic models for survival analysis. In pursuing this goal, the article makes the following contribution: First, we provide empirical results concerning the performance of DeepHit (Lee et al., 2018), a recently proposed DL model for single and competing risks, in a mortgage modelling case study. We find DeepHit to perform highly competitively compared to econometric approaches and an alternative ML approach, which extends the random forest algorithm to a competing risk setting (Ishwaran et al., 2014). To our best knowledge, this is the first evaluation of DL and random survival forests in credit risk management. Second, to shed light on the internal mechanisms of the DL model, we introduce an approach to extract feature importance scores from DeepHit through extending permutation-based feature importance (Fisher et al., 2019) to a

CONTACT Stefan Lessmann stefan.lessmann@hu.berlin.de School of Business and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.

Supplemental data for this article is available online at https://doi.org/10.1080/01605682.2020.1838960.

Table 1. Overview of previous studies using survival models for credit scoring.

	Dataset				Survival model[a]					
					Statistical		ML		DL	
Study	Number of observations	Number of variables	Timespan	Mortgage setting	Single risk	Competing risks	Single risk	Competing risks	Single risk	Competing risks
Narain (1992)	1242	7	≤1992		x					
Deng (1997)	1,489,372	16	1976–1983	x		x				
Banasik et al. (1999)	50,000	>7	1994–1997		x	x				
Stepanova and Thomas (2001)	11,500	16	≤2001		x					
Ciochetti et al. (2002)	2589	>9	1974–1990	x		x				
Stepanova and Thomas (2002)	50,000	16	≤2002		x	x				
Baesens et al. (2005)	15,000	14	≤2002		x	x	x	x		
Bellotti and Crook (2009)	>200,000	>11	1997–2005		x					
Cao et al. (2009)	25,000	1	2004–2006		x					
Deng and Liu (2009)	103,462	26	1998–2003	x		x				
Tong et al. (2012)	27,527	14	≤2012		x	x				
Zhang and Thomas (2012)	27,278	21	1987–2003		x					
Dirick et al. (2017)	271,040	6–31	≤2015		x	x				
Dirick et al. (2019)	20,000	7	2004–2014		x					
This study	600,000	16–31	1999–2017	x	x	x	x	x	x	x

[a]Our distinction between single and competing risks needs some clarification. Several studies model competing risks by estimating cause-specific models for each risk and treating the occurrence of other events as censored. Examples include Baesens et al. (2005), Banasik et al. (1999), Ciochetti et al. (2002), and Stepanova and Thomas (2001). This approach neglects the fact that the occurrences of competing risks are not independent (Fine & Gray, 1999). To our knowledge, this is the first credit risk article that employs survival models that account for competing risks being mutually exclusive. However, for Table 1, we follow prior work and code studies that consider multiple events as competing risk studies.

competing risk modelling setting. Finally, the article contributes towards understanding the robustness of competing risk models over time. We consider periods before and after the financial crisis and test how well models that were estimated in good economic conditions predict under stressed conditions.

The article is organised as follows. Section 2 elaborates on survival and competing risk modelling. Section 3 describes the experimental design. Section 4 discusses empirical results and Section 5 concludes the article. Additional results and details on the study setup to enable replication are available in an online companion, which interested readers can obtain from Supplementary material.

2. Methodology

The section covers the background of survival analysis and describes the modelling techniques used in the article.

2.1. Foundations and notation

In competing risk modelling, we observe $i = 1, \ldots, n$ individuals, which we represent by a p-dimensional vector $x_i \in \mathbb{R}^p$. The p covariates comprise, for example, borrower- and loan-specific information as well as macroeconomic variables. Each individual may experience an event from a set $\{1, \ldots, K\}$ until a specific time point t_i. We define such an event, k_i, as the default or prepayment of a loan. If no event is experienced until t_i, an instance is called censored, meaning that the observed survival time t_i is smaller than the true survival time T_i. Censoring can occur due to an active loan not experiencing an event until the moment of data collection or due to a terminated loan not having experienced an event over the entire loan period. Note that x_i is not restricted to data available at loan origination but can also contain time-varying covariates, the value of which change during the loan period (e.g. Bellotti & Crook, 2009).

The origin of competing risk modelling is a dataset $\mathcal{D} = \{(t_i,\ k_i,\ x_i)\ \}_{i=1,\ldots,n}$. Based on $\mathcal{D}$, risk models predict event entry probabilities over time conditioned on the covariates. In the reminder, we first assume $K = 1$ and extend to the general case at a later stage. We also restrict our explanations to a continuous-time setting, as they remain conceptually equivalent when switching to discrete-time. For notational convenience, we drop the index i in formulas that do not require distinguishing between individuals.

The distribution of the continuous random time-variable t is characterised through the survival function $S(t) = 1 - F(t)$, which captures the probability of not observing the event until t. Here, $F(t)$ denotes the cumulative distribution function $F(t) = P(T \leq t) = \int_0^t f(u)du$. Modelling $S(t)$ through a (semi-)parametric model is a common approach in survival analysis. One typically assumes $S(0) = 1$ and $S(\infty) = 0$, implying that every individual is expected to experience the event given enough time. This is not plausible in a credit risk context as most loans are repaid. Mixture cure models distinguish instances that never experience the event and those that experience the event although not necessarily in the observation period (e.g. Tong et al., 2012).

The hazard function $h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$ gives an alternative target for survival analysis. It quantifies the probability of experiencing an event

at an infinitesimally short moment given that the individual has survived until t. Modelling the hazard function or the survival function is equivalent because $h(t) = \frac{f(t)}{S(t)} = -\frac{\partial \ln(S(t))}{\partial t}$, with $f(t)$ denoting the probability density function of the random variable t (e.g. Kleinbaum & Klein, 2012).

2.2. Cause-specific cox model

The semi-parametric Cox proportional hazards model is a popular approach for survival analysis. This model assumes that all individuals share the same baseline hazard function $h_0(t)$, which can be any non-negative function of time. The parametric part of the model is introduced by multiplying the baseline hazards with a term containing the covariates, resulting in the hazard function conditional on x :

$$h(t|x) = h_0(t) * \exp\left(\sum_{k=1}^{p} \beta_k x_k\right). \tag{1}$$

The assumption of all individuals sharing the same baseline hazard function implies that the hazard ratio of any two individuals i and j is constant over time:

$$\frac{h_i(t|x_i)}{h_j(t|x_j)} = \exp\left(\sum_{l=1}^{p} \beta_j (x_{i,l} - x_{j,l})\right), \tag{2}$$

which is a questionable assumption in many applications. The covariates exerting a linear, additive effect is another assumption of the model and requires modellers to design nonlinear effects and interactions manually.

The Cox model was designed for a single event setting. One can expand the model to a multiple risk setting by estimating a hazard function for each event individually (e.g. Ciochetti et al., 2002). Thereby, the respective competing risks are considered as censored, i.e. k is set to zero when a competing event $k' \neq k$ occurs. Such a cause-specific hazard function (3) is equal to the original hazard formula applied to each event type k (e.g. Kleinbaum & Klein, 2012):

$$h_k^{cs}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t,\ \mathcal{K} = kT \geq t)}{\Delta t},\ k = 1, \ldots K \tag{3}$$

However, the cause-specific approach contradicts the definition of competing risks in which the occurrence of one event precludes the occurrence of the other events, and, therefore, leads to biased estimates.

2.3. Fine-Gray model

A statistical model for competing risks was introduced by Fine and Gray (1999). They replace the cause-specific hazard function by the subdistribution hazard function, which denotes the probability of experiencing event k at an infinitesimally small moment of time t, given that the individual has neither experienced event k nor any competing event $j \neq k$, thereby implementing the definition of competing risks correctly:

$$h_k^{sd}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t,\ \mathcal{K} = kT \geq t \cup (T < t \cap \mathcal{K} \neq k))}{\Delta t},\ k = 1, \ldots, K \tag{4}$$

Similarly, the cumulative incidence function, $CIF(t)$, denotes the incidence of an occurrence of an event whilst taking the occurrence of competing events into account (Austin & Fine, 2017).

2.4. Random survival forest

Extending the well-known random forest algorithm, Ishwaran et al. (2014) introduce random survival forests (RSF) for competing risks. The goal of their model is to estimate the cumulative incidence function, $CIF(t)$, for competing risks. The algorithm follows the same logic as random forest. It draws B bootstrap samples from a dataset and grows a competing risk tree from each sample. In doing so, RSF also borrows the random subspace mechanism and randomly selects $M \leq p$ candidate variables when branching a node (Breiman, 2001). To identify the split variable and the corresponding threshold, RSF uses Gray's log-rank splitting rule for competing risks. Iteratively applying the splitting rule to each node, a tree is grown to full size provided that each terminal node contains more than $n_0 > 0$ instances.

Each survival tree provides an estimate of the cumulative incidence function $CIF_k^b(t|x)$, where b indexes the bootstrap samples. RSF averages these estimates to obtain the final estimated cumulative incidence function. Since the focus of this article is on DL, we refer readers to Ishwaran et al. (2014) for a detailed explanation of the estimation of $CIF_b(t)$. Unlike the above survival models, random forest-based approaches do not rely on parametric assumptions and promise autonomous modelling of non-linearities and interaction effects amongst covariates.

2.5. DeepHit

DeepHit (DHT) is a multi-layered neural network, which architecture is tailored to survival analysis with competing risks (Lee et al., 2018). The modelling target is to estimate the cumulative incidence function for all events. To achieve this, DHT draws inspiration from multi-task learning (Collobert & Weston, 2008). It incorporates a subnetwork that is shared by all risk types and K subsequent cause-specific subnetworks. Figure 1 depicts the architecture of DHT for the two risks considered here.

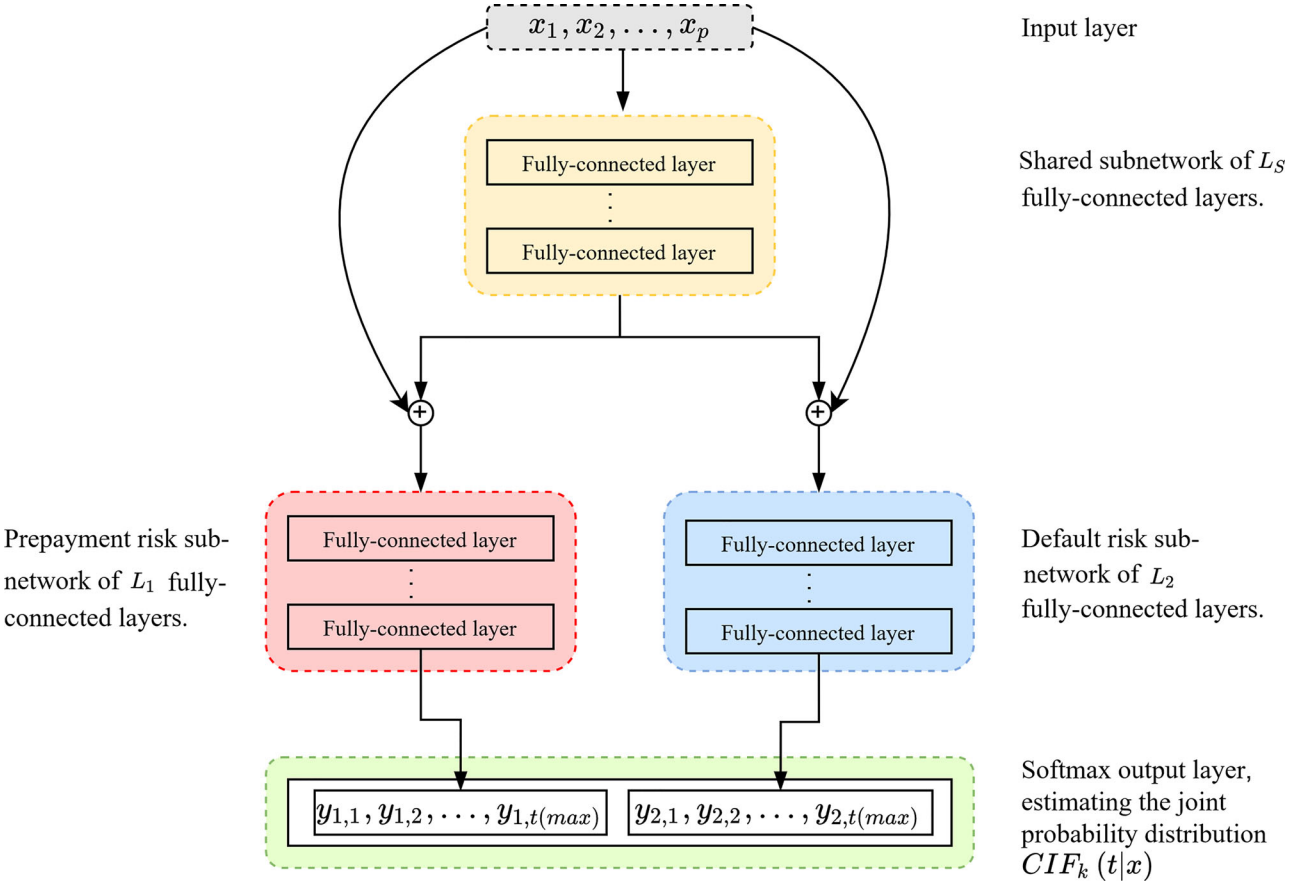

Figure 1. Multilayered architecture of the DeepHit network for competing risk modelling based on Lee et al. (2018).

Covariate values first enter a shared subnetwork of L_s fully connected layers. This subnetwork outputs a vector $f_s(x)$, capturing a latent representation of the data, that is common to the K competing risks. The latent representation is taken as an input for each cause-specific subnetwork, which in turn consists of L_{CS} fully connected layers. Each cause-specific subnetwork also receives the original covariates as an additional input via residual connections. Passing both the latent representation $f_s(x)$ and the original covariates to the cause-specific subnetworks enables these to learn risk-specific relationships amongst covariates whilst still having access to the latent representation of the covariates common to all risk types.

The output layer of the DeepHit model is constructed by stacking the output layers from the cause-specific subnetworks. The output layer uses the softmax function to convert hidden layer output scores into probabilities. More formally, the DHT output is given by $y = [y_{1,1}, \ldots, y_{1,t_{max}}, \ldots, y_{K,1}, \ldots, y_{K,t_{max}}]$ and corresponds to the estimated joint probability distribution for the K events given x (Lee et al., 2018):

$$\begin{bmatrix} P(k=1,\ t=1|x), \ldots, P(k=1,\ t=t_{max}|x), \ \ldots, \\ P(k=K,\ t=1|x), \ldots P(k=K,\ t=t_{max}|x) \end{bmatrix} \tag{5}$$

Based on the estimated joint probability distribution, the cumulative incidence function can be estimated as follows:

$$CIF_k(t|x) = P(T \le t, \mathcal{K} = k|x) = \sum_{t^*=1}^{t} P(T = t^*, \mathcal{K} = k|x) = \sum_{t^*=1}^{t} y_{k,t^*}, k = 1, \ldots, K \tag{6}$$

To train connection weights, DHT minimises a loss function specifically designed for estimating the cumulative incidence function based on right-censored data. It consists of two loss terms and a weighting term β :

$$\begin{aligned} &\mathcal{L} = \mathcal{L}_1 + \beta * \mathcal{L}_2, \ \text{with} \\ &\mathcal{L}_1 = -\sum_{i=1}^{N} \left[\mathbf{1}(k_i \neq 0) * \log\left(y^i_{k_i,t_i}\right) \right. \\ &\qquad \left. + \mathbf{1}(k_i = 0) * \log\left(1 - \sum_{k=1}^{K} F_k(t_i x_i)\right) \right] \\ &\mathcal{L}_2 = \sum_{k=1}^{K} \sum_{i \neq j} A_{k,i,j} * \exp\left(\frac{-(F_k(t_i x_i) - F_k(t_i x_j))}{\sigma}\right) \end{aligned} \tag{7}$$

The term $\mathcal{L}_1$ is a log-likelihood loss. If an instance experiences an event k_i at time t_i, the network output for that instance $y^i_{k_i,t_i}$ should be close to one. The higher the respective $y^i_{k_i,t_i}$ of the uncensored instance i, the less impact it has on increasing $\mathcal{L}$. DHT uses an indicator function to filter the censored cases and applies the *log*-function on the predicted output, which leads to lower predicted values

being increasingly penalised. If an instance i is censored at the time t_i, the ideal probability estimate would assign a probability of zero for experiencing any event until the observed time point t_i. Deviations from this ideal estimate are penalised by the second part of $\mathcal{L}_1$; again using an indicator function to filter uncensored instances and applying the log-function on the predicted cumulative survival probabilities summed up over all events. The effects are then aggregated across all instances.

The term $\mathcal{L}_2$ is a ranking loss based on the concept of concordance. An instance i that experiences event k_i at time t_i should be assigned to a higher cumulative risk for event k_i until time t_i than an instance j that has not experienced any event until t_i. The relation between two instances can only be tested for those pairs that satisfy the condition of one instance having experienced an event and the other instance not having experienced any event for a longer period than the first instance. The indicator function $A_{k,i,j} = \mathbf{1}(k_i = k,\ t_i < t_j)$ filters incomparable pairs. A positive difference between the estimated cumulative probabilities between two comparable instances i and j at time t_i, $F_k(t_i|x_i) - F_k(t_i|x_j)$, indicates that the model has estimated the different cumulative incidence functions coherently. By implementing the difference between the cumulative probabilities within an exponential, the loss function increases particularly in those cases in which the estimated difference is negative. The effects are then aggregated across all acceptable pairs and event types.

In summary, $\mathcal{L}_2$ finetunes the network by assessing cumulative probabilities estimated at different time points and comparing them amongst comparable pairs of instances, whereas $\mathcal{L}_1$ uses the event or censoring times for each instance separately. Note that $\mathcal{L}_2$ includes a scaling term σ, which was set to 0.1 in this study. Lee et al. (2018) also incorporate additional coefficients α_k, $k = 1, ..., K$ to weight ranking losses caused by each event type differently. In our context, these coefficients would facilitate a weighting of the importance of default and prepayment risks. Whilst being a potentially interesting feature for mortgage risk modelling, we do not consider risk weighting in this initial evaluation of DHT and set $\alpha_K = 1 \forall k$. This also avoids bias in comparisons of DHT to survival models that do not support risk-specific weights.

2.6. Variable importance estimation with machine learning models

ML models conceal the feature-target relationship, which they estimate mode data. This is a major concern in banking practice. Regulatory frameworks such as the Basel Capital Accord or GDPR demand explainable models that reveal how feature values are mapped into risk predictions. One step towards interpretability consists of extracting feature importance scores. Model-specific ways to calculate importance scores have been developed for some learning algorithms including RSF (Ishwaran, 2007). We make use of the model-agnostic paradigm of permutation importance (e.g. Fisher et al., 2019) and provide an implementation for DHT.

Permutation importance assesses the decrease of model performance after corrupting one feature. The larger the decrease the more important that feature. Compared to removing a variable from a model, which requires model re-estimation, destroying the predictive information within a feature whilst keeping it in the model is more efficient. We follow Ishwaran (2007) and corrupt a variable by adding random noise. More specifically, we train a DeepHit model on the original data and evaluate its performance on a hold-out test set. For each feature l, we then estimate its variance σ_l^2, $l = 1, ..., p$, from the training data, draw noise values from a normal distribution $\mathcal{N}(0, \sigma_l^2)$, and add these values to feature l in the test set. Next, we re-apply DHT to the adulterated test set and capture the difference in model performance, which we assess using the time-dependent concordance index (see Section 3) for each event individually. The decrease in performance gives the importance score of feature l. We repeat this approach p times to calculate one score for every feature. The resulting scores clarify on which features DHT relies the most when generating predictions and facilitate comparing the relative strength of a feature through the lens of a model across features.

3. Experimental setup

3.1. Data

The dataset consists of roughly 600,000 publicly available single-family US mortgages with loan terms between 15 and 30 years issued by Freddie Mac between 1999 and 2017.[1] The target variable indicates whether an observation is censored or experiences a default or prepayment event. Default is defined as a loan turning into 3-month delinquency for the first time. Prepayment describes the moment of a loan being repaid completely and unexpectedly. The annual distributions of event occurrences and censoring differ substantially across the years. Especially recent years show low event rates. The online companion details event rates per year in supplementary information Table S1. The data also includes a time-variable, which captures the time between loan origination and event

Table 2. Overview of explanatory variables.

	Variable	Explanation
Loan-level variables	int.rate	Initial interest rate
	orig.upb	Original unpaid balance
	fico.score	Initial FICO score
	dti.r	Initial debt-to-income ratio
	ltv.r	Initial loan-to-value ratio
	bal.repaid	Current repaid balance in percent
	t.act.12m	Number of times not being delinquent in the last 12 months
	t.del.30d.12m	No. of times being 30 days delinquent in the last 12 months
	t.del.60d.12m	No. of times being 60 days delinquent in the last 12 months
Macroeconomic variables	hpi.st.d.t.o	Difference of house price index (HPI) between loan origination date and today (state-level)
	hpi.zip.o[a]	HPI at origination date (zip-level)
	hpi.zip.d.t.o[a]	Difference of HPI between origination date and today (zip-level)
	ppi.c.FRMA	Current prepayment incentive, i.e. difference between interest rate and current fixed rate mortgage average
	TB10Y.d.t.o	Difference of 10-year treasury bill rate between origination date and today
	FRMA30Y.d.t.o	Difference of 30-year fixed rate mortgage average between origination date and today
	ppi.o.FRMA	Prepayment incentive at origination date, i.e. difference between interest rate and fixed rate mortgage average at origination date
	equity.est[a]	Estimated home equity
	hpi.st.log12m	HPI – 12-month log return (state-level)
	hpi.r.st.us	Ratio of HPI between state-level and US-level today
	hpi.r.zip.st[a]	Ratio of HPI between zip-level and state-level today
	st.unemp.r12m	Unemployment rate today – 12-month log return (state-level)
	st.unemp.r3m	Unemployment rate today – 3-month log return (state-level)
	TB10Y.r12m	Current 10-year treasury bill rate – 12-month return
	T10Y3MM	Yield between 3-month and 10-year treasury bill rates today
	T10Y3MM.r12m	Yield between 3-month and 10-year treasury bill rates today – 12-month return

[a]This variable is not available for all periods and only used in some of the experiments.

occurrence or censoring in months. We observe two forms of censoring: an active loan not experiencing an event until the time point of observation, or a loan being event-free for the maximum observation period.

The dataset comprises 25 explanatory variables. Nine covariates capture loan-level information. Five of these capture loan characteristics recorded at origination. The other four covariates capture loan characteristics recorded during the previous 12 months of a loan and were constructed based on the Freddie Mac monthly performance data. Additionally, we construct 16 time-varying covariates that capture macroeconomic information based on data retrieved from the Federal Reserve Economic Data, the National Conference of State Legislatures, the US Bureau of Labor Statistics, the Federal Housing Finance Agency, and Freddie Mac. Table 2 depicts the variables and offers a brief explanation.

The US government launched several programs in response to the financial crisis, which enabled borrowers to renegotiate their loans when becoming delinquent. Altering the definition of a default event, these programs break the feature-response relationship, which implies that data before and after 2009 are not directly comparable. Therefore, governmental programs impose a natural partitioning of the data, creating a subset from 1999 to 2009 and one from 2010 to 2017. We refer to these subsets as "dataset 1" and "dataset 2", respectively. The share of defaulted loans is substantially lower in the second subset due to the new options for renegotiation and the definition of default. The portion of censored observations also differs between the two datasets, as censoring naturally decreases the more time has passed since loan origination. The financial crisis further partitioned dataset 1 into a pre-crisis dataset from 1999 to 2006 and a crisis dataset from 2007 to 2009. We make use of the structural breaks in the data when designing the experiments to assess survival models under different conditions.

3.2. Experimental design

The empirical analysis consists of two parts, denoted as Analysis 1 and Analysis 2. The former focuses on prediction performance and the latter on feature importance. Both parts emphasise DHT and compare it to benchmarks. We design a battery of (sub-) experiments to assess survival models from different angles and test their robustness. Figure 2 offers a graphical summary of our experimental design. The online companion provides further information in Table S2 of the supplementary information.

The first three experiments of Analysis 1 use pre-crisis data and different sets of covariates. The goal is to appraise the predictive power of different types of variables and the degree to which survival models cope with dimensionality. The first two experiments work with loan-level (Exp. 1.1) and macroeconomic (Exp. 1.2) variables, respectively. Exp. 1.3 uses all available covariates.

The second experiment (Exp. 2) estimates models on pre-crisis data (1999–2006) and assesses

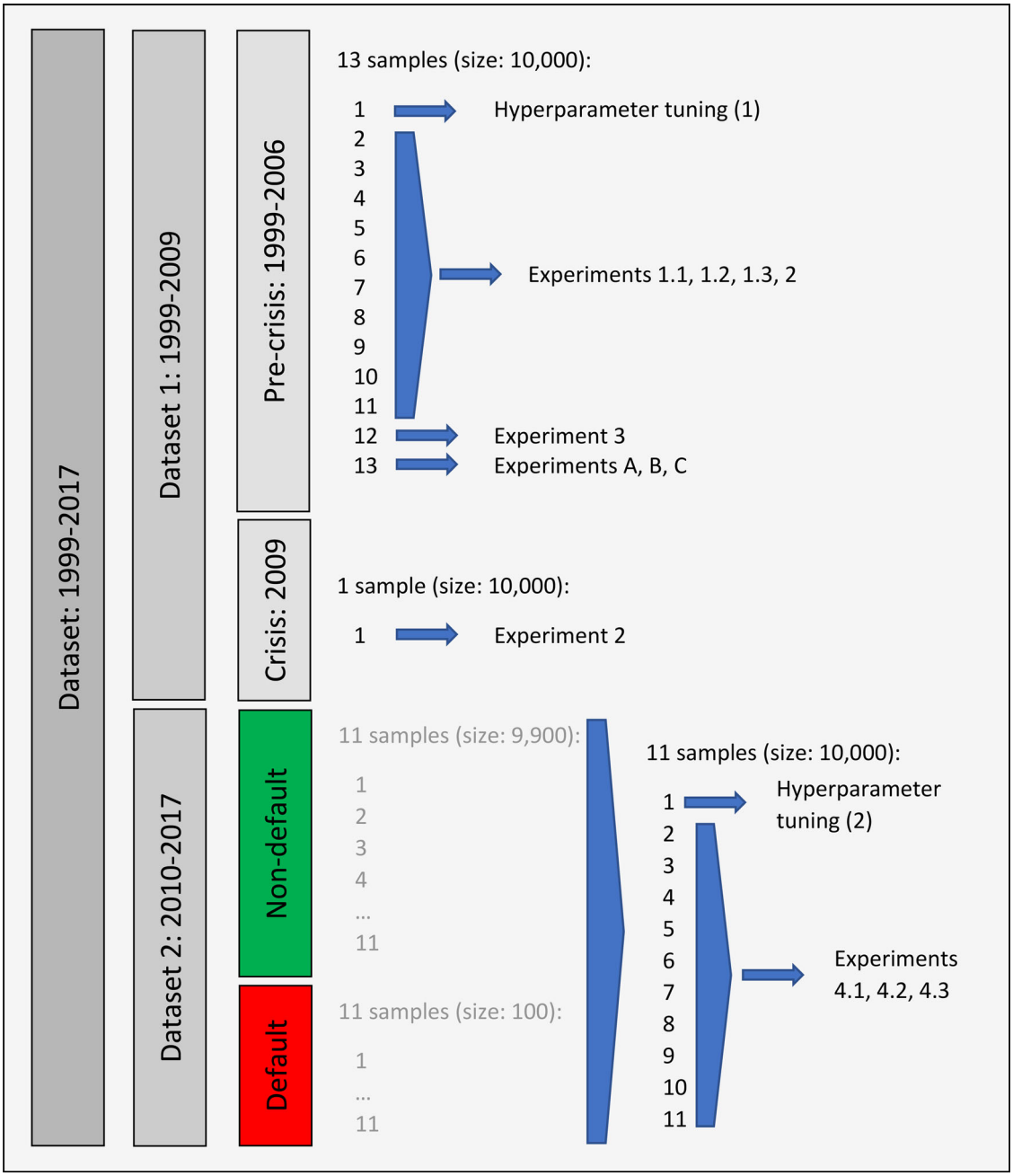

Figure 2. Organisation of the dataset to design (sub-)experiments for assessing survival models before and during the financial crisis and in recent years.

predictive performance in 2009. Assuming the training and test set to exhibit different structural relationships due to the financial crisis, Exp. 2 facilitates exploring models' resilience towards structural breaks between the training and test data and clarifies their robustness towards exceptional conditions during out-of-time validation. Using data from 2009 for the test set, we let model predict into the peak of the financial crisis shortly after the bankruptcy of Lehman Brothers on 15 September 2008. To make the modelling task even harder, we use only macroeconomic covariates.

A third experiment (Exp. 3) uses the same covariates as Exp. 1.1 but reduces the training data to 500 observations. This is to appraise the degree to which ML-based models depend on the size of the training set. We require the data to include both event types (e.g. to tune meta-parameters of ML/DL models) and found the value of 500 instances to be the smallest feasible sample size for the experiment.

To test model performance on more recent data and expose the models to a dataset with a different distribution of event occurrences, we replicate the experiments Exp. 1.1, 1.2, and 1.3 using the data from 2010 to 2017. We label the corresponding experiments as Exp. 4.1, 4.2, and 4.3. Note that the set of features differs slightly between the two streams of experiments because some variables from the set of macroeconomic variables are not available in the recent data (see Table 2).

To evaluate variable importance in Analysis 2, we compare the feature importance scores produced

from DHT using the approach proposed in Section 2.6. We re-use the variable subsets from Exp. 1.1, 1.2, and 1.3 for corresponding tests and label experiments as Exp. A, B, and C. We also estimate feature importance scores for the post-crisis dataset (Dataset 2). Interested readers find corresponding results together with a sensitivity analysis of how importance scores vary across pre- and post-crisis data in the online companion (see supplementary information). We also compare our approach to compute DHT feature importance to RSF importance scores based on Ishwaran (2007).

Implementing the experiments involved several decisions related to data partitioning, meta-parameter tuning, choice of packages/libraries for different algorithms, etc. The online companion of the article provides a comprehensive description of such choices to facilitate replication of the study. In a nutshell, we augment the temporal splitting of the data for individual experiments by random subsampling to train and test models over different folds each of which has a size of 10,000 instances (Exp. 2 being an exception). RSF proved computationally demanding and prohibited the consideration of larger subsamples. Given the subsamples, we train and test models using roughly 80% and 20% of the data, respectively, in a cross-validation design, whereby we reserve onefold for tuning algorithmic meta-parameters via grid-search and random search for RSF and DHT, respectively (see Figure 2).

3.3. Performance evaluation

To assess the predictive performance of survival models, we use the time-dependent concordance index (Antolini et al., 2005), which is based on the idea that an instance i experiencing event k_i at time t_i should be assigned to a higher cumulative risk for event k_i at that time point than an instance j that has not experienced any event until t_i. The concordance index for event k and time point t is then obtained by calculating the portion of pairs that satisfy $F_k(t_i x_i) > F_k(t_i x_j)$ amongst all acceptable pairs:

$$C_k^{td}(t) = \frac{\sum_{i \neq j} A_{k,i,j} * \mathbf{1}(F_k(t_i x_i) > F_k(t_i x_j))}{\sum_{i \neq j} A_{k,i,j}}$$

Randomly guessing cumulative risks for each instance would lead to a concordance index of 0.5 on average. Increasing values indicate higher predictive power, with a value of 1 being the highest attainable value and indicating that all acceptable pairs have been ranked correctly.

We consider the concordance index at 24, 48, and 72 months after loan issuance for each event type. Recall that we obtain multiple estimates for the concordance index due to subsampling the data. To assess models, we examine the average concordance index (across subsamples) for the above periods and an overall average (across the three periods) per model. We report these values for each risk individually and calculate a total average concordance index per model and experiment through averaging the risk-specific indices.

4. Empirical results

4.1. Model performance comparison

Table 3 reports the results of the survival model prediction comparison (Analysis 1) at the three different time points per event and the corresponding cause-specific and total means. For the reader's convenience, we multiply the concordance index by 100 to ease readability. We use boldface to highlight the single model that gives the overall best performance (i.e. highest concordance index across both risks $ØC$). In addition, we use italic font to identify the models with the highest concordance index for prepayment $ØC_1$ and default $ØC_2$, respectively. The concordance indices of Table 3 represent averages, which we compute across 10 random train/test splits. Detailed results for individual test sets are available in the supplementary information. Based on the results from the 10 repetitions, we test whether DHT performs significantly better than a benchmark using a pairwise unequal variances t test. The last column of Table 3 reports the corresponding results. An asterisk depicts p values less than five percent and two asterisks p values less than 1%.

Analysis 1 provides strong evidence in favour of DHT. It achieves the highest overall concordance index in every experiment. Table 3 also evidences the superiority of ML models. We observe the best results for the cause-specific and overall means of the concordance index for DHT or RSF but never for a statistical model. Rather, DHT consistently outperforms CSC and FGR significantly. Occasionally, DHT also performs significantly better than RSF, whereas RF consistently outperforms the statistical benchmarks. Amongst the statistical models, we find the Fine-Gray model (FGR) to outperform the cause-specific Cox model (CSC) across most settings and periods. Considering the two risks, we observe higher performance for default predictions than for prepayment predictions across all models and experiments. In the majority of cases, survival models perform better when estimating event occurrences at earlier time points (e.g. $ØC_{[1,2]}(24)$) than later time points (e.g. $ØC_{[1,2]}(78)$).

When comparing the $ØC$-results from Exp. 1.1 and Exp. 1.2, all models except CSC perform better with loan-level data than with macroeconomic variables. The difference is less pronounced for DHT.

Table 3. Results of Analysis 1: Period-specific, event-specific, and overall concordance for mortgage prepayment (risk 1) and default (risk 2) across experimental settings. (An asterisk depicts p values less than 5% and two asterisks p values less than 1%.)

Experiment 1.1	$C_1(24)$	$C_1(48)$	$C_1(72)$	$C_2(24)$	$C_2(48)$	$C_2(72)$	$ØC_1$	$ØC_2$	$ØC$	
CSC	78.12	52.20	52.01	96.01	76.46	61.92	60.78	78.13	69.45	**
FGR	81.92	81.56	81.56	96.10	96.02	95.22	81.68	95.78	88.73	*
RSF	87.26	81.07	76.83	99.32	99.19	99.11	81.72	*99.20*	90.46	
DHT	87.49	83.75	77.88	98.74	98.09	97.00	*83.04*	97.94	**90.49**	
Experiment 1.2	$C_1(24)$	$C_1(48)$	$C_1(72)$	$C_2(24)$	$C_2(48)$	$C_2(72)$	$ØC_1$	$ØC_2$	$ØC$	
CSC	79.11	66.77	59.51	77.68	71.73	70.11	68.46	73.17	70.82	**
FGR	79.59	77.41	77.19	74.64	72.58	71.41	78.06	72.87	75.47	**
RSF	89.20	79.06	70.73	75.90	71.86	71.82	79.66	73.20	76.43	**
DHT	94.80	93.13	86.47	93.99	90.35	75.77	*91.47*	*86.70*	**89.09**	
Experiment 1.3	$C_1(24)$	$C_1(48)$	$C_1(72)$	$C_2(24)$	$C_2(48)$	$C_2(72)$	$ØC_1$	$ØC_2$	$ØC$	
CSC	79.47	58.42	52.16	96.35	70.79	56.43	63.35	74.53	68.94	**
FGR	83.92	82.88	82.60	96.57	95.87	94.96	83.14	95.80	89.47	**
RSF	91.27	82.22	74.29	99.48	99.27	99.20	82.60	*99.32*	90.96	**
DHT	93.92	92.99	91.85	98.52	97.70	96.62	*92.92*	97.61	**95.27**	
Experiment 2	$C_1(24)$	$C_1(48)$	$C_1(72)$	$C_2(24)$	$C_2(48)$	$C_2(72)$	$ØC_1$	$ØC_2$	$ØC$	
CSC	58.47	52.84	52.92	67.48	62.72	61.50	54.74	63.90	59.32	**
FGR	49.96	53.61	52.81	68.41	65.29	64.39	52.12	66.03	59.08	**
RSF	79.79	73.83	67.59	61.23	47.61	47.60	*73.74*	52.15	62.94	**
DHT	65.16	66.59	59.17	82.18	85.47	85.50	63.64	*84.38*	**74.01**	
Experiment 3	$C_1(24)$	$C_1(48)$	$C_1(72)$	$C_2(24)$	$C_2(48)$	$C_2(72)$	$ØC_1$	$ØC_2$	$ØC$	
CSC	79.77	58.61	50.24	75.58	68.99	55.58	62.87	66.72	64.79	**
FGR	81.97	81.57	80.06	93.16	92.34	92.14	81.2	92.54	86.87	*
RSF	86.23	79.41	76.08	98.34	98.39	98.37	80.57	*98.36*	89.47	
DHT	85.80	84.38	83.79	95.1	96.28	92.77	*84.66*	94.72	**89.69**	
Experiment 4.1	$C_1(24)$	$C_1(48)$	$C_1(72)$	$C_2(24)$	$C_2(48)$	$C_2(72)$	$ØC_1$	$ØC_2$	$ØC$	
CSC	82.09	79.15	69.41	99.28	96.72	96.61	76.88	97.54	87.21	**
FGR	81.64	78.93	69.70	98.32	98.86	97.99	76.76	98.39	87.57	**
RSF	89.35	84.27	73.89	99.81	99.40	98.38	82.50	*99.20*	90.85	**
DHT	88.98	85.54	83.00	99.00	99.05	97.56	*85.84*	98.54	**92.19**	
Experiment 4.2	$C_1(24)$	$C_1(48)$	$C_1(72)$	$C_2(24)$	$C_2(48)$	$C_2(72)$	$ØC_1$	$ØC_2$	$ØC$	
CSC	83.26	80.90	73.93	74.92	68.64	73.10	79.37	72.22	75.79	**
FGR	83.32	81.14	74.56	69.74	64.23	71.16	79.67	68.38	74.02	**
RSF	89.85	83.34	74.72	69.77	71.86	75.16	82.64	72.26	77.45	**
DHT	89.61	85.96	82.92	81.18	77.64	74.29	*86.17*	*77.71*	**81.94**	
Experiment 4.3	$C_1(24)$	$C_1(48)$	$C_1(72)$	$C_2(24)$	$C_2(48)$	$C_2(72)$	$ØC_1$	$ØC_2$	$ØC$	
CSC	86.96	83.46	72.85	99.14	98.99	97.68	81.09	98.60	89.85	**
FGR	86.46	84.35	75.47	98.50	98.91	97.77	82.09	98.39	90.24	**
RSF	93.50	87.02	74.35	99.76	99.47	98.79	84.96	*99.34*	92.15	**
DHT	94.18	91.86	90.35	99.39	98.32	95.29	*92.13*	97.67	**94.90**	

Survival model abbreviations have the following meaning: CSC, cause-specific Cox; FGR, Fine and Gray; RSF, random survival forest; DHT, DeepHit.

Corresponding experiments with more recent data (i.e. Exp. 4.1 and 4.2) confirm loan-level variables to provide more information than macroeconomic variables.

Examining the $ØC$-performances between Exp. 1.1 and Exp. 1.3, we observe model performance to improve when combining loan-level information with variables that characterise macroeconomic conditions. Only the CSC model exhibits a small decrease in the overall mean concordance index, which might come from this model being affected by the higher dimensionality in Exp. 1.3 compared to Exp. 1.1. DHT appears especially successful in distilling additional information from the macroeconomic variables over and above what is contained in the loan-level covariates. The concordance indices of DHT are substantially larger in Exp. 1.3 compared to Exp. 1.1. When examining corresponding results of Exp. 4.1 versus 4.3, we also observe a trend of macroeconomic variables providing additional information. The cause-specific and overall mean concordance indices $ØC_1$, $ØC_2$, and $ØC$ are higher in Exp. 4.3 compared to Exp. 4.1 in the majority of cases. Furthermore, the cause-specific mean concordance indices reveal that the importance of macroeconomic variables is generally higher when predicting prepayment, whereas performance improvements are often marginal for default risk. Considering the results from more recent data (i.e. Exp. 4.1 and Exp. 4.3), we even observe a small performance decrease of DHT in $ØC_2^{4.3}$ compared to $ØC_2^{4.1}$. We attribute the higher value of macroeconomic variables in prepayment risk prediction to the fact that the loan-level characteristics in our data already facilitate accurate modelling of default risks. That is, the values of $ØC_2$ are consistently higher than the corresponding values of $ØC_1$ in Table 3.

In Table 3, comparing the panel of Exp. 1.2 to that of Exp. 2 facilitates drawing conclusions related to the robustness of survival models when

predicting into periods with adverse economic climate. Recall that in Exp. 2, we re-use the models trained in Exp. 1.2 and apply these to the crisis dataset for out-of-time validation. Training a survival model with data up to the year 2006 and using that model to generate predictions for 2009 sets the model to a difficult task. As expected, $ØC$-performances decrease in Exp. 2 compared to corresponding results in Exp. 1.2. Specifically, computing the relative differences between the concordance indices for individual periods and averaging over the resulting differences for $ØC_1$ and $ØC_2$, we find that $ØC^{CSC}$ decreases by 16.0%, $ØC^{FGR}$ by 21.3%, $ØC^{RSF}$ by 18.1%, and $ØC^{DHT}$ by 16.1%. The decrease in performance is more pronounced for prepayment prediction. Only RSF shows an opposite tendency in that the relative decrease in performance between Exp. 1.2 and Exp. 2 is much higher for default prediction than for prepayment prediction. It is difficult to appreciate the observed magnitude of performance deterioration. The financial crisis had a substantial impact on the economy and the stressed housing market. Therefore, decreases of the concordance index of roughly 20% appear plausible but should be taken as case-based evidence, which applies to the specific data, variables, and survival models studied here. More generally, predictive models that rely on past data are vulnerable to changes in the conditions under which they have been estimated; including the distribution of covariates and the relationship between covariates and the target variable. Forecasting risks under extreme economic conditions such as those during the financial crisis or recently under the Covid-19 pandemic is especially difficult and cannot be accomplished with empirical models alone. However, for the specific data studied here, we observe no evidence that would question the use of ML-/DL-based survival models in principle. Rather, the default predictions coming from DHT lose the least accuracy in Exp. 2. We observe $ØC_2^{DHT}$ to remain above 84, which is much higher than what we observe for the benchmarks. Whilst we caution against overemphasising results from a single dataset, the empirical findings provide evidence that the DL model is not more vulnerable towards changes in the economic climate, or more generally a shift in data distributions, than alternative models.

Finally, Table 3 sheds light on the amount of data required to estimate survival models. To that end, we compare the results of Exp. 3 to corresponding results in Exp. 1.1, which differ only in the amount of data we use for model training (500 observations in Exp. 3 c.f. 10,000 observations in Exp. 1.1 per fold). According to Table 3, the decrease in the concordance index due to less training data, which we calculate in the same way as above, amounts to 6.7% for $ØC^{CSC}$, 2.1% for $ØC^{FGR}$, 1.1% for $ØC^{RSF}$, and 0.9 for $ØC^{DHT}$. Again, the two ML-based approaches give strong results and do not appear to require more training data than statistical survival models in the form of CSC and FGR. Clearly, this finding requires external validation using different datasets. However, given limited experience with ML-based survival models (see Table 1) and noting that data availability may be a concern in credit risk management, especially in non-retail settings, it is appealing to observe small-sized portfolios to not rule out ML per se. Of course, sample sizes of 500 observations may still be difficult to obtain for specialised portfolios and the reason that prohibited testing smaller sizes in the article, namely ensuring a sufficient number of events, may be a serious problem in practice, for example in low-default portfolios. We argue that corresponding "scarce-data" settings are not the right application field for ML and require other solutions. The results of Exp. 3 provide a first reference at which sample sizes one might consider ML and DL survival models. Concerning low-default portfolios, we refer interested readers to supplementary information, which examines whether the small number of defaults in our post-crisis data impedes predictive modelling. Rerunning experiments with the strongest imbalance and applying the SMOTE algorithm to mitigate class skew, we find all survival models do not benefit substantially from data rebalancing, which suggests that the results of Table 3 are robust towards a potential imbalance effect.

4.2. Variable importance estimation with DeepHit

The opaqueness of ML models is a well-known concern amongst practitioners and regulators. To gain insight into the feature-target relationship that DHT models embody, we use the approach proposed in Section 2.6 for measuring variable importance in DHT. The main body of the article exemplifies the experiments performed in this part (Analysis 2) and summarises relevant findings. Detailed results can be found in the online companion of the article (see Tables S7–S14 in the supplementary information and Tables S23–S44 in the supplementary information). Table 4 reports feature importance scores for DHT in Experiments A, B, C aggregated over the two risks of default and prepayment. The individual rankings underneath the correlation analysis are available in the online companion.

The differences of the concordance index before and after corrupting a variable through adding random noise reveal that DHT bases its predictions to

Table 4. Feature importance scores and ranking for DHT in the pre-crisis period 1999–2006 when using only loan-level covariates (Exp. A), only macroeconomic covariates (Exp. B), and both of these together (Exp. C).

Experiment A			Experiment B			Experiment C		
Variable	$\Delta\emptyset C$	Rank	Variable	$\Delta\emptyset C$	Rank	Variable	$\Delta\emptyset C$	Rank
int.rate	−10.18	4	hpi.st.d.t.o	−41.22	1	int.rate	−7.15	18
orig.upb	−8.93	6	hpi.zip.o	−29.35	5	orig.upb	−6.78	19
fico.score	−4.26	7	hpi.zip.d.t.o	−33.64	2	fico.score	−3.96	22
dti.r	−1.63	9	ppi.c.FRMA	−13.44	10	dti.r	−2.98	25
ltv.r	−3.15	8	TB10Y.d.t.o	−9.37	12	ltv.r	−6.31	20
bal.repaid	−34.03	1	FRMA30Y.d.t.o	−20.22	7	bal.repaid	−23.44	6
t.act.12m	−11.08	3	ppi.o.FRMA	−9.21	13	t.act.12m	−8.7	15
t.del.30d.12m	−9.78	5	equity.est	−4.04	14	t.del.30d.12m	−4.11	21
t.del.60d.12m	−20.72	2	hpi.st.log12m	−32.88	3	t.del.60d.12m	−13.86	8
			hpi.r.st.us	−30.47	4	hpi.st.d.t.o	−39	1
			hpi.r.zip.st	−19.98	8	hpi.zip.o	−29.47	4
			st.unemp.r12m	−2.62	16	hpi.zip.d.t.o	−30.28	3
			st.unemp.r3m	−3.13	15	ppi.c.FRMA	−12.33	9
			TB10Y.r12m	−16.88	9	TB10Y.d.t.o	−9.67	14
			T10Y3MM	−21.08	6	FRMA30Y.d.t.o	−10.23	13
			T10Y3MM.r12m	−10.47	11	ppi.o.FRMA	−7.79	17
						equity.est	−8.59	16
						hpi.st.log12m	−36.3	2
						hpi.r.st.us	−26.72	5
						hpi.r.zip.st	−10.48	12
						st.unemp.r12m	−3.09	24
						st.unemp.r3m	−3.09	23
						TB10Y.r12m	−16.26	7
						T10Y3MM	−10.92	11
						T10Y3MM.r12m	−11.3	10

a large extent on repayment information. Considering the leftmost panel of Table 4, the current repaid balance (bal.repaid) is the most important variable in the model. Examining the next most important variables points to a caveat of the proposed feature importance estimation method. Variables that count the number of months in which a borrower is not delinquent, or delinquent for more than 30 or 60 days receive high scores in Table 4. These variables are correlated. One may argue that the degree to which DHT anchors its predictions on one variable out of a set of correlated variables is arbitrary. For example, one might be concerned that the importance of the variable *t.del.30d.12m* is underestimated because of its high correlation with *t.del.60d.12m*. The former receives roughly half the importance score of the latter. In this regard, it is important to emphasise that the proposed approach estimates the importance of a feature to a model. This is different from estimating the strengths of the relationship between a feature and the actual target variable. Multicollinearity is a general impediment to permutation-based feature importance (Fisher et al., 2019). Although feature correlation does not impact the predictive performance of DHT, which follows from Table 3, addressing multicollinearity and redundancy amongst variables in a data preparation step is useful to enhance the insights from permutation-based feature importance analysis.

Results from Experiment B in Table 4 suggest the house price index-related variables to be especially important. Such variables also remain most important when transitioning from Experiment B to Experiment C. We observe that the magnitude of the feature importance scores for one variable differs moderately when moving from Experiment A or Experiment B to Experiment C. For example, our variable *int.rate*, capturing the initial interest rate, receives an importance score of −10.18 when using only loan-level variables and a score of −7.15 in Experiment C when using the full set of variables. We consider the stability of the importance scores as a sign that the proposed method is robust. A more comprehensive robustness check is performed in the supplementary information considering, amongst others, several different noise intensities. Corresponding results also indicate the robustness of the proposed approach to calculate feature importance scores. As a final check, we compare the feature importance scores extracted from DHT using our method to RF-based feature importance. We focus on dataset 1 because Table 3 suggests that the predictive performance of DHT and RSF is comparable across data subsets and periods. Therefore, using either dataset 1 or dataset 2 appears sufficient. The advantage of dataset 1 is that it incorporates a few more features (see Table 2). Detailed results at the level of individual features are available in the online companion.

Table 5 reveals a positive correlation between the feature importance rankings of DHT and RSF. Typically, the correlation is high, meaning that DHT and RSF rank features in a similar order. A similar ranking implies that both models agree on which variables are most important and that the features govern model predictions in a similar way. This also suggests that DHT and RSF capture similar patterns in the data for estimating the feature-target relationship. Results from the online companion confirm this view in that we often observe DHT and RSF rely on the same set of top-

Table 5. Rank correlation (Spearman's ρ) between the feature importance rankings of DHT and RFS in the pre-crisis period 1999-2006 when using only loan-level covariates (Exp. A), only macroeconomic covariates (Exp. B), and both of these together (Exp. C).

	Experiment A	Experiment B	Experiment C
Prepayment	0.92	0.74	0.73
Default	0.87	0.94	0.33
Total	0.99	0.90	0.63

ranked features. The only notable exception is Experiment C, where the importance of variables for default prediction differs substantially between DHT and RSF. Table S46 in the supplementary information reveals that this comes from DHT favouring macroeconomic variables related to the house price index, whereas the RSF ranking includes more loan-related variables in the top ranks. However, the corresponding ranking for the prepayment event does not display this pattern. Overall, the reported results as well as the many additional results of the online companion suggest that the proposed approach to calculating feature importance for DHT is viable and helps to shed light on how DHT uses features to form forecasts.

5. Conclusion

The article aimed at contributing to the empirical literature on survival models in credit risk management. We applied new modelling approaches to a large dataset of US mortgages in a competing risk setting. Previous studies have modelled competing risks in a simplified manner by applying a standard, single risk survival model for each risk individually. This approach neglects the fact that competing risks are mutually exclusive. The article introduced three techniques that avoid this simplification and promise a more suitable treatment of competing risks. Our literature analysis (Table 1), suggests that each of these techniques, FGR, RSF, and DHT, are new to the credit scoring literature and originally tested here. The study is also the first to use DL for survival modelling in credit scoring. Given the concern about black-box models in the community, another contribution of the article is to extend approaches for feature importance estimation to the DHT model and to demonstrate how the corresponding estimates shed some light on how DHT uses variables to make predictions. Whilst still far from achieving regulatory compliance, feature importance scores may be considered a first step towards readying advanced ML models for risk modelling practices. They can also help with building trust in black-box models amongst risk analysts in that the estimated importance scores can be checked against domain knowledge. For example, one experimental setting in Section 4.2 shows how DHT weighted house price index-related variables higher than other macroeconomic variables, which seems plausible for the specific dataset and a mortgage risk model.

The empirical results have implications for credit scoring practice and research. From a managerial point of view, the study supports decision-makers in judging the potential and industry-readiness of ML-based survival models. We find DHT and RSF to model default and prepayment risks substantially more accurately than established techniques. Especially DHT delivered excellent prediction performance. Based on this observation, analysts working in model validation could consider DHT or RSF as benchmark models, for example, when validating an IFRS9 model. Considering that the literature advocates the merit of survival models over simpler classification models for many years (e.g. Banasik et al., 1999) and provided regulatory frameworks do not prohibit the use of advanced ML technology, which could be the case for non-banking financial institutions, the models considered here could even be employed for credit decisioning. The study also addressed two reservations, data insensitivity and robustness, a practitioner might have against ML models. First, the observed results suggest that neither DHT nor RSF requires magnitudes more training data than statistical survival models. Second, when training an ML model on data from favourable economic conditions and using that model to predict risks in a stressed economy (i.e. the crisis dataset), we observed the accuracy of DHT and RSF to decrease. However, the magnitude of the performance deterioration was similar to that observed for the statistical benchmarks. This suggests that the ML approaches are not less robust towards changes in the economy. In summary, the tested ML approaches displayed features – high predictive accuracy, robustness, and a moderate appetite for training data – that appeal to baking practice and indicate that they are worth exploring.

Form an academic point of view, the article introduces a new methodology and adds empirical insight to the literature on survival analysis in credit scoring. Most notably, we find evidence against the practice of modelling competing risks as independent events. The CSC model can be considered the de facto standard in prior research. We observe FGR, RSF, and DHT to consistently outperform this approach and recommend that future work on competing risks should use one of these more powerful techniques.

Note

1. Source: http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.page.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Funded by Deutsche Forschungsgemeinschaft in thescope of IRTG 1792.

References

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning. *Risks*, *6*(2), 38–20. https://doi.org/10.3390/risks6020038

Antolini, L., Boracchi, P., & Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine*, *24*(24), 3927–3944. https://doi.org/10.1002/sim.2427

Austin, P. C., & Fine, J. P. (2017). Practical recommendations for reporting Fine-Gray model analyses for competing risk data. *Statistics in Medicine*, *36*(27), 4391–4400. https://doi.org/10.1002/sim.7501

Baesens, B., Van Gestel, T., Stepanova, M., Van den Poel, D., & Vanthienen, J. (2005). Neural network survival analysis for personal loan data. *Journal of the Operational Research Society*, *56*(9), 1089–1098. https://doi.org/10.1057/palgrave.jors.2601990

Banasik, J., Crook, J. N., & Thomas, L. C. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society*, *50*(12), 1185–1190. https://doi.org/10.1057/palgrave.jors.2600851

Bellotti, T., & Crook, J. N. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, *60*(12), 1699–1707. https://doi.org/10.1057/jors.2008.130

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cao, R., Vilar, J. M., & Devia, A. (2009). Modelling consumer credit risk via survival analysis. *Statistics and Operations Research Transactions*, *33*(1), 3–30.

Ciochetti, B. A., Deng, Y., Gao, B., & Yao, R. (2002). The termination of commercial mortgage contracts through prepayment and default: A proportional hazard approach with competing risks. *Real Estate Economics*, *30*(4), 595–633. https://doi.org/10.1111/1540-6229.t01-1-00053

Cohen, B. H., & Edwards, G. A. J. (2017, March). The new era of expected credit loss provisioning. *BIS Quarterly Review*, 39–56.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning [Paper presentation]. Proceedings of the 25th International Conference on Machine Learning (ICML'2008), Helsinki, Finland (pp. 160–167). ACM.

Deng, Y. (1997). Mortgage termination: An empirical hazard model with a stochastic term structure. *The Journal of Real Estate Finance and Economics*, *14*(3), 309–331. https://doi.org/10.1023/A:1007758412993

Deng, Y., & Liu, P. (2009). Mortgage prepayment and default behavior with embedded forward contract risks in China's housing market. *The Journal of Real Estate Finance and Economics*, *38*(3), 214–240. https://doi.org/10.1007/s11146-008-9151-1

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. https://doi.org/10.1080/07350015.2016.1260471

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*(6), 652–665. https://doi.org/10.1057/s41274-016-0128-9

Eder, B., (2019). A survey on the estimation of expected credit losses for IFRS 9 [Paper presentation]. The 26th Conference on Credit Risk and Credit Control (CRC XVI), Edinburgh, Scotland.

Fine, J. P., & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, *94*(446), 496–509. https://doi.org/10.1080/01621459.1999.10474144

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81.

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, *1*(0), 519–537.

Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J., & Lau, B. M. (2014). Random survival forests for competing risks. *Biostatistics (Oxford, England))*, *15*(4), 757–773. https://doi.org/10.1093/biostatistics/kxu010

Kleinbaum, D. G., & Klein, M. (2012). *Survival analysis: A self-learning text*. Springer.

Lee, C., Zame, W. R., Yoon, J., & Schaar, M. v d. (2018). DeepHit: A deep learning approach to survival analysis with competing risks [Paper presentation]. The 32nd AAAI Conference on Artificial Intelligence (AAAI'18), New Orleans, Louisiana.

Lessmann, S., Baesens, 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*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, *28*(1), 161–170. https://doi.org/10.1016/j.ijforecast.2011.01.006

Narain, B. (1992). Survival analysis and the credit-granting decision [Paper presentation]. Proceedings of Credit Scoring and Credit Control (CRC'92), Edinburgh, Scotland (pp. 109–121). Oxford: OUP.

Stepanova, M., & Thomas, L. C. (2001). PHAB scores: Proportional hazards analysis behavioural scores. *Journal of the Operational Research Society*, *52*(9), 1007–1016. https://doi.org/10.1057/palgrave.jors.2601189

Stepanova, M., & Thomas, L. C. (2002). Survival analysis methods for personal loan data. *Operations Research*, *50*(2), 277–289. https://doi.org/10.1287/opre.50.2.277.426

Tong, E. N. C., Mues, C., & Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research*, *218*(1), 132–139. https://doi.org/10.1016/j.ejor.2011.10.007

Yao, X., Crook, J., & Andreeva, G. (2017). Enhancing two-stage modelling methodology for loss given default with support vector machines. *European Journal of Operational Research*, *263*(2), 679–689. https://doi.org/10.1016/j.ejor.2017.05.017

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. https://doi.org/10.1016/j.ijforecast.2010.06.002