JOURNAL OF THE OPERATIONAL RESEARCH SOCIETY
JORS
Thomas Archibald, Jonathan Crook

Journal of the Operational Research Society

ISSN: 0160-5682 (Print) 1476-9360 (Online) Journal homepage: https://www.tandfonline.com/loi/tjor20

Random survival forest for competing credit risks

Halina Frydman & Anna Matuszyk

JOURNAL OF THE OPERATIONAL RESEARCH SOCIETY
https://doi.org/10.1080/01605682.2020.1759385

ORIGINAL ARTICLE

Random survival forest for competing credit risks

Halina Frydman[a] and Anna Matuszyk[b]

[a]Department of Technology, Operations, NYU Leonard N Stern School of Business, New York, New York, USA; [b]Institute of Finance, Warsaw School of Economics, Warsaw, Poland

ABSTRACT
Random survival forest for Competing Risks (CR Rsf) is a tree-based estimation and prediction method. The applications of this recently proposed method have not yet been considered in the extant credit risk literature. The appealing features of CR Rsf compared to the existing competing risks methods are that it is nonparametric and has the ability to handle high-dimensional data. This paper applies CR Rsf to the financial dataset which involves two competing credit risks: default and early repayment. This application yields two novel findings. First, CR Rsf dominates, in terms of prediction accuracy, the state of art model in survival analysis-Cox proportional hazard model for competing risks. Second, ignoring the competing risk event of early repayment results in an upwardly-biased estimate of the cumulative probability of default. The first finding suggests that CR Rsf may be a useful alternative to the existing competing risks models. The second has ramifications for the extant literature devoted to the estimation of the probability of default in cases where a competing risk exists, but is not explicitly taken into account.

ARTICLE HISTORY
Received 16 October 2019
Accepted 19 April 2020

KEYWORDS
Machine learning; brier score; competing credit risks; cox hazard model; cumulative incidence function of default

1. Introduction

Random survival forest for Competing Risks (CR Rsf) is a tree-based estimation and prediction method recently proposed in Ishwaran et al. (2014), where it was shown through simulations to provide accurate predictions of the cumulative probability function of each risk. However, except for a single medical application in Ishwaran et al. (2014), CR Rsf has not been applied to the data, in particular not to the credit risks data. The purpose of this paper is three-fold. The first is to introduce CR Rsf to the finance literature concerned with credit risk. The second purpose is to compare the default prediction performance of CR Rsf with the state of art model in survival analysis, namely the Cox proportional hazard model for competing risks (CR Cox). We do this comparison by applying both methods to the data set consisting of 7,874 car leases with 34 time-fixed covariates for each lease (CL data) granted by a Polish financial institution, in which two competing risks are default and early repayment of a lease. It shows that CR Rsf dominates CR Cox in terms of prediction accuracy as measured by integrated Brier score, a widely used measure of prediction accuracy. This finding suggests that CR Rsf provides a useful alternative to the existing competing risks models. Lastly this work illustrates with the CL data that ignoring the presence of a less important risk, namely of early repayment, results in an upward-biased estimate of the cumulative probability of default. This result has ramifications for the existing literature devoted to the estimation of the probability of default in cases where a competing risk or risks exist, but are not explicitly considered.

To provide the context for our discussion of CR Rsf, we first briefly review the parametric and semi-parametric competing risks models that have been proposed in the literature on credit risk. Typically, credit risk data are on personal or companies' loans, with default on a loan treated as a primary risk, and early repayment as a secondary risk. In the earlier literature (Stepanova and Thomas (2002), Baesens et al (2005), Chancharat et al. (2010)), CR Cox was prominently featured and discussed independently or in comparison with other competing risks models such as logistic regression and neural networks. In these papers, the cumulative probability function of default, referred to in the competing risks literature as the cumulative incidence function of default (CIF), was estimated assuming that all loans present in the sample just before the time of default, were at risk of default.

Later, Watkins et al. (2014) argued that the loans that mature should not be considered at a risk of default or early repayment. This argument led to the development of the competing risks mixture models for time to default. The mixture model assumes that there is a covariates'- dependent fraction of loans that will never default or be repaid early- these loans mature according to the terms of contract. They

CONTACT Halina Frydman hfrydman@stern.nyu.edu Department of Technology, Operations, NYU Leonard N Stern School of Business, 44 West 4th Street, Suite 8-55, New York, NY 10012, USA

proposed a parametric competing-risks mixture model which was extended to a semiparametric mixture model by Dirick et al. (2015) and Dirick et al. (2017).

As CR Rsf is a non-parametric tree-based estimation procedure, the incorporation of the mixture concept into it is challenging; we intend to consider it in our future work. Thus, similarly to the earlier literature, in our estimation of CR Rsf and CR Cox, we assume that all car leases present in the sample at the time of default are at risk of default.

For estimation of the CIF of default using CR Rsf, we select the covariates for the final model using minimal depth and variable importance (VIMP) criteria, which are selection criteria specific to random survival forests. We describe these selection criteria in more detail below.

The paper is organized as follows. Section 2 contains preliminary material including definitions of the functions of interest in the analysis of competing risks and their estimators. Section 3 describes CR Rsf, its splitting rules, methods for selecting variables and the integrated Brier score. The methods are applied to the CL data in Section 4.

2. Preliminaries

2.1. Data setup

An analysis of time-to-an event when a customer can experience more than one type of credit event is referred to as competing risks analysis. In general, a customer can experience any one of J distinct credit event types. For the i'th customer, we observe (T_i, δ_i, x_i), where $T_i = \min(C_i^0, T_i^0)$. Here T_i^0 is the time to a credit event, and C_i^0 is the censoring time. If $C_i^0 < T_i^0$ then a customer is said to be right censored at C_i^0, meaning that a customer hasn't experienced any credit event by the time C_i^0, which is the last time that he/she was observed. We let δ_i^0 be the indicator of an event type, $\delta_i^0 \in \{1, ..., J\}$ and $\delta_i = \delta_i^0 I(T_i^0 < C_i^0)$, where $I(A) = 1$ if event A occurred $I(A) = 0$, otherwise. Thus, when $\delta_i = 0$, the customer is right censored at C_i^0; otherwise $T_i^0 < C_i^0$ and if $\delta_i = j$ with $j \in \{1, ..., J\}$, the customer is said to have experienced a type j credit event at time T_i^0.

Finally, in a sample of n customers, $x_i = (x_{i1}, ...x_{ip}), 1 \le i \le n$, is a vector of p characteristics (covariates) of an i'th customer observed when a customer enters into a contract with a lending institution. If the contract is for a car lease, these could include car age, car price, length of a lease and a ratio of down payment to car price. Throughout we assume that observations on different customers are independent of each other, and that the censoring times are independent of the covariates, the event times and the event type. We also assume that T_i^0 is continuous and there is a maximum length of follow-up time τ.

2.2. Definitions

In the competing risks framework the functions of interests are: the cumulative probability function of event j, referred to as event j-specific cumulative incidence function (CIF) and an event j-specific hazard function. To simplify notation, in this subsection we suppress the dependence of these functions on the i'th customer's covariate vector x_i. The event j-specific CIF is defined as the probability that a type j event occurs at or before time t:

$$F_j(t) = P(T^0 \le t, \delta = j) = \int_0^t P(T^0 = u, \delta = j)du \equiv \int_0^t f_j(u)du, \tag{1}$$

and the event j-specific hazard function at time u is

$$\begin{aligned} h_j(u) &= \lim_{\Delta u \to 0} \frac{P(u \le T^0 < u + \Delta u, \delta = j | T^0 \ge u)}{\Delta u} \\ &= \frac{1}{P(T^0 \ge u)} \lim_{\Delta u \to 0} \frac{P(u \le T^0 < u + \Delta u, \delta = j)}{\Delta u} \\ &= \frac{1}{P(T^0 \ge u)} f_j(u), j = 1, 2, ..., J \end{aligned} \tag{2}$$

where $h_j(u)$ represents the instantaneous rate of a type j event occurring at time u in the presence of other competing events, $f_j(u)$ is the subdensity of T^0 corresponding to the j'th event, and $S(u) = P(T^0 \ge u)$ is a survival function. From (2)

$$f_j(u) = h_j(u)P(T^0 \ge u) = h_j(u)S(u), \tag{3}$$

where

$$S(u) = P(T^0 \ge u) \tag{4}$$

is a survival function. Combining (1) with (3) gives a useful expression for the j'th event-specific CIF

$$F_j(t) = \int_0^t h_j(u)P(T^0 \ge u)du. \tag{5}$$

The estimators of $S(u)$ and $F_j(t)$ and $h_j(u)$ are presented in the next section.

2.3. Nonparametric estimators

To describe the nonparametric estimators of $S(u)$ in (4) and $F_j(t)$ in (5), let $0 = t_0 < t_1 < t_2 < ... < t_m$ be $m < n$ distinct and ordered failure times in the sample $(T_{i,} \delta_i)_{1 \le i \le n}$. Further, let $m(t) = \max(k : t_k \le t)$, and $d_j(t_k) = \sum_{i=1}^n I(T_i = t_k, \delta_i = j)$ be the number of type j events occurring at $t_k, N_j(t) = \sum_{i=1}^n I(T_i \le t, \delta_i = j)$ the number of type j events occurring by time t, and let $Y(t_k)$ be the number of customers at risk at time t_k. The estimator of $h_j(u)$ in (2) is defined on $t_k, 1 \le k \le m$, as $\hat{h}_j(t_k) = d_j(t_k)/Y(t_k)$, and is equal to zero otherwise. It can be interpreted as the probability of failure at time t_k. Then the Kaplan-Meier estimator of survival function is

$$\hat{S}(t) = \prod_{k=1}^{m(t)} (1 - \hat{h}_j(t_k)), \qquad (6)$$

and j event-specific CIF, $F_j(t)$, is estimated using Aalen-Johansen estimator:

$$\hat{F}_j(t) = \sum_{k=1}^{m(t)} \hat{S}(t_{k-1}) \frac{d_j(t_k)}{Y(t_k)}, \qquad (7)$$

where by (6)

$$\hat{S}(t_{k-1}) = \prod_{l=1}^{k-1} (1 - \frac{d(t_l)}{Y(t_l)}), k \geq 1.$$

3. Random survival Forest for competing risks

3.1. Binary competing risks trees

We start with the description of a binary survival tree, and then indicate the required modifications for competing risks trees. Survival trees are binary trees grown by recursive splitting of tree nodes. A tree is grown starting at the root node, which is the top of the tree comprising all the data. The root node is split into two daughters: a left and right daughter nodes, using the splitting rule that maximizes survival difference between the daughter nodes. The process is repeated in a recursive fashion for each subsequent node. This way, the tree pushes dissimilar cases apart. Eventually, as the number of nodes increases, and dissimilar cases become separated, each node in the tree becomes homogeneous and is populated by cases with similar survival function.

The construction of competing risks tree essentially proceeds in the same way as for survival tree except there is number of choices of a splitting rule and they involve the hazard functions or cumulative incidence functions rather than survival function. The splitting rules are discussed in Section 3.3.

3.2. CR Rsf algorithm

CR Rsf is a fully nonparametric method which uses binary trees for the estimation of event-specific cumulative incidence functions and survival probability function. We assume that our data is divided into training and testing samples. The training sample is used to grow the competing risks forest and testing sample is used for prediction. A competing risk tree is grown using an independent bootstrap sample of the training data. Eventually, the survival tree reaches saturation point when no new daughters can be formed because of the criterion that each node must contain a minimum of $n_0 > 0$ unique cases. The most extreme nodes in a saturated tree are called terminal nodes. We summarize the steps required for constructing a competing risks forest noting that $S(t|x)$ and $F_j(t|x)$ are the survival function and j-specific CIF for the subject with a covariate vector x from the training sample.

Step 1. Using training sample, draw B bootstrap samples with replacement and of the same size as the training sample.
Step 2. Grow a competing risk tree for each bootstrap sample. At each node of the tree, randomly select $K \leq p$ candidate variables along with two random split points. The node is split using the candidate variable that provides the most distinct daughters as described in Section 3.3.
Step 3. Grow the tree to full size under the constraint that a terminal node should have no less than $n_0 > 0$ unique observations.
Step 4. Calculate $\hat{F}_{j,b}(t|x), (1 \leq j \leq J)$ and $\hat{S}_b(t|x)$ for the b'th tree.
Step 5. Take the average of each estimator over the B trees to obtain the ensemble estimates: $\bar{F}_j(t|x)$, $(1 \leq j \leq J)$ and $\bar{S}(t|x)$.

3.2.1. Final estimates

Step 5 calculates the final estimate of CIF and survival function for a customer x from a training sample as follows. Because of the binary nature of survival tree, customer x falls into only one terminal node for each tree, but can occur more than one time in that node. Let $c_{x,b}$, be the number of times customer x occurs in bootstrap sample b, and $\nu_b(x)$ denote the indices for customers that are in the same terminal node as the subject with x. For simplicity, we still use t_k's to denote the distinct and ordered uncensored times of the customers in $\nu_b(x)$ and $m(t) = \max\{k : t_k < t\}$. For the terminal node which contains x in the b'th tree we define the quantities for an event $j : d_{j,b}(t_k) = \sum_{l \in \nu_b(x_i)} c_{l,b} I$ $(T_l = t_k, \delta_l = j)$ is the number of type j events at time $t_k, d_b(t_k) = \sum_j d_{j,b}(t_k), Y_b(t_k) = \sum_{l \in \nu_b(x_i)} c_{l,b} I$ $(T_l > t_k)$ is the number of customers at risk at t_k. In terms of these quantities, the estimates from the b'th bootstrap sample of the survival function and the CIF for customer x are

$$\begin{aligned}\hat{S}_b(t|x) &= \prod_{k=1}^{m(t)} (1 - \frac{d_b(t_k|x)}{Y_b(t_k|x)}), \hat{F}_{j,b}(t|x) \\ &= \sum_{k=1}^{m(t)} \hat{S}_b(t_{k-1}|x) \frac{d_{j,b}(t_k|x)}{Y_b(t_k|x)}. \qquad (8)\end{aligned}$$

Note that all customers in the same terminal node with x have the same estimated survival function and the same estimated CIFs.

The ensemble estimates of the survival and CIFs are averages of those quantities obtained from each of the B competing risks trees, namely

$$\bar{S}(t|x) = \frac{1}{B}\sum_{b=1}^{B} \hat{S}_b(t|x), \bar{F}_j(t|x) = \frac{1}{B}\sum_{b=1}^{B} \hat{F}_{j,b}(t|x). \quad (9)$$

3.2.2. Predictions

For a customer with covariate x_i from the testing sample, drop it down the b'th tree, it eventually falls into a terminal node. The b'th tree prediction of survival function and j'th event-specific CIF are given by $\hat{S}_b(t|x_i)$ and $\hat{F}_{j,b}(t|x_i)$ computed as in (8), and the ensemble prediction of survival function and j'th event-specific CIF are given by $\bar{S}(t|x_i)$ and $\bar{F}_j(t|x_i)$ computed as in (9).

3.3. Splitting rules

In tree growing, the splitting rule is used to select the best split of a given node into a left (l) and a right (r) daughter node. We consider two risk– j specific hypotheses tests and the corresponding splitting rules for competing risks:

- Generalized log-rank test

$$H_0 : h_{j,l}(t) = h_{j,r}(t), \text{ for all } t \le \tau \quad (10)$$

  and
- Gray's test

$$H_0 : F_{j,l}(t) = F_{j,r}(t), \text{ for all } t \le \tau \quad (11)$$

Suppose that the proposed split for a node is of the form $x \le c$ and $x > c$ for a continuous predictor x. For (10), the test statistic is denoted by $L_j^{LR}(x, c)$, and for (11) by $L_j^{G}(x, c)$, where the superscripts LR and G refer to the log-rank test and Gray's tests respectively. The best split for (10) and (11) corresponds to the covariate x and the value of c that maximize $|L_j^{LR}(x, c)|$ and $|L_j^{G}(x, c)|$ respectively. For (10) it means that the best split maximizes the difference between the estimated j risk-specific hazard functions in the daughter nodes; for (11) it means that it maximizes the difference between the estimated j risk-specific CIFs in the daughter nodes. The explicit expressions for the test statistics are given in Ishwaran et al (2014). The risk-j specific log-rank splitting rule is useful when the main interest is in detection of variables that influence the risk-j specific hazard. However, if the main interest is in identifying variables that directly affect cumulative incidence function for j'th risk, the better choice is Gray's splitting rule. We will be using default-specific Gray's splitting rule in the application section. By default-specific we mean the rule, which considers prediction of the CIF of default more important than the prediction of the CIF of early repayment. But, if the aim is to treat the two risks on equal footing, one can consider composite splitting rule; for the detailed discussion of the event-specific and composite splitting rules, see Ishwaran et al (2014).

3.4. Prediction performance-Brier score

To measure the prediction performance of CR Rsf, we use the Brier score (BS) and an integrated Brier score (IBS). BS is the squared difference between the observed and predicted outcome. The Brier score for event j-specific CIF at a given time $t > 0$ is defined as

$$BS_j(t) = E\Big[\big(I(T_i^0 \le t, \delta = j) - \bar{F}_j(t|x_i)\big)^2\Big], \quad (12)$$

where $I(T_i^0 \le t, \delta = j)$ is an observed outcome and $\bar{F}_j(t|x_i)$ is the ensemble predicted CIF for this outcome, discussed in Section 3.2. Here, expectation is taken with respect to the data (x_i) of a customer i belonging to the testing data set. When there is no censoring, and D_M is the testing data set of size M, the empirical version of (12) is

$$\widehat{BS}_j(t) = \frac{1}{M}\sum_{i \in D_M} \Big\{I(T_i^0 \le t, \delta_i = j) - \bar{F}_j(t|x_i)\Big\}^2,$$

However, when there is right-censoring the empirical version of (12) takes a more complicated form: we first define the inverse of the probability of censoring weights

$$\hat{\omega}_i(t) = \frac{I(T_i \le t, \delta_i \ne 0)}{\hat{G}(T_i)} + \frac{I(T_i > t)}{\hat{G}(t)}, \quad (13)$$

and then

$$\widehat{BS}_j(t) = \frac{1}{M}\sum_{i \in D_M} \hat{\omega}_i(t)\big\{I(T_i \le t, \delta_i = j) - \bar{F}_j(t|x)\big\}^2, \quad (14)$$

where $\hat{G}(t)$ denotes the Kaplan-Meier estimate of the censoring distribution G. For an illuminating exposition of Brier score leading to (14), see Graf et al. (1999) and Gerds and Schumacher (2006)

Note that Brier score as defined in (14) corresponds to splitting the data once into the training and testing samples. However, if one data set is used to estimate the model and predict its performance there is a danger of overfitting. To avoid the overfitting, we use the bootstrap cross-validation method to split the original data set of size N (denoted by D_N) into B bootstrap training samples D_b and corresponding test samples $M_b = D_N\ D_b, (b = 1, \dots, B)$, The models are then trained on the bootstrap samples and evaluated on the test samples. The bootstrap cross-validation estimate of

Table 1. Description of covariates.

Predictor variables	Levels/ categories	Range	Total ($n = 7874$)
Factors			
Car type	vehicle = 1/ other = 0		7447/427
New/used car	new = 1/ used = 0		7390/484
Phone	yes = 1/ no = 0		6740/1134
Customer before	yes = 1/ no = 0		2636/5238
Continuous Median, IQR			
Company's net assets		0-9996	218.85 [0; 7159.8]
Down payment/car price		0-0.45	0.2 [0.1;0.3]
Company's annual turnover last year		0-9981	1380.00[277.87; 3247.027]
Company's annual income last year		0-9919	139.88 [40.44; 77.83]
Car price		13.82-926.8	70.73 [47.22; 107.31]
Residual value		0-237.9	0.9 [0.54; 2.56]
Residual value/car price		0-0.678	0.01 [0.0099; 0.01]
Annual costs of the company		0-180	0 [0; 0]
Discrete			
Company's age		0-95	8 [3;12]
# of all active previous contracts		0-15	0 [0; 0]
# of all previous contracts		0-67	0 [0;1]
# of all bad previous contracts		0-7	0 [0;0]
# of all completed contracts		0-66	0 [0;1]
Car age		0-11	0 [0; 0]
# of instalments		24-60	36 [36; 48]
# of employees		0-9641	6 [1;21]
Length of cooperation		0-15	0 [0; 1]
Qualitative with the reference category			
Legal form of the company (ref.: Freelancer)	Ltd. and partnership, Joint stock, Other		4104, 328, 39
Geographical Region (ref.: Region E)	A, B, C, D		1702, 2062, 1633, 1422
Branch (ref.: Not specified)	Services, Construction, Sales, Production, Health, Other		1143, 368, 629, 255, 283, 1135

The continuous covariates, except D/C = Down payment/car price, are in 1000s of PLN. D/C is a fraction. Company's age, Car's age and Length of cooperation are in years. IQR is an inter-quartile range.

the prediction error at time t is calculated by averaging over the test data sets:

$$\mathrm{BootCvErr}(j,t) = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{M_b}\sum_{i\in M_b}\hat{\omega}_i(t)\Big\{I(T_i\le t,\delta_i=j)-\hat{F}_{j,b}(t|x_i)\Big\}^2 \tag{15}$$

For the CL data, we use $B = 1000$ bootstrap samples obtained without replacement, each of size $M_b = 5275$, which is 67% of all observations, and estimate prediction error using (15) Then the integrated empirical Brier score is

$$IBS_j(\tau) = \frac{1}{\tau}\int_0^{\tau}\mathrm{BootCvErr}(j,t)dt, \tag{16}$$

which is a summary measure of the prediction error for event $j-$specific CIF.

3.5. Variable selection in CR rsf

3.5.1. Variable importance

Variables can be selected by filtering on the basis of their variable importance (VIMP). The VIMP for a variable x is calculated as follows. By the standard bootstrap theory, whenever we draw a bootstrap sample with replacement from the training data set we leave out about 37% of observations. These left out observations are referred to as out-of-bag (OOB) ones and the observations in the bootstrap sample as in bag ones. To calculate VIMP for a variable x, drop each OOB observation down its in-bag competing risks tree. Whenever a split for x is encountered, assign a daughter node randomly. The event j-specific CIF from each such tree is calculated and averaged. The VIMP for x is the prediction error for the original ensemble event j-specific CIF (obtained when each OOB observation is just dropped down its in-bag competing risks tree), subtracted from the prediction error for the new ensemble obtained using randomizing x assignments (Breiman (2001), Ishwaran (2007)). The prediction errors are computed using Integrated Brier score. Large importance values indicate variables with predictive ability, whereas zero or negative values identify nonpredictive variables.

3.5.2. Minimal depth

Minimal depth (Ishwaran et al (2010), Ishwaran et al. (2011)) is an alternative method for assessing the importance of the variables. It uses the inspection of the forest construction to rank variables. It assumes that variables with high impact on the prediction are those that most frequently split nodes nearest to the root node, where they partition the largest samples. Within each tree, node levels are numbered based on their relative distance to the root of the tree (with the root at 0). Minimal depth identifies important variables by averaging the depth

```
                         Sample size: 7874
                    Number of deaths: 350
                     Number of trees: 1000
           Forest terminal node size: 30
       Average no. of terminal nodes: 136.149
No. of variables tried at each split: 6
              Total no. of variables: 34
       Resampling used to grow trees: swr
    Resample size used to grow trees: 7874
                            Analysis: RSF
                              Family: surv
                      Splitting rule: logrank *random*
       Number of random split points: 2
                          Error rate: 8.49%
```

Figure 1. Basic output for Rsf based on car leases data when default is considered to be the only risk.

of the first split for each variable over all trees within the forest. In general, to select variables according to VIMP, we examine the VIMP values, looking for some point along the ranking where there is a large difference in VIMP measures. Given minimal depth is a quantitative property of the forest construction, Ishwaran et al. (2010) also derive an analytic threshold for evidence of variable impact. A simple threshold rule uses the mean of the minimal depth distribution, classifying variables with minimal depth lower than this threshold as important in forest prediction. Note that the minimal depth is non-event specific criterion whereas VIMP can be both event-specific and non-event specific. In the application that follows we will use event-specific VIMP.

4. Application to car-leasing data

4.1. Description of the data

The data comes from a Polish financial institution, and consists of 7,874 car leases containing information about the repayment status for each lease. The status of the lease belongs to one of the following categories: default, paid off earlier than stipulated by the contract (early repayment), paid off according to the contract's terms (matured lease), and still being repaid. The first two categories are the competing risks, the last corresponds to the right-censored leases and, as discussed in the Introduction, the matured lease is also treated as right-censored observation with respect to the competing risks categories.

The customers were small and medium-size enterprises which applied for the lease between December 2010 and March 2014. The lease contracts were signed for a period from 24 to 60 months. Among 7,874 customers, there were 350 (4.45%) defaults and 143 (1.43%) early repayments. We consider the remaining 7,381 customers as being right-censored. Of those, 2,219 (28.18%) repaid their leases on time. The data for each customer included 34 time-fixed covariates, which are listed together with their descriptive statistics in Table 1. As noted by a referee, our sample is imbalanced as it consists of only 493 events and 7,381 right-censored observations. In section 4.3 we show, in the competing risks setting, that making the sample more balanced deteriorates the prediction performance of both Rsf and Cox model.

Table 2. VIMP and md of the covariates in Rsf when default is considered to be the only risk. Only predictors with md < threshold value = 8.237 are included in the table.

Rsf for default			
#	Variables:	Md	VIMP
1	Company's net assets	4.060	0.062
2	# of instalments	2.700	0.055
3	Down payment/car price	2.720	0.054
4	# of all bad previous contracts	2.020	0.030
5	Company's age	4.040	0.017
6	Customer before	5.260	0.015
7	Company's annual turnover last year	4.340	0.013
8	Company's annual income last year	3.980	0.012
9	Ltd. and partnership	5.320	0.010
10	Length of cooperation	5.680	0.008
11	# of employees	5.310	0.005
12	Car age	4.340	0.004
13	Car price	5.340	0.004
14	Residual value	5.390	0.003
15	# of all previous contracts	6.850	0.003
16	Services branch	7.500	0.003
17	Phone	6.460	0.003
18	Region 1	6.500	0.003
19	Residual value/car price	4.830	0.002
20	# of all completed contracts	6.940	0.002
21	Region 3	6.520	0.002
22	Other branch	6.410	0.001
23	Firm joint stock	8.220	0.001
	Threshold	8.327	

4.2. Random survival Forest for predicting default

We start by considering default as the only risk. In this simple setting, we illustrate different aspects of fitting Rsf for default to the CL data. Figure 1 presents the basic output for Rsf obtained from randomForestSRC, version 2.9.1, the R package for implementing Rsf (and CR Rsf) created by Ishwaran and Kogalur (2013). It shows the default parameters and rules for building Rsf: 1000 trees are used to build a forest, each terminal node should have at least 30 observations, the number of randomly chosen variables for splitting is $6 \approx \sqrt{34}$, the number of randomly chosen split points $= 2$, the splitting rule used is logrank, and swr-bootstrap samples are obtained with replacement. Finally, the error rate of 8.49% refers to an internal error, which is computed using out-of-bag (OOB) ensembles. Its computation is detailed in Appendix A.

To select predictors for Rsf, we first use the minimal depth (md) threshold which is 9.639. This eliminates 11 covariates that fall above the threshold level, leaving 23 covariates. These are sorted in Table 2 from the largest to the lowest VIMP value.

Table 3. VIMP and md of the covariates in default-specific CR Rsf with early repayment as a competing risk. Only covariates with md $<$ threshold value $= 9.375$ are included in the table.

	CR Rsf for default		
#	Variables:	md	VIMP
1	# of instalments	2.390	0.065
2	Company's net assets	3.430	0.064
3	Down payment/car price	2.560	0.053
4	# of all bad previous contracts	2.190	0.029
5	Company's age	3.850	0.018
6	Company's annual turnover last year	4.460	0.015
7	Customer before	4.600	0.014
8	Company's annual income last year	3.730	0.012
9	Ltd.&partnership)	4.720	0.009
10	# of employees	4.370	0.009
11	Length of cooperation	5.180	0.007
12	Car price	4.930	0.005
13	Residual value	5.080	0.005
14	Car age	4.670	0.004
15	# of all active contracts	7.130	0.004
16	Residual value/car price	4.360	0.003
17	# of all previous contracts	5.280	0.003
18	# of all completed contracts	5.170	0.003
19	Region C	5.930	0.002
20	Region A	5.900	0.002
21	Services branch	5.780	0.002
22	Phone	5.650	0.001
23	Other branch	5.510	0.001
24	New/used car	7.170	0.001
25	Firm joint stock	6.810	0.001
26	Construction branch	7.290	0.001
	Threshold	**9.375**	

4.3. CR Rsf for predicting default in the presence of early repayment

We next fit the default-specific CR Rsf, which treats early repayment as a secondary risk. Gray's rule is used for splitting a node into daughter nodes. To select predictors, we see from Table 3 that the md threshold is 9.375, which eliminates eight covariates that fall above the threshold. The 26 remaining covariates are sorted in Table 3 by the default-specific VIMP values from the largest to the smallest.Comparing the VIMP columns in Tables 2 and 3, we see that the 11 most important covariates are the same in Rsf and CR Rsf, and appear in the same order, though, as the respective columns show, their VIMP values differ. Both tables and the ones that follow were obtained using rfsrc function from the R package randomForestSRC.

We compare the prediction performance of the two models, Rsf with 23 covariates and CR Rsf with 26 covariates, by computing their integrated Brier scores using (16). The IBS is 0.026 for each model, so their predictive performance is the same. We also compared prediction performance of Rsf and CR Rsf using for each only the covariates with VIMP>0.01, that is, using the 8 most important covariates according to VIMP measure in each, and obtained the same Brier score curves and integrated Brier scores as before. This shows that the variables with VIMP≤ 0.01 had no effect on predicting default. The integrated Brier scores are computed over an interval (0.46 months) as no defaults occurred after the 46th month. The 8 covariates with their md and VIMP measures for Rsf and CR Rsf are presented in Tables 4 and 5 respectively. Though both tables have the same sets of covariates, their VIMP orderings as well as their md values are different.

Table 4. Results for Rsf with eight covariates. All covariates have VIMP > 0.01.

#	Variables:	Md	VIMP
1	# of all bad previous contracts	1.307	0.062
2	# of instalments	1.379	0.053
3	Down payment/car price	1.595	0.052
4	Company's net assets	2.404	0.043
5	Company's age	2.632	0.023
6	Company's annual income last year	2.878	0.019
7	Company's annual turnover last year	3.241	0.017
8	Customer before	3.407	0.011
	Threshold	6.290682	

Even though Rsf for default and default-specific CR Rsf have the same predictive power, we demonstrate, using the 8 variable model for each, that their default predictions may be different. Consider a new hypothetical car lease with all covariates taking on median values shown in Table 1, except for the binary variable" Customer before", which is set to 0. In Figure 2, we plot the predicted cumulative probability of default as a function of the duration of a car lease estimated by each model. In Rsf for default, the predicted cumulative probability of default is computed as 1-predicted survival probability; in CR Rsf, this is the predicted CIF of default. We see that at each lease duration, the predicted cumulative probability of default for Rsf is above the one for CR Rsf. It shows that ignoring a competing event (in our case, early repayment) results in an upward-biased predicted probability of the primary credit risk (in our case, default). As pointed out by a referee, the estimated cumulative probability of an event of interest will always be upward-biased when a competing risk or risks are ignored by being treated as right-censored observations. The referee's argument is stated in Appendix B. The plots in Figure 2 were obtained using the R-package pec surveyed by Mogensen et al. (2012) and maintained by Gerds (2018).

We next show how we can assess the effect of a single covariate on the CIF of default using as an example covariate Down payment/Car price (D/CP). In Figure 3, we plot the predicted CIF of default when D/CP $= 0$, 0.2, and 0.4, keeping the other variables constant at the values described above. We see that the predicted CIF of default when D/CP $= 0.4$ is smaller at each duration than the CIF when D/CP $= 0.2$, which in turn is smaller than the CIF when D/CP $= 0$. As one would expect, this implies that the larger the ratio of down payment to car

Table 5. Results for the default-specific CR Rsf with eight covariates. All covariate have VIMP > 0.01.

#	Variables:	Md	VIMP
1	# of instalments	1.398	0.060
2	Company's net assets	2.461	0.055
3	Down payment/car price	1.680	0.052
4	# of all bad previous contracts	1.209	0.043
5	Company's age	2.634	0.024
6	Company's annual turnover last year	3.235	0.020
7	Customer before	3.358	0.018
8	Company's annual income last year	2.912	0.013
	Threshold	7.424943	

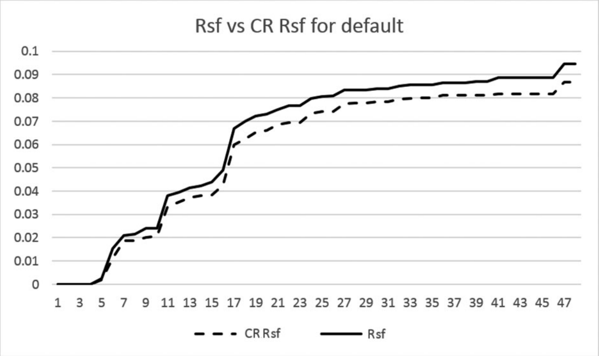

Figure 2. Predicted cumulative probability of default by the 8 variables Rsf (solid line) and by 8 variables CR Rsf (dashed line). Duration of a lease, on horizontal axis, is in months. The solid plot was obtained using plotPredictSurvProb and dashed one using plotPredictEventProb from the R package **pec**.

price, the smaller the cumulative probability of default.

4.4. Comparison of CR Rsf and CR Cox

Here, we compare the prediction performance of default-specific CR Rsf and default-specific CR Cox using CL data. More precisely, we compare how well the two methods predict the CIF of default. There are two ways to estimate the CIF of an event in CR Cox. The first estimates the cause specific hazard function of each of the competing events and derives from them the estimates of the CIFs. The second, known as the Fine-Gray method (Fine & Gray 1999), directly estimates the CIF of default as a function of predictors. Here, we use the Fine-Gray method, as our interest is in assessing the effects of covariates on the CIF. The first method does not allow for such assessment; it is used when one is interested in assessing the influence of covariates on the hazard rate function. For the accessible exposition of the two methods see e.g. Haller et al. (2013) Sections 1-3.1.

To select covariates for CR Cox, we employed backward selection combined with AIC, as described in Kuk and Varadhan (2013). This resulted in a CR Cox model with eight covariates, which are listed together with their sub-distribution hazard ratios and standard errors in Table 6. Three of the selected covariates were log-transformed because they were right skewed. To assess the influence of a covariate on the CIF of default, we vary covariate D/CP while keeping the other continuous covariates at their median values and two binary covariates at a value of one. We that the coefficient of D/CP is negative, which implies that the plot of the CIF of default for a larger value of D/CP will be below the one with a smaller value of D/CP. This is illustrated in the left panel of Figure 4, which shows the expected ordering of the CIFs, and also large differences between the CIFs at different values of D/PC, which reflect the large absolute value of the coefficient of D/CP.

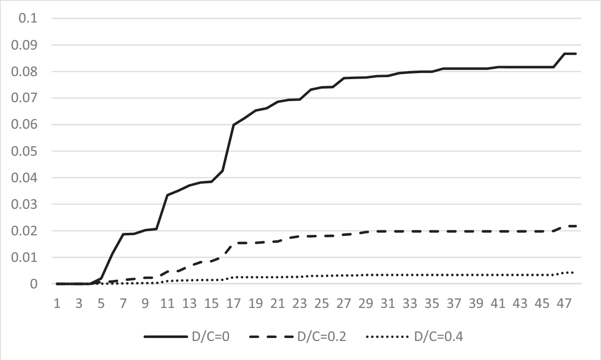

Figure 3. Predicted CIFs of default by the 8 variables CR Rsf when D/CP = 0 (solid line), D/CP = 0.2 (dashed line) and D/CP = 0 (dotted line). Duration of a lease, on horizontal axis, is in months. The plots were done using plotPredictEventProb from **pec**.

To compare the prediction performance of default-specific CR Rsf and default-specific CR Cox, both with eight covariates, we show in Figure 5 the plots of Brier score for the reference, CR Cox and CR Rsf models. The reference plot corresponds to the Brier score for the nonparametric estimate of the CIF of default shown in (7). We see that the reference plot, as expected, is above the Brier score plots for both ensemble methods, and that at each point in time the Brier score for CR Rsf is lower than that for CR Cox, showing that CR Rsf dominates CR Cox in terms of predictive accuracy. Consequently, the IBS for CR Rsf, 0.0267, is smaller than the IBS for CR Cox model, which is 0.03.

We now investigate how the prediction accuracy of CR Rsf and CR Cox is affected by balancing our sample. To this end, we create four new samples which include $493(= 350 + 143)$ events, and $10\%, 25\%, 50\%$ and 75% of randomly selected right-censored observations from all 7,381 right-censored observations. The IBS of CR Rsf and CR Cox estimated from these new samples are shown in Table 7 together with the results for complete sample of $7,381 + 493 = 7,874$ observations discussed above. We observe that the more balanced the sample is the worse is the prediction performance of both CR Rsf and CR Cox. The results in Table 7 also show, that for each sample size, CR Rsf makes more accurate predictions than CR Cox.

Table 6. Results for default-specific CR Cox model. They were obtained using the wrapper function selectFGR and the function selectCox from **pec**.

	Coef	exp(coef)	se(coef)	z	p-value
Log company's annual turnover last year	−2.28631	1.02e-01	0.477058	−4.79	1.6e-06
Log company's net assets	0.60145	1.82e + 00	0.284660	2.11	3.5e-02
Car price	0.00246	1.00e + 00	0.000679	3.63	2.8e-04
Log # of all bad previous contracts	2.43064	1.14e + 01	0.137388	17.69	0.0e + 00
Down payment/car price	−10.32305	3.29e-05	0.634431	−16.27	0.0e + 00
Customer before	−0.76519	4.65e-01	0.185073	−4.13	3.6e-05
New/used car	−0.56544	5.68e-01	0.205710	−2.75	6.0e-03
Log length of cooperation	−0.64125	5.27e-01	0.131675	−4.87	1.1e-06

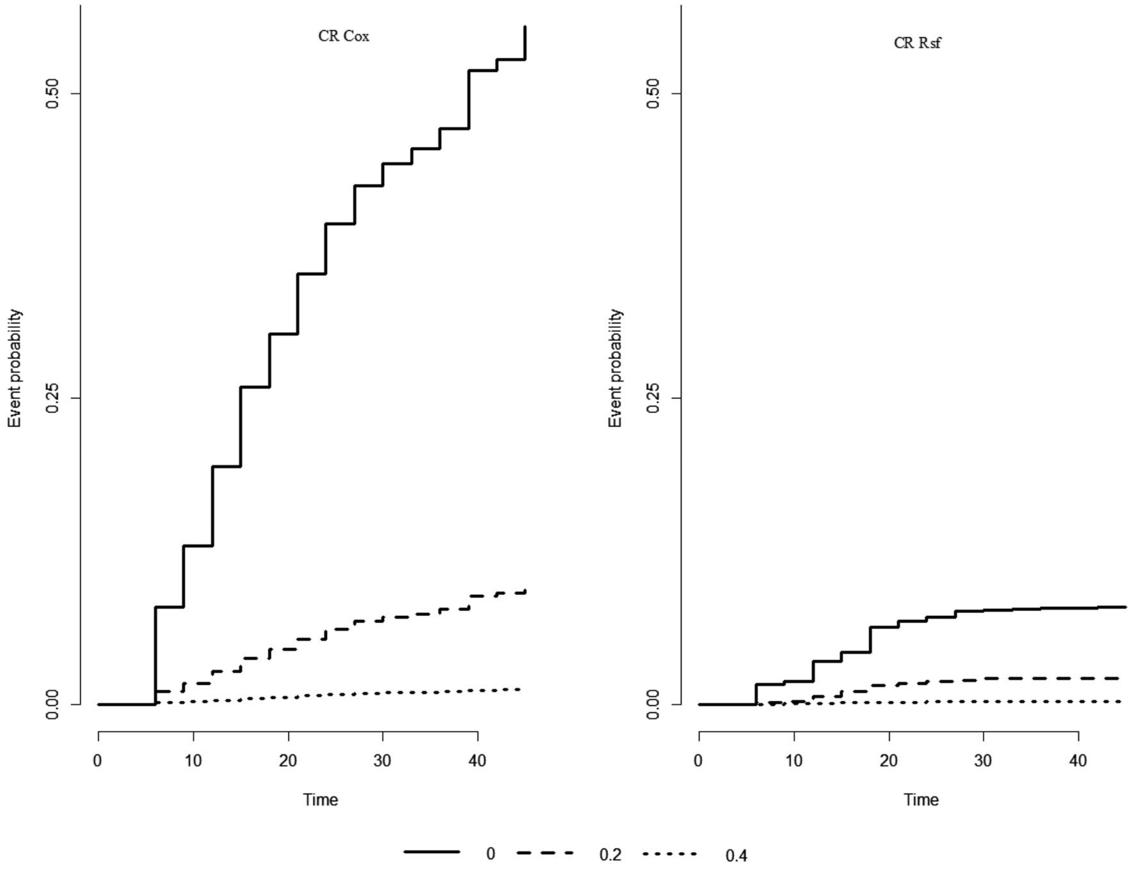

Figure 4. Predicted CIF for the 8 variables CR Cox (left panel) and CR Rsf (right panel) when Down payment/car price = 0 (solid curve), =0.2 (dashed curve), and =0.4 (dotted curve). Duration of a lease, on horizontal axis, is in months. The plots were obtained with function plotpredictEventProb from **pec**.

Finally, we illustrate the difference between CR Rsf and CR Cox's default predictions. We note that the two methods share the covariate down payment/car price (D/CP). We consider three hypothetical car leases with covariates taking on constant values, as described before, except D/CP, which takes a value of 0 for the first lease, 0.2 for the second and 0.4 for the third. We see from Figure 4 that when down payment is 0.4, or 40% of the car price, both methods predict the CIF of default to be essentially 0 by the end of the 46th month, when D/CP = 0.2 or 20%, the predicted CIF is just above.0.02 for CR Rsf, and much larger, above 0.1, for CR Cox. But when D/CP = 0, the CIF of CR Cox is dramatically larger over the range of 7-46 months compared to the CIF of CR Rsf.

The comparison of CR Rsf with CR Cox based on the CL data shows that the two models select different sets of optimal covariates; that CR Cox makes more extreme predictions; and, importantly, that the predictions of CR Cox are less accurate as measured by the IBS. Thus, it suggests that CR Rsf may be a useful alternative to the competing risks models considered in literature on credit risk.

5. Conclusion

In this paper, we have introduced Random survival forest for competing risks to the literature on credit risk; we presented the basic concepts in modeling competing risks, described the algorithm for constructing CR Rsf, and explained how CR Rsf makes predictions. We showed, with the data on car leases, that CR Rsf dominates CR Cox-the most popular model in survival analysis-in terms of the prediction accuracy. In addition, we demonstrated, that even

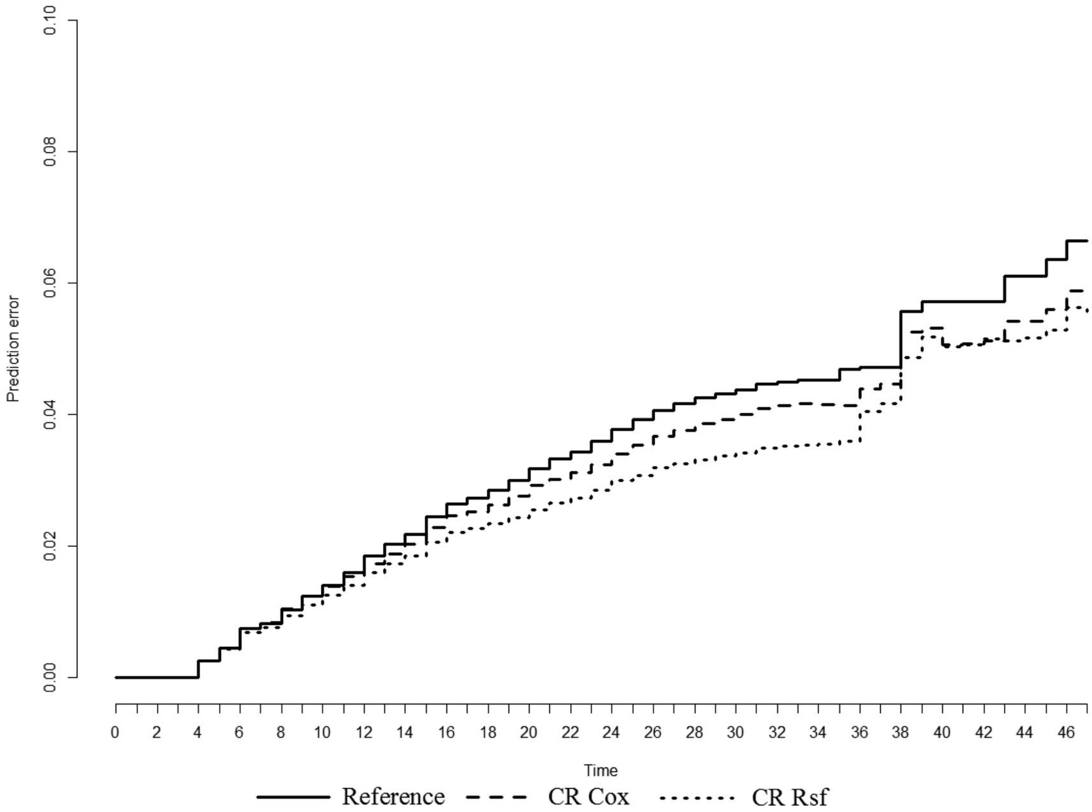

Figure 5. Brier score curves for Reference model (solid line), CR Cox (dashed line) and CR Rsf (dotted line). The plots were obtained with command crps and function ibs from pec.

Table 7. The IBC for CR Rsf and CR Cox models as functions of the percentage of randomly selected right-censored observations.

Percentage	IBS-Rsf	IBS-Cox
10	0.116	0.128
25	0.074	0.080
50	0.048	0.052
75	0.038	0.040
100	0.0267	0.03

when the percentage of primary credit events (defaults) is small and the percentage of secondary credit events (early repayments) is even smaller, ignoring the secondary events in the analysis, results in the overestimation of the cumulative default probability. An interesting goal for future research would be to treat leases (or other financial instruments), which mature, as non-susceptibles, and incorporate them as such into the CR-Rsf estimation framework.

Acknowledgements

We are grateful to Thomas Gerds for his informative responses to our many questions concerning the R-package pec. We thank the referees for their helpful comments, which improved the paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Anna Matuszyk received Fulbright Scholarship from the Polish-U.S. Fulbright Commission for spending academic year 2017-2018 at Stern School of Business and doing research in credit risk area. Otherwise, this research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

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

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

Chancharat, N., Tian, G., Davy, P., McCrae, M., & Lodh, S. (2010). Multiple state of financially distressed companies: Tests using a Competing-Risks Model. *Australian Accounting Business and Finance Journal*, *4*, 27–44.

Dirick, L., Claeskens, G., & Baesens, B. (2015). An Akaike information criterion for multiple event mixture cure models. *European Journal of Operational Research*, *241*(2), 449–457. https://doi.org/10.1016/j.ejor.2014.08.038

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

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

Gerds, T. A. (2018). *pec: Prediction error curves for risk prediction models in survival analysis*. R package "pec". Version 2018.07.26.

Gerds, T. A., & Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, *48*(6), 1029–1040. https://doi.org/10.1002/bimj.200610301

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, *18*(17–18), 2529–2545. https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5

Haller, B., Schmidt, G., & Ulm, K. (2013). Applying competing risks regression models: An overview. *Lifetime Data Analysis*, *19*(1), 33–58. https://doi.org/10.1007/s10985-012-9230-8

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, *1*(0), 519–537. https://doi.org/10.1214/07-EJS039

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*, *15*(4), 757–773. https://doi.org/10.1093/biostatistics/kxu010

Ishwaran, H., & Kogalur, U. B. (2013). *randomForestSRC: Random Survival Forests for Survival, Regression and Classification.*, R package version 3.6.4.

Ishwaran, H., Kogalur, U. B., Chen, X., & Minn, A. J. (2011). Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining*, *4*(1), 115–132. https://doi.org/10.1002/sam.10103

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., & Lauer, M. S. (2010). High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, *105*(489), 205–217. https://doi.org/10.1198/jasa.2009.tm08622

Kuk, D., & Varadhan, L. (2013). Model selection in competing risks regression. *Statistics in Medicine*, *32*(18), 3077–3088. https://doi.org/10.1002/sim.5762

Mogensen, U. B., Ishwaran, H., & Gerds, T. A. (2012). Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software*, *50*(11), 1–23. https://doi.org/10.18637/jss.v050.i11

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

Watkins, J. G. T., Vasnev, A. L., & Gerlach, R. (2014). Multiple event incidence and duration analysis for credit data incorporating nonstochastic loan maturity. *Journal of Applied Econometrics*, *29*(4), 627–648. https://doi.org/10.1002/jae.2329

Appendix A

We compute an internal error rate using the “out-of-bag observations”. The “out-of-bag” (OOB) observations for the b’th tree are those that are left out of a bootstrap sample. In fact about 37% of observations in the original sample are out-of-bag observations for each bootstrap sample. The OOB estimate of $F_j(x_i)$ can be obtained by using each of the B trees in which observation x_i is out-of-bag. It is given by

$$\bar{F}_j^{\mathrm{00B}}(t|x_i) = \frac{1}{|\mathcal{O}_i|}\sum_{b\in\mathcal{O}_i}\hat{F}_{j,b}(t|x_i),$$

where $\mathcal{O}_i \subset \{1,\ldots,B\}$ is the index set of trees for which x_i is out of bag. The OOB estimate of the integrated brier score for event j is given by

$$\widehat{IBS}_j(\tau) = \int_0^{\tau}\frac{1}{n}\sum_{i=1}^{n}\hat{\omega}_i(t)\{I(T_i \le t, \delta_i = j) - \bar{F}_j^{\mathrm{00B}}(t|x)\}^2,$$

where $\hat{\omega}_i(t)$ are the weights defined in (13).

Appendix B

Rsf considers only one risk (default). As a result, the CIF is a cumulative distribution function $F(t)$ with the property $\lim_{t\to\infty}F(t) = 1$. On the other hand, CR Rsf considers more than one risk (default and early repayment). In case of J competing risks, CIF for an event j, $F_j(t) = P(T^0 \le t, \delta = j)$, is not equal to a cumulative distribution function. The following equality holds: $F(t) = \sum_{j=1}^{J} F_j(t)$ and $\lim_{t\to\infty}F_j(t) = P(\delta = j)$. As a result, at any time point CIF for event j is lower than $F(t)$ calculated for the single risk assuming that other competing risks are censored. To sum up, for any competing risks data set the relation observed in Figure 2 will hold.