Technological and Economic Development of Economy
ISSN: 2029-4913 / eISSN: 2029-4921
2021 Volume 27 Issue 1: 96–119
https://doi.org/10.3846/tede.2020.13997

A DYNAMIC CREDIT SCORING MODEL BASED ON SURVIVAL GRADIENT BOOSTING DECISION TREE APPROACH

Yufei XIA[1*], Lingyun HE[2], Yinguo LI[1], Yating FU[3], Yixin XU[3]

[1]*Business School, Jiangsu Normal University, Xuzhou, PR China*
[2]*School of Economics and Management, China University of Mining and Technology, Xuzhou, PR China*
[3]*Sino-Russian College, Jiangsu Normal University, Xuzhou, PR China*

Received 29 April 2020; accepted 11 October 2020

Abstract. Credit scoring, which is typically transformed into a classification problem, is a powerful tool to manage credit risk since it forecasts the probability of default (PD) of a loan application. However, there is a growing trend of integrating survival analysis into credit scoring to provide a dynamic prediction on PD over time and a clear explanation on censoring. A novel dynamic credit scoring model (i.e., SurvXGBoost) is proposed based on survival gradient boosting decision tree (GBDT) approach. Our proposal, which combines survival analysis and GBDT approach, is expected to enhance predictability relative to statistical survival models. The proposed method is compared with several common benchmark models on a real-world consumer loan dataset. The results of out-of-sample and out-of-time validation indicate that SurvXGBoost outperform the benchmarks in terms of predictability and misclassification cost. The incorporation of macroeconomic variables can further enhance performance of survival models. The proposed SurvXGBoost meanwhile maintains some interpretability since it provides information on feature importance.

Keywords: credit scoring, survival analysis, survival gradient boosting decision tree, probability of default, consumer loan, machine learning.

JEL Classification: C53, D81, D14, G17.

Introduction

Since the global financial crisis, risk management has achieved much prominence and become a primary focus of both academia and industry. Among the various types of risks in financial institutions, credit risk, which is defined as the potential loss when a counterparty fails to meet his/her obligation, is regarded as the largest risk that financial institutions faces (Apostolik et al., 2009). The Basel Accord proposed several credit risk parameters to quantify credit risk. Financial institutions are also allowed to employ internal ratings-based

*Corresponding author. E-mail: *6020180093@jsnu.edu.cn*

(IRB) methods, which means that financial institutions can build their own quantitative models to estimate these risk parameters as Basel Accord II instructed. Although advanced IRB approach is prohibited for some types of assets in Basel Accord III, it can still be used to estimate key risk parameters for retailing portfolio. Among the credit risk parameters, probability of default (PD) has received much attention from banks and researchers since it supports decision making in consumer loans and the calculation of regulatory capital requirement (Crook et al., 2007). Credit scoring, defined as an empirical model-based prediction on the undesired behaviour of a potential borrower (Lessmann et al., 2015), is a popular measure to estimate PD in practice.

Credit scoring mainly includes three sequential stages: pre-modelling, modelling, and post-modelling. During the first phase, feature selection (Maldonado et al., 2017), wavelet analysis (Hung, 2019) and data transformation (Han & Ge, 2017) are utilized to provide a representative dataset. In the post-modelling stage, model validation (e.g., misclassification cost and profit-based evaluation measures) (Lohmann & Ohliger, 2019; Xia et al., 2017b), probability calibration (Bequé et al., 2017), credit rating migration (Huang et al., 2020; Liang et al., 2016) and interpretability (Munkhdalai et al., 2019) have been considered in prior studies. The pursuit of accurate model is the central task of modelling stage because a minor improvement of credit scoring model may incur enormous economic benefits (Finlay, 2011). Despite credit scoring models are established by clustering algorithms in some cases (Lim & Sohn, 2007), they are routinely built using classification methods. Statistical models (e.g., logistic regression and generalized additive models) are initially used for credit scoring modelling. Despite their transparency and easy-to-implementation, statistical methods hold some strong assumptions that are far from reality (e.g., linear separability or normal distribution of input data). Consequently, statistical methods are not comparable with machine learning methods in terms of predictability as suggested in several comprehensive comparative analysis (Baesens et al., 2003; Lessmann et al., 2015). Popular machine learning approaches used in credit scoring include decision tree (DT), support vector machine (SVM), artificial neural network (ANN), evolutionary algorithms, among others (Huang et al., 2007; Ong et al., 2005; West, 2000). Inspired by the famous "no free lunch theorem" (Wolpert & Macready, 1997), there is a growing trend that ensemble learning, which combines the predictions of multiple models, is extensively introduced to credit risk assessment mainly due to its superior predictive accuracy as shown in several comparative studies (Finlay, 2011; Wang et al., 2012). Random forest (RF) is even advocated as industry benchmark in credit scoring (Lessmann et al., 2015). Among the ensemble models, gradient boosting decision tree (GBDT) and its variant algorithms have been applied as a homogeneous ensemble model itself (Ma et al., 2018; Xia et al., 2020a, 2018b) or as a critical component of heterogeneous ensemble structure (Xia et al., 2018a).

However, these classification approaches are typically cross-sectional models, which share a number of drawbacks that can be further overcome by survival models. Survival models have a long history and was initially applied in medical research (Klein & Moeschberger, 2006). Over the past few decades, the application of survival models into credit risk assessment is becoming research hotspot (Malik & Thomas, 2010; Stepanova & Thomas, 2002; Tong et al., 2012). The advantages of survival models applied in credit scoring are:

1. Survival models can provide a dynamic PD prediction overtime (e.g., 12-months or lifetime PDs of loan portfolio as required in CECL and IFRS 9). A dynamic PD means that the financial institutions can precisely adjust the capital charge and collection strategy during the payment of loans;
2. Survival models can offer a reasonable explanation on censoring, which contributes to a realistic and practical credit scoring model (Leow & Crook, 2016);
3. Survival models can be easily incorporated with time-dependent covariates (e.g., macroeconomic or behavioural covariates) (Bellotti & Crook, 2009).

Recent advancements with regard to applying survival models to credit risk assessment mainly concern on building accurate survival models. One path to achieve this goal is improving statistical models. Time-dependent covariates and coefficients are examined in-depth for survival function and partial likelihood function to consider the time-varying effects in credit scoring (Dirick et al., 2019; Djeundje & Crook, 2018, 2019; Leow & Crook, 2016). Another solution concerns on the fusion of survival models with machine learning. By far, survival models have been integrated with ANN (Baesens et al., 2005) and RF (Wang et al., 2018). In a benchmark study, Dirick et al. (2017) compared several classical survival models used in credit scoring and revealed that Cox PH-based models provided a good performance especially in combination with spline methods. However, few studies have combined survival model with GBDT-based techniques despite GBDT showed superiority relative to classical classifiers when using cross-sectional data (Xia et al., 2017a). Table 1 includes a selection of research applying survival models to credit scoring, which shows that limited studies have incorporated survival analysis into machine learning. We aim to overcome this gap in this paper. Validated on a large dataset of consumer loans, we develop SurvXGBoost, a survival gradient boosting decision tree approach that combines XGBoost (Chen & Guestrin, 2016) and survival analysis, to provide an accurate and dynamic prediction on PD. The experimental results also illustrate the efficiency of our proposal.

We make three contribution in this paper to prior literature. First, we establish a novel method (i.e., SurvXGBoost) to predict PD overtime. SurvXGBoost is a modified Cox proportional hazard (PH) model (Cox, 1972) whereas departs from the prototype by relaxing PH assumption and allowing for non-linearity for covariates. To the best of knowledge, survival gradient boosting decision tree approaches have not been applied to credit risk assessment. Second, as shown in Table 1, four out of seven studies have considered only a small number of macroeconomic variables in modelling. Macroeconomic variables can reflect the sudden change of economy (Kartal, 2020; Sukharev, 2020) and the economic uncertainty (Liu et al., 2019), which are expected to affect the borrowers' ability and wiliness to pay (Zhang & Thomas, 2012) and therefore determine the PD. Thus, we further enhance the predictability of SurvXGBoost model by extracting information from the principle components of 1,042 monthly macroeconomic variables. Finally, the out-of-sample (OOS) validation that frequently used in existing studies may theoretically contain future information in the training set and thus over-estimates the model performance. Instead of fixed training and test set used in prior studies, we compare model performances under both OOS and out-of-time (OOT) validation to examine the external validity of empirical comparisons.

Table 1. Literature table

Authors	Data	Model	Macroeconomic variables	Validation	Evaluation measure
Tong et al. (2013)	A consumer loan dataset from a major UK bank	Mixture cure model	None	OOS	AUC, KS, H measure
Bellotti and Crook (2013)	Three large datasets from 1999 to 2006	Discrete linear survival model	9 macroeconomic variables	OOS	Log-likelihood ratio
Leow and Crook (2016)	A credit card dataset from 2002 to 2011	Time-varying Cox PH	12 macroeconomic variables	OOS	Compare model parameters
Dirick et al. (2017)	Five datasets containing personal loans and small enterprises loans	AFT, Cox PH, Cox PH with splines, Mixture cure, Mixture cure with multiple events	None	OOS	AUC, MAE, MSE, FV
Leow and Crook (2018)	A credit card dataset from 2005 to 2010	Multistate delinquency models	10 macroeconomic variables	OOS	Discrepancy measure
Wang et al. (2018)	A P2P lending dataset from 2013 to 2015	Mixture random forests	None	OOS	AUC, KS, H measure
Djeundje and Crook (2019)	A credit card dataset from 2002 to 2011	Time-varying Cox PH	Index of production, consumer confidence, FTSE index, and unemployment rate	OOT and OOS	AUC and misclassification cost

Note: AFT – Accelerated failure time model, OOS – Out-of-sample validation, OOT – Out-of-time validation.

The remaining part is structed as follows. Section 1 introduces the preliminaries. In Section 2, the proposed SurvXGBoost model. Section 3 explains experimental setup, including the data, model validation, and evaluation measures are discussed in details. Section 4 compares and analyses the experimental result. Finally, conclusions and future research are discussed in the last section.

1. Preliminaries

1.1. Standard survival analysis

Survival analysis is used to model the time of a certain event (e.g., default or survival). Since the event distribution is usually modelled as a continuous function of time, we define survival function as the probability of not having encountered the event until a specific time t, namely

$$S(t) = P(T > t) = \int_t^{\infty} f(s)ds, \tag{1}$$

where $f(s)$ is the probability density function. A close concept to survival function is hazard function

$$h(t) = \frac{f(t)}{S(t)} = \lim_{\tau \to 0} \frac{P(t \le T \le t + \tau | T > t)}{\tau}, \tag{2}$$

which measures the event rate at time t conditional on survival until t. Once the hazard function is acquired, one can retrieve the survival function through the cumulative hazard $H(t)$ by:

$$S(t) = e^{-H(t)} = e^{-\int_0^t h(s)ds}. \tag{3}$$

In real-world data, a proportion of censored samples exists, which means non-default loan applications exist by the time of data collection. Under this circumstance, early payment and fully paid loans are interpreted as censored, the true event time of which is unknown. Instead of observing the true event time T^*, we can only observe right-censored event time $T = \min\{T^*, C\}$, where C is the censored time. Moreover, the status indicator $\delta = \mathbf{1}(T = T^*)$, where $\mathbf{1}(\cdot)$ is the indicator function. Specifically, $\delta = 1$ denotes the occurrence of default event, and $\delta = 0$ indicates an early or fully paid loan. For the i-th sample in the dataset, let x_i denote the covariates and T_i imply the observed during time. The likelihood for right-censored data is represented as follows:

$$L(\theta; T_i, \delta_i, x_{ii=1}^n) = \prod_i f(T_i | x_i, \theta)^{\delta_i} S(T_i | x_i, \theta)^{1-\delta_i} = \prod_i h(T_i | x_i, \theta)^{\delta_i} \exp\left[-H\left((T_i | x_i, \theta)\right)\right], \tag{4}$$

where θ is the set of parameters. Eq. (4) can be optimized by maximum-likelihood approach over functional space of S and parameter space of θ, whereas this is usually intractable when no prior form is specified on hazard function. The recent extensions on standard survival analysis mainly focus on the specification on the survival function, hazard function, or cumulative hazard. We will subsequently introduce two types of modified survival analysis model, namely Cox PH model and random survival forests model.

1.2. Cox Proportional hazard model

The Cox PH model is widely used in survival analysis. It develops a semi-parametric specification on the hazard function described in Eq. (2):

$$h(t | x, \theta) = h_0(t) \exp(\theta^T x), \tag{5}$$

where $h_0(t)$ is a non-parametric baseline hazard, and $\exp(\theta^T x)$ denotes a parametric relative risk function. Due to the semi-parametric form, the function cannot be directly optimized using the maximum-likelihood approach. In this paper, the cumulative baseline hazard (i.e., $H(t)$ described in Eq. (3)) is estimated by Breslow estimator:

$$\hat{H}_0(t) = \sum_{T_i \le t} \Delta \hat{H}_0(T_i), \tag{6}$$

where $\Delta\hat{H}_0(T_i) = \delta_i / \sum \exp\left(\hat{\theta}^T x\right)$. The baseline hazard $h_0(t)$ can therefore be estimated by smoothing the increments (i.e., $\Delta\hat{H}_0(T_i)$). The parametric part, namely the relative risk function, is fitted by maximizing the Cox partial likelihood, which is expressed as follows:

$$L_{cox}\left(\theta; T_i, \delta_i, x_{ii=1}^n\right) = \prod_i \left(\frac{\exp\left(\theta^T x_i\right)}{\sum_{j \in R_i} \exp\left(\theta^T x_j\right)} \right)^{\delta_i}, \tag{7}$$

where R_i denotes the set of samples that not censored before time T_i. The Cox PH model has been extensively applied to the survival analysis in credit scoring. Some recent studies modified the conventional Cox PH model to handle time-varying covariate or time-varying coefficient (Bellotti & Crook, 2009; Djeundje & Crook, 2018, 2019; Leow & Crook, 2016). Specifically, the hazard function described in Eq. (5) are modified as follows:

$$h\left(t|x,\theta\right) = h_0\left(t\right) exp\left(\theta^T x\left(t\right)\right); \tag{8}$$

$$h(t \mid x,\theta) = h_0\left(t\right) exp\left([\theta(t)]^T x\right). \tag{9}$$

Although the two models are closer to reality than the prototype, it remains one major drawback, namely the parametric forms of relative risk function. The universal approximation property of ANN provides an alternative to non-parametric relative risk function (Baesens et al., 2005), but abundant evidences have shown that a single ANN is not comparable to ensemble models in credit risk assessment (Lessmann et al., 2015). To further enhance the predictability of survival models, we aim to introduce tree-based ensemble methods into survival analysis.

1.3. Random survival forests

RF is a popular non-parametric tree-based ensemble algorithm proposed by Breiman (2001). In survival setting, random survival forests (RSF) (Ishwaran et al., 2008) make predictions on time-to-event by combining the results of multiple survival trees. The general procedure of RSF is as follows:

Step 1. Take samples from training set via bootstrap method.

Step 2. Train a survival tree based on the samples in the training set. For each node of the survival tree, RSF selects the splitting variable randomly. The splitting threshold is determined by a certain criterion such as log-rank test.

Step 3. Continue to split the nodes until a stopping criterion is reached.

Step 4. Aggerate the information of all the survival trees to obtain the risk prediction of RSF.

Let *h* denote the terminal node of the survival tree. Different from the conventional forms illustrated in Subsection 1.1, the cumulative hazard and survival function for terminal node *h* are estimated using Nelson–Aalen and Kaplan–Meier estimators, respectively:

$$\tilde{H}_h(t) = \sum_{t_{j,h} \le t} \frac{\tilde{d}_{j,h}}{\tilde{Y}_{j,h}}; \tag{10}$$

$$\tilde{S}_h(t) = \prod_{t_{j,h} \le t} \left(1 - \frac{\tilde{d}_{j,h}}{\tilde{Y}_{j,h}}\right), \tag{11}$$

where $\tilde{d}_{j,h}$ and $\tilde{Y}_{j,h}$ denote the number of default samples and individual at risk at time $t_{j,h}$. Given a new sample with feature $\boldsymbol{X}$, $\boldsymbol{X}$ would be assigned into a unique terminal node h. The cumulative hazard and survival function for $\boldsymbol{X}$ can be calculated as

$$\hat{H}_h(t \mid \boldsymbol{X}) = \tilde{H}_h(t); \tag{12}$$

$$\hat{S}_h(t \mid \boldsymbol{X}) = \tilde{S}_h(t). \tag{13}$$

The ensemble cumulative hazard and survival function are determined by averaging the tree estimators in all the survival trees. Let $\hat{H}_i(t \mid \boldsymbol{X})$ and $\hat{S}_i(t \mid \boldsymbol{X})$ denote the cumulative hazard and survival function of the i-th survival tree. The ensemble estimators are computed as

$$\bar{H}(t \mid \boldsymbol{X}) = \frac{1}{N} \sum_{i=1}^{N} \hat{H}_i(t \mid \boldsymbol{X}); \tag{14}$$

$$\bar{S}(t \mid \boldsymbol{X}) = \frac{1}{N} \sum_{i=1}^{N} \hat{S}_i(t \mid \boldsymbol{X}). \tag{15}$$

2. SurvXGBoost model

GBDT is a member of boosting algorithms, which combines multiple weak learners into a storing one by an additive manner (Friedman, 2000). XGBoost (Chen & Guestrin, 2016) is an advanced GBDT-based approach that provide superior performance in credit risk assessment (Xia et al., 2017a). For a given dataset $D = \{(x_i, y_i)\}$, $i = 1,2,...,n$, GBDT-based techniques use K additive functions to make predictions on the target variable:

$$\hat{y}_i = F(x_i) = \sum_{k=1}^{K} f_k(x_i), \tag{16}$$

where $f(x) = \omega_{q(x)}$ is the space of classification and regression trees. To determine the set of $f(x)$ in Eq. (16), SurvXGBoost aims to minimize the objective function L_{obj} below:

$$L_{obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k), \tag{17}$$

where $l(\cdot,\cdot)$ is a convex loss function that measures the difference between the true value y_i and the prediction $\hat{y}_i$. Based on the likelihood defined in Eq. (4), the loss function of SurvXGBoost for right-censored survival data is defined as $l(\omega) = -\delta_i \left[F_\omega(x_i) - \log \sum_{j \in R_i} e^{F_\omega(x_j)} \right]$. $\Omega(\cdot) = \gamma T + \frac{1}{2}\lambda\omega$ herein is a regularization term. γ is a regularization hyper-parameter. T

denotes the number of splits, and λ is a L2 regularization term. Let y_i^t denote the prediction of the *i*-th sample at the *t*-th model. SurvXGBoost attempts to optimize Eq. (17) by adding a new base learner f_t:

$$L_{obj}^t = \sum_{i=1}^{n} l\left(y_i, y_i^{t-1} + f_t\left(x_i\right)\right) + \Omega\left(f_t\right). \tag{18}$$

In SurvXGBoost, the optimal new base leaner is approximated by Newton-Raphson method rather than the gradient descent method used in conventional GBDT since the use second-order gradient information usually provides a quick approximation. After removing the constant term, the Eq. (18) can be simplified to the following form at the *t*-th iteration:

$$\tilde{L}_{obj}^t \approx \sum_{i=1}^{n} [g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right)] + \Omega\left(f_t\right), \tag{19}$$

where g_i and h_i represents the first- and second-order gradient of the *i*-th sample, respectively. Once the optimal base learner is acquired, it is added to the prior functions following Eq. (16) to finish an iteration. Moreover, XGBoost also makes some engineering optimization to build a scalable GBDT approach. For example, a histogram-based approximation algorithm is developed to quickly optimize Eq. (19). Moreover, XGBoost supports distributed learning and GPU computing which accommodates the application of big data.

3. Experimental setup

3.1. Data

To clarify the superiority of SurvXGBoost, an experiment is conducted based on a real-world consumer loan dataset. The dataset is derived from a consumer loan transactions of a major P2P lending platform in the U.S. This dataset consists of 226,148 loan applications which were issued between January 2009 and December 2013.

The dataset comprises a variety of application variables, which can be roughly categorized into three types, namely loan characteristics, borrower's creditworthiness, and borrower's solvency. In this dataset, we define the loan status "reject" as 120 days or more past due. The status "fully paid" means that the loan is paid before or at the duration time. The summary statistic of the time-to-event, loan status, and the features are displayed in Table 2, which shows that the class distribution is imbalanced in this dataset.

Moreover, macroeconomic variables are added to reflect the dynamics of business circle on payment of retailing loans. Instead of the limited number of macroeconomic variables used in prior studies (Bellotti & Crook, 2009; Dirick et al., 2019; Djeundje & Crook, 2019), we applied 1,042 macroeconomic variables to provide a comprehensive description on economic conditions. All these macroeconomic variables are monthly series. Principal component analysis is applied to convert the original set of macroeconomic variables into high-performance variables since multicollinearity occurs between macroeconomic variables. The principle components (PCs) of macroeconomic variables are subsequently employed as features of survival analysis. Table 3 displays the top five influential macroeconomic variables for each PC. According to the influential macroeconomic variables, the PCs are named as labour force indicator, employment indicator, price indicator, recession indicator, technological diffusion and advancement indicator, and inventory indictor, respectively.

Table 2. Summary statistics of the dataset

	Type	Min	Max	Mean	S.D.
Loan status	Categorial	–	–	0.15	–
Loan characteristics					
Funded amount	Numeric	1000	35000	13800.94	8066.58
Term	Categorial	–	–	41.83	–
Interest rate	Numeric	0.05	0.26	0.14	0.04
Borrower's creditworthiness					
Credit grade	Categorial	–	–	11.61	–
No. of delinquency	Numeric	0	29	0.22	0.67
Total number of credit lines	Numeric	2	105	24.15	11.16
Borrower's solvency					
Employment length	Numeric	0	11	5.98	4.07
Home ownership		0	3	1.67	0.62
Annual income	Numeric	4000	7141778	71709.82	54011.22
Income verification status		0	2	1.11	0.87
DTI ratio	Numeric	0	34.99	16.44	7.57
Revolving utilization rate	Numeric	0	1.40	0.57	0.24
Macroeconomic variables					
PC1	Numeric	–9.08	1.50	–0.80	2.48
PC2	Numeric	–5.02	1.21	–4.07	0.63
PC3	Numeric	–7.40	3.17	1.99	1.40
PC4	Numeric	–2.06	1.37	0.47	0.90
PC5	Numeric	–2.71	3.10	0.38	0.74
PC6	Numeric	–3.25	2.02	–0.04	1.23

3.2. Model validation

The initial task of model validation lies on the splitting of training and test set. In prior studies, a fixed training and test set were determined based on a certain time threshold and the survival models were built only once using the training set and made predictions on test set (Bellotti & Crook, 2009; Dirick et al., 2019; Djeundje & Crook, 2018, 2019). However, such a validation approach arouses concerns in that it may provide unreliable experiment results especially for small dataset (Lessmann et al., 2015). As a result, we use two types of validation approaches in this paper: regarding the OOS validation, a five-fold cross-validation (CV) is performed for 50 times and the average performance is used to evaluate models. Concerning the OOT validation, we apply a sliding-window method, the basic idea of which is described in Figure 1. The "window" is specified as one year and the test set comprises samples in the corresponding year. The training set includes samples before the window. Once the models are built and evaluated, the window slides into the next year. Sliding-window method does not stop until the samples in the final year of issuing date have been employed as test set.

Table 3. Top five macroeconomic variables for each principle components

Ranking	PC1: labor force indicator	PC2: employment indicator	PC3: price indicator
1	Civilian Labor Force Participation Rate: High School Graduates, No College, 25 years and over	Employment Level: 25 to 54 years	Producer Price Index by Commodity for Chemicals and Allied Products: Chlorine, Sodium Hydroxide, and Other Alkalies
2	Labor Force Participation Rate: White	All Employees, Construction	Sweden / U.S. Foreign Exchange Rate
3	Small Time Deposits - Total (Seasonally Adjusted)	Nonfarm Private Construction Payroll Employment (Not Seasonally Adjusted)	Manufacturers: Inventories to Sales Ratio (Seasonally Adjusted)
4	Small Time Deposits - Total (Not Seasonally Adjusted)	Total Construction Spending: Commercial (Not Seasonally Adjusted)	Real Trade Weighted U.S. Dollar Index: Other Important Trading Partners, Goods
5	Civilian Labor Force Participation Rate: 20 years and over, White Men	Employment Rate: Aged 25-54: All Persons for the United States (Percent, Monthly, Seasonally Adjusted)	Total Business: Inventories to Sales Ratio
Ranking	PC4: recession indicator	PC5: technological diffusion and advancement indicator	PC6: inventory indictor
1	Domestic Auto Inventories	Chicago Fed National Activity Index: Diffusion Index	Retailers: Inventories to Sales Ratio
2	OECD based Recession Indicators for the United States from the Peak through the Trough	San Francisco Tech Pulse (Percent Change at Annual Rate, Monthly, Seasonally Adjusted)	Total Business Inventories (Seasonally Adjusted)
3	Other Separations: Total Nonfarm (Rate, Monthly, Seasonally Adjusted)	Business Tendency Surveys for Manufacturing: Confidence Indicators: Composite Indicators: OECD Indicator for the United States	Manufacturers' Inventories (Not Seasonally Adjusted)
4	OECD based Recession Indicators for the United States from the Period following the Peak through the Trough	San Francisco Tech Pulse (Percent Change from Year Ago, Monthly, Seasonally Adjusted)	Other Separations: Total Nonfarm (Rate, Monthly, Not Seasonally Adjusted)
5	All Employees: Mining and Logging: Oil and Gas Extraction	San Francisco Tech Pulse (Percent Change, Monthly, Seasonally Adjusted)	Consumer Price Index: OECD Groups: All Items Non-Food and Non-Energy for the United States

The proposed SurvXGBoost model is compared with three non-parametric survival models, namely the Cox PH model, RSF, and survival GBDT model (Chen et al., 2013). The Cox PH model is a state-of-the-art benchmark that has been extensively considered in existing literature (Bellotti & Crook, 2009; Wang et al., 2018). We have also employed a time-varying Cox PH model following Bellotti and Crook (2013) to capture the time-varying effects of

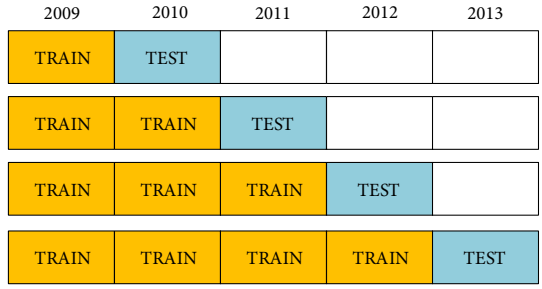

Figure 1. An illustration of out-of-time validation

some features. The number of trees is determined as 200 for RSF. To optimize the hyper-parameters in GBDT and XGBoost, a Bayesian hyper-parameter tuning approach is performed following Xia et al. (2017a). Bayesian hyper-parameter optimization is a type of sequential model-based optimization, which means it pre-sets the number of iterations for optimization. During each iteration, Bayesian hyper-parameter optimization builds a surrogate probability model of a fitness function and determines the optimal hyper-parameters for the surrogate. The hyper-parameters are afterwards used to the real fitness function and returns the corresponding results. The hyper-parameters and the corresponding results are then used to update the surrogate model. The iteration continues unless the number of iterations is reached. Specifically, the five-fold CV Concordance Index (C-index) of training set is employed as the fitness function of survival GBDT and SurvXGBoost. The surrogate probability model is established based on the Tree Parzen Estimator following Bergstra et al. (2011). The hyper-parameters considered in this paper is summarized in Table 4.

Table 4. The definitions and searching spaces of hyper-parameters for survival GBDT and SurvXGBoost

Model	Hyper-parameter	Function	Searching space
Survival GBDT	Number of iterations	The number of iterations of boosting process	[50, 500]
	Maximum depth	Maximum depth of a single CART	[3, 6]
	Subsample rate	The fraction of samples used for training a single CART	[0.6, 1]
	Learning rate	It shrinks the contribution of each CART	0.01
SurvXGBoost	Number of iterations	The number of iterations of XGBoost	[50, 500]
	Maximum depth	Maximum depth of a single base learner	[3, 6]
	Subsampling rate	The fraction of samples used for training a single base learner	[0.6, 1]
	Learning rate	It shrinks the contribution of each base learner	0.01
	Column sampling rate	The fraction of features used for training a single base learner	[0.6, 1]
	Gamma	Minimum loss reduction required to make a further partition	[0, 5]

3.3. Evaluation measures

We employ four popular evaluation measures to examine model performance in terms of discriminative ability and label prediction. The discriminative ability implies the model's capability of distinguishing between default and non-default borrowers. Following the instruction of CECEL and IFRS 9, the 12-month PD of survival models is evaluated. The evaluation metrics consist of:

(1) *C-index*, which is a popular performance metric that quantifies the quality of rankings for survival models. It is computed as the fraction of concordant pairs divided by the number of possible evaluation pairs. The range of C-index is $\left[0.5,1\right]$, where 0.5 indicates a random guess and 1 indicates a perfect model.

(2) *Area under the ROC curve* (*AUC*), which evaluates the quality of model's prediction irrespective what decision threshold is determined. AUC has been regarded as a frequently used metric in evaluating credit scoring models in prior literature (Bequé & Lessmann, 2017; Xia et al., 2017a). AUC measures the entire two-dimensional area under the ROC curve. Following Huang and Ling (2005), AUC is calculated as follows for a binary classification:

$$\text{AUC} = \frac{S_0 - n_0\left(n_0+1\right)/2}{n_0 n_1}, \tag{20}$$

where n_0 and n_1 denote the number of non-default and default loans in test set, respectively. $S_0 = \sum rank_j$ is the rank of probability predications of j-th default loans.

(3) *H measure*, which is proposed by Hand (2009) to overcome the inconsistent misclassification costs that potentially assumed in AUC. Specifically, AUC potentially assumes that the misclassification cost is dependent on the classifiers rather than the datasets, which is far from reality. Thus, Hand (2009) advocated to employ a beta distribution to fit a cost weight function. In a further analysis of Hand and Anagnostopoulos (2014), the optimal parameter of beta distribution is discussed, and in this paper, we follow the suggestion of Hand and Anagnostopoulos (2014) to use a beta (2, 2) distribution. Recent credit scoring studies have introduced H measure as an efficient alternative to AUC when evaluating models (Ala'raj & Abbod, 2016; He et al., 2018).

(4) *Misclassification cost*. Since cost-sensitivity usually occurs in credit scoring, accuracy can seldom provide an overall evaluation on the label prediction (Shen et al., 2020). As a result, we follow Lohmann and Ohliger (2019) to evaluate the capability of label prediction by misclassification cost defined as follows:

$$\text{Misclassification cost} = \frac{\sum_{i=1}^{n}\left\{\alpha\left[status_i = 1 \,|\, \widehat{status}_i = 0\right] + \left[status_i = 0 \,|\, \widehat{status}_i = 1\right]\right\}}{n}, \tag{21}$$

where $\left[\cdot\right]$ is the Iverson bracket. α herein is the cost parameter and we determine it as 5 in this paper. The misclassification cost further raises an issue on the decision threshold, which can dramatically affect the label prediction. The samples with higher PDs than decision threshold will be rejected and the remaining loan applications will

be granted. In a highly imbalanced dataset, credit scoring model tends to predict all the applications as the majority class (usually non-default) and thus lacks the capability to discriminate risky ones (Sahin et al., 2013). Cost-sensitive learning is a solution to imbalanced dataset, which can be roughly divided into the direct method and indirect ones (Shen et al., 2020; Xia et al., 2017b). The direct cost-sensitive learning methods design models that are cost-sensitive in themselves, whereas the indirect methods transform the cost-insensitive models into cost-sensitive one by sampling or thresholding. The sampling technique means balancing the class distribution in training set, and the thresholding indicates adjusting the decision threshold. In this paper, we employ the thresholding technique due to its easy-to-implementation and popularity. Specifically, we determine the decision threshold as the fraction of good and risky applications in training set as advocated by Bequé and Lessmann (2017) and Xia et al. (2020b).

4. Experimental results

4.1. Out-of-sample validation

The results in Table 5 show the average performance of SurvXGBoost and benchmarks around the evaluation measures. The standard deviations of performance are described in brackets and the best-performing model for each evaluation metric is highlighted in bold. Performances that are significantly inferior to the best model at a 95% confidence level with respect to a paired t-test are denoted in underlines. Table 5 exhibits several important findings.

First, the superiority of machine learning algorithms is explicitly demonstrated. The proposed SurvXGBoost performs significantly better than the benchmark models for both discriminative capability and misclassification cost. Moreover, survival GBDT and RSF are also marginally better than Cox PH models. These results are similar with those in Chen et al. (2013) and further imply that the predictive ability of machine learning algorithms is superior to statistical ones in most cases. This finding advocates the extension of machine learning approaches to real-world credit risk evaluation.

Table 5. Performance of SurvXGBoost and the benchmark models

Model	C-index	AUC	H measure	Misclassification cost
Cox PH	0.6521 (0.0025)	0.6728 (0.0028)	0.0949 (0.0030)	0.6580 (0.0055)
Cox PH (time-varying)	0.6524 (0.0029)	0.6680 (0.0031)	0.0892 (0.0032)	0.6628 (0.0063)
RSF	0.6513 (0.0032)	0.6772 (0.0031)	0.0998 (0.0035)	0.6510 (0.0072)
Survival GBDT	0.6566 (0.0027)	0.6793 (0.0030)	0.1017 (0.0034)	0.6496 (0.0058)
SurvXGBoost	**0.6573** (0.0026)	**0.6808** (0.0029)	**0.1035** (0.0033)	**0.6484** (0.0055)

Note: the best-performing model for each evaluation measure is highlighted in bold. The value in brackets is the standard deviation. The underlines imply a significant difference between the corresponding model and the best-performing one.

Second, when comparing among the parametric survival models, time-varying Cox PH model provides better results in terms of C-index and misclassification cost than the original model. This implies that the time-varying Cox PH model quantifies the non-linear relationship between covariates and default time to some extent. However, the inferior performance of time-varying Cox PH model relative to the non-parametric survival models indicates that the parametric form proposed by Bellotti and Crook (2013) can only capture parts of the non-linear effects.

Finally, when comparing among non-parametric models, Table 5 reveals that a combination of GBDT-based model improves model performance relative to RSF. This result is similar with those reported in Xia et al. (2017a) and Xia et al. (2020) in case of classification credit scoring models. Future research can explore applying other advanced GBDT-based methods into survival credit scoring models.

4.2. Out-of-time validation

The OOT validation starts from the year of 2010, implying that the samples issued before 2010 are employed as training set. The loan transactions issued during the year of 2010 are employed as test set. The window slides into the next quarter until the last quarter of the dataset is reached. Figures 2 to 5 display the heatmaps of C-index, AUC, H measure, and misclassification cost for the models, respectively. In these figures, the columns represent the models and the rows display the year.

When making a horizontal comparison, the machine learning variants of survival analysis provide promising results. The SurvXGBoost provides the best performance on all evaluation metrics except H measure. This is in line with those revealed in the previous subsection and again demonstrates the superiority of the proposed SurvXGBoost. RSF and survival GBDT achieve the best performance when the data is limited. Concretely, when samples in 2011 is utilized as test set, the two models achieve the best AUC, H measure and misclassification cost.

When making a vertical comparison, Figures 2 to 5 reveal that model performance varies in different years. Although survival models provide unsatisfying performance when training data is limited, their performance does not necessarily improve when number of training sample grows. A possible explanation on this phenomenon lies on that the characteristics of loan applications show very different patterns over the year of 2010 to 2013. Further investigation is required for this argument.

4.3. The effects of macroeconomic variables

The main goal of this subsection is to examine whether model performance is improved after adding macroeconomic variables. Moreover, we aim to see whether the determinants of default and time-to-event differ. Thus, a Cox PH model (abbreviated as Model 1) and logistic regression model (abbreviated as Model 2) are established with all the covariates listed in Table 2.

The fitted coefficients for the two models are reported in Table 6. A few important findings can be revealed from Table 6. Concretely, we can observe that loan characteristics, borrowers' creditworthiness and solvency are powerful determinants of time-to-default.

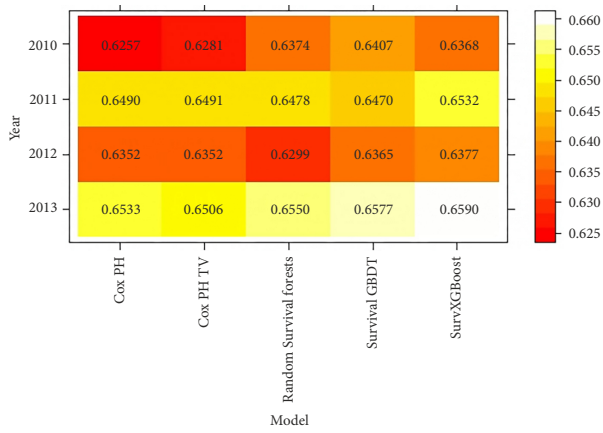

Figure 2. Out-of-time validation results of C-index

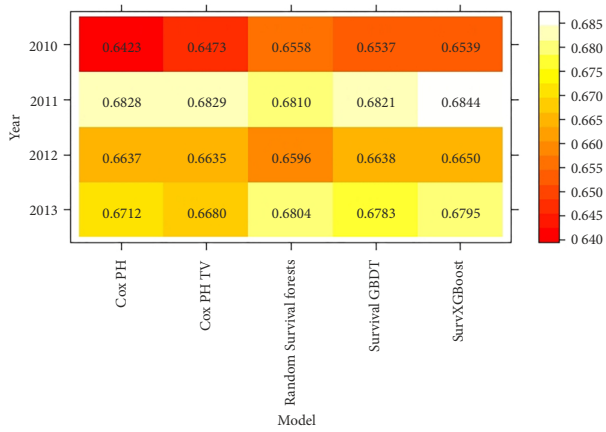

Figure 3. Out-of-time validation results of AUC

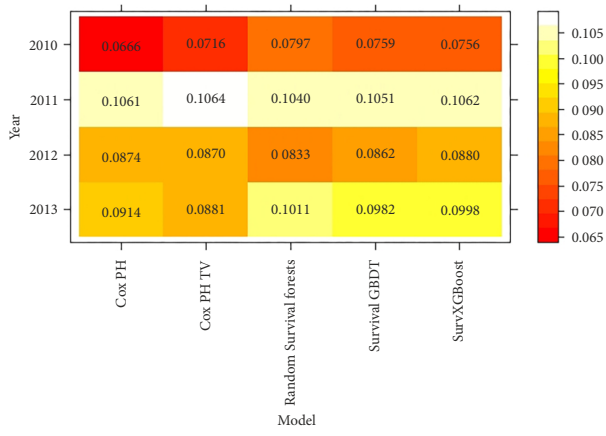

Figure 4. Out-of-time validation results of H measure

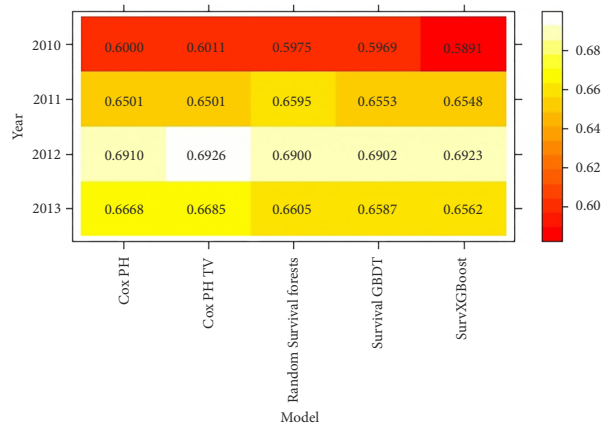

Figure 5. Out-of-time validation results of misclassification cost

Table 6. Parameter estimates for Cox PH and logistic regression model

	Model 1: Cox PH model		Model 2: logistic regression model	
	Est.	*p*-val	Est.	*p*-val
Loan characteristics				
Funded amount	0.19590	0.00000	0.27281	0.00000
Term	–0.03442	0.02439	0.39629	0.00000
Interest rate	3.10000	0.00000	2.98617	0.00000
Borrower's creditworthiness				
Credit grade	–0.85400	0.00000	–0.72478	0.00001
No. of delinquency	–0.98050	0.00003	–0.21077	0.41608
Total number of credit lines	0.27410	0.00000	–0.20787	0.00130
Borrower's solvency				
Employment length	–0.07123	0.00000	–0.05475	0.00109
Home ownership	–0.13580	0.00000	–0.16920	0.00000
Annual income	–33.64000	0.00000	–38.82243	0.00000
Income verification status	–0.00371	0.80150	–0.00399	0.80628
DTI ratio	0.19730	0.00000	0.45685	0.00000
Revolving utilization rate	–0.28480	0.00000	–0.04250	0.29225
Macroeconomic variables				
PC1	–0.00224	0.88052	–0.02947	0.07560
PC2	–0.08298	0.00498	–0.09046	0.00609
PC3	–0.04303	0.06983	–0.03210	0.22528
PC4	–0.07547	0.00100	–0.05918	0.02112
PC5	0.05761	0.00004	0.07053	0.00000
PC6	0.00371	0.47953	0.00744	0.20597
Intercept	–	–	–3.02939	0.00000

This finding is in parallel with those revealed in Wang et al. (2018) and Dirick et al. (2019). Moreover, the fitted coefficients for Cox PH model and logistic regression exhibit very different patterns. Concretely, the variables of No. of delinquency, revolving utilization rate, and price indicator can hardly affect loan status significantly whereas they are significant determinants of time-to-default. This encourages a deep investigation on the determinants of time-to-default in future research. Finally, macroeconomic variables, especially indicators concerning employment, price, recession, and technological diffusion and advancement, are powerful determinants in survival analysis. This further encourages us to examine the predictability of survival models when incorporating macroeconomic variables.

The comparisons of the evaluation measures are presented in Table 7. The models are benchmarked against models that employ the same loan characteristics, borrowers' creditworthiness and solvency variables without macroeconomic variables. Both the results of OOS and OOT validation are reported to give a comprehensive description of the models. Concerning the results of OOS validation, a comparison between Tables 5 and 7 shows that the performance of benchmark models is improved after adding macroeconomic variables to survival models in most cases, therefore partially demonstrating the effectiveness of macroeconomic variables in credit risk assessment. According to the famous five Cs of credit, character, capacity, capital, collateral, and conditions are key factors to predict borrower's PD. The former four characteristics have been attached much attention whereas the macroeconomic condition is not frequently considered in credit risk modelling. Future research should be performed on this topic.

However, the OOS validation may over-estimate the effects of macroeconomic variables since it includes future information of macroeconomic variables in training set. Thus, we also report the OOT results in Table 7. This table reveals that macroeconomic variables can enhance model performance for OOS validation in most cases. The comparison between OOS and OOT validation indicates that in-time modelling can capture a larger share of the variation in the training set and lead to higher C-index, H measure, and AUC than those in OOT validation. Considering the fact that OOT validation is closer to real-world modelling process whereas it gathers limited attention in concerning studies, this finding highlights the necessity of OOT validation in model comparisons.

Heterogeneity is also witnessed for model performance under OOT validation. Concretely, for benchmark Cox PH model, the incorporation of macroeconomic variables does not better off the model performance. These results are in line with expectations since the macroeconomic variables may have non-linear effects on time-to-event. For example, Figure 6 shows the estimates of time-dependent coefficients for the six PCs of macroeconomic variables of Cox PH model, where the solid vertical lines indicate the coefficients of zero and the dashed lines represent the fixed coefficients of Cox PH model. This figure illustrates a rapid change in the sign and the magnitude of the coefficients for macroeconomic variables. The Cox PH model can hardly capture the non-linear effects of macroeconomic variables and thus lead to inferior performance than non-parametric survival models.

From Table 7 we can also observe that the ranks of models hold the same when macroeconomic variables are not included. SurvXGBoost becomes the best-performing model under OOS and OOT validations, which confirms the robustness of the proposed method.

Table 7. Results of models under out-of-sample and out-of-time validation when macroeconomic variables excluded

Model	C-index	AUC	H measure	Misclassification cost
Out-of-sample validation (*macroeconomic variables excluded*)				
Cox PH	0.6518 (0.0032)	0.6728 (0.0030)	0.0946 (0.0032)	0.6583 (0.0055)
Cox PH (time-varying)	0.6522 (0.0030)	0.6679 (0.0032)	0.0891 (0.0032)	0.6631 (0.0066)
RSF	0.6510 (0.0032)	0.6770 (0.0031)	0.0998 (0.0034)	0.6501 (0.0070)
Survival GBDT	0.6566 (0.0031)	0.6792 (0.0029)	0.1016 (0.0033)	0.6500 (0.0060)
SurvXGBoost	**0.6572** (0.0032)	**0.6807** (0.0030)	**0.1034** (0.0034)	**0.6485** (0.0057)
Out-of-time validation (*macroeconomic variables excluded*)				
Cox PH	0.6437	0.6699	0.0925	0.6476
Cox PH (time-varying)	0.6416	0.6667	0.0892	0.6512
RSF	0.6415	0.6682	0.0907	0.6489
Survival GBDT	0.6455	0.6700	0.0912	0.6510
SurvXGBoost	0.6462	0.6707	0.0922	0.6484

Note: the best-performing model for each evaluation measure is highlighted in bold. The value in brackets is the standard deviation. The underlines imply a significant difference between the corresponding model and the best-performing one. For out-of-time validation, only the average values are reported.

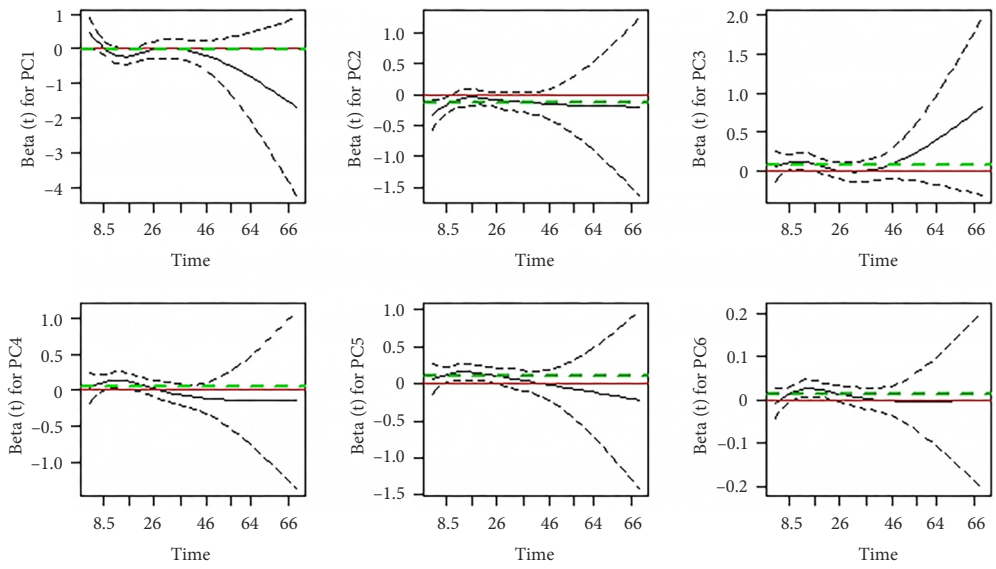

Figure 6. Time-dependent coefficients for PCs of macroeconomic variables of Cox PH model

4.4. Interpretability

Lack of interpretability may hinder the mangers' wiliness to employ complex credit models. Moreover, transparent models are required by regulators in many regions or countries. Since CART is employed as the base learner of SurvXGBoost, one can plot all the base models in a graph so that the process of decision making is clear to the users. However, it is a tough work to take hundreds of base models into consideration. Thus, the proposed SurvXGBoost maintain some interpretability whereas it is not so interpretable relative to parametric ones.

Nevertheless, we can explain the proposed SurvXGBoost model by figuring out the important features. XGBoost provides several feature importance measures to describe how important a feature is during modelling. In this paper, we select *relative gain* measure since it directly evaluates the relative contribution of a certain feature selected as the splitting variable to the model. A higher relative gain indicates a more important feature when generating base models. Figure 7 shows the feature importance measured by relative gain for the 50×5-fold CV. The error bars in Figure 7 indicate the confidence intervals. As shown in Figure 7, interest rate, annual income, and credit grades account for the top three important features in the modelling of SurvXGBoost. On the contrary, some features concerning loan characteristic, borrowers' creditworthiness and solvency can hardly be used for splitting nodes. The macroeconomic variables, especially the labour force and employment indicators also play important roles in building SurvXGBoost models.

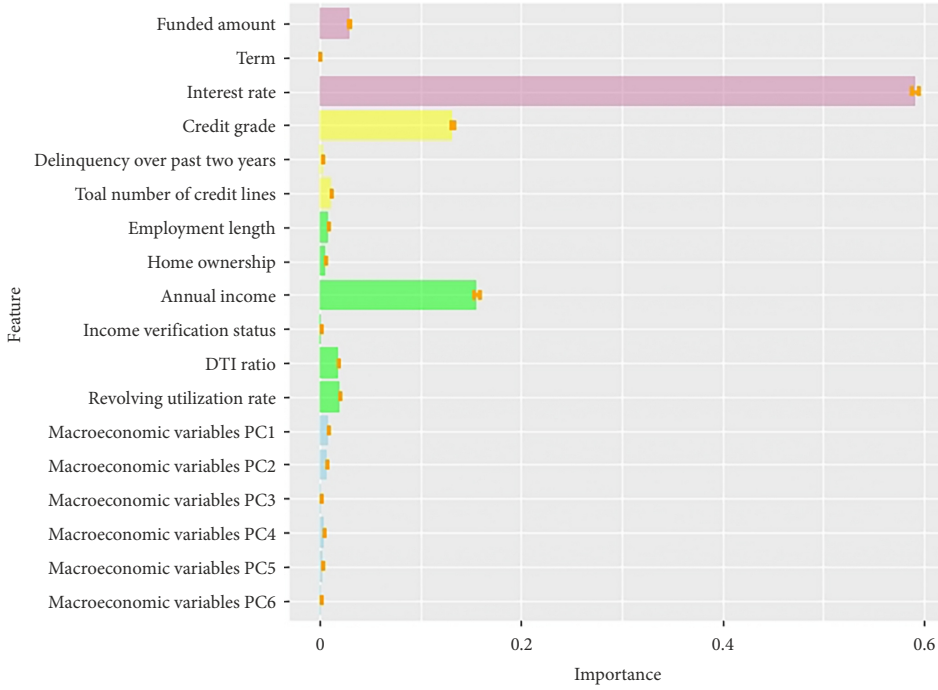

Figure 7. Feature importance of SurvXGBoost model

Conclusions and future research

Credit risk is a major type of risk that financial institution encounters. To quantify the credit risk of a portfolio, financial institutions have developed several risk parameters, among which PD is a major concern. Credit scoring is a common method to predict PD. In the modelling stage of credit scoring, ensemble models which combines the predictions of multiple models have shown their superiority in predictability. Moreover, survival analysis which can provide dynamic predictions on PD over time has been considered as an alternative to common classification algorithms. Thus, we develop a novel SurvXGBoost model which integrates a state-of-the-art ensemble model (i.e., XGBoost) and survival analysis. Our proposal is compared with several benchmark survival models on a large real-world consumer loan dataset. To further enhance model performance of survival models, information extracted from the principle components of 1,042 macroeconomic variables are integrated with the original features which include loan characteristics, borrower's creditworthiness and solvency. The model performance on predictability and misclassification cost are compared under OOS and OOT validation. For OOS validation, the proposed SurvXGBoost model outperforms the benchmarks significantly in all the evaluation measures. Concerning OOT validation, SurvXGBoost is marginally better than the benchmarks in most cases. The information extracted from macroeconomic variables can improve the predictability of survival models, which confirms the relationship between business circle and time-to-event. The model performance under OOT validation is worse than those in OOS validation, thereby confirming the fact that OOS validation may include future information of macroeconomic variables in modelling and therefore lead to over-estimated performance. It is thus recommended that OOT validation should be emphasized in the comparison of credit models. We also found that SurvXGBoost can maintain some interpretability by disclosing the feature importance.

Regarding the directions of future research, the proposed SurvXGBoost can only handle right-censored data at present. One may further extend SurvXGBoost to support other type of censored data. Moreover, the interpretability of non-parametric survival model requires further exploration. Maybe the SurvLIME algorithm can be applied into the explanation of complex survival models in credit risk assessment. The integration of other efficient GBDT-based algorithms into survival analysis is also an interesting research direction.

Acknowledgements

We are grateful for the Research Support Project for Doctoral Degree Teachers of Jiangsu Normal University (18XWRX021), the Project of Philosophy and Social Science Research in Colleges and Universities in Jiangsu Province (2020SJA1018), the National Natural Science Foundation of China (71874185), and National Social Science Foundation of China (15BTJ033).

Author contributions

Yufei Xia and Lingyun He conceived the study and were responsible for the design and development of the data analysis. Yating Fu and Yixin Xu were responsible for data collection

and analysis. Yating Fu and Yinguo Li were responsible for data interpretation. Yufei Xia and Yinguo Li wrote the first draft of this article. Yufei Xia revised the draft.

Disclosure statement

The authors declare no competing financial, professional, or personal interests from other parties.

References

Ala'raj, M., & Abbod, M. F. (2016). Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*, *104*, 89–105. https://doi.org/10.1016/j.knosys.2016.04.013

Apostolik, R., Donohue, C., & Went, P. (2009). *Foundations of banking risk: an overview of banking, banking risks, and risk-based banking regulation* (Vol. 507). John Wiley & Sons Incorporated.

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

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*(6), 627–635. https://doi.org/10.1057/palgrave.jors.2601545

Bellotti, T., & Crook, J. (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

Bellotti, T., & Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting*, *29*(4), 563–574. https://doi.org/10.1016/j.ijforecast.2013.04.003

Bequé, A., Coussement, K., Gayler, R., & Lessmann, S. (2017). Approaches for credit scorecard calibration: an empirical analysis. *Knowledge-Based Systems*, *134*, 213–227. https://doi.org/10.1016/j.knosys.2017.07.034

Bequé, A., & Lessmann, S. (2017). Extreme learning machines for credit scoring: an empirical evaluation. *Expert Systems with Applications*, *86*, 42–53. https://doi.org/10.1016/j.eswa.2017.05.050

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). *Algorithms for hyper-parameter optimization* [Conference presentation]. 25th Annual Conference on Neural Information Processing Systems. Granada, Spain.

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

Chen, T., & Guestrin, C. (2016). *Xgboost:* A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chen, Y., Jia, Z., Mercola, D., & Xie, X. (2013). A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Computational and Mathematical Methods in Medicine*, *2013*, Article 873595. https://doi.org/10.1155/2013/873595

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B* (*Methodological*), *34*(2), 187–202. https://doi.org/10.1007/978-1-4612-4380-9_37

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, *183*(3), 1447–1465. https://doi.org/10.1016/j.ejor.2006.09.100

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

Djeundje, V. B., & Crook, J. (2018). Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards. *European Journal of Operational Research*, *271*(2), 697–709. https://doi.org/10.1016/j.ejor.2018.05.040

Djeundje, V. B., & Crook, J. (2019). Dynamic survival models with varying coefficients for credit risks. *European Journal of Operational Research*, *275*(1), 319–333. https://doi.org/10.1016/j.ejor.2018.11.029

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, *210*(2), 368–378. https://doi.org/10.1016/j.ejor.2010.09.029

Friedman, J. H. (2000). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, *29*, 1189–1232. https://doi.org/10.1214/aos/1013203451

Han, L., & Ge, R. (2017). Wavelets analysis on structural model for default prediction. *Computational Economics*, *50*(1), 111–140. https://doi.org/10.1007/s10614-016-9584-1

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, *77*(1), 103–123. https://doi.org/10.1007/s10994-009-5119-5

Hand, D. J., & Anagnostopoulos, C. (2014). A better Beta for the H measure of classification performance. *Pattern Recognition Letters*, *40*, 41–46. https://doi.org/10.1016/j.patrec.2013.12.011

He, H., Zhang, W., & Zhang, S. (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems with Applications*, *98*, 105–117. https://doi.org/10.1016/j.eswa.2018.01.012

Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, *33*(4), 847–856. https://doi.org/10.1016/j.eswa.2006.07.007

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299–310. https://doi.org/10.1109/TKDE.2005.50

Huang, Z., Jiang, T., & Wang, Z. (2020). On a multiple credit rating migration model with stochastic interest rate. *Mathematical Methods in the Applied Sciences*, *43*(12), 7106–7134. https://doi.org/10.1002/mma.6435

Hung, N. T. (2019). Equity market integration of China and Southeast Asian countries: further evidence from MGARCH-ADCC and wavelet coherence analysis. *Quantitative Finance and Economics*, *3*(2), 201–220. https://doi.org/10.3934/QFE.2019.2.201

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3), 841–860. https://doi.org/10.1214/08-AOAS169

Kartal, M. T. (2020). The behavior of Sovereign Credit Default Swaps (CDS) spread: evidence from Turkey with the effect of Covid-19 pandemic. *Quantitative Finance and Economics*, *4*(3), 489–502. https://doi.org/10.3934/QFE.2020022

Klein, J. P., & Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Leow, M., & Crook, J. (2016). The stability of survival model parameter estimates for predicting the probability of default: Empirical evidence over the credit crisis. *European Journal of Operational Research*, *249*(2), 457–464. https://doi.org/10.1016/j.ejor.2014.09.005

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

Liang, J., Zhao, Y., & Zhang, X. (2016). Utility indifference valuation of corporate bond with credit rating migration by structure approach. *Economic Modelling*, *54*, 339–346. https://doi.org/10.1016/j.econmod.2015.12.002

Lim, M. K., & Sohn, S. Y. (2007). Cluster-based dynamic scoring model. *Expert Systems with Applications*, *32*(2), 427–431. https://doi.org/10.1016/j.eswa.2005.12.006

Liu, Y., Zheng, Y., & Drakeford, B. (2019). Reconstruction and dynamic dependence analysis of global economic policy uncertainty. *Quantitative Finance and Economics*, *3*(3), 550–561. https://doi.org/10.3934/QFE.2019.3.550

Lohmann, C., & Ohliger, T. (2019). The total cost of misclassification in credit scoring: A comparison of generalized linear models and generalized additive models. *Journal of Forecasting*, *38*(5), 375-389. https://doi.org/10.1002/for.2545

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, *31*, 24–39. https://doi.org/10.1016/j.elerap.2018.08.002

Maldonado, S., Bravo, C., López, J., & Pérez, J. (2017). Integrated framework for profit-based feature selection and SVM classification in credit scoring. *Decision Support Systems*, *104*, 113–121. https://doi.org/10.1016/j.dss.2017.10.007

Malik, M., & Thomas, L. C. (2010). Modelling credit risk of portfolio of consumer loans. *Journal of the Operational Research Society*, *61*(3), 411–420. https://doi.org/10.1057/jors.2009.123

Munkhdalai, L., Wang, L., Park, H. W., & Ryu, K. H. (2019). Advanced neural network approach, its explanation with LIME for Credit scoring application. In N. Nguyen, F. Gaol, T. P. Hong, & B. Trawiński (Eds.), *Lecture notes in computer science: Vol. 11432. Intelligent information and database systems* (pp. 407–419). Springer. https://doi.org/10.1007/978-3-030-14802-7_35

Ong, C.-S., Huang, J.-J., & Tzeng, G.-H. (2005). Building credit scoring models using genetic programming. *Expert Systems with Applications*, *29*(1), 41–47. https://doi.org/10.1016/j.eswa.2005.01.003

Sahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, *40*(15), 5916–5923. https://doi.org/10.1016/j.eswa.2013.05.021

Shen, F., Wang, R., & Shen, Y. (2020). A cost-sensitive logistic regression credit scoring model based on multi-objective optimization approach. *Technological and Economic Development of Economy*, *26*(2), 405–429. https://doi.org/10.3846/tede.2019.11337

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

Sukharev, O. S. (2020). Economic crisis as a consequence COVID-19 virus attack: risk and damage assessment. *Quantitative Finance and Economics*, *4*(2), 274–293. https://doi.org/10.3934/QFE.2020013

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

Wang, G., Ma, J., Huang, L., & Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems*, *26*(2), 61–68. https://doi.org/10.1016/j.knosys.2011.06.020

Wang, Z., Jiang, C., Ding, Y., Lv, X., & Liu, Y. (2018). A novel behavioral scoring model for estimating probability of default over time in Peer-to-Peer lending. *Electronic Commerce Research and Applications*, *27*, 74–82. https://doi.org/10.1016/j.elerap.2017.12.006

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, *27*(11), 1131–1152. https://doi.org/10.1016/s0305-0548(99)00149-5

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893

Xia, Y., He, L., Li, Y., Liu, N., & Ding, Y. (2020a). Predicting loan default in peer-to-peer lending using narrative data. *Journal of Forecasting*, *39*(2), 260–280. https://doi.org/10.1002/for.2625

Xia, Y., Liu, C., Da, B., & Xie, F. (2018a). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications*, *93*, 182–199. https://doi.org/10.1016/j.eswa.2017.10.022

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017a). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, 225–241. https://doi.org/10.1016/j.eswa.2017.02.017

Xia, Y., Liu, C., & Liu, N. (2017b). Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electronic Commerce Research and Applications*, *24*, 30–49. https://doi.org/10.1016/j.elerap.2017.06.004

Xia, Y., Yang, X., & Zhang, Y. (2018b). A rejection inference technique based on contrastive pessimistic likelihood estimation for P2P lending. *Electronic Commerce Research and Applications*, *30*, 111–124. https://doi.org/10.1016/j.elerap.2018.05.011

Xia, Y., Zhao, J., He, L., Li, Y., & Niu, M. (2020b). A novel tree-based dynamic heterogeneous ensemble method for credit scoring. *Expert Systems with Applications*, *159*, Article 113615. https://doi.org/10.1016/j.eswa.2020.113615

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