XGBoost-based Survival Analysis in Business Risk Prediction

Ying-ying Li
School of Mathematics and Computing Science, Guilin University of Electronic Technology
Guilin, China
xxxxiliying@gmail.com

Zeng-min Xu*
School of Mathematics and Computing Science, Guilin University of Electronic Technology
Guilin, China
zengminxu@gmail.com

Cong Feng
Shenzhen New Generation Information Technology Institute Co., Ltd.,
Shenzhen, China
fengcong@eee.hku.hk

You Jiang
Shenzhen New Generation Information Technology Institute Co., Ltd.,
Shenzhen, China
xyy.2012@foxmail.com

***Abstract*—Business risk of enterprises have occurred frequently in economy area recently. Although many companies and scholars have built enterprise early warning system by binary classification or multi-class research, traditional machine learning models have poor time explanation, as they excessively pursue the prediction performance with complex machine learning approaches or deep learning models, may lead to some economic paradoxes of important risk factors, and deviates from the original intention of risk prediction. Therefore, we establish a novel non-linear survival analysis method, which not only provides a qualitative analysis of the key factors in business data, but also improves the prediction performance of XGBoost-based models. Impressive experiments have been conducted on the CSMAR database, results show the outperformance of our method compared to other approaches.**

Keywords—survival analysis, machine learning, cox proportional hazard model

I. INTRODUCTION

In the global economic downturn cycle, enterprises are facing increasing pressure on their business and financial conditions. Some enterprises would fall into financial difficulties and business crises in the harsh economic environment. The business risk of enterprise may cause heavy losses for related investors, which destroys the livelihood of the country. This destructive power would spread to the healthy development of the whole society. Therefore, how to effectively predict the business risk of enterprises and help them make scientific decisions, becomes a research topic for enterprises scholars and data analysis engineers[1-2].

In the study of modelling algorithms related to business risk, almost all of the literature attributes prediction objectives to either binary classification or multi-classification problems. Generally, objective function optimization of classifier algorithm is the main trend for improving prediction. Typically, in order to monitor these indicators and build dynamic early warning systems, business risk models are constructed to explore the risk indicators affecting operations. **However, pursuing classifier performance may go against the actual requirement of risk prediction, and yield the exact opposite economic conclusion.**

On the one hand, business management is highly influenced by economic cycles, and business indicators vary widely across economic cycles, indicating a lack of consistency in the association of business indicators with risk events over time. But the classification algorithm pursues the prediction of absolute risk, which might lead to poor temporal explanation of the model in time[3-4]. On the other hand, traditional classification algorithms are concerned with the emergence or otherwise of business risk, but the practical requirements for risk prediction also involves uncovering important risk indicators and their risk attributes.

Therefore, this paper presents a novel non-linear survival analysis method to predict business risk. Since survival analysis defines risk as a probability related to time and influencing factors, it can learn knowledge from the risk precursor environment. Considering the problem of nonlinear interactions of features, decision trees are introduced to calculate the loss functions in survival analysis, for improving the model performance while retaining the interpretability of risk variables. Therefore, we optimize the loss function by survival analysis with a decision tree framework, combining the non-linear relationship of risk variables.

II. RISK PREDICTION MODEL

A. Construction of the Risk Indicator System

The development of a company is influenced by various factors such as its own management, investment structure and the macroeconomic environment. As a result, most studies usually select some important indicators from the financial statements of a company to determine its current state of development. The indicators that affect business operations are referred to here as risk indicators, and the risk indicators are selected from the financial, stock market and corporate management levels of the companies. To obtain the best subset of features, this paper will remove redundant information through conditional judgments of correlation and redundancy analysis.

1) Covariance Diagnostics

Covariance diagnosis constructs a linear regression equation between indicators and then calculates a Variance Inflation Factor (VIF) to test the extent to which an indicator can be linearly replaced by other indicators. Removing indicators with a large degree of co-linearity can be used to remove redundant information. The steps for calculating VIF are shown below.

Step1: Constructing linear regression equations between indicators. Any one of the indicators will be used as the test variable. Construct a linear regression equation with indicator Z_j as the explained variable and other indicators as

*Corresponding author

National Nature Science Foundation of China (61862015), Science and Technology Project of Guangxi (AD21220114), Guangxi Key Research and Development Program (AB17195025)

978-1-6654-9144-0/22/$31.00 ©2022 IEEE

the explanatory variables, where are the parameters $a_0,...,a_m$ to be estimated for the regression equation. The equation is expressed as

$$Z_j = a_0 + a_1 Z_1 + ... + a_{j-1} Z_{j-1} + a_{j+1} Z_{j+1} + ... + a_m Z_m \quad (1)$$

$a_0, a_1, ..., a_m$ are parameters of the regression equation (1), and here they are estimated by the least method.

Step2: Calculate the value of VIF.

$$VIF_j = \frac{1}{1 - R_j^2} \quad (2)$$

The formula for R_j^2 is

$$R_j^2 = \frac{\sum_{i=1}^{n} (\hat{z}_{ij} - \bar{z}_j)^2}{\sum_{i=1}^{n} (z_{ij} - \bar{z}_j)^2} \quad (3)$$

$\hat{z}$ is the estimated value of the indicator *Zj* and $\bar{z}$ is the average of the indicator *Zj*. The higher VIF_j value of Z_j, the closer value of R_j^2 approximates to 1, which means that the indicator Z_j can be linearly represented and replaced by other indicators. When the *VIFj* value of *Zj* exceeds 10, it means that more than 90 percent of the information of *Zj* can be obtained from the linear combination of other variables, and it is generally considered to delete *Zj*.

Step3: Obtain the optimal attribute indicator system.

2) Information Gain method

Information Gain theory means to accurately describe the relevant characteristics of knowledge with precise values and use them to measure the importance of risk indicators, with the aim of eliminating less important indicators. Here, The formula for calculating information entropy $H(P)$ is

$$H(P) = -\sum_{i=1}^{m} P(X_i) * \ln P(X_i) \quad (4)$$

$P(X) = |X_i| / |U|, i = 1,2,...,m$ represents the cardinality of a set *U*. The importance of the indicator *a* is calculated from equation (5)

$$S_A(a) = |H(A) - H(A - \{a\})| \quad (5)$$

The steps for calculating $S_A(a)$ are shown below.

Step1: Calculate the information entropy $H(A)$ and $H(A - \{a\})$ of risk Indicator System $A = \{a_1, a_2, ..., a_n\}$ according to equation (4).

Step2: Apply equation (5) to calculate $S_A(a_i)$ and delete the indicator corresponding to $S_A(a_i) = 0$.

Step3: Obtain the optimal attribute indicator system.

B. XGBoost-based Survival Analysis in business risk prediction

This article introduces the Cox proportional hazard (Cox or CPH) model [5] in survival analysis to achieve dynamic forecasting of business risk. The Cox model is essentially a semi-parametric regression model that measures the impact of multiple risk factors on survival status and survival time. The linear model is good at learning and explaining the economic significance of risk indicators, but the model has quite strict limitations as it is not applicable to the learning of complex non-linear relationships between risk indicators and risk ratios. In order to retain the good features of the Cox model and improve on its shortcomings, many scholars have widely introduced some machine learning into survival analysis. Ishwaran applied the Random Forest (RF) to the optimization scheme of survival analysis under the condition that the Cox does not satisfy the proportional hazard assumption, and final Random Survival Forest (RSF) achieved better performance than the Cox model[6]. Ridgeway introduced Cox into the Gradient Boosting Model (GBM) [7] to realize the complex non-linear relationship.

In this paper, the EXSA method implemented by Pei Liu [8] which optimizes Cox with eXtreme Gradient Boost[9] (XGBoost) , is applied to the prediction of business risk for improving prediction accuracy. Since XGBoost is based on Gradient Boosting Decision Tree, it can achieve computational improvements and in-process tuning. **Compared to other integrated learning algorithms, XGBoost has performed better in many classification problems and regression problems**.

Assume that the training set is $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}, x_i \in X \subseteq R^m, y_i \in Y \subseteq R$, while X, Y denote the input and output space. The objective function of XGBoost takes the following form.

$$L^{(t)} \cong \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \Omega(f_t) \quad (6)$$

$\Omega(f_t)$ refers to complexity of the objective function *f* at moment *t*. To achieve quick optimization of the objective, XGBoost introduces second-order approximation into its loss function, with a loss function as in equation (7), where $g_i = \frac{\partial l(y_i, \hat{y}_i^{t-1})}{\partial \hat{y}_i^{t-1}}$ and $h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{t-1})}{\partial \hat{y}_i^{t-1}}$ are first-order and second-order.

$$L^{(t)} \cong \sum_{i=1}^{n} \left(l\left(y_i, \hat{y}_i^{t-1}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right) + \Omega(f_t) \quad (7)$$

Optimization of Cox using XGBoost means that the loss function of Cox is used as the iterative objective of XGBoost, for learning the parameters that satisfy the minimum value of the XGBoost-based loss function. Cox is composed of covariates, state variables, and survival time. The covariates are the risk factors that influence the status of individual enterprise. The state variables are used to mark whether the individual enterprise is in a risky state at certain point of time.

Risk data for enterprises is represented $\{(X_j(t), T_j, \delta_j) | j = 1,...,n\}, X_j \in R^m$,where n refers to number of enterprises; X_j refers to risk indicator; m is the dimension of risk indicators; $T_i \in R^+$ is the last observed time of enterprise; $\delta_i = 1$ indicates that a company already deep in risky situations, $\delta_i = 0$ indicates that the company has not yet experienced business risk. Assume that $X = (X_1, X_2,..., X_P)^T$ is the risk indicator system. In the process of predicting business risk, this paper uses the multivariate Cox as follow.

$$h(t|X) = h_0(t)\exp[\beta X(t)] = h_0(t)\exp[\beta_1 X_1(t) + ... + \beta_p X_p(t)] \quad (8)$$

$h_0(t)$ denotes the Cumulative risk ratio for the past period; $\exp[\beta X(t)]$ denotes the risk ratio constituted by the combination of risk factors related to the company's own operating conditions; $\beta = (\beta_1, \beta_2,..., \beta_p)$ denotes the regression coefficient of risk indicators. The parameters β are usually estimated using a partial likelihood function, which uses the category of the listed company's operating conditions and survival time as dependent variables, denoted as $Y = (\delta, t)$, with the aim of finding the one that maximizes the partial likelihood function. In this paper, the Efron estimator[13] is chosen as the parameter estimation method.

$$L = \prod_{t \in D} \frac{\sum_{k \in q(t)} \exp(\hat{y}_j)}{\prod_{l=1}^{C_t} \left[\sum_{j \in R(t)} \exp(\hat{y}_j) - \frac{l-1}{C_t} \sum_{k \in q(t)} \exp(\hat{y}_j) \right]} \quad (9)$$

$\hat{y}_i = f(x_i) = \beta x_i$ refers to log-hazard ratio of individual enterprise. To optimize the Cox model and simplify calculation, the L_E obtained by the negative logarithm of L will be used as the learning target of XGBoost, i.e., let $y = L_E$ in equation (7).

$$L_E = \sum_{t \in D} \left\{ \sum_{j \in q(t)} \left[\ln\left(\sum_{j \in R(t)} e^{y_j} - \omega_j * \sum_{k \in q(t)} e^{y_j} - y_j \right) \right] \right\} \quad (10)$$

Where $\omega_j = {}^{l-1}/_{C_t}, l = 1,...,C_t$ is assigned unique weight of j in $q(t)$, $q(t)$ is the set of all firms that can be observed at time t, $R(t)$ is the set of firms that fall into business risk at time t. $\sum e^{y_j}_{k \in q(t)}$ measures the sum of hazard ratio of all firms at time t, while $\sum e^{y_j}_{j \in R(t)}$ measures the sum of the hazard ratio of some firms in business risk at time t.

The next step is to import L_E into the equation (6) and derive its first-order derivative g_i and second-order derivative h_i

$$g_i = \begin{cases} e^{y_i} \left\{ \left[\sum_{t \le T} \alpha(t) \right] - \beta(T_i) \right\} - 1 & , \delta_i = 1 \\ e^{y_i} \sum_{t \le T} \alpha(t) & , \delta_i = 0 \end{cases} \quad (11)$$

$$h_i = \begin{cases} g_{i|\delta_i=1} - \left(e^{y_i}\right)^2 \left\{ \left[\sum_{t \le T} \phi(t) \right] - \sigma(T_i) \right\} + 1 & , \delta_i = 1 \\ g_{i|\delta_i=0} - \left(e^{y_i}\right)^2 \sum_{t \le T} \phi(t) & , \delta_i = 0 \end{cases} \quad (12)$$

The relevant variables in equations (11), (12) are calculated as follows

$$\alpha(t) = \sum_{j \in q(t)} \frac{1}{\left[\sum_{j \in R(t)} e^{y_j} - \omega_j * \sum_{k \in q(t)} e^{y_j} \right]} \quad (13)$$

$$\beta(t) = \sum_{j \in q(t)} \frac{\omega_j}{\left[\sum_{j \in R(t)} e^{y_j} - \omega_j * \sum_{k \in q(t)} e^{y_j} \right]} \quad (14)$$

$$\varphi(t) = \sum_{j \in q(t)} \frac{1}{\left[\sum_{j \in R(t)} e^{y_j} - \omega_j * \sum_{k \in q(t)} e^{y_j} \right]^2} \quad (15)$$

$$\sigma(t) = \sum_{j \in q(t)} \frac{\left[1 - (1 - \omega_j)^2 \right]}{\left[\sum_{j \in R(t)} e^{y_j} - \omega_j * \sum_{k \in q(t)} e^{y_j} \right]^2} \quad (16)$$

In this way, the loss function of the Cox is regarded as a new learning target of the XGBoost algorithm, g_i and h_i are the new first and second order derivatives in the loss function of XGBoost. The iterative process of the XGBoost algorithm is the optimization procedure of the Cox model.

III. EMPIRICAL ANALYSIS

The definition of business risk directly determines the label generation rules and the sample size of data set. It also has a significant impact on the accuracy and stability of the model. In most domestic studies, listed companies marked as "ST" are selected as high-risk samples. According to the China Securities Regulatory Commission, there are clear rules for identifying a listed company that is marked as "ST", which occurs sometime after the risk event. In this paper, we focus on those companies being marked as "ST".

This paper uses quarterly data of listed companies in Shanghai and Shenzhen A-shares from 2010 to 2020 as the research sample, with sample data obtained from the CSMAR[11] database. For the definition of survival time, the starting point of observation is the listing date of the enterprise, the end point of observation for enterprises with business risks is the first time they are marked as "ST" or "*ST", and the end point of observation for normal enterprises is 31 March 2020. The data are processed as follows: according to the 2012 version of industry classification rules issued by CSRC, listed companies in the financial industry and real estate industry are excluded; to ensure data continuity, listed companies that have been established for less than 3 years or have been "ST" or "*ST" for less than 3 years are excluded; to get rid of outliers, each risk variable is Winsorized by 1%.

TABLE I. INDICATORS TABLE

Indicators	VIF	Information Gain
Asset-liability Ratio(TDR)	4.206	0.003
Current Ratio	105.986	0.002
Quick Ratio	102.974	0.001
Return on Total Assets(ROTA)	1.729	0.002
Operating Profit Margin (OPM)	1.026	0.003
Earning per Share (EPS)	1.695	0.004
Current Asset Turnover	10.694	0.001
Inventory Turnover	1.000	0.000
Total Assets Turnover	10.500	0.001
Total Assets Growth Rate	1.001	0.0006
Tobins-Q	2.961	0.001
Ownership Concentration	6.954	0.001

Given the risk indicators are mentioned in authoritative institution reports and international advanced researches, coupled with the comprehensive consideration from the perspective of economics, multiple indicators reflecting solvency, profitability, operating capacity, development capacity, and corporate governance are selected at the level of enterprise financial status, constituting a multidimensional risk indicator system for the exploration of enterprise business risk prediction as shown in Table I.

Using above equations to calculate the VIF and Information Gain of each indicator, then Quick Ratio, Current Asset Turnover, and Inventory Turnover are removed from the risk indicator system based on the calculation results in Table I. The remaining 9 indicators are applied for risk modeling and business risk prediction.

The survival function curves obtained from the Cox are shown in Fig. 1. It can be seen that the Cox shows consistency between the survival functions estimated in the training and test set, which do not show large differences due to the random division of the data. The difference between the two survival curves' bottom may owe to sample variability. Timeline is the length of time a company has to be listed. Across the timeline, there is no overall interval where the probability of risk plummets significantly.

9 indicators and their coefficients are shown in Fig. 2. The importance degree of each indicator is proportional to the absolute value of indicator coefficient. Two main conclusions can be drawn from the Fig. 2.

TABLE II. COMPARISON OF MODEL ACCURACY

Model	Training set	Test set
Cox	90.0%	89.1%
XGBoost-Cox	97.2%	93.0%

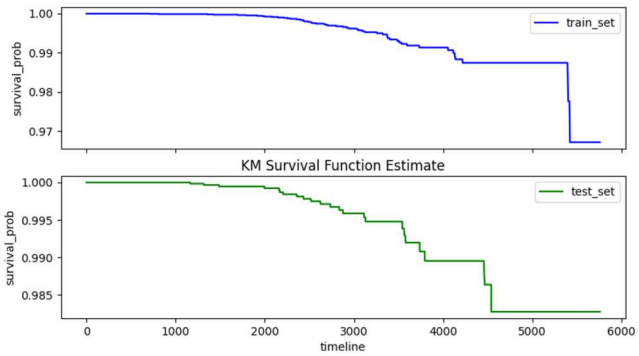

Fig. 1. Survival function curve

ROTA is the indicator that has the most significant impact on operational risk, followed by Total Assets Turnover and EPS. The coefficients of these three indicators are all negative, indicating that the greater the value of these three indicators, the lower the degree of operational risk, all of which are protective factors. In order to reduce the business risk, it is necessary to focus on the size of net profit, as well as operating income to ensure the enterprise profitability.

Both TDR and Tobins-Q can explain the asset's market value of individual enterprise, whose coefficients are positive risk factors. This indicates that TDR and Tobins-Q are negative to business risk, i.e. business risk decreases as these two coefficients' value increase. Since Asset-liability ratio is critical on enterprise business, debt management to be a key factor of business risk.

Table II shows the model prediction accuracy of Cox and Cox optimized by XGBoost(i.e. XGBoost-Cox in Table II). It can be clearly seen that the optimized model is more accurate, and it is more capable for learning non-linear features.

Meanwhile, in order to observe the performance of different models at some point of time, the Time-Dependent-AUC curve is shown in Fig. 3. It can be clearly seen that the Optimization of Cox (i.e. XGBoost-Cox in Fig. 3) has higher accuracy, higher mean AUC, and it outperforms the RSF and GBM. The Cox model outperforms the Optimization of Cox at certain points in the early stages, but it does not perform as well as the Optimization of Cox overall in the later stages. In summary, Optimization of Cox has higher discrimination ability to predict business risk and identify high-risk companies than other models.

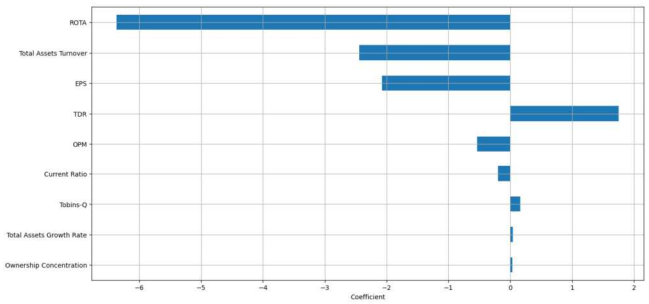

Fig. 2. Indicator coefficients and importance degree of indicators

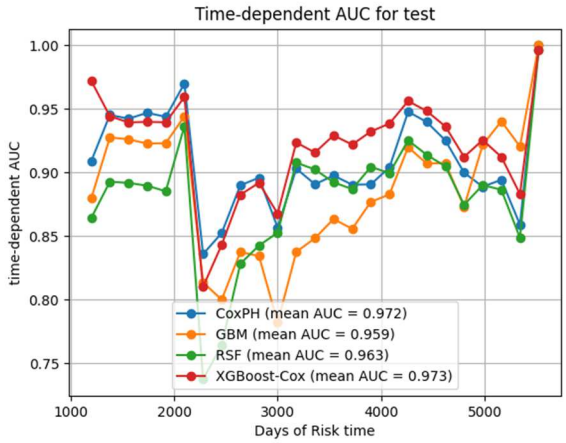

Fig. 3. Time-Dependent-AUC curve

IV. CONCLUSIONS

This article uses the Cox to conduct an empirical study on Shanghai and Shenzhen A-share listed companies in China, and also analyzes the key factors affecting business risk. Our XGBoost-based survival analysis in business risk prediction can overcome the poor nonlinear learning ability of existing model. Using Cox as a new learning objective of XGBoost algorithm and formulating specific Efron loss function based on risk indicator system, can enhance the learning ability of XGBoost and effectively improve the prediction performance of the model. We select Cox, RSF, and GBM for comparison, experiment results show that the XGBoost-Cox has the highest prediction accuracy. This algorithm can realize nonlinear machine learning of survival analysis and improve the prediction accuracy of hazard ratio.

REFERENCES

[1] LIANG Longyue, LIU Bo, "Research on Financial Risk Early Warning of Listed Companies Based on Text Mining," Computer Engineering and Applications, 2022, vol.58(04), pp.255-266.

[2] ZHANG Lu, LIU Jiapeng, TIAN Dongmei, "Application of Stacking-Bagging-Vote multi-source information fusion model for financial early warning," Journal of Computer Applications, 2022, vol.42(01), pp.280-286.

[3] LI Hongxi, Song Yu, "Dynamic Financial Early Warning Model Based on Time-Dependent Cox Regression and Empiricial Study," Operations Research and Management Science, 2020, vol.29(08), pp.177-185.

[4] BAO Xinzhong, "Early Financial Warning Based on PSO K-means Clustering Algorithm and Rough Set Theory," Journal of Systems & Management, 2012, vol.21(04), pp.1005-2542.

[5] Hemant Ishwaran, Udaya B Kogalur, "Consistency of random survival forests," Statistics & probability letters, 2010, vol.80, pp.1056-1064.

[6] Ridgeway G, "The State of Boosting," Computing Science & Statistics, 1999, pp.172-181.

[7] Liu P, Fu B, Yang S X, et al., "Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer," IEEE Transactions on Biomedical Engineering, 2020, vol.68(1), pp.148-160.

[8] Chen T, Guestrin C, "XGBoost: A scalable tree boosting system," 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp.785-794.

[9] CHEN Yiyun, "Financial Distress Prediction for Listed SME Based on the Text Information," Operations Research and Management Science, 2022, vol.31(04), pp.136-143.

[10] China Economic and Financial Research Database (CSMAR). Research Database on Business Distress of Listed Companies in China, https://www.gtarsc.com/, 2021-11-11.

[11] LI Shujin, JI Xiaojia, "The Application of the Lasso-Cox Model in Personal Credit Risk Assessment," Resource Development & Market, 2021, vol.37(2), pp.129-135.

[12] TAN Chunping, QIN Xuezhi, SHANG Qin, et al, "A Financial Distress Relief Method Based on Perpetual Bond Replacement and Equity Reiinancing," Journal of Systems & Management, 2022, vol.31(03), pp.467-475.