2020 IEEE 22nd Conference on Business Informatics (CBI)

Bankruptcy Prediction Using Survival Analysis Technique

Yuri Zelenkov

Abstract — Currently, there is an extensive set of bankruptcy prediction models, but almost all of them are classification based, i.e., they allow to estimate the posterior probability that a particular firm will fail, given its financial characteristics. The expected time to failure is not considered explicitly. On the other hand, there is a survival analysis that deals with the time of the occurrence of the event of interest (while this event may not occur during observation). However, despite its popularity in the medical and technical sciences, survival analysis is relatively rarely used in predicting financial failure. Even when it is applied, most authors use the simplest form of a model. The goal of our work is to evaluate the applicability of survival analysis to bankruptcy prediction. We compare a few state-of-art statistical and machine learning models using a real dataset. Our findings confirm that survival analysis allows (1) to extract from given data valuable information regarding the dynamics of risks and (2) to estimate the impact of features.

Index Terms— Financial computer applications, modeling and prediction, survival analysis, machine learning.

1 Introduction

THE prediction of business failure plays an essential role both in economics and society. The loss resulted from bankruptcies leads to a violation of the stability of the business environment, so it becomes a particularly challenging and important issue for business actors to estimate the sustainability of partners, customers, and financial institutes.

Currently, there is an extensive set of bankruptcy prediction models [1],[2], but almost all of them are classification based, i.e., they allow to estimate the posterior probability that a particular firm will fail, given its financial characteristics. The expected time to failure is not considered explicitly. For example, if a classification model is based on data taken one year before failure, the output from the model is the posterior probability that a particular firm will fail within one year. Decisions based on these probabilities may not be in time to prevent the failure that would occur in much less than one year [3].

On the other hand, there is a survival analysis that deals with the time of the occurrence of the event of interest (while this event may not occur during observation). However, despite its popularity in the medical and technical sciences, survival analysis is relatively rarely used in predicting financial failure. For example, Aziz and Dar (2006) [1] in the review of bankruptcy prediction models listed 12 types of classification models (from discriminant analysis and logit to case-based reasoning, neural networks, and rough sets), but do not mention survival analysis. According to this publication, the prevalent techniques are multiple discriminant analysis and logistic regression; more than 50% of works reviewed dedicate to these two models. Authors of a recent review published in 2018 [2] identified eight popular tools that include two statistical techniques (multiple discriminant analysis and logistic regression) and six machine learning models (neural network, support vector machines, rough sets, case-based reasoning, decision tree, and genetic algorithm). As we can conclude from this, survival analysis is not in the focus of researchers dealing with financial failure prediction.

The goal of our work is to evaluate the applicability of survival analysis (SA) to bankruptcy prediction.

SA models, as well as classification methods, can be divided into two main categories: statistical and based on machine learning (ML). Historically, first statistical SA models appeared in the early 70, while machine learning SA models are the results of recent research. There is a lot of research confirming that ML models outperform statistical ones in classification and regression tasks and, in particular, in classification-based bankruptcy prediction (e.g. [4]). Some publications also present similar results regarding the superiority of ML applications in various areas of survival analysis [5]. However, despite these results, most authors of bankruptcy prediction methods, even applying SA, use the simplest statistical models [6],[7].

Thus, we compare a few state-of-art statistical and machine learning SA models using a real dataset on 2457 Russian companies, 280 of which were going to bankruptcy in the one year after reporting.

The last but not least goal of our research is to evaluate available open-source software tools for Python and R languages that implement different models of SA.

The rest of the paper is organized as follows. After introducing the SA models used in our research, we briefly review publications that apply survival analysis to the financial failure problem. Next, we present a dataset and evaluate SA models, highlighting the valuable information that can be obtained using each of them. After it, we discuss the use of SA models in the classification task. Finally, we present some scenarios of how the SA technique can be used for bankruptcy prediction.

- *Yuri Zelenkov is with the Faculty of Business and Management, National Research University Higher School of Economics, 100100, Russia, Moscow, Myasnitskaya 20. E-mail: yzelenkov@hse.ru.*

2378-1971/20/$31.00 ©2020 IEEE
DOI 10.1109/CBI49978.2020.10071

2 LITERATURE REVIEW

2.1 Survival Analysis Models

In general, the survival analysis problem formulates as follows. Suppose there are n subjects, each with m covariates, denoted by$x_i = (x_{i1}, x_{i2}, \dots, x_{1m})$, for $i = 1, \dots, n$. For each subject, there is also a pair of variables (t_i, e_i), where t_i denotes the time when the event happens, and e_i is an indicator representing whether the subject fails ($e_i = 1$) or not ($e_i = 0$). In the last case, the subject is the right-censored one; it means that we have no information about the failure of this subject, except that it did not fail yet at the observation time.

The classification problem can be represented as

$$g(x) = \operatorname{sign}(p(x, \theta) - \tau),$$

where $p(x, \theta)$ is the discriminant function, θ is the vector of parameters determined by the training sample, τ is the threshold. The equation $p(x, \theta) = \tau$ defines a margin, i.e., $p(x, \theta)$ is the probability that the event will happen someday.

Survival models can take censored data into account and incorporate this uncertainty; this allows predicting the probability that an event happens at a particular time.

Let T is the time from financial data publication to company bankruptcy, and $f(t)$ is a probability density function of T. Cumulative distribution function $F(t) = \mathbb{P}(T < t)$ gives us the probability that the bankruptcy occurred before t. In other words, $F(t)$ defines the proportion of firms with the time to bankruptcy less than t.

Survival function $S(t) = 1 - F(t) = \mathbb{P}(T \geq t)$ gives us the probability that the failure has not occurred by the time t.

Hazard function $h(t)=f(t)/S(t)$ is the rate at which event happens in the surviving firms at given time t, i.e., it is a measure of risk: the higher the hazard between times t_1 and t_2, the higher the risk of failure in this time interval.

Wang et al. (2019) [5] argue that survival analysis models can be classified into two main categories: statistical methods and ML-based methods. The main difference between them is that the former focus more on characterizing the distributions of the event times and the parameter estimation by estimating the survival curves; in contrast, the latter focus primarily on the prediction of event occurrence at a given time.

In a set of statistical methods, the authors [5] distinguished non-parametric methods (e.g., Kaplan – Meier model), semi-parametric (e.g., Cox's regression), and parametric (e.g., Accelerated Failure Time model).

According to the non-parametric Kaplan-Meier method, estimation $\hat{S}(t)$ of $S(t)$ can be obtained as

$$\hat{S}(t) = \prod_{i;t_i<t} \frac{n_i - d_i}{n_i},$$

here $d_i$are the subjects for which event is occurred at time t and n_i is the subjects at risk of bankruptcy prior to time t.

Cox's proportional hazard model [8] presumes that the log-hazard of an individual object is a linear function of its covariates $\eta(x_i) = \exp(\sum_{k=1}^{m} \omega_k x_{ik})$ and a population-level baseline hazard $h_0(t)$ that changes over time, i.e.

$$h(t|x_i) = h_0(t)\eta(x_i),$$

here ω_k are the coefficients to determine. According to this model, the only the baseline hazard depends on time, the partial hazard is a time-invariant scalar factor that only increases or decreases the baseline hazard.

Aalen's additive regression [9] is an alternative to Cox's model. It allows investigating the effect of covariates on survival since the hazard rate is a linear function of the covariates with time-varying coefficients:

$$h(t|x_i) = h_0(t) + h_1(t)x_{i1} + \dots + h_m(t)x_{im}.$$

While semi-parametric models do not specify the time component of the hazard function, parametric models assume that its distribution is known. One of the most popular accelerated failure time models is based on Weibull distribution:

$$h(t|x_i) = \beta\lambda(\lambda t)^{\beta-1}, \qquad \lambda = \alpha \exp(\vec{x}_i \cdot \vec{\omega})$$

with α, β and $\vec{\omega}$ the coefficients to find.

The main challenge facing machine learning methods in SA is the difficulty of dealing appropriately with censored information and the time estimation of the model [5].

Wang et al. (2019) review the adaptation of four machine learning models to survival analysis, namely decision trees, Bayesian methods, artificial neural networks, and support vector machines [5].

Tree-based methods adaptively partition the covariance space into regions by setting a threshold for each feature. The partitioning of the covariate space creates "bins" of observations that are assumed to be approximately homogeneous. However, the original tree-based method can neither consider the censored information in the model. So, the primary difference between a survival tree and the standard decision tree is in the choice of splitting criterion.

Wang et al. (2019) list a few splitting criteria used for survival trees. One of them uses the measurement of the node deviance on the local full likelihood estimation [10]. This survival tree can be used as a base model in the Random Forest ensemble [11] when each estimator trains on bootstrapping samples drawn randomly from the given dataset with a random subset of covariates. The ensemble response is an averaging of estimators' predictions.

The prevalent model of Random Survival Forest is proposed in [12]. The node is split using the covariate that maximizes the survival difference between daughter nodes. The tree is growing to full size under the constraints that a terminal node should have no less than $d_0 > 0$ unique deaths. Then cumulative hazard function (CHF) is computed for each tree and averaged to obtain the ensemble CHF. Finally, the algorithm uses out-of-bag data to calculate prediction error for the ensemble CHF. The CHF estimate is the Nelson-Aalen estimator

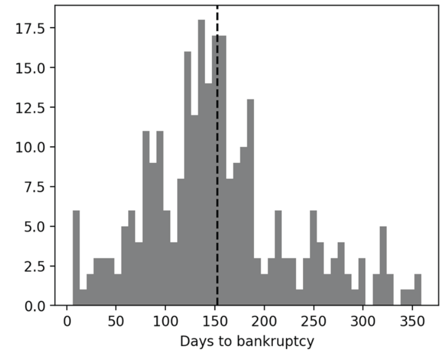


Fig. 1. Distribution of days to failure since the publication of the financial report.

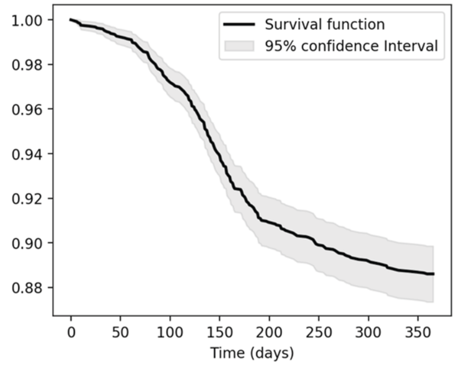


Fig. 2. Kaplan-Meier estimation of the survival function.

$$\widehat{H}(t) = \sum_{i;t_i<t} \frac{d_i}{n_i}.$$

Both of the described approaches include the specifics of right-censored data directly in the algorithm for the model training. Yet another possible approach is to look at survival analysis as a ranking problem. In that case, instead of modeling the probability that an event will occur, the model should predict whether the object has a high or low risk of experiencing the event. Such an approach opens the way to use support vector machines (SVM) for survival analysis. Currently, the applications of both linear and kernel SVM's to SA are presented [13],[14].

Multi-task logistic regression (MTLR) [15] proposes yet another way to adapt the machine learning models to censored data. It is a series of logistic regression models built on different time intervals to estimate the probability that the event of interest happened within each interval. This approach was extended in [16] by using neural networks as part of the original MTLR design that should help work with nonlinear elements in the data.

2.2 Bankruptcy Prediction on the Base of Survival Analysis

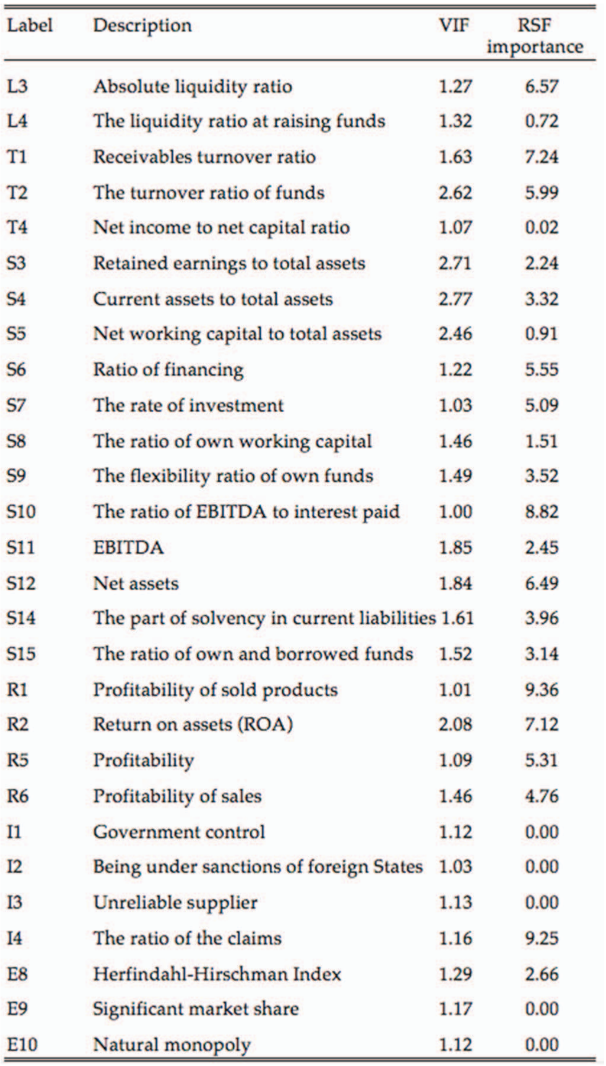
In this Section, we will review some publications that apply the survival analysis to the bankruptcy prediction problem.

Authors of one of the earliest publications in this scope [3] used Cox's model to analyze 130 failed and 334 non-failed US banks on the base of 21 financial ratios. They noted that to be useful as a part of an early warning system, Cox's model must be able to discriminate between sound banks and those likely to fail. To check the classification ability of Cox's model, they compared it to that of multiple discriminant analysis (MDA). The findings show that, statistically, neither model dominates; Cox's model has a significantly lower type I error while discriminant analysis possesses a lower type II error rate.

Type I error is defined to be the misclassification of a failed subject as non-failed (false negatives), and a type II error is defined to be the misclassification of a non-failed subject as failed (false positives). Note that the balance between these two types of errors is a critical issue of bankruptcy prediction since there is no reasonable basis for claiming that one kind of error can lead to higher losses than another.

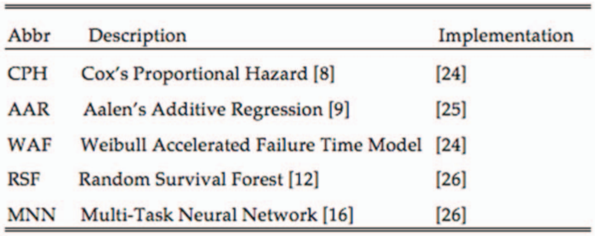
The Cox proportional hazard model is still one of the

TABLE 1
Covariates Used for Bankruptcy Prediction



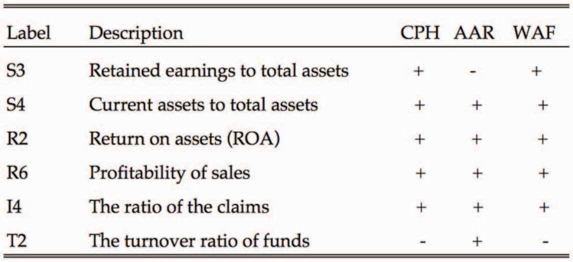
Label	Description	VIF	RSF importance
L3	Absolute liquidity ratio	1.27	6.57
L4	The liquidity ratio at raising funds	1.32	0.72
T1	Receivables turnover ratio	1.63	7.24
T2	The turnover ratio of funds	2.62	5.99
T4	Net income to net capital ratio	1.07	0.02
S3	Retained earnings to total assets	2.71	2.24
S4	Current assets to total assets	2.77	3.32
S5	Net working capital to total assets	2.46	0.91
S6	Ratio of financing	1.22	5.55
S7	The rate of investment	1.03	5.09
S8	The ratio of own working capital	1.46	1.51
S9	The flexibility ratio of own funds	1.49	3.52
S10	The ratio of EBITDA to interest paid	1.00	8.82
S11	EBITDA	1.85	2.45
S12	Net assets	1.84	6.49
S14	The part of solvency in current liabilities	1.61	3.96
S15	The ratio of own and borrowed funds	1.52	3.14
R1	Profitability of sold products	1.01	9.36
R2	Return on assets (ROA)	2.08	7.12
R5	Profitability	1.09	5.31
R6	Profitability of sales	1.46	4.76
I1	Government control	1.12	0.00
I2	Being under sanctions of foreign States	1.03	0.00
I3	Unreliable supplier	1.13	0.00
I4	The ratio of the claims	1.16	9.25
E8	Herfindahl-Hirschman Index	1.29	2.66
E9	Significant market share	1.17	0.00
E10	Natural monopoly	1.12	0.00

TABLE 2
SURVIVAL ANALYSIS MODELS



Abbr	Description	Implementation
CPH	Cox's Proportional Hazard [8]	[24]
AAR	Aalen's Additive Regression [9]	[25]
WAF	Weibull Accelerated Failure Time Model	[24]
RSF	Random Survival Forest [12]	[26]
MNN	Multi-Task Neural Network [16]	[26]

TABLE 3
SIGNIFICANT VARIABLES IN STATISTICAL MODELS



Label	Description	CPH	AAR	WAF
S3	Retained earnings to total assets	+	-	+
S4	Current assets to total assets	+	+	+
R2	Return on assets (ROA)	+	+	+
R6	Profitability of sales	+	+	+
I4	The ratio of the claims	+	+	+
T2	The turnover ratio of funds	-	+	-

most popular tools for analysis of survival rates for financial failures [6],[7]. However, the authors of work [17] use a model based on a Weibull distribution of failure time. They found that, for their data, a duration model identifies more significant variables than does the logit model.

We should also note that sometimes researchers propose own modifications of hazard models for bankruptcy prediction. Among recent publications, the work [18] tests Altman's z-score, contingent claims, and discrete hazard models introduced in [19]. The authors show that this hazard model is best suited for the UK market. The z-score and the contingent claims-based model are miscalibrated; in contrast, hazard models have average default probabilities that are closer to observed default rates. Similar results are presented in [20] for the Japan market.

In the conclusion of this brief review, it should be noted that some authors studying the effect of time on the probability of bankruptcy use other time-dependent models. For example, in [21] a model based on Markov chains was introduced, and in [22] self-organizing maps are used to represent the trajectories of firms in time.

3 EXPERIMENT SETUPS

3.1 Dataset

We use an extended version of the dataset presented in [23]. It contains data on 2457 Russian companies, 280 of which went bankrupt a year after reporting for 2014. Fig. 1 presents the distribution of days to failure since the publication of financial reports, the dashed line corresponds to mean time to failure. The minimum number of days passed before the official declaration of bankruptcy is 6, the maximum is 359, and the average is 153. Fig. 2 shows the estimation of survival function according to the Kaplan-Meier model. As follows from Fig. 1, the most significant number of companies have declared bankruptcy in the interval from 75 to 175 days after reporting, this period corresponds to a sharp decrease in the likelihood that failure will not occur in the set of observed firms (Fig. 2).

The original dataset contains 55 covariates that reflect various aspects of the firm activity, namely financial ratios, micro, and macroeconomics indicators, etc. Detail description of all covariates is presented in [23], here we will discuss only those that significant for models analyzed. Before training the models, we performed the feature selection using the Variance Inflation Factor (VIF) with threshold 3. After this procedure, we get 28 significant covariates used for further experiments (cf. Table 1).

3.2 Models

Table 2 presents the list of survival models used in our research. We analyze three statistical models: Cox's Proportional Hazard (CPH) regression, Aalen's Additive Regression (AAR), and Weibull Accelerated Failure Time Models (WAF), also as a two machine learning models Random Survival Forest (RSF) and Multi-task Neural Network (MNN) that is an adaptation of Multi-task Logistic Regression [15],[16].

Since our goal is also to test open-source packages that implement various techniques of SA, we checked a few Python and R libraries. As a result, we selected CPH and WAF implemented in `Lifelines` for Python [24], since this library provides tools for analysis of the significance of regression coefficients and AAR implementation in `survival` package for R [25] for the same reason. To run machine learning models, we used the software code of `PySurvival` library for Python [26].

To compare the SA-based approach with a classification technique that is more traditional for bankruptcy prediction tasks, we also used a few classifiers implemented in the `scikit-learn` library [27]. We checked Logistic Regression (LR) as a baseline model, and the ensembles such as AdaBoost (AB), Gradient Boosting (GB), Random Forest (RF), and Bagging Classifier (BC). Note that GB's loss function was defined as deviance (i.e., logistic regression) with probabilistic outputs since gradient boosting with exponential loss is identical to the AdaBoost algorithm.

3.3 Metrics

The most popular metric used in survival analysis is the concordance index (CI) that generalizes the area under the ROC curve (AUC) to take into account censored data. It represents the model's ability to correctly provide a reliable ranking of survival times based on individual risk scores [28].

Similarly to the AUC, CI = 1 corresponds to the best model prediction, and CI = 0.5 represents a random prediction.

4 RESULTS

4.1 Comparision of Survival Analysis Models

The first step in the design of the prediction model is the selection of its architecture, and the next one is tuning of hyperparameters.

TABLE 4
10-FOLDS CROSS-VALIDATION: CONCORDANCE INDEX

Model	Mean	Std
CPH	0.806	0.029
AAR	0.797	0.032
WAF	0.803	0.031
RSF	0.840	0.029
MNN	0.798	0.035

For statistical models, it is essential to select just those covariates that are significant, i.e., their impact on the target variable is confirmed by the analytical testing. Table 3 lists covariates whose coefficients are statistically significant (p-value less than 0.005) for CPH, AAM, and WAF models. Sign '+' marks the significance of the corresponding covariate for the relevant model.

As we can see, for each kind of regression, only five covariates are significant; this immensely simplifies the final model without notable reduction of the concordance index and other performance metrics (e.g., log-likelihood).

Note, the sets of significant covariates do not wholly match for all models. It reflects the fact that models are derived on the base of different hypotheses on the relationship between covariates and survival function.

The primary hyperparameters of the Random Survival Forest model are the number of trees M, maximal depth of the tree d, and the number of features F to consider when looking for the best split. Using a grid search procedure, we found that combination $M = 100$, $d = 10$, and $F = \sqrt{N}$, where N is the number of features, provides the best results. The values of all other parameters were set by default as in the PySurvival software library.

The primary hyperparameters that determine the behavior of the neural multi-task regression model (MNN) are the number of subdivisions of the time axis (bins), the structure of the Neural Network, and the learning rate. We should note that this model is very sensitive to these parameters, especially the learning rate. Using a grid search procedure, we determined that the optimal number of bins is 24, the learning rate is 1E-18. As a base model, we used the single-layer Neural Network with 25 neurons in the hidden layer with the ReLU activation function.

To compare all models listed in Table 2, we performed 10-fold cross-validation. On each iteration, the dataset was split on ten folds, nine of whose were used for model training and last for testing. Table 4 shows the average values and standard deviations of the concordance index for all ten iterations.

As we can see, the machine learning model (RSF) outperforms statistical ones. To check the significance of this result, we conducted additional tests as proposed in [29].

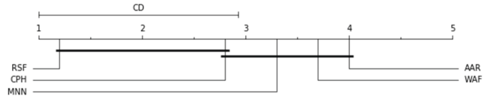


Fig. 3. Comparison of SA models performance against each other with the Nemenyi test. Groups of methods that are not significantly different (at =0.05) are connected with a solid black line.

First, we conducted the Friedman test to compare the overall performance of models on different folds obtained in the cross-validation procedure. Results obtained (F_F = 19.44, the corresponding p-value is 6E-4, and the critical value of χ^2 distribution is 9.49) confirm that the null hypothesis of the equivalent performance of all algorithms should be rejected at =0.05.

If the null hypothesis is rejected, we can proceed with a post-hoc Nemenyi test [29]. Fig. 3 displays the results of this test for =0.05. Differences in performance between models whose average ranks on cross-validation folds are further than a critical distance (CD) are statistically significant. The obtained value of the CD is 1.929. In Fig. 3, models whose differences in performance are not statistically significant are connected with a solid line.

As follows from the data presented in Fig. 3, according to the results of experiments, two groups of models are distinguished, models inside one group have statistically comparable results. The first group includes machine learning algorithm RSF and statistical model CPH, while RSF has an advantage within this group. The second group consists of the models CPH, MNN, WAF, and AAR; the CPH in this group shows the best results.

4.2 Comparision with Classification Models

How we stated above, classification as a subpart of supervised learning is the primary method of bankruptcy prediction research. In this section, we discuss how to build a robust classifier based on survival models.

As follows from the discussion in Section 2.1, the primary goal of survival analysis is to model hazard function. So, SA models can predict hazards for new objects. However, the hazard function $h(t)$ is rarely used in its original form. Most of the time, the time axis is subdivided into K parts, and the risk score of sample x is calculated as

$$r(x) = \sum_{i=1}^{K} h(t_i, x).$$

So, to convert the predicted risk of firm x to the posterior probability that it will fail, we have to determine the threshold τ. Thus, the probability of fail of firms with risk scores below the threshold will be zero. For firms with the risk scores above the threshold, the probability of bankruptcy is one. Mathematically, this is equivalent to the classification problem

$$g(x) = \text{sign}(r(x) - \tau).$$

Here $g(x)$ is the prediction of the classifier, $g(x) = 1$ for bankrupts and $g(x) = -1$ otherwise.

TABLE 5
10-FOLDS CROSS-VALIDATION: ROC AUC

Model	Mean	Std
LR	0.692	0.062
AB	0.777	0.037
BC	0.624	0.046
GB	0.646	0.049
RF	0.609	0.052
RSF	0.776	0.028
CPH	0.740	0.030

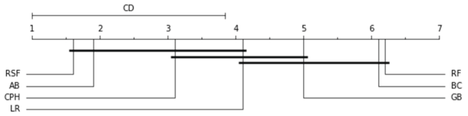


Fig. 4. Survival analysis and classification models performance on classification tasks against each other with the Nemenyi test. Groups of methods that are not significantly different (at =0.05) are connected with a solid black line.

This approach allows transforming survival analysis models to the classification models. Note that the reverse transformation of classifiers to SA models is impossible since the classifiers do not consider censoring data and time.

To determine threshold τ, we realized a simple search procedure that should maximize the ROC AUC metric on training data. The obtained value is used later to convert the predicted risk to a classification label for new objects.

Using this approach, we compared the best survival models (RSF and CPH) with standard classifiers listed in Section 3.2. We also used the 10-folds cross-validation procedure. Since the dataset is highly unbalanced, we used ROC AUC as a score.

For each classification model, we performed a simple hyperparameter tuning. The number of estimators (decision trees) was set to 200 for each ensemble. In addition, class weights were set in accordance with the imbalance ratio for those models that support this function. For boosting methods, we also tuned the learning rate. We had set the stump (decision tree with depth = 1) as a base model for AdaBoost, and maximal depth of tree = 3 for Gradient Boosting.

Obtained results are presented in Table 5. These results show that machine learning survival models remarkably outperform all standard classification models except AdaBoost in this task. One possible reason for this is a high imbalance of data since it is known that standard classifiers do not cope very well with this situation.

We also conducted the Friedman and Nemenyi tests for these data. Results (F_F = 45.5 at the p-value 4E-8 and the critical value of χ^2 distribution 12.59) confirm that the results of models are statistically different at =0.05.

Fig. 4 presents the results of the Nemenyi test. Critical distance (CD) is 2.85. As we can see, Random Survival Forest and AdaBoost demonstrate comparable performance in the classification tasks. Two statistical models (Logistic Regression from the classification model set and Cox regression from survival analysis set) also statistically belong to the best group; however, they show the worst results.

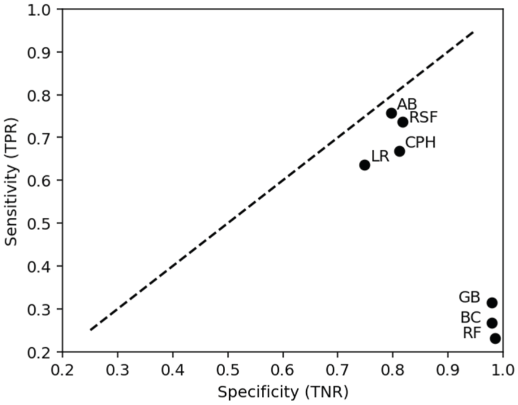


Fig. 5. Mean values of sensitivity and specificity of classification models in 10-folds cross-validation testing.

Note, according to the Nemenyi test, the Random Survival Forest is slightly ahead of AdaBoost. It is due to the fact the predictions of the RSF on various cross-validation folds are more stable; this is evidenced by the low value of variation (see Table 5).

As we noted above, the balance of false positives (Type II error) and false negatives (Type I error) predictions is essential in bankruptcy prediction tasks. To estimate it, we can use sensitivity and specificity metrics

$$Sensitivity = \frac{TP}{TP + FN},$$

$$Specificity = \frac{TN}{FP + TN}.$$

Here TP is the number of true positives (i.e., bankrupts correctly identified as bankrupts), FN is the number of false negatives (bankrupts incorrectly identified as healthy firms), TN and FP are true negatives and false positives correspondingly. Thus, sensitivity is the probability of positive labeling given that the firm is going to bankruptcy or True Positive Rate (TPR); and specificity is the probability of negative labeling given that the firm is well or True Negative Rate (TNR).

Figure 5 shows the average values of sensitivity and specificity obtained for classification models during the 10-fold cross-validation testing. The dashed line corresponds to equal values of TPR and TNR.

Presented data show that AdaBoost provides the most balanced combination of TPR and TNR. All other models that statistically comparable with AdaBoost (RSF, CPH, LR), also produce more or less acceptable results. Ensemble models like GB, RF, and BC have very low sensitivity; it means that they ill distinguish firms with financial failures. It is the effect of the high imbalance of data; these models optimize to predict the majority class.

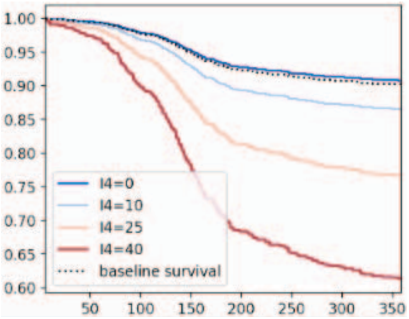


Fig. 6. Impact of I4 covariate on baseline survival function according to Cox's Proportional Hazard Model.

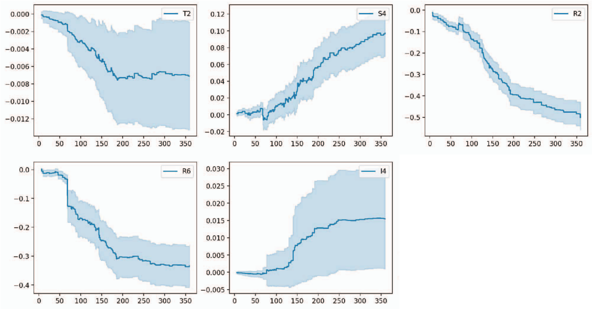


Fig. 7. Mean value and 95% confidence interval of time-varying coefficients according to Aalen's Additive Regression.

5 Discussion

Survival analysis gives yet another look at the bankruptcy prediction problem and can add valuable information for decision making. In this Section, we will discuss some scenarios based on SA.

First, it is the ability to investigate the individual impact of covariates on hazard, risks, and survival function. First of all, this can be based on statistical models since they allow evaluating the significance of covariates (cf. Table 3).

Interestingly, among more or less popular financial ratios often used in bankruptcy prediction research (S3, S4, R2, R6, T2), all models select the covariate I4 that is the ratio of the number of claims to the firm to the number of claims of the firm to other companies. To our knowledge, such a variable is rarely used in the analysis of bankruptcy indicators.

Since survival regression models based on different assumptions on the nature of covariates coefficients, it opens additional ways to analyze the impact of covariates on the hazard function. Cox's model implies that there is a baseline time-dependent hazard function. Time-invariant coefficients of covariates only increase or decrease this baseline. Fig. 6 illustrates this for I4 covariate. As we can see, I4=0 corresponds to baseline survival, and growing of I4 quickly reduces the probability of firm survival. We can conclude from this that the number of claims to the firm can be used as a very sensitive indicator of financial failure.

According to Aalen's additive regression, covariates coefficients are time varying. Fig. 7 shows the dynamics of coefficients in time. Note that its value of the I4 coefficient sharply grows after day 153 that is the average time between financial reporting and failure. The values of the other two coefficients (T2 and R6) sharply decrease until this moment, and then remain unchanged. The impact of S4 more or less stably increases all the time, and the value of R2 decreases. It gives additional insight into how each variable impacts the survival function. Monitoring of the mutual behavior of these variables also can be used as an indicator of oncoming bankruptcy.

The Random Survival Forest model also makes it possible to assess the importance of covariates. Table 1 lists this data in column 'RSF importance.' RSF estimation of covariates' importance differs comparing to statistical models. Five most essential covariates in order of importance reducing are

- R1- Profitability of sold products.
- I4 - The ratio of the claims.
- S10 - The ratio of EBITDA to interest paid.
- T1 - Receivables turnover ratio.
- R2 - Return on assets (ROA).

Note that RSF also ranks I4 as a critical indicator. Besides, all the models considered the covariate R2 highly.

It is curious that all models, including the RSF, exclude from consideration characteristics of the company, reflecting its relationship with the Government. These are the variables I1, I2, I3, E9, E10. Although they have a low VIF value, they do not have the discriminatory ability. Perhaps this can be explained by the fact that companies closely associated with the Government always receive the support that helps them to avoid bankruptcy.

The next useful tool of survival analysis is the ability to predict the values of survival or hazard functions at a given time for both individual firms and a group of firms of interest. It may be of interest not only for individual companies evaluating suppliers and consumers but also for regulators assessing the impact of decisions taken on the business environment. From this point of view, it is crucial to have a method that can model survival/hazard functions with a high level of accuracy.

Our results presented in the previous Session show that Random Survival Forest dominates other SA models. To demonstrate its ability to predict survival function, we split the dataset used above into two parts in the proportion of 0.67:0.33. We trained the RSF model on the first part of data, and next compared the prediction with actual survival function derived from the second part. Fig. 8 present the results. As we can see, the RSF model produces a good prognosis in such conditions, the corresponding value of the concordance index is 0.827.

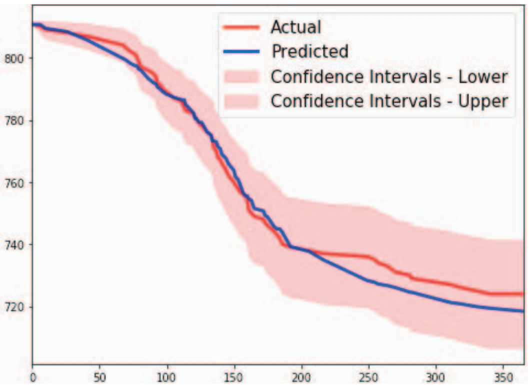


Fig. 8. Predicted and actual survival function.

The data presented in Fig. 8 can be transformed into risk scores, as described in Section 4.2. Next, firms can be distributed to different risk groups, see an example in Fig. 9. On the base of this analysis, various strategies can be elaborated to act with these groups (for example, for suppliers ranking).

The last but not least issue that we should discuss here is the readiness of open source software that implements SA models for industrial applications. In our opinion, all software libraries used in this work [24],[25],[26] have enough mature status and ready for application. Each of them has its strengths and weaknesses, but all are also in intensive development.

6 Conclusions

To summarize all of the above survival analysis is a handy tool that can be applied to bankruptcy prediction problem. It is especially true for machine learning models, such as the RSF.

Survival analysis models not only allow to reasonably accurately identify potential bankruptcy but assess the dependence of risks on times using censoring data.

What is especially valuable, quite mature free tools are currently available, which could potentially reduce the cost of introducing SA models into industrial operation.

References

[1] M.A. Aziz and H.A. Dar, "Predicting Corporate Bankruptcy: Where We Stand?" *Corporate Governance*, vol. 6, no. 1, pp. 18-33, 2006, doi: 10.1108/14720700610649436.

[2] H.A., Alaka, L.O. Oyedele, H.A. Owolabi, V. Kumar, S.O. Ajayi, O.O. Akinade and M. Bilal, "Systematic Review of Bankruptcy Prediction Models: Towards a Framework for Tool Selection," *Expert Systems with Applications*, vol. 94, pp. 164-184, 2018, doi: 10.1016/j.eswa.2017.10.040.

[3] W.R. Lane, S.W. Looney and J.W. Wansley, "An Application of the Cox Proportional Hazards Model to Bank Failure," *Journal of Banking and Finance*, vol. 10 pp. 511-531, 1986.

[4] F., Barboza, H. Kimura and E. Altman, "Machine Learning Models and Bankruptcy Prediction," *Expert Systems with Applications*, vol. 83, pp. 405-417, 2017, doi: 10.1016/j.eswa.2017.04.006.

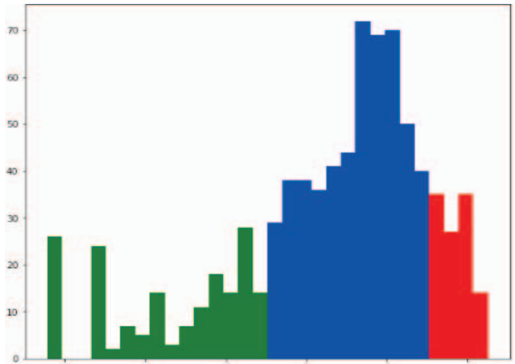


Fig. 9. Distribution of firms by bankruptcy risk group.

[5] P. Wang, Y. Li and C.K. Reddy, "Machine Learning for Survival Analysis: A Survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, Article 110, 2019, doi: 10.1145/3214306.

[6] A. Beretta and C. Heuchenne, "Variable Selection in Proportional Hazards Cure Model with Time-Varying Covariates, Application to US Bank Failures," *Journal of Applied Statistics*, vol. 46, no. 9, pp. 1529-1549, 2019, doi: 10.1080/02664763.2018.1554627.

[7] R.C. Cox, R.K. Kimmel and G. Wang, "Proportional Hazards Model of Bank Failure: Evidence from USA," *Journal of Economic & Financial Studies*, vol. 5, no. 3, pp. 35-45, 2017, doi: 10.18533/jefs.v5i03.290.

[8] D.R. Cox, "Regression Models and Life Tables," *Journal of Royal Statistical Society: Series B*, vol. 34, no. 2, pp. 187-200, 1972.

[9] O.O. Aalen, "A Linear Regression Model for the Analysis of Life Times," *Statistics in Madicine*, vol. 8, no. 8, pp. 907-925, 1989, doi: 10.1002/sim.4780080803.

[10] M. LeBlanc and J. Crowley, "Relative Risk Trees for Censored Survival Data," *Biometrics*, vol. 48, pp. 411-425, 1992, doi: 10.2307/2532300.

[11] T. Therneau and E. Atkinson, "An Introduction to Recursive Partitioning Using RPART Routines," https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf. 2019.

[12] H. Ishwaran, U.B. Kogalur, E.H. Blackstone and M.S. Lauer, "Random Survival Forests," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841-860, doi: 10.1214/08-AOAS169.

[13] V. Van Belle, K. Pelckmans, J. Suykens and S. Van Huffel, "Support vector machines for survival analysis," *In Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*, pp. 1-8, 2007.

[14] S. Pölsterl, N. Navab, and A. Katouzian. "Fast training of support vector machines for survival analysis." *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pp. 243-259, 2015, doi: 10.1007/978-3-319-23525-7-15.

[15] C.N. Yu, R. Greiner, H.C. Lin and V. Baracos, "Learning patient-specific cancer survival distributions as a sequence of dependent regressors," *In Advances in Neural Information Processing Systems 24 (NIPS 2011)*, pp. 1845-1853, 2011.

[16] S. Fotso, "Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework," arXiv:1801.05512, 2018.

[17] S.H. Lee and J. L. Urrutia, "Analysis and Prediction of Insolvency in the Property-Liability Insurance Industry: A Comparison of Logit and Hazard Models," *The Journal of Risk and Insurance*, vol. 63, pp. 121–130, 1996, doi: 10.2307/253520.

[18] J. Bauer and V. Agarwal, “Are Hazard Models Superior to Traditional Bankruptcy Prediction Approaches? A Comprehensive Test,” *Journal of Banking & Finance*, vol. 40, pp. 432-442, 2014, doi: 10.1016/j.jbankfin.2013.12.013.

[19] T. Shumway, “Forecasting Bankruptcy More Accurately: A Simple Hazard Model,” *The Journal of Business*, vol. 74, no. 1, pp. 101-124, 2001, doi: 10.1086/209665.

[20] S. Tian and Y. Yu, “Financial Ratios and Bankruptcy Predictions: An International Evidence,” *International Review of Economics & Finance*, vol. 51, pp. 510-526, 2017, doi: 10.1016/j.iref.2017.07.025

[21] D. Duffie, L. Saita and K. Wang, “Multi-Period Corporate Default Prediction with Stochastic Covariates,” *Journal of Financial Economics*, vol. 83, no. 3, pp. 635-665, 2007, doi: 10.1016/j.jfineco.2005.10.011.

[22] P. Du Jardin, “Bankruptcy Prediction Using Terminal Failure Processes,” *European Journal of Operational Research*,” vol. 242, no. 1, pp. 286-303, 2015, doi: 10.1016/j.ejor.2014.09.059.

[23] Y. Zelenkov, E. Fedorova and D. Chekrizov, “Two-step classification method based on genetic algorithm for bankruptcy forecasting,” *Expert Systems with Applications*, vol. 88, pp. 393-401, 2017, doi: 10.1016/j.eswa.2017.07.025.

[24] C. Davidson-Pilon. *Lifelines*, Zenodo, https://github.com/CamDavidsonPilon/lifelines, 2020, doi: 10.5281/zenodo.3677104.

[25] T. Therneau and T. Lumley, *survival: Survival Analysis*, https://CRAN.R-project.org/package=survival, 2019.

[26] S. Fotso et al., *PySurvival}: Open source package for Survival Analysis modeling*, https://www.pysurvival.io/, 2019.

[27] F. Pedregosa et al. "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, Vol. 12, no. pp. 2825-2830, 2011.

[28] H. Uno, T. Cai, M.J. Pencina, R.B. D'Agostino, and L. J. Wei. "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data." *Statistics in Medicine*, vol. 30, no. 10, pp. 1105-1117, 2011, doi: 10.1002/sim.4154.

[29] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” *Journal of Machine Learning Research*, vol. 7, pp.1-30, 2006.

Yuri Zelenkov received his first degree (Ph.D.) in mathematics at St. Petersburg State University, Russia (1998), and the second one (Doctor of Science) in computer science at South Ural State University, Chelyabinsk, Russia (2014). Currently, he is with the Faculty of Business and Management, National Research University Higher School of Economics, Moscow, Russia. His research interests include strategic IT management, organization knowledge management, and machine learning applications.