ARGUMENTA OECONOMICA
No 1 (44) 2020
PL ISSN 1233-5835

Aneta Ptak-Chmielewska, Anna Matuszyk**

APPLICATION OF THE RANDOM SURVIVAL FORESTS METHOD IN THE BANKRUPTCY PREDICTION FOR SMALL AND MEDIUM ENTERPRISES

Credit risk is considered to be a key risk in banking activity. The statistical and data mining models used during the assessment process of the SMEs' credit risk are mainly based on the financial data sourced from the financial statements. However, in the case of small and medium enterprises (SMEs), the non-financial factors seem to play a significant role when assessing the credit risk and this is the reason why the most frequently used ones will be discussed. The purpose of this paper was to check whether the inclusion of the non-financial factors (such as the age of the company, branch, location, legal form and number of employees) improves the prediction of the credit risk model. The combination of non-financial factors and financial ratios will be presented. During the model building process, the Random Survival Forests (RSF) method was applied. The results of the model were compared with those received using the single semiparametric Cox regression survival model. In the analysis the authors used a data sample consisting of 806 companies, including 312 bankruptcies, provided by financial institutions operating in the Polish market. Random Survival Forests provided not only better results but also more stable ones than the semiparametric Cox regression survival model.

Keywords: random survival forests, credit risk, bankruptcy probability, classification
JEL Classifications: G33, C1, C51
DOI: 10.15611/aoe.2020.1.06

1. INTRODUCTION

The last financial crisis affected the SMEs sector in different countries at different levels and strength. Even in the EU, some economies suffered less in comparison with others. SMEs represent the backbone of the economy of every country, therefore they need a prediction model easily adaptable to their characteristics.

The traditional approach in modeling the bankruptcy of SMEs was mostly based on simple discriminate analysis and mainly financial factors (Altman 1968, Beaver 1966). The authors extended this approach by trying new methods like Random Survival Forests (RSF), as well as considering

∗ Warsaw School of Economics.

additional non-financial variables that allow the segmentation of companies (size, age, region, legal form).

The paper expands the existing literature on empirical research by developing the bankruptcy prediction model using the non-parametric technique RSF, and comparing the performance with the semiparametric Cox regression survival model. Moreover, the authors apply financial and non-financial predictors to the models in order to check the influence of the region, size and legal form.

The research was inspired by the results of the previous studies undertaken for the Polish market. The models predicting bankruptcy were built using mainly the linear discriminant analysis. The data sets used in the former studies were quite small (fewer than 100 observations). Financial ratios were used as the main determinants in models. New approaches like RSF applied by Ishwaran and Kogalur (2007) seem promising. To the best of the authors' knowledge, the only application for SMEs default prediction was made by Fantazzini and Figini (2009).

All of this motivated the authors to use the machine learning technique, in this case RSF, and compare it with the survival Cox regression model,and also extending the group of financial ratios by adding the non-financial, macroeconomic and geographical factors.

The following hypotheses were tested:

(1) The inclusion of the non-financial factors in the bankruptcy risk model for SMEs increases the prediction of the model.
(2) The application of the nonparametric Random Survival Forests model improves the learning performance (effectiveness) of the model compared to a single survival semiparametric model.

The remainder of the paper is structured as follows. The next section briefly reviews the literature and previous empirical research. Section 3 gives an overview of the data set used and methods applied. The results of the empirical analysis are presented in Section 4, And Section 5 concludes with the discussion.

2. LITERATURE REVIEW

One of the first comparisons of the Random Survival Forests with another method was carried by Fantazzini and Figini (2009), who found that the RSF model outperformed the logit model for the in-sample, while in the case of the out-of-sample the results were the opposite. The authors want to extend this approach by comparing the RSF model with the semiparametric

proportional hazard model. The built model's performance will not be based on the in- and out-of-samples but on the out-of-bag error (OOB).

Modina and Pietrovito (2014) identified that both the capital structure and interest expenditure of SMEs play a more important role than the economic characteristics while specifying the determinants of company's default probability. The approach presented by Andreeva et al. (2014) combines the use of Generalized Extreme Value (GEV) regression. Additionally, the authors compared two different ways of treating the missing values, namely multiple imputation and the Weights of Evidence (WoE) approach. According to the results obtained, the GEV approach outperforms the logistic regression, where in the case of missing values WoE showed better results. Gordini (2014) proposed the usage of a genetic algorithm for predicting SMEs bankruptcy. Using Italian data consisting of 3100 manufacturing SMEs, three models were built and compared, namely: genetic algorithms, logistic regression and SVM. According to the results, the GA model outperformed the latter two, both considering the default prediction accuracy as well as in reducing the type II error. In order to identify defaulted SMEs, Calabrese et al. (2015) investigated the binary regression accounting-based model. The results obtained suggest that their approach outperformed the classical logistic regression model for different default horizons.

Kalak and Hudson (2016), using the data coming from the US market, built four discrete-time duration-dependent hazard models for SMEs, micro, small, and medium companies that became insolvent between 1980 and 2013. The authors indicated that there are significant differences between micro and small firms and these categories should be considered separately when building the credit risk models. This finding was confirmed by Gupta et al. (2015) who suggested that separate models for micro firms are desired. Sohn et al. (2016) analyzed the results of applying a fuzzy logistic regression and comparing the obtained results to typical logistic regression. The undertaken approach outperformed the logistic regression model.

Very limited literature has been dedicated to the application of survival models to bankruptcy predictions of Polish enterprises. Markowicz (2012) used the survival models for enterprises' liquidation prediction in one of the regions in Poland. Ptak-Chmielewska (2016) applied nonparametric, parametric and semiparametric models, also for correlated data based on a small sample of SMEs in Poland.

The authors found no applications and no empirical examples for the comparison of RSF with the survival model, especially for SMEs bankruptcy or default prediction.

3. DATA AND METHODS

Random Survival Forests, being closely patterned after Random Forests, naturally inherits many of its good properties. It is user-friendly because only three, fairly robust, parameters need to be set (the number of randomly selected predictors, the number of trees grown in the forest, and the splitting rule to be used). It is highly data-adaptive and virtually model- assumption-free. This last feature is especially helpful in survival analysis. Standard analyses often rely on restrictive assumptions such as proportional hazard models.

Moreover, with such methods there is always a concern whether the association between predictors and hazards was modeled appropriately, and whether or not the non-linear effects or higher order interactions for predictors should be included. Such problems are handled automatically within the Random Forests approach.

The algorithm used by `randomSurvivalForest` (R) is described as follows (Ishwaran and Kogalur 2007):

1. Draw `ntree` bootstrap samples from the original data.
2. Grow a tree for each bootstrapped data set. At each node of the tree randomly select `mtry` predictors (covariates) for splitting on. Split on a predictor using a survival splitting criterion. A node is split on that predictor which maximizes survival differences across daughter nodes.
3. Grow the tree to full size under the constraint that a terminal node should have no less than `nodesize` unique deaths.
4. Calculate an ensemble cumulative hazard estimate by combining information from the `ntree` trees. One estimate for each individual in the data is calculated.
5. Compute an out-of-bag (OOB) error rate for the ensemble derived using the first b trees, where b = 1, …, `ntree`.

In the semiparametric model (the Cox proportional hazards model) only the regression part is parametrically specified (interaction between processes), while the time distribution is not parametrically specified (nonparametric approach). It is assumed that the continuous variable T means the time until the occurrence of the event. For the Cox regression model the hazard function is given by:

$$h(t \mid x_1,...,x_k) = h_0(t)\exp(\alpha_1 x_1 + ... + \alpha_k x_k)$$

where: $h_0(t)$ – base hazard, parametrically non-specified function of time, $x_1, x_2, \ldots, x_k$ – explanatory variables.

Cox (1972) proposed using the partial maximum likelihood method to estimate the semiparametric models. In this approach, the integrity function is divided into two parts, the first one containing only the parameters and the second one containing the parameters and the hazard function as well.

The main advantage of the Cox model is the assessment of the variables influence on the process without the necessity of base hazard $h_0(t)$ specification. The main disadvantage of the Cox model is the hazard proportionality assumption (Blossfeld and Rohwer 2002). This assumption requires that for each pair of individuals in any time the hazard rate is fixed. This problem may be solved by including additional time dependent variables. For checking the proportionality assumption, the easy way is to include the interaction with time. The significance of these parameters confirms that the proportionality assumption is violated. In this case the model is called the non-proportional hazards Cox regression model. The results of Cox model estimation are the parameters describing the influence of explanatory variables on the probability of event occurrence and on the base hazard. The main advantage of the Cox regression model is that apart from the question about "if" one asks the question about "when" the event occurs (default). It is possible to include censored information about the customer. There is no need for the fixed time observation period for default observation (like in logistic regression). The results provide the "dynamic" prediction of probability of the event. It is possible to include the macroeconomic changes in the model (time-varying variables), however some disadvantages exist. There is a strong proportionality assumption that must be verified before estimating the model. This paper used the proportional hazards tests and diagnostics based on weighted residuals (Grambsch and Therneau 1994). All the assumptions used in regression models, namely: normality assumption, noncollinearity assumption, etc. are in force. This model is non-resistant to missing data, and all the observations with missing information are excluded. There is a need for the information about the exact time of the event (in this case default) which is sometimes not available to obtain in practice.

This research used a sample consisting of 806 SMEs, including 312 bankrupted. Financial Statements (FS) were available for 2010 for 'good' enterprises and from 2008 to 2010 for the 'bad' ones. The bankruptcies covered the period of 2010-2012. Random variable T (duration) is time in months starting from the date of FS till event (bankruptcy). Ending time is bankruptcy or censoring if the event did not occur within a 24-month period from the FS date.

Table 1

Financial ratios used in the analysis (calculated at the date of FS – start of the observation period)

Ratio	Name	Formula
w1	current liquidity	$\frac{\text{current assets}}{\text{short term liabilities}}$
w2	quick ratio	$\frac{\text{current assets} - \text{inventory} - \text{prepayments}}{\text{short term liabilities}}$
w3	liquidity	$\frac{\text{cash}}{\text{short term liabilities}}$
w4	capital share in assets	$\frac{\text{current assets} - \text{short term liabilities}}{\text{total assets}}$
w5	gross margin	$\frac{\text{gross profit / loss on sales}}{\text{operating expenses}}$
w6	operating profitability of sales	$\frac{\text{profit / loss on operating activities}}{\text{total revenues}}$
w7	operating profitability of assets	$\frac{\text{profit / loss on operating activities}}{\text{total assets}}$
w8	net profitability of equity	$\frac{\text{net profit / loss}}{\text{equity}}$
w9	assets turnover	$\frac{\text{total revenues}}{\text{total assets}}$
w10	current assets turnover	$\frac{\text{total revenues}}{\text{current assets}}$
w11	receivables turnover	$\frac{\text{total revenues}}{\text{receivables}}$
w12	inventory turnover	$\frac{\text{total revenues}}{\text{inventory}}$
w13	capital ratio	$\frac{\text{equity}}{\text{total liabilities}}$
w14	coverage of short-term liabilities by equity	$\frac{\text{equity}}{\text{short term liabilities}}$
w15	coverage of fixed assets by equity	$\frac{\text{equity}}{\text{fixed assets}}$
w16	share of net financial surplus in total liabilities	$\frac{\text{net profit / loss} + \text{amortisation} + \text{interests}}{\text{total liabilities}}$

Source: prepared by the authors.

All the calculations were performed in R package (`randomForestSRC`: Random Forest for Survival, Regression and Classification: Version 3.4.1).

In order to assess the financial situation of the companies, the authors selected 16 financial ratios covering a wide range of the enterprise's activity aspects (see Table 1), andadditionally included all the available non-financial factors (see Table 2). Non-financial factors allow for segmentation according to size, region, legal form and age of the enterprise. Inclusion of the non-financial factors enriches the analysis because the majority of empirical applications in the literature is limited only to the financial ratios and financial situation of the enterprise. The addition of the non-financial factors can be considered as a supplementary tool in the risk assessment (Fantazzini and Figini 2009, p. 41).

Table 2

Non-financial factors used in the analysis

Name	Attributes/categories
Sector of activity	Equal proportion of companies from sectors: Production, Trade and Services. This variable was dichotomized and reference category was set to Services (lowest risk of bankruptcy).
Cluster of regions	16 regions grouped into 3 clusters according to bankruptcy rate ("low risk", "average risk", "high risk") and dichotomized. Reference category was set to "high risk" group.
Legal form	Group1: limited liability company and group2: joint stock company, limited partnership company, other (e.g. cooperative, association, etc.). Reference category was set to group1.
Age of the company	Variable on ratio scale (age in completed years at the start of the observation period).
Number of employees	Variable on ratio scale (number of employed workers on the date of FS).

Source: prepared by the authors.

For model accuracy prediction the Concordance Error Rate was used (C-Harrell's concordance index (Harrell et al. 1982). C-Harrell's index does not depend on choosing a fixed time for evaluation of the model and takes into account the censoring of individuals. Concordance is defined as *Pr(agreement)* for any two randomly chosen observations, where in this case agreement means that the observation with the shorter survival time of the two also has the larger risk score. The predictor (or risk score) will often be the result of a Cox model or other regression.

For continuous covariates, concordance is equivalent to Kendall's tau, and for logistic regression it is equivalent to the area under the ROC curve.

The value of 1 signifies perfect agreement, 0.6-0.7 is a common result for survival data, 0.5 is agreement that is no better than chance, and 0.3-0.4 is the performance of some stock market analysts.

The computation involves all $n(n-1)/2$ pairs of data points in the sample. For the survival data, however, some of the pairs are incomparable. For instance a pair where the first one has a censored value. One does not know whether the first survival time is greater than or less than the second one. Among the observations that are comparable, pairs may also be tied on survival time (but only if both are uncensored) or on the predictor. The final concordance rate is defined as follows:

$$C = \frac{agree + tied/2}{agree + disagree + tied},$$

$$Error = 1 - C.$$

By default the concordance only counts ties in *x*, treating tied survival times as incomparable; this agrees with the AUC calculation used in logistic regression.

As shown by Fantazzini and Figini (2009), the accuracy of simple models like logistic regression is surprisingly good compared to more sophisticated and complicated models. According to Fantazzini and Figini (2009), opinion differences in performance may be swamped by other sources of uncertainty that generally are not considered in the classical supervised classification.

4. RESULTS OF THE MODELS ESTIMATION

4.1. RSF Model for financial ratios only

As the base model, the RSF model was chosen and estimated including only the financial variables (in our case ratios); 4 out of 16 financial indicators were randomly chosen,1000 trees with a minimum node size of 3 observations were specified. A *logrank* test was chosen for splitting in the construction of trees. The error rate for this model was 22.11% (Output 1).

The variable importance measure is the difference in the out-of-bag error rate when the variable is randomly permuted compared to the out-of-bag[1]

[1] Out-of-bag means that each bootstrap sample leaves out about 37% of the examples. These left-out examples can be used to form accurate estimates of important quantities. For instance, they can be used to give much improved estimates of node probabilities and node error rates in trees. They can also be used to give nearly optimal estimates of errors.

error rate without any permutation (Fantazzini and Figini 2009, p. 32). The high positive value indicates informative variables. Among predictors in this model only 7 ratios (w16, w14, w13, w6, w8, w7, w1) had the importance value substantially larger than others (see Table 3).

Table 3

RSF for financial ratios only, variables importance

Variable	Importance	Relative importance
w16	0.0351	1.0000
w14	0.0209	0.5955
w13	0.0203	0.5788
w6	0.0148	0.4227
w8	0.0112	0.3198
w7	0.0112	0.3185
w1	0.0111	0.3152
w3	0.0097	0.2750
w15	0.0093	0.2636
w2	0.0085	0.2415
w4	0.0082	0.2342
w12	0.0075	0.2139
w11	0.0066	0.1885
w10	0.0039	0.1123
w9	0.0029	0.0832
w5	0.0017	0.0484

Source: prepared by the authors in R.

4.2. RSF Model for all variables

In the next step, the RSF model included additionally the non-financial factors such as: sector, legal form, employment, age, and cluster of regions. The RSF consisted of 1000 trees with a minimum node size of 3 observations. A *logrank* test was used for splitting in the trees construction process. The addition of non-financial variables improved the model's prediction quality by decreasing the error rate to 21.46%.

However, the variables' importance in this model did not change. The same financial ratios are the most important in splitting as in the previous model with only financial ratios (see Table 4). Among the non-financial factors the most important was employment (size of the company).

Table 4

RSF for financial ratios and non-financial variables, variables importance

Variable	Importance	Relative Importance
w16	0.0279	1.0000
w13	0.0212	0.7592
w14	0.0168	0.6002
w6	0.0144	0.5174
w7	0.0109	0.3904
w1	0.0105	0.3776
w8	0.0104	0.3732
w4	0.0081	0.2888
w15	0.0076	0.2737
w2	0.0074	0.2641
w3	0.0071	0.2530
w12	0.0066	0.2371
w11	0.0050	0.1780
employment	0.0032	0.1151
w5	0.0031	0.1096
w10	0.0028	0.0997
w9	0.0028	0.0991
legal_form	0.0027	0.0976
age	0.0011	0.0410
region_low_risk	0.0007	0.0264
production	0.0006	0.0204
trade	0.0003	0.0105
region_medium_risk	0.0001	0.0037

Source: prepared by the authors in R.

4.3. Cox regression survival model – proportionality assumption

To compare the results of RSF with the traditional approach, the Cox regression semiparametric model was estimated. It is important to check for proportionality assumption in this model. To verify the assumption of proportionality, a diagnostic test was chosen based on the correlation coefficient proposed by Grambsch and Therneau (1994), who propose a formal test analogous to plotting a function of time versus scaled Schoenfeld residuals and comparing the slope of a regression line to zero. In total, as well as for each variable separately, this test confirmed the proportionality assumption (see Table 5).

Table 5

Cox regression model – proportionality assumption

Variable	rho	chi-square	p-value
w1	-0.01406	0.09039	0.7637
w2	0.04232	1.23385	0.2667
w3	-0.04685	3.02352	0.0821
w4	0.02178	0.07804	0.7800
w5	0.05248	0.90721	0.3409
w6	-0.08110	1.10865	0.2924
w7	0.06666	0.84235	0.3587
w8	0.07100	1.40287	0.2362
w9	0.01562	0.21440	0.6433
w10	0.00861	0.03003	0.8624
w11	0.02564	0.19932	0.6553
w12	-0.00733	0.05644	0.8122
w13	-0.09742	1.49369	0.2216
w14	0.03400	0.50148	0.4788
w15	-0.03780	1.45912	0.2271
w16	0.01380	0.23025	0.6313
trade	-0.07440	1.86414	0.1721
production	0.02212	0.16276	0.6866
region_low_risk	0.00617	0.01213	0.9123
region_medium_risk	0.08239	2.27831	0.1312
legal_form	-0.05964	1.10583	0.2930
age	0.03705	0.55731	0.4553
employment	0.00223	0.00177	0.9665
GLOBAL	NA	27.64292	0.2296

Source: prepared by the authors in R.

4.4. Cox regression survival model – only financial ratios

Due to the inclusion of the interval indicators to the model, the interpretation of the hazard ratio is difficult. At the level of 0.05, ten financial ratios were statistically significant (based on the Wald test). Only six ratios were not significant, namely: w4 (capital share in assets), w5 (gross margin), w8 (net profitability of equity), w11 (receivables turnover), w12 (inventory turnover) and w16 (share of net financial surplus in total liabilities). However, this model has lower quality compared to RSF. The error rate is higher – 32% (see Table 6).

Table 6

Cox regression model for financial ratios only – results of estimation

Variable	coef	exp(coef)	se(coef)	Wald (z)	Pr(>|z|)
w1	**0.0286**	**1.0290**	**0.0130**	**2.1970**	**0.0280**
w2	**-0.1065**	**0.8990**	**0.0235**	**-4.5270**	**0.0000**
w3	**0.0887**	**1.0930**	**0.0257**	**3.4470**	**0.0006**
w4	-0.0597	0.9421	0.0706	-0.8450	0.3980
w5	0.0213	1.0220	0.0113	1.8900	0.0588
w6	**-0.4127**	**0.6619**	**0.0725**	**-5.6900**	**0.0000**
w7	**0.1101**	**1.1160**	**0.0422**	**2.6090**	**0.0091**
w8	0.0026	1.0030	0.0027	0.9720	0.3312
w9	**-0.0795**	**0.9236**	**0.0330**	**-2.4110**	**0.0159**
w10	**0.0667**	**1.0690**	**0.0147**	**4.5280**	**0.0000**
w11	0.0000	1.0000	0.0001	0.0690	0.9448
w12	0.0048	1.0050	0.0026	1.8250	0.0680
w13	**-0.1517**	**0.8592**	**0.0542**	**-2.8000**	**0.0051**
w14	**0.0148**	**1.0150**	**0.0038**	**3.8510**	**0.0001**
w15	**0.0013**	**1.0010**	**0.0003**	**3.7510**	**0.0002**
w16	-0.0744	0.9283	0.0431	-1.7280	0.0840

Source: prepared by the authors in R.

4.5. Cox regression survival model – only variables with high importance in RSF

Using only financial ratios important in RSF (7 ratios: w16, w14, w13, w6, w8, w7, w1) decreases the predictive power of the model. The error rate increases up to 37.9% (see Table 7).

Table 7

Cox regression model for financial ratios with high importance in RSF – results of estimation

Variables	coef	exp(coef)	se(coef)	Wald (z)	Pr(>|z|)
w16	0.0074	1.0074	0.0094	0.7860	0.4321
w14	**0.0039**	**1.0040**	**0.0020**	**1.9590**	**0.0501**
w13	**-0.1883**	**0.8283**	**0.0326**	**-5.7760**	**0.0000**
w6	**-0.3513**	**0.7037**	**0.0695**	**-5.0530**	**0.0000**
w8	0.0047	1.0047	0.0027	1.7190	0.0856
w7	**0.0956**	**1.1003**	**0.0342**	**2.7980**	**0.0051**
w1	0.0013	1.0013	0.0037	0.3540	0.7235

Source: prepared by the authors in R.

4.6. Cox regression survival model – all variables

In the final step the Cox regression model was estimated using all variables: financial ratios and non-financial factors. Among the non-financial factors, only three were significant in this model: legal form, cluster of regions, and sector. The results of the Cox regression survival model estimation confirmed that higher bankruptcy ratio in the region contributes to the higher risk of the company's bankruptcy. The region where the company operates also seems to be a significant factor. The legal form of the company and sector of activity is important. Companies in trade and production sectors are characterized by higher risk of bankruptcy compared to the services sector, reference category (see Table 8).

Table 8

Cox regression model for financial ratios and non-financial variables – results of estimation

Variables	coef	exp(coef)	se(coef)	Wald (z)	Pr(>|z|)
w1	**0.0298**	**1.0300**	**0.0134**	**2.2200**	**0.0264**
w2	**-0.1067**	**0.8988**	**0.0249**	**-4.2910**	**0.0000**
w3	**0.0857**	**1.0890**	**0.0272**	**3.1530**	**0.0016**
w4	-0.1028	0.9023	0.0745	-1.3790	0.1678
w5	0.0203	1.0210	0.0111	1.8300	0.0673
w6	**-0.3893**	**0.6775**	**0.0741**	**-5.2570**	**0.0000**
w7	**0.0889**	**1.0930**	**0.0440**	**2.0210**	**0.0433**
w8	0.0019	1.0020	0.0027	0.6890	0.4908
w9	**-0.0845**	**0.9190**	**0.0330**	**-2.5610**	**0.0105**
w10	**0.0723**	**1.0750**	**0.0137**	**5.2570**	**0.0000**
w11	0.0000	1.0000	0.0001	0.1520	0.8788
w12	0.0052	1.0050	0.0027	1.9190	0.0549
w13	-0.1087	0.8970	0.0572	-1.9020	0.0571
w14	**0.0154**	**1.0150**	**0.0039**	**3.9130**	**0.0001**
w15	**0.0013**	**1.0010**	**0.0004**	**3.4910**	**0.0005**
w16	-0.0796	0.9235	0.0439	-1.8140	0.0696
trade	0.2785	1.3210	0.1526	1.8250	0.0679
production	**0.3108**	**1.3650**	**0.1489**	**2.0880**	**0.0368**
region_low_risk	**-1.1840**	**0.3060**	**0.2756**	**-4.2970**	**0.0000**
region_medium_risk	**-0.5085**	**0.6014**	**0.1493**	**-3.4070**	**0.0007**
legal_form	**0.1560**	**1.1690**	**0.0445**	**3.5060**	**0.0005**
age	0.0023	1.0020	0.0058	0.3990	0.6897
employment	-0.0001	0.9999	0.0002	-0.5640	0.5726

Source: prepared by the authors in R.

CONCLUSIONS

The typical ratios that are calculated on the basis of the information coming from the balance sheet are those that specify the financial strength of the company (debt-to-equity ratio) and the activity ratios showing how well the company manages its operating cycle. These ratios can provide an insight into the operational efficiency of the company. Fantazzini and Figini (2009) used the set of 16 financial ratios: supplier target, outside capital structure, industrial rights ratio, liquidity ratio, debt ratio, equity ratio, tied-up capital (financial leverage), short-term over long-term debt, tax over sales, provisions over sales, personnel expenses over sales, depreciation over sales, net income over total assets, equity over debt, short-term debt ratio, and interest income over total assets. Among the financial ratios only four were important in the RSF model, namely: personnel expenses over sales, net income over total assets, supplier target days and depreciation over sales (Fantazzini and Figini, 2009).

In the analysis the RSF has a lower concordance error comparing to the single Cox PH survival model. Important ratios in the RFS models were as follows:

- current liquidity,
- operating profitability of sales,
- operating profitability of assets
- net profitability of equity,
- capital ratio,
- coverage of short-term liabilities by equity,
- share of net financial surplus in total liabilities.

Employment was the most important ratio among the non-financial ones.

According to the results, the RSF model outperformed the Cox proportional model. RSF seems to be a promising technique in default prediction models. RSF models are more flexible compared to the Cox PH models because there is no need to test the PH assumption. The only limitation is the lack of parametric specification of the ratios' influence on the bankruptcy.

The literature on the application of the RSF method in models forecasting the bankruptcy of companies is limited. Most studies were focused on the application of other techniques, i.e. Generalized Extreme Value (GEV) regression (Andreeva et al. 2014), genetic algorithms, logistic regression and SVM (Gordini 2014), binary regression accounting-based model (Calabrese et al. 2015), and fuzzy logistic regression (Sohn et al. 2016). To the best of

the authors' knowledge, there were no applications and comparisons of the RSF model with the survival model for SMEs bankruptcy or default prediction in the literature. Admittedly, Fantazzini and Figini (2009) built an RSF model, but compared its results with a logit model and found that the RSF model outperformed the logit model only for the in-sample,while the out-of-sample logistic model performed better. The authors did not use the out-of-sample and in-sample approach but the out-of-bag error (OOB).

In future research the authors would like to extend their approach and make a comparison of the models built using different methods, like logistic regression, neural networks, SVM and others. The authors would also like to check the results on different samples and a wider range of the variables.

REFERENCES

Altman, E. I., *Financial ratios, Discriminant analysis and the prediction of corporate bankruptcy*, "Journal of Finance", Vol. 23, No. 4 (Sep., 1968), pp. 589-609, 1968.

Andreeva, G., Calabrese, R., Osmetti, S. A., *A comparative analysis of UK and Italian small businesses using Generalised Extreme Value models*, https://arxiv.org/pdf/1412.5351.pdf, 2014.

Beaver, W. H., *Financial Ratios as Predictors of Failure*, "Journal of Accounting Research", Vol. 4, Issue Empirical Research in Accounting: Selected Studies, pp. 71-111, 1966.

Blossfeld, H. P., Rohwer, G., *Techniques of Event History Modeling. New Approaches to Causal Analysis.* Lawrence Elbaum Associates Publishers, London 2002.

Calabrese, R., Marra, G., Osmetti, S. A., *Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model*, "Journal of the Operational Research Society", Vol. 67, Issue 4, 2015.

Cox, D. R., *Regression Models and Life-Tables*, "Journal of the Royal Statistical Society. Series B (Methodological)", Vol. 34, No. 2., pp. 187-220, 1972.

Fantazzini, D., Figini, S., *Random survival forests models for SME credit risk measurement, Methodology and Computing in Applied Probability*, Vol. 11, Issue 1, pp. 29–45, 2009.

Gordini, N., *A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy*, "Expert Systems with Applications", Vol. 41, Issue 14, pp. 6067-6536, 2014.

Grambsch, P., Therneau, T., *Proportional hazards tests and diagnostics based on weighted residuals,* "Biometrika", Vol. 81, pp. 515-26, 1994.

Gupta, J., Gregoriou, A., Healy, J., *Forecasting bankruptcy for SMEs using the hazard function: A review of quantitative finance and accounting*, Vol. 45, Issue 4, pp. 845-869, 2015.

Harrell, F. E. Jr, Califf, R. M., Pryor, D .B., Lee, K. L., Rosati, R. A., *Evaluating the yield of medical tests, "*Journal of the American Medical Association", Vol. 247, pp. 2543-2546, 1982.

Ishwaran, H., Kogalur, U. B., *Random Survival Forests for R, R News*, Vol. 7/2, October 2007, pp. 25-31, 2007.

Kalak, E. I., Hudson, R., *The Effect of Size on the Failure Probabilities of SMEs: An Empirical Study on the US Market Using Discrete Hazard Model* (November 26, 2015), "International Review of Financial Analysis", Vol. 43, No. 1, 2016.

Markowicz, I., *Statystyczna analiza żywotności firm* [*Statistical analysis of firms survival*]. Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, 2012.

Modina, M., Pietrovito, F., *A default prediction model for Italian SMEs: the relevance of the capital structure,* "Applied Financial Economics", Vol. 24, issue 23, pp. 1537-1554, 2014.

Ptak-Chmielewska, A., *Determinanty przeżywalności mikro i małych przedsiębiorstw w Polsce* [*Determinants of survival of micro and small enterprises in Poland*]. OW SGH, Warsaw 2016.

Sohn, S. Y., Kim, D. H., Yoon, J. H., *Technology credit scoring model with fuzzy logistic regression*, "Applied Soft Computing", Vol. 43, pp. 150-158, 2016.

Received: December 2018, revised: May 2019