Prediction of company bankruptcy in the context of changes in the economic situation

Barbara Pawełek[1]

Abstract

The papers on the prediction of company bankruptcy emphasise that there is no uniformity of a research set, including time uniformity of financial data. The economic conditions of business activities change over time. Therefore, the researchers have made attempts to study the impact of changes in the economic environment on the results of the prediction of company bankruptcy. The studies assume that the level of bankruptcy threat depends not only on the assessment of the financial condition of a given company, but it may also change depending on the economic situation in a given country or industry. The purpose of the work is to compare the predictive accuracy of selected bankruptcy prediction models built on the basis of financial data derived from various years, with the predictive accuracy of a corresponding aggregate model consisting of models built on the basis of data referring to individual years. The analysis uses the financial data of companies operating in the industrial processing sector in Poland in the years 2005-2008. The research covered selected data classification methods such as: a classification tree, *k*-nearest neighbours algorithm, support vector machine, a neural network, random forests, bagging, boosting, naive Bayes, logistic regression and linear discriminant analysis. The predictive accuracy of the constructed models was assessed on the basis of a test set. The following measures were used: sensitivity and specificity. The calculations were performed in the R programme.

Keywords: *company bankruptcy, economic situation, prediction, predictive accuracy*

JEL Classification: C380, C520, G330

1 Introduction

Corporate bankruptcy is a phenomenon characteristic for the market economy. Due to the social and economic impact of company bankruptcy, effective prediction of company bankruptcy risk is of constant interest for politicians, business practitioners and researchers. The issues related to corporate bankruptcy, including the possibility of predicting bankruptcy risk, have been widely discussed in the economic literature. Works on the methodology for forecasting company bankruptcy are being continued. The works comprise both developing methods applied traditionally in this research area (e.g. Hauser and Booth, 2011) and proposing the application of new methods, including Data Mining (e.g. Pawełek and Grochowina, 2016; Virág and Nyitrai, 2014).

[1] Corresponding author: Cracow University of Economics, Department of Statistics, 27 Rakowicka Street, 31-510 Kraków, Poland, e-mail: barbara.pawelek@uek.krakow.pl.

In papers on corporate bankruptcy prediction special attention is paid to, among other things, the problem of the homogeneity of the research set, including time-related homogeneity of financial data (e.g. Pawełek and Pociecha, 2012). Predictions of corporate bankruptcy risk are made using methods based primarily on data derived from financial statements of enterprises. If it is not possible to compile a large enough set of data for a single year then the database is built using financial data from several years. Economic conditions in which companies operate change over time. Therefore, attempts are made to study the impact of changes in the economic environment on the outcomes of corporate bankruptcy predictions, and the paper is part of this research trend. In such research it is assumed that the degree of corporate bankruptcy risk depends not only on the assessment of the financial standing of a given business but may also change depending on the economic situation in a given country or sector of economy. The literature on the subject offers proposals that take into account changes of the economic situation in bankruptcy prediction models. One proposal involves the addition of dummy variables that identify the analysed years to the set of the model explanatory variables (e.g. Pawełek et al., 2016). Another proposal points to the relevance of the incorporation of selected macroeconomic ratios in the set of explanatory variables of the model used to forecast bankruptcy (e.g. De Leonardis and Rocci, 2014).

The presented empirical research assumes that the economic situation in a given country influences the financial situation of companies. On the other hand, the evaluation of the financial situation of companies is based on the values of financial ratios. Changes of financial ratios are caused partially by changes in the economic environment of companies, including changes in the economic situation. Thus, it has been assumed that the information concerning the year of the financial statements is the link between the financial ratios and the economic situation. Partial models were built on the basis of financial data for individual years and then combined to form an aggregate model. This was intended to take account of the changes in the economic situation in the process of predicting company bankruptcy. Then, the predictive accuracy of the aggregate model was compared with the predictive accuracy of the model developed on the basis of financial data from various years. It aimed at checking whether the time-related heterogeneity of financial data, being the basis for the development of contemplated corporate bankruptcy prediction models, affects their predictive accuracy.

The aim of the paper is to present the results of comparing the predictive accuracy of selected company bankruptcy prediction models based on financial data from various years with the predictive accuracy of the corresponding aggregate model composed of models built on the data concerning individual years. The added value of the paper lies in the presentation and comparison of outcomes obtained using selected data classification methods to predict bankruptcy in the case of companies operating in the industrial processing sector in Poland on the basis of the financial data from 2005-2008[2], where such data were analysed in aggregate and for particular years. In addition, it describes the results of the verification of the hypothesis stating that values of a selected predictive accuracy measure for a given method obtained as a result of the application of two research approaches to the problem of time-related heterogeneity of financial data are derived from populations with the same averages. The calculations were made in R software using mainly ‘rminer’ (Cortez, 2015), ‘stats’ (R Core Team, 2015) and ‘HH’ packages (Heiberger, 2016).

2 Data and research procedure

The empirical research was carried out on the basis of the set of 5,920 companies operating in the industrial processing sector in Poland. Among the analysed objects there were 123 companies which the court had declared bankrupt in 2007-2010 (bankrupt companies). The financial data were related to the years 2005-2008 i.e. they were derived from financial statements published two years prior to the declaration of bankruptcy. There were 5,797 ‘healthy’ enterprises in the database i.e. companies continuing business in 2007-2010 (non-bankrupt companies). When selecting ‘healthy’ companies for the data set, the authors were guided, among other things, by their main area of business and the availability of financial data for the years 2005-2008. The data were downloaded from the Emerging Markets Information Service (https://www.emis.com/pl). Company bankruptcy risk was predicted two years in advance.

In view of the aim of the paper, the input data set was analysed in aggregate and for particular years. Four sets were separated which contained financial data from particular years in the period 2005-2008. The data from 2005 were related to 1,203 companies including 16 bankrupt companies; from 2006, to 1,368 companies including 17 bankrupt companies; from 2007, to 1,663 companies including 63 bankrupt companies; and from 2008, to 1,686 companies including 27 bankrupt companies.

On the basis of each of the four separated data sets concerning particular years, balanced research sets were created[3]. The sets comprised, respectively, 32 companies for 2005 (S_{05} set),

[2] The period between 2005 and 2008 includes the beginning of the worldwide financial crisis (e.g. Pawełek at al., 2016).

[3] The research presented in this paper is pilot research; similar deliberations will be carried out for non-balanced sets.

34 companies for 2006 (S_{06} set), 126 companies for 2007 (S_{07} set) and 54 companies for 2008 (S_{08} set). The selection of 'healthy' companies for the research sets was made at random[4]. Each of the four balanced research sets was divided at random (30 times each) into the training part (⅔ of all companies) and the test part (⅓ of all companies)[5]. The fifth balanced research set was obtained as a result of combining the four balanced sets concerning particular years from the period 2005-2008. This set included 246 companies and contained financial data from the years 2005-2008 ($S_{05\text{-}08}$ set). The training and test parts for this set were created by combining the training and test parts of the four sets related to particular years. The procedure used to compile the research sets and to divide them into training and test parts aimed at ensuring the comparability of the outcomes obtained on the basis of the $S_{05\text{-}08}$ set and after aggregating the outcomes obtained on the basis of the S_{05}, S_{06}, S_{07} and S_{08} sets.

The research entailed a dummy variable which equalled '1' in the case of companies that had been declared bankrupt in 2007-2010 and '0' for 'healthy' companies. In addition, the research involved an analysis of 32 financial ratios divided into groups: liquidity ratios (4 variables), liability ratios (10 variables), profitability ratios (7 variables) and operating efficiency ratios (11 variables).

The research comprised ten data classification methods i.e. a classification tree (rpart), *k*-nearest neighbours algorithm (knn), support vector machine (svm), a neural network (mlp), random forests (randomForest), bagging, boosting, naive Bayes (naiveBayes), logistic regression (lr) and linear discriminant analysis (lda)[6]. The methods chosen for the analysis include both traditional methods of bankruptcy prediction (e.g. a classification tree, logistic regression) and methods that are becoming increasingly popular in corporate bankruptcy risk predictions (e.g. support vector machine, random forests). Calculations were made using selected functions of 'rminer' package from R environment. In functions performing the contemplated methods, the default settings of this package were left unchanged[7].

[4] In further research, random selection of 'healthy' companies for the research sets will be repeated several times in order to assess the generality of conclusions formulated on the basis of the findings of the pilot research.

[5] The proportions for the division of the set into the training and test parts and for the number of division repetitions were assumed arbitrarily.

[6] Abbreviated names of the methods that were used in the tables are given in parentheses.

[7] Such a solution, due to the aim of the research, was considered to be justified. At the adopted level of the generality of deliberations there is no need to introduce changes in function parameters. In further analyses being the continuation of the research and dedicated to a particular data classification method, it is worth considering changes of the values of some parameters of functions to values dedicated to the analysed phenomenon of corporate bankruptcy.

Based on the outcomes obtained for each of the contemplated methods, the companies were classified into two groups: companies at risk of bankruptcy in two years (‘B’ group) and companies not at risk of bankruptcy in two years (‘NB’ group).

The evaluation of the predictive accuracy of the analysed methods was based on the values of the following classification accuracy measures for companies from the test set:

- sensitivity (the percentage of bankrupt companies classified in the ‘B’ group);
- specificity (the percentage of ‘healthy’ companies classified in the ‘NB’ group).

The hypothesis stating that values of a selected measure of the classification accuracy of a given method, calculated on the basis of the test sample created from $S_{05\text{-}08}$ set or generated after aggregating the outcomes obtained on the basis of test samples from S_{05}, S_{06}, S_{07} and S_{08} sets originating from populations with the same averages, was verified using univariate ANOVA. In order to verify whether the assumptions regarding the normality of the distribution of a variable in groups and the equality of variances for a variable in groups were observed, the Shapiro-Wilk test (Shapiro and Wilk, 1965) and the Brown-Forsythe test (Brown and Forsythe, 1974) were applied. In the case of non-compliance with such assumptions, the Kruskal-Wallis test (Kruskal and Wallis, 1952) was used.

3 Results of the empirical research

As a result of the calculations, 30 values of sensitivity and specificity measures were separately obtained for each of the ten contemplated methods. These values indicate the classification accuracy of objects from the test samples for a given method. The mean values of sensitivity and specificity measures will become the basis for the evaluation of the predictive accuracy of contemplated methods applied on $S_{05\text{-}08}$ set or S_{05}, $S_{06,}$ S_{07} and S_{08} sets.

At the first stage of the research the authors checked whether the values of the selected measure of the classification accuracy of a given method, calculated on the basis of the test sample created from $S_{05\text{-}08}$ set or generated after aggregating the outcomes obtained on the basis of test samples from S_{05} S_{06}, S_{07} and S_{08} sets, originate from populations with the same averages. The rejection of the null hypothesis (at the adopted significance level) stating the equality of averages will support the conclusions regarding the advantage (in terms of the adopted criterion) of one of the two approaches used in company bankruptcy predictions in the case of using financial data from several years.

As such, the authors checked whether the assumptions of the parametric univariate ANOVA concerning the normality of the distribution of a variable in groups and the equality of variances for a variable in groups were met. The Shapiro-Wilk test was applied in the case of each of the ten contemplated methods and each of the two analysed classification accuracy measures for objects from the test sample. Table 1 presents *p*-value in the Shapiro-Wilk test for the group of bankrupt companies (the sensitivity measure) and Table 2 – for the group of non-bankrupt companies (the specificity measure). At the significance level of 0.10, the obtained outcomes indicate the failure to meet the assumption about the normality of the distribution of both measures only in a few cases. For the sensitivity measure the null hypothesis (α = 0.10) is rejected for the classification tree and the naive Bayes method in the case of based on the $S_{05\text{-}08}$ set. For the specificity measure this is the case for the *k*-nearest neighbours algorithm and the naive Bayes method in the case of based on the S_{05} S_{06}, S_{07} and S_{08} sets.

Method	**Shapiro-Wilk test**		**Brown-Forsythe test**	**ANOVA**	**Kruskal-Wallis test**
	$S_{05\text{-}08}$	S_{05}–S_{08}			
rpart	0.0124	0.3322	0.2986	0.3581	0.2531
knn	0.2718	0.2734	0.5172	0.3609	0.3191
svm	0.7772	0.5448	0.4893	0.0132	0.0189
mlp	0.3657	0.7462	0.1544	0.1833	0.2150
randomForest	0.3439	0.6207	0.6501	0.7757	0.7945
bagging	0.4851	0.2156	0.8516	0.0576	0.0549
boosting	0.5835	0.5603	0.7305	0.7668	0.8001
naiveBayes	0.0000	0.1124	0.0216	0.0000	0.0000
lr	0.4112	0.6241	0.4728	0.0020	0.0015
lda	0.1102	0.4836	0.8773	0.0129	0.0063

Table 1. *p*-value in the Shapiro-Wilk test, the Brown-Forsythe test, ANOVA and the Kruskal-Wallis test for the sensitivity measure.

The fulfilment of the assumption about the equality of variances for a variable in the groups was verified using the Brown-Forsythe test. Table 1 presents *p*-value in the Brown-Forsythe test for the sensitivity measure and Table 2 – for the specificity measure. Test results (α = 0.10) indicate the non-satisfaction of the assumption about the equality of variances for measures only for the naive Bayes method.

Due to the fact that in the majority of analysed cases the assumptions regarding the normality of the distribution of a variable in groups and the equality of variances for a variable in groups were met, the parametric univariate ANOVA was used. The outcomes for the sensitivity measure are shown in Table 1 and for the specificity measure – in Table 2. The

Kruskal-Wallis test was also applied to be able, in cases when at least one of the assumptions of parametric ANOVA is not fulfilled, to comment (at the adopted significance level) on the equality of the averages (Tables 1 and 2).

Method	**Shapiro-Wilk test**		**Brown-Forsythe test**	**ANOVA**	**Kruskal-Wallis test**
	$S_{05\text{-}08}$	S_{05}–S_{08}			
rpart	0.1344	0.4679	0.7015	0.8501	0.8121
knn	0.8908	0.0007	0.2969	0.3754	0.6329
svm	0.2537	0.1528	0.2931	0.0034	0.0083
mlp	0.4768	0.2663	0.6945	0.9382	0.9941
randomForest	0.9415	0.2943	0.3226	0.6525	0.6232
bagging	0.3515	0.2226	0.4718	0.5709	0.5313
boosting	0.1102	0.4101	0.3335	0.4760	0.7321
naiveBayes	0.1136	0.0853	0.0017	0.0000	0.0000
lr	0.3329	0.4121	0.2993	0.0640	0.0468
lda	0.4341	0.7826	0.1157	0.0000	0.0000

Table 2. *p*-value in the Shapiro-Wilk test, the Brown-Forsythe test, ANOVA and the Kruskal-Wallis test for the specificity measure.

On the basis of the outcomes of the tests ($\alpha = 0.10$), only in a few cases did the null hypothesis have to be rejected in favour of the alternative hypothesis stating that values of a given measure of the classification accuracy, calculated on the basis of the test sample created from $S_{05\text{-}08}$ set or generated after aggregating the outcomes obtained on the basis of test samples from S_{05}, S_{06}, S_{07} and S_{08} sets, originate from populations with different averages. Such was the case for the sensitivity measure in the support vector machine, bagging, the naive Bayes method, the logistic regression and the linear discriminant analysis. It means that only for these five methods is it possible to identify a research approach supporting a statistically significant improvement of the classification accuracy of bankrupt companies from the test set. For the specificity measure there were four cases found in which, at the significance level of 0.10, the null hypothesis had to be rejected. This was the case for the support vector machine, the naive Bayes method, the logistic regression and the linear discriminant analysis. Hence, in the case of these four methods, one can point to the relevance of one of the two analysed research approaches for the improvement of the classification accuracy of non-bankrupt companies from the test set.

At the second stage of the research, values of basic descriptive measures for the sensitivity and specificity measures on the test set were calculated for each of the ten contemplated data classification methods. Table 3 below presents only the values of the arithmetic mean and standard deviation.

	Sensitivity measure				**Specificity measure**			
Method	**Mean**		**Stand. deviation**		**Mean**		**Stand. deviation**	
	$S_{05\text{-}08}$	S_{05}–S_{08}	$S_{05\text{-}08}$	S_{05}–S_{08}	$S_{05\text{-}08}$	S_{05}–S_{08}	$S_{05\text{-}08}$	S_{05}–S_{08}
rpart	0.698	0.671	0.119	0.105	0.663	0.668	0.104	0.095
knn	0.656	0.674	0.081	0.069	0.688	0.671	0.077	0.071
svm	0.695^{**}	0.635	0.085	0.097	0.684^{***}	0.618	0.087	0.080
mlp	0.642	0.670	0.088	0.070	0.641	0.639	0.086	0.075
randomForest	0.711	0.705	0.079	0.075	0.708	0.716	0.075	0.063
bagging	0.742^{*}	0.704	0.075	0.078	0.715	0.725	0.064	0.068
boosting	0.701	0.695	0.075	0.073	0.676	0.691	0.086	0.071
naiveBayes	0.262	0.479^{***}	0.098	0.134^{**}	0.934^{***}	0.765	0.039	0.108^{***}
lr	0.673^{***}	0.607	0.085	0.074	0.664^{*}	0.628	0.069	0.081
lda	0.658^{**}	0.617	0.066	0.056	0.717^{***}	0.597	0.065	0.086

In ANOVA: * – p-value < 0.10, ** – p-value < 0.05, *** – p-value < 0.01.

In the Brown-Forsythe test: * – p-value < 0.10, ** – p-value < 0.05, *** – p-value < 0.01.

Table 3. Mean and standard deviation values for the sensitivity and specificity measures.

On the basis of the values contained in Table 3, one can conclude that only in the case of three out of ten contemplated methods did the inclusion of the S_{05}, S_{06}, S_{07} and S_{08} sets in the analysis result in an increase in the arithmetic mean of the sensitivity measure against the mean value derived on the basis of the $S_{05\text{-}08}$ set. These were the following methods: the *k*-nearest neighbours algorithm, the neural network and the naive Bayes method. However, the outcomes of the tests applied in the first phase of the research make it possible to recognise the advantage of the approach based on the S_{05}, S_{06}, S_{07} and S_{08} sets as statistically significant ($\alpha = 0.10$) only in the case of the naive Bayes method. On the other hand, the superiority of the approach based on the $S_{05\text{-}08}$ set can be recognised ($\alpha = 0.10$) in the case of the support vector machine, bagging, the logistic regression and the linear discriminant analysis. The remaining differences between the arithmetic mean values should be approached with caution as the outcomes of the tests ($\alpha = 0.10$) did not provide grounds for rejecting the null hypothesis stating the equality of averages in populations.

In the case of the specificity measure, basing the analysis on the $S_{05\text{-}08}$ set for the six contemplated methods resulted in a higher arithmetic mean than in the case of basing it on the S_{05}, S_{06}, S_{07} and S_{08} sets. However, the test outcomes make it possible to recognise as statistically significant ($\alpha = 0.10$) the advantage of the approach based on the $S_{05\text{-}08}$ set only in the case of four methods, namely the support vector machine, the naive Bayes method, the logistic regression and the linear discriminant analysis.

Conclusions

Based on the results of the research, one can conclude that the predictive accuracy of certain methods of company bankruptcy prediction may an increase as a result of applying the research approach in which the fact that the data are from various years is ignored, as compared with the approach taking into account the time-related heterogeneity of financial data. Such a situation occurred in the case of the support vector machine, the logistic regression and the linear discriminant analysis – both in the group of bankrupt companies and in the group of non-bankrupt companies. On the other hand, for the naive Bayes method the aforementioned regularity was observed only in the group of non-bankrupt companies. The advantage of the approach based on the data concerning individual years over the solution based on data originating from various years was revealed in the case of the naive Bayes method, but only in the group of bankrupt companies. However, for half of the analysed corporate bankruptcy prediction methods, at the significance level of 0.10, the advantage (in terms of the adopted criterion) of one of the two analysed approaches to the problem of the time-related heterogeneity of financial data has not been proven. Thus, research on the impact of changes in the economic situation on company bankruptcy prediction results should be continued.

In further research the authors intend primarily to repeat the analysis for unbalanced sets, apply v-fold cross-validation method and expand the set of measures of the predictive accuracy of the analysed methods.

Acknowledgements

Publication was financed from the funds granted to the Faculty of Management at Cracow University of Economics, within the framework of the subsidy for the maintenance of research potential.

References

Brown, M. B., & Forsythe, A. B. (1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association, 69(346),* 364–367.

Cortez, P. (2015). *rminer: Data Mining Classification and Regression Methods*. R package version 1.4.1. URL: http://CRAN.R-project.org/package=rminer.

De Leonardis, D., & Rocci, R. (2014). Default Risk Analysis via a Discrete-time Cure Rate Model. *Applied Stochastic Models in Business and Industry, 30(5),* 529–543.

Hauser, R. P., & Booth, D. (2011). Predicting Bankruptcy with Robust Logistic Regression. *Journal of Data Science, 9(4),* 565–584.

Heiberger, R. M. (2016). *HH: Statistical Analysis and Data Display: Heiberger and Holland*. R package version 3.1-31. URL: http://CRAN.R-project.org/package=HH.

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association, 47(260),* 583–621.

Pawełek, B., & Grochowina, D. (2016). The random subspace method in the prediction of the risk of bankruptcy of companies in Poland. In: Malina, A., Węgrzyn, R. (eds.), *Knowledge – Economy – Society. Challenges and Development of Modern Finance and Information Technology in Changing Market Conditions,* Foundation of the Cracow University of Economics, Kraków, 25–34.

Pawełek, B., & Pociecha, J. (2012). General SEM Model in Researching Corporate Bankruptcy and Business Cycles. In: Pociecha, J., Decker, R. (eds.), *Data Analysis Methods and its Applications,* C.H. Beck, Warsaw, 215–231.

Pawełek, B., Pociecha, J., & Baryła, M. (2016). *Dynamic Aspects of Bankruptcy Prediction Logit Model for Manufacturing Firms in Poland*. In: Wilhelm, A. F. X., Kestler, H. A. (eds.), *Analysis of Large and Complex Data,* Studies in Classification, Data Analysis, and Knowledge Organization, Switzerland: Springer, 369–382.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika, 52(3/4),* 591–611.

Virág, M., & Nyitrai, T. (2014). The application of ensemble methods in forecasting bankruptcy. *Financial and Economic Review, 13(4),* 178–193.