DATA ANALYSIS OF THE FINANCIAL INDICATORS OF POLISH COMPANIES

Anna BICEKOVÁ, Ľudmila PUSZTOVÁ
Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic, tel. +421 55 602 4220,
e-mail: anna.bicekova@tuke.sk, ludmila.pusztova.2@tuke.sk

ABSTRACT

This article aims to present the issue of the company´s bankruptcy and defines which financial indicators affects and can accurately detect the financial health of the company and thus better predict the emergence of potential bankruptcy. Currently, these methods include mainly modern techniques from the data mining area. For the practical application of this approach to predict the future state of the company, were used the financial indicators of Polish companies. We used the most suitable algorithms for predicting bankruptcies – decision trees that provide simple results interpretation. The analytical process is managed by the CRIPS-DM methodology, which offers a description of the important steps needed to solve this task. Part of the article constitutes an analysis of the current state, which presents solutions to this problem by other authors. Analysing available data we found that the most effective financial indicators are Attr27[profit on operating activities / financial expenses], Attr34 [operating expense/total liabilities] and Attr41 (total liabilities/[(profit on operating activities + depreciation)(12*365)]). The model that best-predicted bankruptcy was the C5.0 decision tree algorithm.*

Keywords: *Data mining, bankruptcy prediction, financial indicators, CRISP-DM methodology*

1. INTRODUCTION

The issue the company´s bankruptcy is an essential part of economic life. For any kind of corporations in the economy, is important continuous growth, development, and keeping the market position. All these aspects can expose unfavourable decisions that cause problems, and in the worst case, can lead to a gradual decline of the company. An example is the many financial crises – the best known in 2008 [1] that influenced to known and unknown companies and banks around the world.

Information technology and resources is a necessity for the survival of economic subjects in the market. The use of data analysis to predict a company´s aspect is a key way to prevent this situation. Nowadays, the use of acquired knowledge from important data using modern techniques is popular. Therefore, the focus of the topic in this area was motivating for the solution of the selected issue. Data from financial indicators and appropriate prediction methods can show a company what status may occur in the future and prepare specific measures.

To determine the financial health of the company serves *financial indicators* acquired from the financial statements of the company. The historical development of bankruptcies is dividing into stage *before* and *after 1966* [2]. During this period, the authors reached various conclusions as to which financial indicator affects and can accurately detect the financial health of the company and thus better predict the emergence of potential bankruptcy.

Before 1996 was found that among the important ratios of the financial indicators when assessing corporations are ratios: *equity and liabilities*, and the ratio of *net profit and equity* (by FitzPatrick [3]); *ratio working capital* and *total assets* (by Smith a Winakor [4]); *net working capital ratio* and *total assets*, *short-term liquidity ratio* and *equity ratio to total liabilities* (by Merwina [5]).

After 1966, began to use *dynamic prediction models*, which were able to determine the risk of bankruptcy for each company at any time. Among the important financial indicators included ratio the *net profit* and *total liabilities* (by Beaver [6]).

It is very important for companies (banks or businesses) to track their „own numbers " and to use the resources to help them make better decisions about the future of the company and avoid financial problems.

1.1. State-of-the-art

The bankruptcies problem and their prediction are an interesting and increasingly sought-after area for many experts. They are trying to find the best techniques that could help corporations and investors to estimate the company's market position in the future and avoid the various complications leading to financial problems. In the analysis, we dealt with four works that are directly related to our selected dataset. The latest work deals with evaluating the best model on Chinese company data, which can serve as a warning system for business management before it reaches to bankruptcy.

The work of Chinese scientists [7] was dealt with the issue of bankruptcy prediction with on the same dataset of Polish companies as we used. They decided to solve the problem of distorted data by applying several anomaly detection algorithms. To identify this variation, they used three different models, namely, multiple Gaussian Distribution, One-Class SVM (Support Vector Machine), and Isolated Forest. The purpose of the first experiment was to use anomaly detection methods to determine the best prediction model. In the second experiment, they used four controlled learning models. The available data were divided into a training and test set of 60:40 and used 5-fold cross-validation. The authors also evaluated individual models using performance metrics, ROC (Receiver Operating Characteristics Curve), and AUC (Are Under the Curve) curves. They found that model generated by isolation forest in the methods of detecting deviations had the best results, e.g., in the 1st Year classification case, the isolated forest had a mean of 0.93 and the neural network

of all seven compared models has the lowest mean of 0.84. They pointed out that even models of detecting anomalies can solve the problem of unbalanced and distorted data because models of controlled learning often fail to solve the problem.

The authors of the paper [8] were dealing with the creation of the most appropriate model for predicting bankruptcies, which they used to develop a decision support system. The data used differed in the number of attributes (7 columns) but did not describe the financial indicators, but the qualitative characteristics of the company such as competitiveness, management risk, credibility, industrial risk, financial flexibility, operating risk, and class bankruptcy status. The first six attributes had nominal values - classes with positive, average, and negative levels; these values then transformed into numeric attributes such as 1, 0.5, and 0. Class attribute had two values: non-bankruptcy or bankruptcy. The ratio of the class attribute was balanced, 107 records had a bankruptcy value, and 143 had a non-bankruptcy value. The data were dividing into a training and testing set of 70:30. To find the most accurate model, the authors compared several algorithms that included logistic regression, random forest, naive Bayes, SVM method, and neural networks. They used 10-fold cross-validation to verify the success of each classifier. The algorithms were compared based on four performance metrics, namely success in %, ratio true-positive, and true-negative, and accuracy. The most successful model was creating with SVM algorithm with a success rate of 99.6%, a neural network with a success rate of 98.6% and the third best model was naive Bayes with a success rate of 98.3%. In general, all models had a high percentage of accuracy and success over 90%.

In the study [9] by Maciej Zięba, Sebastian K. Tomczak, Jakub M. Tomczak proposed a new approach to bankruptcy prediction that uses the extreme gradient boosting (XGB) method to teach a set of decision trees. The research aimed to identify the best classification model for each of the five datasets. Consideration was given to 16 classification models, including, e.g., linear discriminant analysis, concealed layer multilayer perceptron, logistic regression, AdaBoost, random forest, and extreme gradient boosting. They used the AUC curve to evaluate the models and 10-fold cross-validation to test the quality of the various training parameter tools. The results of the experiment presented as an average and standard deviation for each of the five classification cases. In addition to comparing the results between previous models, they focused on the results of the XGBE, XGB, and EXGB models, which are an extension of the XGB model. Using the Wilcoxson p-value test, the authors evaluated that the best classification model is EXGB with the highest average values - 0.959, for XGB comparison it was 0.945 or random forest with 0.851, from among all 16 classification models.

The publications "A general introduction to data analytics" [10] deals with a general data analysis for students or enthusiasts and is written as a guide on how to proceed with your projects. As one example, they used Polish company data and followed the CRISP-DM methodology. During the data preparation phase, they solved the problem of disproportion between classes, so they decided to delete more than 800 lines (records) with class 0, i.e., non-bankruptcy companies. Subsequently, the attribute values were normalized, and the missing values in the attributes were replaced by the average of the column values.

In the modelling phase, the authors chose three models. The K-nearest neighbour with k=15 and used the Euclidean distance as a measure of distance, the data was divided into a training and testing set in 70:30 ratio and the selected attributes for modelling were Attr6, Attr11, Attr24, Attr27, and Attr60. Also, they used the decision tree algorithm C4.5 and the random forest algorithm in which were generated 500 trees. They used 10-fold cross validation for all three algorithms. The obtained results showed that the random forest algorithm had the best accuracy - 98.47%. For comparison, the C4.5 algorithm had 98.30% accuracy and k-NN method 98 21% accuracy.

Analysing the current state, we have encountered many studies with a central issue of solving the problem of bankruptcies and their prediction. Many of them have used our data, so after analysis, we can compare our results. Data mining techniques used in these works are different, most often decision trees, neural networks, SVM method, but also random forest or k-nearest neighbour model.

1.2. Used methods

For the proper management of the analytical task, have been developed various methodologies to provide an overview of the whole life cycle of the task. The best-known is the **CRISP-DM** (Cross-Industry Standard Process for Data Mining) **methodology**, which practically used since 1996. The use of this methodology is primarily focusing on large-scale projects that are managed faster, more efficiently, and with less funding. Provides step by step instructions, but their order is not accurate. If necessary, you can go back to previous steps and repeat certain actions. This model is an idealized sequence of events, and according to [10] it consists of the following 6 phases:

Problem understanding - the first and most important phase of the process. In this step, it is important to understand the problem from a *business point of view* and in *terms of data mining*. In the introduction, it is necessary to evaluate the current situation of the problem, which can be used to identify factors, means, or constraints that could affect the overall project outcome. Also, it is necessary to set success criteria need to evaluate the solution.

Data understanding - describes selected data and their necessary information, such as the number of available records, column names and their description, range of values (minimum and maximum), average values in each attribute, types of individual data (numeric, binary). *Data quality* is determined, i.e., their consistency and the possibility of occurrence of missing values. The last task is to analyse data using *statistical analysis* and simple visualizations that provide information about the interrelationships between attributes, the distribution of key variables, or various simple statistics [10].

Data preparation - includes activities that lead to the creation of the resulting data set adequate to the specified task goal. In *selecting data* is important to choose those that subsequently used for analysis and modelling. The selection applies to both - attributes and records. *Data cleaning* depends on the quality of available data (which

may be noisy, inconsistent, contain missing/empty values). In the case of missing values, it is necessary to add or remove them. For example, the role of *data construction* is to transform attribute values, generate new records, and attributes. *Data integration* is performing when data comes from different tables and needs to be merged.

Modelling - applying the best and most suitable techniques for modelling to prepared data. Since most methods have different data requirements, interaction with the preparation phase is necessary. After creating the models, it is important to evaluate the models using various criteria that are presenting in the problem understanding phase.

Evaluation - aims to evaluate models in general from the business point of view. It focuses on verifying that the objective goals have been done. In this step, it is necessary to focus on failed tasks that have been neglected during modelling. Further steps are also taken - either the decision to terminate the project or go to the next phase.

Deployment - if in the evaluation phase was decided to continue of project, it is important the acquired results prepare to clear form for the recipient. The outcome of the phase is the overall assessment of the project, the problems encountered, or the potential steps that could still be developing within the framework of the problem.

In our analytical task, we applied the *decision tree technique* to the selected data sample in modelling phase. In practice are used different types of decision tree algorithms, such as C4.5, C5.0, CART, or random forest (RF), which differ mainly in their way of improving model accuracy or by splitting record sets in modelling.

To evaluate each model we used *quality metrics* [11], namely *accuracy* (measures the proportion of correctly classified positive cases to all cases), *success rate* (the proportion of correctly classified cases to all), *error* (the proportion of incorrectly classified cases to all), *sensitivity* (the proportion of correctly classified positive cases to all positive cases), *specificity* (proportion of correctly classified negative cases to all negative cases) and *AUC value* (quantifies the overall ability of the model to distinguish between correct and misclassified cases).

In the data preparation phase, we also applied the *attribute selection method* [12], which by reducing the number of dimensions enabled better data handling. Specifically, we used the *PCA method, LASSO method*, and the c*orrelation between attributes*.

PCA (Principal component analysis) is a method suitable for datasets with many attributes and correlation. It provides the selection of those attributes that offer the most information and are linearly independent of each other [13]. This method initially looks for components that are eigenvectors representing the direction of the greatest data dispersion. Each eigenvector has the corresponding eigenvalue, and the most beneficial one is a component with an eigenvalue greater than 1.

LASSO (Least absolute shrinkage and selection operator) [14] is a regression method. The method applies a penalty process and ensures that important attributes for modelling have a non-zero value. An important factor is a parameter λ, which controls the strength of the penalty. If the value λ is higher, the more attributes will be zero, and the dimensions will be reducing

For reducing, the number of attributes can also use *correlation of dependency relationships*. It based on various statistical tests. Dependence is comparing between two attributes and the correlation coefficients can be used to determine the correlation strength, which may be weak ($c < 0.5$), medium ($0.5 \leq c < 0.8$) or strong ($0.8 \leq c$).

2. ANALYTICAL PROCESS

The analytical process was performed using the programming language R by the CRISP-DM methodology.

2.1. Problem understanding

The business goal of this task was to help the investor decide whether or not to choose a company for their investment plans. For investor is important to know in what condition the company will be located after three, four, or five years based on the current financial indicators.

The goal from the data mining point of view was to find a model that, based on the financial indicators of Polish companies, would be able to predict bankruptcy (1) or non-bankruptcy (0) of the company. For creating this model, we used a specific data mining task - *classification*. The individual classification models were initially generated on the training set and subsequently evaluated on the testing set. In addition to the metrics mentioned above, we also used a pivot table to determine the models, which identified the number of bankruptcy / non-bankruptcy companies (predicted and actual value).

2.2. Data understanding

The data sample used for this task comes from the Machine Learning Repository [15] describing the financial indicators of Polish companies from 2000 to 2013. The data was dividing into five sets based on the bankruptcy prediction period. Each set contained a different number of records (companies), the same number and meaning of attributes, and different values of each attribute. The number of attributes was 65 - the first 64 were ratio indicators, and the last was the target class indicating the company status, i.e., 0 as the "non-bankruptcy" company and 1 as the "bankruptcy" company.

In table 1 are several ratio indicators with their range of values for the dataset 1st Year.

Table 1 Ratio indicators for the dataset 1st Year

Ratio indicators	Range of values	
	max	min
net profit/total assets	-256,89	94,28
total liabilities/total assets	-72,162	441,5
working capital / total assets	-440,5	1
current assets / short-term liabilities	0	1017,8
retained earnings / total assets	-397,89	303,67
sales/total assets	0	3876,1
gross profit / short-term liabilities	-23,207	331,46
(net profit + depreciation) / total liabilities	-21,793	612,88
(total liabilities - cash) / sales	-149,07	152 860
operating costs / total liabilities	-280,26	884,2
profit on sales / total assets	-169,47	445,47
total sales / total assets	0	3876,1

profit on operating activities / sales	-701,63	2156,8
net profit / inventory	-256,230	5986,8
short-term liabilities / total assets	0	441,5
working capital	-800 470	4 398 400
long-term liabilities / equity	-327,97	119,58
sales / inventory	0	2 137 800
sales / receivables	0	21 110
(short-term liabilities * 365) / incomes	0	25 016 000
sales / short-term liabilities	0	1 042,2
sales / fixed assets	0	294 770
(short-term liabilities * 365) / cost of products sold	0	453,96
constant capital / total assets	-440,55	1 099,5
total liabilities / [(operating income + depreciation) * (12 * 365)]	-77,791	813,14
(short-term assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)	-315,37	126,67

The first set - **1st Year** - contained financial ratios from the first year of the prediction period and reflected the bankruptcy status after five years. The dataset included 7027 companies (records), of which 6756 were non-bankruptcy, and 271 were bankruptcy.

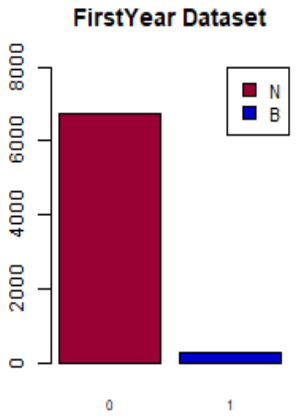

Fig. 1 A distribution graph for the target attribute of the 1st Year dataset

The second set – 2nd Year - contained financial ratios from the second year of the prediction period and reflected the bankruptcy status after four years. The dataset included 10 173 companies (records), of which 9773 were non-bankruptcy, and 400 were bankruptcy.

The third set – 3rd Year - contained financial ratios from the third year of the prediction period and reflected the bankruptcy status after three years. The dataset included 10 503 companies (records), of which 10 008 were non-bankruptcy, and 495 were bankruptcy.

The fourth set – 4th Year - contained financial ratios from the fourth year of the prediction period and reflected the bankruptcy status after two years. The dataset included 9 792 companies (records), of which 9 277 were non-bankruptcy, and 515 were bankruptcy.

The fifth set – 5th Year - contained financial ratios from the fifth year of the prediction period and reflected the bankruptcy status after one year. The dataset included 5 910 companies (records), of which 5 500 were non-bankruptcy, and 410 were bankruptcy.

At this phase, we also found *dependencies between* individual numerical attributes with each other. Since the dependence values were different, we focused mainly on the values of dependencies <0.8; 1> and <-0.8; -1>. The highest dependencies that were common to all five data sets were, for example, dependencies:

Attr1: *net profit / total assets* and **Attr7**: *EBIT/ total assets,*

Attr7: *EBIT/total assets* and **Attr14**: *(gross profit + interest) / total assets,*

Attr2: *total liabilities / total assets* and **Attr10**: *equity / total assets,*

By statistical analysis we found, that data had many missing values. For example, Attr21 *(sales (n) / sales (n-1))*, Attr27 *(profit on operating activities / financial expenses),* Attr45 *(net profit / inventory)* and Attr60 *(sales / inventory)* had the highest number of missing values. However, Attr37 *((current assets - inventories) / long-term liabilities)* contained the highest number of missing values in **all five data sets.**

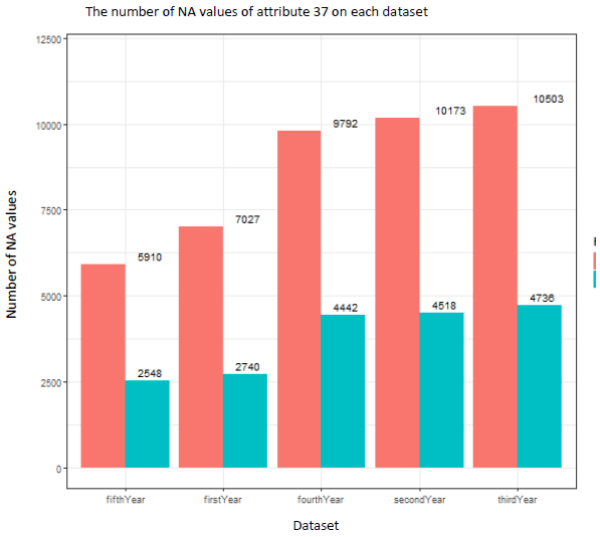

Fig. 2 Distribution graph of missing values in attribute Attr37 of each dataset

2.3. Data preparation

Initially, we decided to join the individual datasets into one set, since the number and type of attributes were the same in each set. By joining them, we got 43,405 records. Then we solved the problem of missing values. We decided to delete the attribute with the highest number of missing values - Attr37. In attributes with a lower number of missing values, we replaced the values in two ways, namely using the k-NN method (with k = 4) and the average of the given attribute values. Similarly, the authors of the study [10] have also chosen this method.

In this step, we also used the methods of selecting attributes - PCA and LASSO.

PCA - from the analysis of the principal components, we found how much percentage of the information get from the individual components. Most often, the highest

percentage appeared in the first component; the remaining percentages were dividing into the remaining components.

On the dataset where we replaced the missing values by k-NN method, we achieved up to 93% of the information for the first 25 components. **The first component (A)** represented **15.5%** and the **second component (B) 13.8%,** so we focused on selecting attributes from the first two components. On the set where we replaced the missing values with the average of the attributes, component (A) represented **15.6%** and component (B) **14%.**

LASSO - the attributes generated by this method were selected based on non-zero values λ, that deviates from zero in the positive or negative direction. Attributes with the highest non-zero values deviate in a positive direction were: Attr48: *EBITDA (profit on operating activities) / total assets* and Attr22: *profit on operating activities / total assets.*

After using these methods, the number of selected attributes by feature selection methods on sets with different replace missing values are in the next table.

Table 2 The number of selected attributes by feature selection methods

k-NN	Correlation	46
	PCA	23
	LASSO	34
average	Correlation	37
	PCA	23
	LASSO	34

2.5. Modeling

In the modelling phase, we applied decision tree algorithms on the created sets. We divided this phase into four experiments in terms of using the type of feature selection method.

Description of the first experiment: a selection of attributes using the LASSO method and use of decision tree algorithms C4.5, C5.0, random forest, and CART. Were created 32 models in different ratios of training and testing set. Due to unbalanced data in the class attribute, we used the sampling method to train the model (under-sampling, over-sampling) but also training without sampling.

- *The best model*: algorithm C4.5, ratio 80/20, success rate 96,35%, important attributes: Attr27, Attr41, Attr34

		Actual value	
		Non-bankruptcy	bankruptcy
Predicted value	Non-bankruptcy	8187	241
	bankruptcy	76	177

Description of the second experiment: a selection of attributes using the PCA method and use of decision tree algorithms C4.5, C5.0, random forest, and CART. Were created 52 models in different ratios of training and testing set. We are using the sampling method as in the first experiment.

- *The best model*: algorithm C5.0, ratio 70/30, success rate 95,33%, important attributes: Attr35, Attr56, Attr34

		Actual value	
		Non-bankruptcy	bankruptcy
Predicted value	Non-bankruptcy	12 370	589
	bankruptcy	24	43

Description of the third experiment: a selection of attributes using correlation coefficients and use of decision tree algorithms C4.5, C5.0, random forest, and CART. Were created 48 models in different ratios of training and testing set. We are using the sampling method as in the first and the second experiment.

- *The best model*: algorithm C5.0, ratio 90/10, success rate 96,5%, important attributes: Attr27, Attr41, Attr34

		Actual value	
		Non-bankruptcy	bankruptcy
Predicted value	Non-bankruptcy	4 128	149
	bankruptcy	3	60

Description of the fourth experiment: use of all attributes without Attr37 generated by decision tree algorithms C4.5, C5.0, random forest, and CART. Were created 28 models in different ratios of training and testing set. We are using no sampling method.

- *The best model*: algorithm random forest, ratio 90/10, success rate 96,71%, important attributes: Attr27, Attr46, Attr34
-

		Actual value	
		Non-bankruptcy	bankruptcy
Predicted value	Non-bankruptcy	4 123	135
	bankruptcy	8	74

2.6. Evaluation

By evaluating all generated models, we received the following interesting findings:

- In terms of accuracy for class 0 (non-bankruptcy companies) had all experiments values above 95,18%.
- The worst results were in models, in which the missing values were replaced by the k-NN method except for the random forest algorithm (if the sampling method was also used).

- The under-sampling method did not affect then resulting model accuracy.
- In the case of the CART algorithm, it was necessary to use a sampling method for each model. Otherwise, the model was over trained.
- The lowest AUC values were in models generated by C4.5 algorithm (which also had the lowest accuracy) and the CART algorithm.
- The highest AUC values were achieved only in models generated by the random forest algorithm.
- The lowest error or highest success rate was achieved in the model generated by the C5.0 algorithm on the set with replaced missing values by the average of attributes.
- Among the most important attributes in the individual partial results of the models were:
    - **Attr27** (profit on operating activities */ financial expenses),*
    - **Attr34** (*operating expense/total liabilities),*
    - **Attr41** (*total liabilities/[(profit on operating activities + depreciation)*(12*365)]).*
- In terms of attribute selection method, models in the 1st, 2nd and 3rd experiments achieved mostly poor accuracy. In the LASSO method on a set with missing values replaced by the average of an attribute for model C4.5, the accuracy was very good compared to the set with missing values replaced with the k-NN method. It was also because the number of attributes was different in both cases. In the model generated by the C5.0 algorithm, in which we selected the attributes using correlations on the set with missing values replaced by the average of attributes, the accuracy results were excellent.
- Models generated on all input attributes in the 4th experiment achieved the best results. This means that good results we also achieved on models without using attribute selection method.

For the best model, concerning all evaluation metrics, we determined the decision tree model C5.0 generated on the set, where the missing values were replaced by the average of the given attributes, and the training and testing set ratio was 90:10. All available attributes were used as input attributes, so without using attribute selection methods. Of the 4,340 companies, 4213 were correctly classified, which means that this model would be able to classify the future state of the company to 97.1%. Basic metrics values:

- **Successful**: 97,07%
- **Classification error**: 2,93%
- **AUC value**: 0,815
- **Sensitivity**: 0,9966
- **Specificity**: 0,4593
- **Accuracy for class 0**: 97,33%
- **Accuracy for class 1**: 87,27%
- **Important attributes**: Attr27, Attr34, Atrr41

We also generated decision rules from this model, such as:

IF (Attr27) *profit on operating activities / financial expense* > 1 096.9 **AND** (Attr34) *operating expenses/total liabilities* <= 0.581 **AND** (Attr56) *(sales – cost of products sold)/sales <= 0.2197* **AND** (Attr9) *sales/total assets* > 0.716 **AND** (Attr9) *sales/ total assets* <= 1.117, **THEN company is in bankruptcy.**

IF (Attr27) *profit on operating activities / financial expense* <= 1 096.9 **AND** (Attr41) *total liabilities/[(profit on operating activities + depreciation)*(12*365)]* <= -0.006 **AND** (Attr58) *total costs/total sales* > 0.975 **AND** (Attr34) *operating expenses/total liabilities* <= 0.011, **THEN company doesn´t in bankruptcy.**

3. CONCLUSION

The prediction of company bankruptcies was one of the main roles of this work to support the decision-making of companies and banks, in solving potential bankruptcy in the future based on the company's financial indicators. Using modern knowledge discovery techniques, we used the financial indicator data of Polish companies to build prediction models to get the best results. The CRISP-DM methodology and its phases were the basis for the efficient handling of all tasks related to the topic of work and its results. Analysis of the current situation has helped us to gain expertise in the field of companies' bankruptcy classification and to point out different views to solve the problem. We compared the results of some case studies with the results of our work.

The results of the individual experiments also partly depended on the data pre-processing. The experiments performed in this work had been different from the use of many data preparing techniques. One of the forms of evaluation of acquired models was also the identification of key attributes (financial conditions) based on which it was the highest possibility to predict bankruptcy.

We have also used several methods for feature selection to identify appropriate modelling attributes (PCA method, LASSO, and correlations) that determined these financial ratios as most important:

- **(Attr24):** *gross profit (in 3 years)/total assets,*
- **(Attr27):** *profit on operating activities / financial expenses,*
- **(Attr34):** *operating expenses / total liabilities,*
- **(Attr41):** *total liabilities/[(profit on operating activities + depreciation)*(12*365)].*

We could compare our analysis with the last case study, which worked with the same data sample. In the data preparation phase, the authors also replaced the missing values with the average of the value of the attribute. In the modelling phase, they also used decision tree algorithm but only on the selected attributes; the random forest algorithm they used on all input attributes as we did. In our experiment, we achieved an accuracy of 14.5% higher. In the mentioned study, accuracy in class 1 was 75.76%, and our result reached 90.26%. Accuracy in class 0 from our

results was 96.73%, and it was about 1.74% lower as in the study. An excellent result in our experiment was a sensitivity metric (TPR) of 99.82%, compared to a lower value of 99.45% in the comparison study.

ACKNOWLEDGMENTS

Research described in the paper was supported by the FEI-2018-52 project funded by the Faculty of Electrical Engineering and Information Technology of the Technical University of Košice, the Agency for Research and Development Support under Contract No. APVV-16-0213 and Scientific Grant Agency MŠVVaŠ SR and SAV, project No.1/0493/16.

REFERENCES

[1] IVASHINA, V. – SCHARFSTEIN, D.: Bank lending during the financial crisis of 2008, Journal of Financial Economics. Volume 97, Issue 3, September 2010, pp. 319-338.

[2] RUDIGER, W. – Jochen, H.: CRISP-DM: Towards a Standard Process Model for Data Mining.

[3] FITZPATRICK, P.: A comparison of ratios of successful industrial enterprises with those of failed companies. The Certified Public Accountant.1932, s.598-731.

[4] SMITH, R. – WINAKOR. A.: Changes in Financial Structure of Unsuccessful Industrial Corporations. Bureau of Business Research, Bulletin No. 51. Urbana: University of Illinois Press.1935.

[5] MERWIN, C.: Financing small corporations in five manufacturing industries. New York: National Bureau of Economic Research.1942, s. 1926-19336.

[6] BEAVER, W.: Financial ratios as predictors of failure. Journal of Accounting Research. 1966, s. 71-111.

[7] FAN, S. – LIU, G. – CHEN, Z.: Anomaly detection methods for bankruptcy prediction. *2017 4th International Conference on Systems and Informatics (ICSAI)*. 2017, s. 1456 – 1460.

[8] NAGARAJ, K. – SRIDHAR, A.: A predictive system for detection of bankruptcy using machine learning techniques. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*. 2015, Vol. 5, No. 1.

[9] ZIEBA, M. – TOMCZAK, S. – TOMCZAK, J. M.: Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems with Applications. 2016, Vol. 58.

[10] MOREIRA, J. – CARVALHO, A. – HORVATH, T.: A general introduction to data analytics. 1. vydanie. 2018. 352 s.

[11] PARALIČ, J.: Objavovanie znalostí v databázach. 1. vydanie. Košice: Elfa, 2003. 150 s.

[12] Machine Learning Plus, Feature Selection – Ten Effective Techniques with Examples https://www.machinelearningplus.com/machine-learning/feature-selection/

[13] CONSTANTIN, C.: Principal component analysis - A powerful tool in computing marketing information, http://webbut.unitbv.ro/BU2014/Series%20V/BULETIN%20V/I-03_CONSTANTIN%20C.pdf

[14] FONTI, V.: Feature selection using LASSO, https://beta.vu.nl/nl/Images/werkstuk-fonti_tcm235-836234.pdf

[15] UCI: Machine Learning Repository. Center for Machine Learning and Intelligent Systems. Polish company's bankruptcy data.

Received May 24, 2019, accepted June 12, 2019

BIOGRAPHIES

Anna Biceková was born on 23.05.1987. In 2014 she graduated (PhD.) at the Faculty of economics, the doctoral study program of Finance at the Technical University of Kosice. Currently, she works as an assistant professor at the Department of Cybernetics and Artificial Intelligence of the Faculty of Electrical Engineering and Informatics at the Technical University of Košice. Her scientific research focused on business Economics, financing of regional self-government, fiscal decentralization, data mining, and data analysis.

Ľudmila Pusztová was born on 16. 02. 1993. In 2017 she graduated (MSc) at the Department of Cybernetics and Artificial Intelligence of the Faculty of Electrical Engineering and Informatics at Technical University in Košice. Since September 2017 she works as PhD. student at the Department of Cybernetics and Artificial Intelligence. The dissertation for PhD study is focused on models and methods of data analysis for the creation of knowledge models from data sources. Currently, she is working on resolve the most critical problem in the case-based reasoning method - adaptation, which is often done manually by the experts in the relevant field.