DE GRUYTER OPEN

ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2018, Vol. 22, No. 2
ISSN 1507-3866; e-ISSN 2449-9994

PREDICTING THE DEFAULT RISK OF COMPANIES. COMPARISON OF CREDIT SCORING MODELS: LOGIT VS SUPPORT VECTOR MACHINES

Natalia Nehrebecka
National Bank of Poland, University of Warsaw, Warsaw, Poland
e-mails: natalia.nehrebecka@nbp.pl; nnehrebecka@wne.uw.edu.pl

DOI: 10.15611/eada.2018.2.05
JEL Classification: C13, G33, C45

Abstract: The aim of the article is to compare models on a train and validation sample, which will be created using logistic regression and Support Vector Machine (SVM) and will be used to assess the credit risk of non-financial enterprises. When creating models, the variables will be subjected to the transformation of the Weight of Evidence (*WoE*), the number of potential predictions will be reduced based on the Information Value (*IV*) statistics. The quality of the models will be assessed according to the most popular criteria such as GINI statistics, Kolmogorov-Smirnov (K-S) and Area Under Receiver Operating Characteristic (AUROC). Based on the results, it was found that there are significant differences between the logistic regression model of discriminatory character and the SVM for the model sample. In the case of a validation sample, logistic regression has the best prognostic capability. These analyses can be used to reduce the risk of negative effects on the financial sector.

Keywords: Basel III, Internal Rating Based System, credit scoring, Support Vector Machines, logistic regression.

1. Introduction and literature review

Credit scoring is a method of assessing a customer's reliability and their capacity to re-pay a liability. At present, given the high number of loan applications and the requirements of banking supervisors, it is a tool every bank needs to utilise. The creation of scoring models is also important because it aims at the effective separation of 'good' customers from those who are behind with repayment of their liabilities, and granting a loan to them might entail financial losses.

Assessment of Credit Risk, and in particular, ensuring the accuracy and reliability of credit ratings by means of validation is of critical importance to many different market participants. The definition of "Credit Risk": traditional (risk of loss due to a debtor's non-payment of a loan (default)); mark-to-market definition (risk of losses

due to a rating-downgrade (i.e. an increased probability of default) or the default of a debtor) [Rossi, Schwaiger, Winkler 2009]. The Basel Committee explains a default event on a debt obligation in the two following ways:

- it is unlikely that the obligor will be able to repay its debt to the bank without giving up any pledged collateral;
- the obligor is more than 90 days past due on a material credit obligation.

The objective of this article is to compare the models developed using logistic regression and Support Vector Machine (SVM) for *WoE* variables on development (training) and the validation sample for the credit risk assessment of nonfinancial companies. In the literature review only Haltuf [2014] compares SVM for *WoE* variables results with the linear kernel after the conversion of variables were very close to logistic regression, but this author draws conclusions based on a training sample. Such a decision might be biased. This feature may have a substantial impact on the results obtained and the conclusions reached on their basis. To avoid this error, in this article SVM results with Gaussian, Laplace kernel are compared with logistic regression for *WoE* variables not only on the training sample but also this method was applied on the validation sample. The subject matter of this article is important and pertinent as there is no consensus made among practitioners regarding the selection of the methods and ways of testing them. The comparison of methods constitutes a significant added value.

The known models for potential insolvency (Altman 1968 on the USA, Keasey and McGuinness 1990 on the UK, Charitou 2004 on Greece, Sheppard 1994 on Canada) were prepared for different macroeconomic conditions, and their use in the Polish environment would be ineffective. The following analysis expands the existing knowledge on the processes associated with the credit risk of Polish companies – can be treated as the first benchmark model. This research is an original concept and high added value as it was performed using data based on Prudential Reporting for over 30,000 nonfinancial companies every year.

According to the author's best knowledge, some variables in the models will be used for the first time with regard to rating systems (e.g. bank-firms relationship, credit period, open credit lines). The results of the empirical research on the financial crises in Asian countries, in Latin America, as well as those regarding the recent financial crisis, indicate that relational financing may allow for better access to finance even when the company has financial problems [Abildgren, Buchholst, Staghøj 2011]. It can be assumed that the stability of the Polish banking sector during the crisis (in 2008-2009, banks in Poland recorded relatively high financial results, bank failures were not observed) the dominant manner of cooperation between banks and enterprises in the form of relationship banking had some influence. This also confirms the significant value of this article.

The ability of the better adaptation of a model to the current situation and the bank's strategy is the core motivation. During the development of the models the variables shall be subject to transformation by Weight of Evidence (*WoE*), the

number of potential predictors shall be reduced based on the statistics of Information Value (IV), and the parameters shall be evaluated using logistic regression and SVM. The quality of such models shall be assessed according to the most popular criteria such as GINI statistics, Kolmogorov-Smirnov (K-S) and Area Under Receiver Operating Characteristic (AUROC). Logistic regression was chosen on account of its well-established position in the scoring literature and repeatedly recognised effectiveness [Anderson 1999], whereas Support Vector Machine was chosen based on the fact that it is based on a classification in which the self-learning algorithm aims at determining the separating hyperplane with a maximum observation margin, belonging in two classes. Many authors indicate that using the Support Vector Machine methodology provides better results compared to logistic regression. Schebesch and Stecking [2005] indicate that SVM provides slightly better results than logistic regression based on data related to loan applications (taking the aforementioned decision does not take into account statistical significance). Belotti and Crook [2009] compared logistic regression and SVM results with those based on data concerning 25,000 credit card users with linear kernel and SVM with nonlinear kernels: polynomial and Gaussian kernel. SVM plus nonlinear kernels provide the worst results, whereas SVM with a linear kernel gives results slightly better than logistic regression. Ghodselahi [2011] proved that SVM with a linear kernel provides better results than logistic regression. Moreover, a comparison with Accuracy Ration by Härdle et al. [2007 and 2009] showed similar conclusions twice in the analysis of 500,000 financial statements of German companies as well as by Lacerda and Moro [2008] by forecasting the bankruptcy of 340,000 Portugese companies.

Most of these works are based on raw data. However, many authors mention that the transformation of raw data using Weight of Evidence (*WoE*) provides more accurate results of Credit Scoring [Sharma 2011; Anderson 2015]. Weight of evidence facilitates the interpretation of results, allows for information shortages modelling and has a greater number of degrees of freedom compared to raw data. The question could thus be raised whether the SVM method provides more accurate results than logistic regression also after applying *WoE* to the variables. It is worth mentioning that Haltuf [2014], containing 6,818 units compared results based on logistic regression and SVM for *WoE* variables based on data (bondora.com). SVM results with linear kernel after conversion of variables were very close to logistic regression. However, SVM results with Gaussian kernel are significantly more effective than logistic regression for *WoE* variables. The author draws conclusions based on a training sample; such a decision might be biased.

The percentage of correct classifications obtained with the use of various methods most often does not differ significantly within one study. This was explained by Lovie and Lovie [1986] as the flat maximum effect, which means that results close to optimal can be achieved in multiple ways with the use of various combinations of variables or parameter estimations. For that reason most methods are able to come close to the optimum solution, but further significant improvements in the model's

efficiency can be achieved by improving the quality of the available data rather than by changing methodology. Therefore it is crucial while selecting the research method to consider all the good and bad points and to choose the method that is most suited to the issue at hand.

The conclusions presented in this article are mainly directed to banking sector employees concerned with identifying the best way to calibrate internal credit risk systems. A detailed presentation and comparison of the different methods allows for a comparison of particular approaches and the selection of the best one. The part of this work that concerns the testing process may be a valuable source of information for validators of risk models.

2. Review of Polish bankruptcy models

The subject matter of bankruptcy of businesses has only been studied by researchers since the 1990s. As Polish companies gained their first experiences related to bankruptcies and the associated problems, the interest of researchers in the topic of bankruptcy and their intent to explain and forecast it has increased. The first Polish model using multidimensional analysis was developed by Mączyńska in 1994.

Another early example of the use of discrimination analysis in the context of the bankruptcy of firms is the study by Pogodzińska and Sojak [1995]. The analysis was conducted on a sample of ten businesses from the then Wrocławskie Province "in whose case there were suspicions that they would go bankrupt in 1993". Six of those businesses did, in fact, go bankrupt, while four continued their operations. Four of the analysed businesses conducted their activities in the industrial sector and two conducted their activities in each of the following sectors: construction, agriculture and retail.

The first important study of the bankruptcy of Polish businesses involved the models prepared by Gajdka and Stos in 1996. The authors estimated a total of four models: the first two based on 40 businesses from different sectors, exactly half of which went bankrupt. The next two models had the same structure, whereby the entities were traded on a stock exchange and conducted activities in the industrial, construction, and retail sectors. The analysis conducted by the authors focused on twenty predefined financial indicators calculated one year prior to their bankruptcy in 1994-1995.

In 1998 Hadasik (Appenzeller) presented very intersting results of her study published in her habilitation dissertation. She analysed a set of models based on companies that in the years 1991-1997, together with their financial reports, filed petitions for bankruptcy with the provincial court in Poznań, Piła, and Leszno, as well as based on companies that continued their operations, which were selected based on their similarity with regard to ownership structure and size. Due to the fact that the financial data of some companies was incomplete the author decided to use a step discrimination analysis on different samples.

Another model based on a sample of Polish businesses is the model prepared by Wierzba in 2000. In his study the author used financial data of 24 businesses that did not face the risk of bankruptcy and of the same number of businesses that were declared bankrupt or initiated an arrangement procedure in the period starting in January 1995 and ending in April 1998. The limit point below which a business is considered to be facing the risk of bankruptcy was determined to be 0.

Another example of the use of discrimination analysis in the area of prediction of the bankruptcy of Polish businesses in 2001 was presented by Hołda. He built his model based on an analysis of 80 businesses conducting operations in sectors classified according to the statistical classification of economic activities in the European Community (no 45 to 74), exactly half of which were businesses that had been declared bankrupt. The time interval of the model is 1993-1996.

In 2003, Gajdka and Stos published the results of their further studies of a bankruptcy prediction model. They worked on a sample of 34 businesses, 17 of which were defined as bankrupt. All the "sound" businesses were quoted on the Stock Exchange for at least three more years. In refining the classification criterion the authors defined bankruptcy as a situation where the liquidation process was initiated due to a bad economic situation, reaching of a court settlement with creditors, or the declaration of a settlement with a bank. The businesses conducted operations in different sectors including light industry, retail, services, and transport. The researchers used twenty financial indicators that were calculated based on the financial statements prepared one year before bankruptcy was declared, which in this case was 1994.

The result of the continuation of the work by Appenzeller (Hadasik) was an article published in 2004 together with Szarzec. Their study was conducted on a sample of 34 publically traded companies facing the risk of bankruptcy and of the same number of similar companies of a good financial status. The risk of bankruptcy was identified based on the filing of at least one petition bankruptcy in a court or the initiation of an arrangement procedure in the period 2000-2003, regardless of their legal consequences.

A frequently cited example of a bankruptcy prediction model is the "Poznań" model, which is the result of the paper by Hamrol, Czajka, and Piechocki published in 2004. The model was developed based on a sample of 100 Polish businesses, half of which faced the risk of bankruptcy.

Prusak conducted a study of the bankruptcy of businesses that used a linear discrimination function. He collected a sample of 40 bankrupt manufacturing companies and the same number of companies that continued their operations. The financial data was taken from the financial statements published one year and two years prior to the bankruptcy (1998-2002).

In his further studies, Prusak developed two more models based on a combined test and validation sample for the first pair of models, which were then subject to

selection as a result of which 140 small and medium-sized businesses were identified, exactly half of which went bankrupt.

One of the most valued and the most relevant examples of studies of the prediction of the bankruptcy of companies is the study by Mączyńska and Zawadzki [2006] conducted at the Institute of Economic Sciences of the Polish Academy of Sciences. The authors selected 40 entities facing the risk of bankruptcy and 40 entities not facing such a risk in the period of 1997-2002 from a sample of the 500 largest companies quoted at the Warsaw Stock Exchange. Out of a group of 45 financial indicators, the number of variables was gradually reduced.

Nehrebecka [2011] proposes an analysis of the changes in the structure of Polish companies with the use of a method based on the Markov chain which would enable forecasts of the composition of the company sector, as well as the average time a company has left until going bankrupt. The findings indicate that across all sectors the longest lifespan is associated with non-specialized exporters. The larger the company, the longer the average age and the average time left until a market exit. This implies that it is important to account for the differences between the subjects studied (active or inactive companies) and thus the changing definitions of the dependent variable.

The empirical studies described in the literature review above are certainly not the complete list of attempts to explain the reasons for the bankruptcy of companies in Poland. Due to the topic of the article, only those models that use one-formula discrimination analysis are described, although the model of concentrations by Sojak and Stawicki [2001], the logit models by Gruszczyński [2003], and the neuron networks models by Korol and Prusak [2005] should also be mentioned. A review of Polish bankruptcy models was presented by Prusak [2005].

3. Data description

The empirical analysis was based on the individual data from different sources (from 2007 to 2016), which are:

Data on banking defaults are drawn from Prudential Reporting (NB300) managed by Narodowy Bank Polski. The Act of the Board of the Narodowy Bank Polski no 53/2011 dated 22 September 2011 concerning the procedure and detailed principles of handing over by banks to the Narodowy Bank Polski, data indispensable for monetary policy, for periodical evaluation of monetary policy, evaluation of the financial situation of banks and the bank sector's risks.

Data on insolvencies/bankruptcies come from a database managed by The National Court Register (KRS), that is the national network of Business Official Register.

Financial statement data sources are AMADEUS (Bureau van Dijk); Notoria OnLine. Amadeus is a database of comparable financial and business information on Europe's biggest 510,000 public and private companies by assets. Amadeus includes

standardized annual accounts (consolidated and unconsolidated), financial ratios, sectoral activities and ownership data. A standard Amadeus company report includes 25 balance sheet items; 26 profit-and-loss account items; 26 ratios. Notoria OnLine standardized the format of financial statements for all companies listed on the Stock Exchange in Warsaw.

The following sectors were removed from the Polish Classification of Activities 2007 sample: section A (agriculture, forestry and fishing), K (financial and insurance activities). For the definition of the total number of obligors the following selection criteria were used:

- the company has been in existence (operating and not liquidated/in liquidation) throughout the entire respective year,
- the company is not in default at the beginning of the year,
- the total exposure reported at least 2 million PLN for each reporting date.

Claims include the following balance-sheet positions: loans and other receivables, debt and equity instruments and remaining receivables. Total exposures – for a bank that is a joint-stock company, state-run bank and a non-associated cooperative bank – mean exposures towards one enterprise in excess of 500,000 PLN.

3.1. Default definition

Article 178 of Regulation (EU) No 575/2013 (Capital Requirements Regulation – CRR) specifies the definition of a default of an obligor that is used for the purpose of the IRB Approach. A default shall be considered to have occurred with regard to a particular obligor when either or both of the following have taken place:

(a) the institution considers that the obligor is unlikely to pay its credit obligations to the institution, the parent undertaking or any of its subsidiaries in full, without recourse by the institution to actions such as realising security;

(b) the obligor is past due more than 90 days on any material credit obligation to the institution, the parent undertaking or any of its subsidiaries. Relevant authorities may replace the 90 days with 180 days for exposures secured by residential or SME commercial real estate in the retail exposure class (as well as exposures to public sector entities). The 180 days shall not apply for the purposes of Article 127.

The dataset, after its initial preparation and while keeping only the observations on which the model can be based, contained 14,191 records for 2016. However the number of observations marked as default was 504 (Table 1). While creating a sample to establish and validate the model, the results of Crone and Finlay's (2012) analysis were taken into account. The proposal for replicating default observations and adding them to all non-default observations was rejected due to the excessive size of the dataset that would have been created as a consequence. The added value arising from the increased number of observations would be insignificant in practical terms, however, extending the calculation time would be significant. For that reason it was decided that all default observations will be added to the non-default firms

which were then randomly reduced to satisfy the condition that defaults accounts for about 20% of the sample. The proportions were established at 20:80 for several reasons. A smaller number of non-default observations drawn would cause difficulty with drawing a proportional sample, whereas a higher number would extend the calculation time while improving the quality of the model only slightly.

Table 1. General statistics

Year	Number of obligors	Thereof Insolvent	There of defaulted with at least one bank	Insolvency rate	Default rate
2007	8 164	54	307	0..66	3.76
2008	9 938	18	507	0.18	5.10
2009	11 494	68	918	0.59	7.99
2010	10 824	37	635	0.34	5.87
2011	11 286	46	619	0.41	5.48
2012	12 302	111	731	0.90	5.94
2013	12 450	80	681	0.64	5.47
2014	12 376	64	624	0.52	5.04
2015	13 091	47	501	0.36	3.83
2016	14 191	31	504	0.22	3.55

* The column total number of obligors shows the number of obligors not in default on 1st of January every year. The column thereof insolvent shows the absolute number of obligors where an insolvency was observed during the year. The column thereof defaulted with at least one bank shows the absolute number of obligors where a default was reported to the credit register for reporting dates within the year. The column insolvency rate shows the relative share of insolvent obligors. The column default rate shows the relative share of obligors where at least one default was reported by on bank in the Prudential Reporting.

Source: own calculation.

Due to the way the data sample was constructed it was tested whether selected non-default firms are proportional for the whole population. For continuous variables, two nonparametric tests, the Wilcoxon-Mann-Whitney test and the Kolmogorov-Smirnov test, were applied to check whether samples are drawn from the same population. Null hypothesis – the samples come from the same population. Also t-test for difference between the sample mean and the population mean was performed. For categorical variables a Pearson test and Population Stability Index (PSI) were used. The PSI coefficient is applied in order to investigate the differences in distribution of the two categorized variables. The higher the value of the coefficient, the greater the statistical distance between the distributions.

The training and validation samples were divided at the ratio of 70:30. This proportion was chosen as an average value between the most popular divisions found in literature, ranging from 60:40 to 80:20.

Based on the literature, the potential default predictors were chosen with the focus on financial indicators. Signals for a deteriorating financial condition of the company are: negative dynamics for revenue, assets and equity, decreasing profits, negative equity, in-creasing indebtedness, problems with financial liquidity, deterio-rating operating efficiency and decreasing investment in tangible assets. Explanatory variables that characterize the company's financial state were constructed such as: turnover dynamics, asset dynamics, equity dynamics, profitability, indebtedness, liquidity and operating efficiency. The analysis included not only the current values of the indicators but also their statistical properties (for example the median) based on different time frames (for example a two-year average).

4. Methodology

In order to construct an indicator which would enable assessing the probability of a company to default, a logistic regression was used. Due to the high number of financial indicators of a company's condition (explanatory variables) in the initial analysis the predicting force of each was determined (Gini coefficient, Information Value Indicator) followed by clustering in order to limit the size of the analysis. Thanks to this variable selection procedure it was possible to avoid the collinearity problem, which was assured by calculating the appropriate Variance Inflation Factor statistics. The model was estimated on categorized variables transformed using the weight of evidence (*WoE*) approach. The *WoE* transformation is often used for the creation of scoring models using logistic regression because such a transformation allows maintaining linear dependence in regard to the logistic function. In addition, *WoE* conveys information on the relative risk associated with each category of the particular variable, with a large negative value indicating a higher risk of default.

$$WoE_i = ln\left(\frac{p_i^{non-default}}{p_i^{default}}\right),$$

where: i – category,

$p_i^{non\text{-}defaults}$ – the percentage of non-default companies that belong to category i,

$p_i^{non\text{-}defaults}$ – the percentage of default companies that belong to category i.

The categorisation was based on the division with the highest information value (*IV*), which measures the statistical Kullback-Leibler distance (H) between the defaults and non-defaults. The *IV* statistic, based on the *WoE*, allows measuring the predicting force of a particular characteristic. The *IV* value depends on the number of categories and division points. The variables for which the *IV* does not exceed 0.1 are assumed to be weak in their relative predicting force, while values exceeding 0.3 bear evidence of a strong discriminating force [Anderson 2007].

$$IV = H\left(q^{non-defaults}||q^{defaults}\right) + H\left(q^{defaults}||q^{non-defaults}\right) = \sum_i \left(p_i^{non-defaults} - p_i^{defaults}\right) WoE_i \quad ,$$

where: q – density function.

The final model was created following the top-down approach. Based on the estimated parameters, weights for particular explanatory variables were determined. As a result, a set of financial indicators allowing to grade companies was obtained and default probabilities were assigned to companies.

It was decided to use two methods for the research – logistic regression (Model 1) and the SVM method (Model 2). The first of those was chosen due to its great popularity not only in academic research but also amongst employees of financial institutions. Model 1, assuming that the entire process is conducted correctly, might constitute a reference for results of Model 2. The SVM method is a very useful tool if there are certain inconsistencies in the data such as irregular data distribution. The aforementioned technique is successfully used when the relation between the score (probability of default) and the variables is not linear. Even if the validation sample involves a selection error, the results of the Support Vector Machine method should be resistant due to the choice of the appropriate parameters *C* (capacity) and *r* (radius) [Härdle et al. 2008]. Moreover, SVM ensures a single unambiguous solution, which makes it stand out among neuron networks that depend on local minimums [Auria, Moro 2008].

The optimum distribution is achieved by developing a learning algorithm minimizing the error function:

$$\frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \, .$$

The limitation might also be presented in the following way:

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0 \, ,$$

where: C is the cost parameter (capacity), w is a coefficient vector, b constant, a ξ_i are a slack variable, and y belong in the set of {–1,+1}, kernel function ϕ transforms inputs to a new plane, x_i are independent variables.

The kernel function turns inputs to a new space of characteristics. Ultimately, the problem comes down to the Lagrange multiplier:

$$L_P = \frac{1}{2} ||w||^2 + \sum_{i=1}^{n} C_i \xi_i - \sum_{i=1}^{n} \alpha_i \{y_i(x_i w + b) - 1 + \xi_i\} - \sum_{i=1}^{n} \mu_i \xi_i \, ,$$

where: $\alpha_i \geq 0$, $\mu_i \geq 0$ – multipliers.

It is worth mentioning that C strongly affects error and this value should be carefully selected given the risk that the model might be overly adapted. The lesser C is, the greater the significance of incorrectly selected values and, as a consequence, the lesser the risk of overestimation for the model.

The general solution of the aforementioned equation is:

$$w(\alpha)^T w(\alpha) = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j {x_i}^T x_j.$$

After replacing the scalar part with the kernel the solution takes the following form:

$$w(\alpha)^T w(\alpha) = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j),$$

where: $K(x_i, x_j)$ – kernel function.

Application of the kernel function is a convenient way to map low data dimensions into higher ones so as to increase the quality of assigning observation to individual groups. The function of the kernel should comply with the Mercer condition, i.e. it should have some scalar characteristics in Hilbert space [Mercer 1990].

A kernel with an anisotropic radial basis (the Gaussian kernel) is one of the most important kernels used in the support vector method:

$$K(x_i, x_j) = \exp\{-\sigma||x_i - x_j||^2\},$$

where σ is the model's parameter (the shape parameter). Another function used in the work is Support Vector Machine with Laplace kernel:

$$K(x_i, x_j) = \exp\{-\sigma|x_i - x_j|\}.$$

5. Results

The research was performed on the sample included companies observed in 2016. In the model the default probability was predicted for a one year horizon.

The research was performed on four models. The first model was estimated using logistic regression (Model I). The greatest weight was assigned to the ROA indicator (16.39%). Profitability ratio is a basic measure that shows the rate of return of assets and equity as well as the effectiveness of the company's operations. It is used in strategic planning and its advantage is the binding effects to inputs. According to Vivet, more profitable companies are also those with a better financial standing. Great weight was assigned to the indicator of Interest due/Total exposure (14.05%). The indicator is associated with solvency. According to a study by Vivet, the higher the interest rate, the higher the debt level. This has a negative impact on the company's profits, which also negatively affects the entire financial situation of the company. The indicator for the borrower-lender relationship (variable – bank-firm relationships) (weight of 11.80%) is the third important variable. The relations between banks and firms in Poland, are identified here with bank-firm relationship. The results of the empirical analysis have shown that Polish companies are eager to build relationships with one bank.

Table 2. Final scorecard – logistic regression

Variables	Weight in the total grade in %	Value		Partial grade
Credit period (Creditors/Operating revenue)*360	6.16%	-INF	36.175	57
		36,175	73,873	34
		73.873	+INF	0
Industry sectors	8.84%	Industry		83
		Construction		0
		Trade		108
		Transport		31
		Other services		43
EBIT (Earnings Before Interest and Tax)	8.04%	-INF	372	0
		372	4696	44
		4696	+INF	78
Bank-firm relationships	11.80%	one bank		99
		two or more banks		0
ROCE (Net Operation Profit / Employed Capital)*100	6.87%	-INF	-4.501	0
		-4.501	12.641	59
		12.641	+INF	85
ROA (Net Income/Total Assets)*100	16.39%	-INF	-10.49	0
		-10.49	1.907	53
		1.907	6.502	91
		6.502	+INF	189
Solvency ratio [(Net Income +Depreciation)/(Short-Term Liabilities + Long-Term Liabilities)]*100	7.96%	-INF	26.221	0
		26.221	54.097	30
		54.097	94.483	50
		94.483	+INF	80
(Interest due/Total exposure)*100 (median of 4 q)	14.05%	-INF	0.016	121
		0.016	0.035	89
		0.035	0.193	44
		0.193	+INF	0
(Bank loans denominated in PLN / Total exposure)*100 (median of 6 q)	9.33%	-INF	6.796	84
		6.796	67.72	44
		67.72	+INF	0
(Open credit lines/Total exposure)*100 (median of 6 q)	10.53%	-INF	1.553	0
		1.553	23.77	25
		23.77	+INF	99
Hosmer-Lemeshow Test			Test statistic	*p*-value
			11,1666	0,1924

Source: own calculation.

The indicator for open credit lines (weight of 10.53%) is the next most important characteristic that can signal potential bankruptcy. It is commonly included in models containing at least one indicator of financial liquidity (e.g. [Görg, Spaliara 2009]).

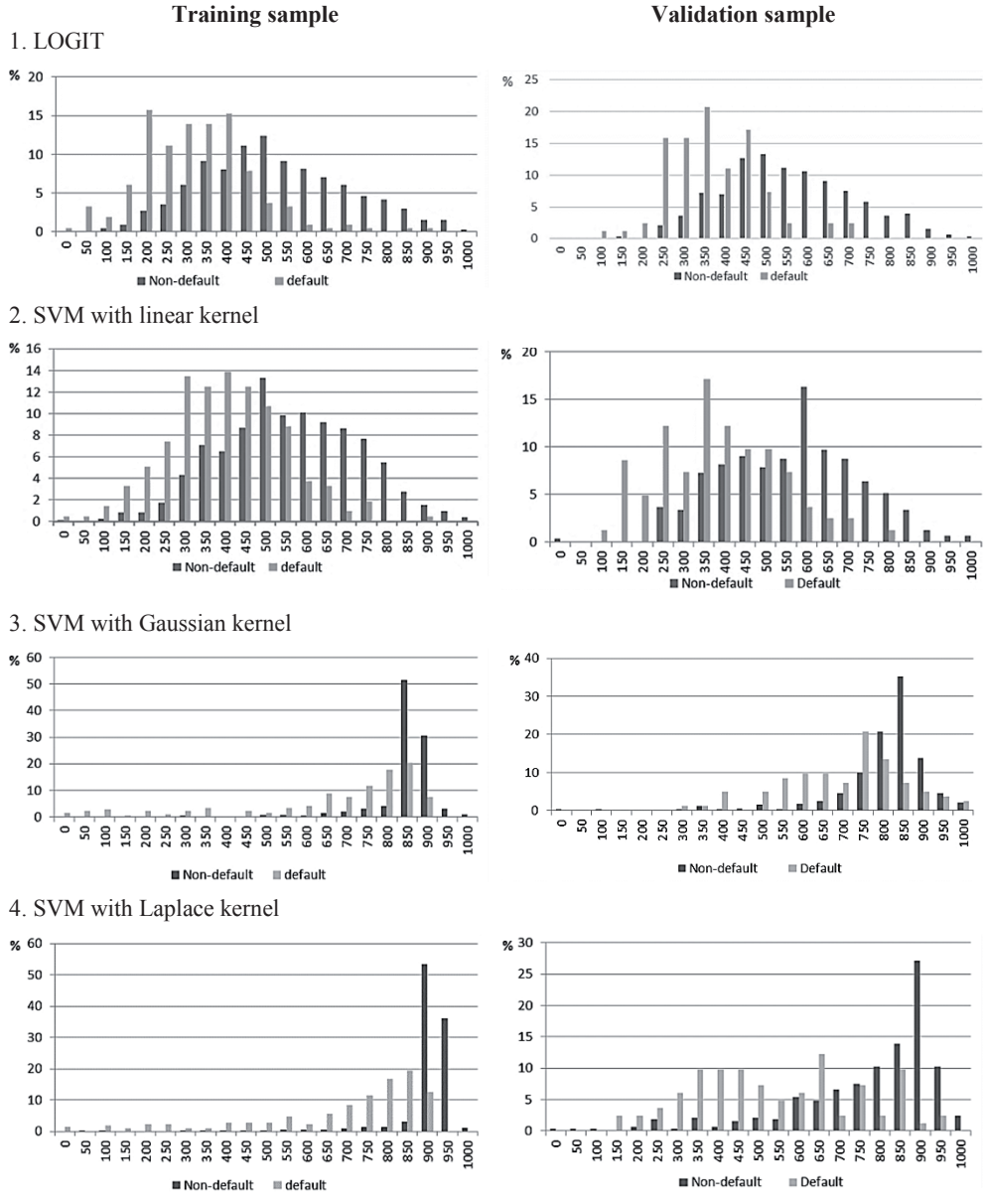

Fig. 1. Distribution of company grades

Source: own calculation.

Logistic regression (Table 2), as the industry standard for the credit scoring, was the implementation of the default based on the Nehrebecka approach (2015), a benchmark model for the purpose of the performance comparison with the Support vector machines. Others models were estimated using SVM: Model II – SVM with Linear kernel, Model III – SVM with Gauss kernel, Model IV – SVM with Laplace kernel.

Based on Model I the distribution of grades across companies which have or have not gone into default shows that the selected indicators (explanatory variables) allow identifying a potential default to a significant degree (Figure 1). The information collected in the database indicated an insignificant risk for companies that obtained more than 600 points which went into default within a year. However for companies with less than 200 points, default was almost certain. Default was predominant among companies from the 200-300 points interval. After that the model was classified using the SVM method for three kernels: Linear, Gaussian and Laplace. The same variables were used as in the case of logistic regression.

Like in the case of logistic regression, when applying SVM with the linear kernel the distribution for the training sample is more regular. Moreover, repetition of the same values is visible for the same score – for instance for 650, 700 for the validation sample. While charts for SVM with the linear kernel were similar for logistic regression, data distribution is different for SVM with the Gaussian kernel. The charts for the training sample and for the validation sample show a more unequal distribution. The distribution after applying SVM for the Laplace kernel resemble the results obtained for SVM with the Gaussian kernel.

Figure 2 presents ROC curves for all the models. In the case of the training sample, SVM with the Laplace kernel has the greatest predictive power, the second greatest being SVM with the Gaussian kernel. The curve for the logistic regression

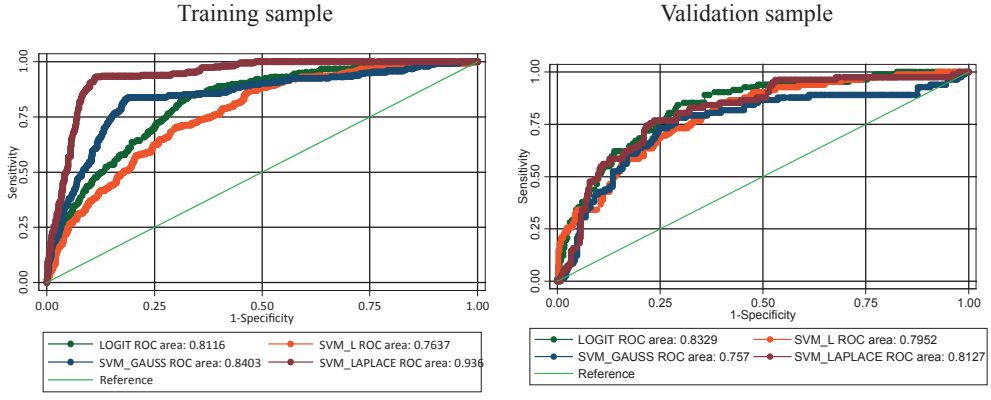

Fig. 2. Distribution of company grades. The comparison of the ROC curves for all model classes

Source: own calculation.

seems to nearly coincide with the SVM curve with the linear kernel. However in the case of the validation sample, logistic regression has the best prognostic capacity. The further places are taken by SVM with the Laplace kernel, SVM with the linear kernel and lastly, SVM with the Gaussian kernel.

Table 3. The statistical tests of the differences in performance among four models

- Training sample

	LOGIT	**SVM** with Linear kernel	**SVM** with **G**aussian kernel	**SVM** with **L**aplace kernel
LOGIT		–0.0479 21.48 0.0000	+0.0287 3.59 **0.0580**	+0.1244 78.80 0.0000
SVM with Linear kernel			+0.0766 21.31 0.0000	+0.1723 117.08 0.0000
SVM with **G**aussian kernel				+0.0957 50.02 0.0000
SVM with **L**aplace kernel				

- Validation sample

	LOGIT	**SVM** with Linear kernel	**SVM** with **G**aussian kernel	**SVM** with **L**aplace kernel
LOGIT		–0.0377	–0.0759	–0.0202
		4.66	6.61	1.80
		0.0309	0.0102	**0.1795**
SVM with Linear kernel			–0.0382	+0.0175
			1.53	0.65
			0.2163	**0.4191**
SVM with **G**aussian kernel				+0.0557
				7.61
				0.0058
SVM with **L**aplace kernel				

Each cell contains three rows: the difference between AUROC of the model in the heading minus AUROC of the model specified on the left-hand side of the table, the respective χ^2 statistics and *p*-value.

Source: own calculation.

The satisfactory quality of the developed models measured using GINI and AUROC value should be higher than 0.6 and 0.8 so as to deem the models as sufficiently good. The differences among the models is summarized in Table 3. For the training sample the AUROC value was highest for SVM with the Laplace kernel and amounted to 0.936, the discriminatory power difference for the models measured with the AUROC statistics (comparison between logistic regression and SVM with the Laplace kernel) was significant and amounted to 12.44% (Table 3). It is worth emphasizing that comparing SVM with the Gaussian kernel and logistic regression with 5% significance level suggesting that there is no significant difference between these two areas.

The entire process was repeated for all the models in order to decrease the risk of adapting the conclusions to the possessed data set. It is also particularly noteworthy that the analysis presented above is vitiated in a certain way as the testing process concerned units of the training sample with which the aforementioned models were tested. The estimates obtained based on a training sample are then verified against the validation sample in order to avoid this error. In the case of the validation sample, the AUROC value was highest for logistic regression and amounted to 0.832, the difference in discriminatory power of the models measured according to the AUROC statistics (when comparing logistic regression and SVM with the Laplace kernel) was insignificant and amounted to 2% (p-value = 0.1795), whereas it was statistically significant in cases of applying SVM method with the Gaussian kernel (p-value = 0.0102) and the linear kernel (p-value = 0.0309) compared to logistic regression (Table 3).

Can the obtained results be deemed as satisfactory for the models? According to Anderson (1999) the minimum acceptable GINI statistics value for the behavioral models is 0.60, whereas the satisfactory result is 0.80. This means that the developed models, based on logistic regression (GINI = 62.3 for the training sample and GINI = 66.5 for the validation sample) and SVM with the Laplace kernel (GINI = 87.2 for the training sample and GINI = 62.5 for the validation sample) are slightly above the minimum value recommended in the literature (Table 4). The question can be asked whether the effectiveness of the models based on available data might be increased? To a small extent, probably yes, but a significant improvement is impossible without a change to the available amount of information. Improvement by way of taking into account additional variables is going to be difficult as all of the available characteristics have already been used.

Another method of measuring the discriminatory power is the Kolmogorov-Smirnov statistics (K-S). This is calculated by measuring the greatest distance between the distribution of increasing non-default customers and a similar distribution of default customers among all the possible values of the score. The result of the statistics calculation method is that the analysis of its value cannot be detached from the broader context. For instance, a model with relatively high K-S values might turn out to be useless for distinguishing customers with a low number of scoring points.

Table 4. Summary statistics of the differences in performance among four models

	LOGIT		SVM with Linear kernel		SVM with Gaussian kernel		SVM with Laplace kernel	
	Training sample	**Validation sample**	Train sample	**Validation sample**	Train sample	**Validation sample**	Train sample	**Validation sample**
GINI	62.3	**66.5**	52.7	**59.0**	68.1	**51.4**	87.2	**62.5**
K-S	51.4	**54.9**	40.3	**45.6**	65.0	**49.1**	81.8	**53.3**
AUROC	0.8116	**0.8329**	0.7637	**0.7952**	0.8403	**0.757**	**0.936**	**0.8127**

Source: own calculation.

For this reason it is good to analyse the entire chart with two distributions so as to assess the effect of the model depending on a chosen cut-off point. The literature states that the statistics value at a level below 20 means a model with a low predictive power, whereas a result higher than 70 is too high and probably means that the model is overly adapted to data, or errors in calculations. In the case of training sample K-S = 51.4 for logistic regression and K-S=81.8 for SVM with the Laplace kernel. In case of validation sample K-S=54.9 for logistic regression and K-S = 53.3 for SVM with the Laplace kernel (Table 4).

6. Conclusion

The objective of this work was to compare credit risk assessment models for non-financial companies using logistic regression and SVM (Support Vector Machine). It turned out that the obtained results were different for the training sample and for the validation sample. In the training sample, statistically significant and more accurate results were obtained using the SVM method with the Laplace kernel. Additionally, all the diagnostic statistics were higher for SVM with the Laplace kernel. It should be emphasized that in the aforementioned model, the K-S statistics reached a result of above 70. According to the specialist literature, such a result is deemed as too high and probably means that the model is overly adapted to the data or that there are errors in calculations. Higher accuracy was obtained for logistic regression with respect to the training sample. All the statistics were significantly higher for logistic regression.

The developed model could be used, for instance, in the process of calculating capital requirements if the conditions imposed by the regulator are fulfilled. Jankwitsch, Pichler and Schwaiger [2007] proved that this use of the model might bring about actual financial gains to a bank. When developing scoring models one needs to realize that a 100% correct classification is impossible to achieve. However, the benefits of using them have been proven many times. Scoring has become a necessary tool for the functioning of a bank and it seems that its significance is going to grow further with time and technological development.

Bibliography

Abildgren K., Buchholst B.V., Staghøj J., 2011, *Bank-firm relationships and the performance of non-financial firms during the financial crisis 2008-09 microeconometric evidence from large-scale firm-level data*, Working Paper, no. 73. Danmarks National Bank.

Akbani R., Kwek S., Japkowicz N., 2004, *Applying support vector machines to imbalanced datasets*, In Machine Learning: ECML 2004. Springer, pp. 39-50.

Anderson R., 1999, *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford University Press, New York.

Anderson R., 2007, *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford: Oxford University Press.

Anderson R., 2015, *Piecewise Logistic Regression: an Application in Credit Scoring*, Credit Scoring and Control Conference XIV, Edinburgh.

Antonowicz P., 2007, *Metody oceny i prognoza kondycji ekonomiczno-finansowej przedsiębiorstw*, Gdańsk.

Appenzeller D., Szarzec K., 2004, *Prognozowanie zagrożenia upadłością polskich spółek publicznych*, Rynek Terminowy, no. 1, pp. 120-128.

Auria L., Moro R., 2008, *Support Vector Machines (SVM) as a Technique for Solvency Analysis*, Discussion Papers, DIW Berlin.

Bellotti T., Crook J., 2009, *Support vector machines for credit scoring and discovery of significant features,* Expert Systems with Applications vol. 36, no. 2, pp. 3302-3308.

Crone S., Finlay S., 2012, *Instance sampling in credit scoring: An empirical study of sample size and balancing*, International Journal of Forecasting, vol. 28.

Gajdka J., Stos D., 1996, *Wykorzystanie analizy dyskryminacyjnej w przewidywaniu bankructwa spółki*, [w:] Duraj J. (red.), *Przedsiębiorstwo na rynku kapitałowym*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Gajdka J., Stos D., 2003, *Ocena kondycji finansowej polskich spółek publicznych w okresie 1998-2001*, [w:] Zarzecki D. (red.), *Czas na pieniądz. Zarządzanie finansami. Mierzenie wyników i wycena przedsiębiorstw*, tom I, Wydawnictwo Uniwersytetu Szczecińskiego, Szczecin.

Ghodselahi A., 2011, *A hybrid support vector machine ensemble model for credit scoring,* International Journal of Computer Applications, no. 17.

Görg H., Spaliara M. E., 2009, *Financial Heath, exports and firm survival: A comparison of British and French firms*, Kiel Institute for the World Economy Working Paper 1568.

Gray R., Owen D., Sopher M.J., 1998, *Setting up a control system for your organization*, Nonprofit World, vol. 16, no 3, pp. 65-76.

Gruszczyński M., 2003, *Modele mikroekonometrii w analizie i prognozowaniu zagrożenia finansowego przedsiębiorstw*, Studia Ekonomiczne nr 34, Wydawnictwo INE PAN, Warszawa.

Hadasik D., 1998, *Upadłość przedsiębiorstw w Polsce i metody jej prognozowania*, Zeszyty Naukowe, Seria II, nr 153, Akademia Ekonomiczna w Poznaniu, Poznań.

Haltuf M., 2014, *Support Vector Machine for Credit Scoring*, University of Economics in Prague.

Hamrol M., Czajka B., Piechocki M., 2004, *Upadłość przedsiębiorstwa – model analizy dyskryminacyjnej*, Przegląd Organizacji, no. 6, pp. 35-39.

Härdle W., Moro R., Schäfer D., 2007, *Estimating Probabilities of Default With Support Vector Machines*, http://edoc.hu-berlin.de/series/sfb-649-papers/2007-35/PDF/35.pdf.

Härdle W., Lee Y.-J., Schäfer D., Yeh Y.-R., 2008, *The Default Risk of Firms Examined with Smooth Support Vector Machines*, SFB 649 Discussion Papers SFB649DP2008-005, Sonderforschungsbereich 649, Humboldt University, Berlin, Germany.

Härdle W., Lee Y.-J., Schäfer D., Yeh Y.-R., 2009, *Variable selection and oversampling in the use of smooth support vector machine for predicting the default risk of companies*, Journal of Forecasting, vol. 28, no. 6, pp. 512-534.

Hołda A., 2001, *Prognozowanie bankructwa jednostki w warunkach gospodarki polskiej z wykorzystaniem funkcji dyskryminacyjnej ZH*, Rachunkowość, no. 5, pp. 306-310.

Jankowitsch R., Pichler S., Schwaiger W., 2007, *Modelling the economic value of credit rating systems*, Journal of Banking & Finance, vol. 31, no. 1.

Korol T., Prusak B., 2005, *Upadłość przedsiębiorstw a wykorzystanie sztucznej inteligencji,* CeDeWu.pl., Wydawnictwo Fachowe, Warszawa.

Kisielińska J., Waszkowski A., *Polskie modele do prognozowania bankructwa przedsiębiorstw i ich weryfikacja*, On line. http://www.wne.sggw.pl/czasopisma/pdf/EIOGZ_2010_nr82_s17.pdf.

Lacerda A., Russ I., Moro A., 2008, *Analysis of the predictors of default for Portuguese firms*, https://www.bportugal.pt/en-US/BdP%20Publications%20Research/WP200822.pdf.

Lovie A., Lovie P., 1986, *The flat maximum effect and linear scoring models for prediction*, Journal of Forecasting, vol. 5, no. 3.

Mączyńska E.,1994, *Ocena kondycji przedsiębiorstwa (uproszczone metody*), Życie Gospodarcze, no. 38, pp. 42-45.

Mączyńska E., Zawadzki M., 2006*, Dyskryminacyjne modele predykcji bankructwa przedsiębiorstw*, Ekonomista. Warszawa.

Mercer J., 1990, *Functions of positive and negative type and their connection with the theory of integral equations*, Philosophical Transactions of the Royal Society of London, no. 209.

Nehrebecka N., 2011, *Wykorzystanie łańcuchów Markowa do prognozowania zmian w strukturze polskich przedsiębiorstw*, Gospodarka Narodowa, no. 10.

Nehrebecka N., 2015, *Approach to the assessment of credit risk for non-financial corporations*, Poland Evidence, Bank for International Settlements.

Pogodzińska M., Sojak S., 1995, *Wykorzystanie analizy dyskryminacyjnej w przewidywaniu bankructwa przedsiębiorstw*, AUNC, Ekonomia XXV, no. 299, Toruń.

Prusak B., 2005, *Nowoczesne metody prognozowania zagrożenia finansowego przedsiębiorstw*, Warszawa.

Rossi S. P.S., Schwaiger M. S., Winkler G., 2009, *How loan portfolio diversification affects risk, efficiency and capitalization: A managerial behavior model for Austrian banks*, Journal of Banking & Finance, Elsevier, vol. 33, no.12, pp. 2218-2226.

Schebesch, K.B., Stecking R., 2005, *Support vector machines for classifying and describing credit applicants: detecting typical and critical regions*, Journal of the Operational Research Society, vol. 56, no. 9, pp. 1082-1088.

Sharma D., 2011, *Evidence in Favor of Weight of Evidence and Binning Transformations for Predictive Modeling*, Social Science Research Network.

Shawe-Taylor J., Cristianini N., 2000, *Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press.

Siarka P., 2011, *Quality measures of scoring models*, Journal of Risk Management in Financial Institutions, London.

Sojak S., Stawicki J., 2001, *Wykorzystanie metod taksonomicznych do oceny kondycji ekonomicznej przedsiębiorstw*, [w:] Bednarski L. (red.), *Zeszyty teoretyczne rachunkowości*, tom 3 (59), Warszawa, pp. 56-67.

Wierzba D., 2000, *Wczesne wykrywanie przedsiębiorstw zagrożonych upadłością na podstawie wskaźników finansowych – teoria i badania empiryczne*, Zeszyty Naukowe nr 9, Wydawnictwo Wyższej Szkoły Ekonomiczno-Informacyjnej w Warszawie, Warszawa, pp. 79-105.

PRZEWIDYWANIE RYZYKA KREDYTOWEGO PRZEDSIĘBIORSTW NIEFINANSOWYCH. PORÓWNANIE MODELI SCORINGOWYCH: REGRESJA LOGISTYCZNA VS *SUPPORT VECTOR MACHINE*

Streszczenie: Celem artykułu jest porównanie modeli na próbie uczącej się i testowej, które powstaną za pomocą regresji logistycznej oraz *Support Vector Machine* (SVM) i posłużą do oceny ryzyka kredytowego przedsiębiorstw niefinansowych. Podczas tworzenia modeli zmienne zostaną poddane transformacji *Weight of Evidence* (*WoE*), liczba potencjalnych predykatorów zostanie zredukowana na podstawie statystyki *Information Value* (*IV*). Jakość modeli zostanie oceniona według najpopularniejszych kryteriów, takich jak statystyka Giniego, Kołmogorowa-Smirnowa (K-S) oraz *Area Under Receiver Operating Characteristic* (AUROC). Na podstawie wyników stwierdzono, iż występują istotne różnice między modelem regresji logistycznej o charakterze dyskryminacyjnym a SVM dla próbki modelowej. W przypadku próby walidacyjnej regresja logistyczna ma najlepszą zdolność prognostyczną. Analizy te można wykorzystać w celu zmniejszenia ryzyka negatywnych skutków dla sektora finansowego.

Słowa kluczowe: Basel III, *Internal Rating Based System*, ryzyko kredytowe, *Support Vector Machines,* regresja logistyczna.