Approach to the assessment of credit risk for non-financial corporations. Poland Evidence

Natalia Nehrebecka
Narodowy Bank Polski (Natalia.Nehrebecka@nbp.pl),
University of Warsaw (nnehrebecka@wne.uw.edu.pl)

Abstract

This paper presents the PD Model and Rating System for non-financial corporations in Poland, developed on the basis of individual data from the following databases: (i) balance-sheet and profit and loss account data from Amadeus (Bureau van Dijk) and Notoria, (ii) credit information from the Prudential Reporting (NB300) managed by Narodowy Bank Polski and (iii) insolvency data from The National Court Register (KRS). These analyses can be used to reduce the risk of adverse effects on the financial sector.

The statistical model is built on logistic regression model, and produces an estimate of the annual Probability of Default (PD) of the assessed company. Models were estimated on categorized variables transformed using the weight of evidence (WoE) approach. The outputs of the Scorecard are used in the PD model. The purpose of the Scorecard is to differentiate between good and bad clients by estimating the Probability of Default (PD) during the following 12 months. Performance measurement purposes and thus model's ability to differentiate between good and bad clients was measured using Gini's coefficient, Kolmogorov-Smirnov statistic, and Information Value.

Keywords: Credit Risk, Scoring Methods, Rating System, Calibration

JEL classification: C190, G210

Contents

Introduction

Among the key activities of the banking sector that constitute the foundations of the financial system in every country, the correct management of assets and liabilities should be distinguished. The impact of the performance of those tasks by banking institutions is crucial for the development of the country's economy, which was evidenced by the financial crisis that begun in 2007 in the United States.

Assessment of Credit Risk, and especially ensuring accuracy and reliability of credit ratings by means of validation is of critical importance to many different market participants (Winkler, 2005). Definition "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). 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.

This article presents a suggestion for an internal credit assessment system. The main risk indicators are described, which demonstrate the financial standing of companies registered in Poland. The article presents both a scoring assessment and a rating system. The purpose of the Scorecard is to differentiate between good and bad firms by estimating the Probability of Default (PD) during the following 12 months. The analysis was performed with the use of a logistic regression on categorized variables transformed using the weight of evidence approach. Scoring methods have been used to create an indicator for grading the companies in the case of defaults. While developing the model, the number of potential predictors was reduced on the basis of Information Value (IV) statistics. The quality of the model is assessed according to the most popular criteria, such as the GINI statistics, the Kolmogorov-Smirnov test (K-S) and Area Under Receiver Operating Characteristic (AUROC). Rating is a risk assessment that determines the ability of a given entity (company) to manage its debt. For many investors, it plays a crucial role while making a decision on committing resources to a given operation. Rating enables the assessment of the financial strength of business entities and the estimation of the risk associated with commercial and credit transactions. It provides an early warning about any possible problems in terms of cash flow and offers a chance to react quickly. The summary presents a migration matrix. Migration matrices are widely applicable to financial risk management. In the credit risk estimation process, the entity is assigned one of several rating classes (statuses), while its assessment (rating) is determined by the Markov chain transition matrix. The likelihood that the borrower is insolvent, i.e. that it transitions into the default state, is read from the migration matrix.

Key Purposes for the Assessment of Credit Risk of Companies by Central Banks:

- keeping track of the (credit risk of the) economy from a macro-economic perspective;
- assessing credit quality of collateral in the context of monetary policy operations;
- assessing and ensuring financial market stability from a macro-prudential perspective.

The first part of the paper presents a review of literature. Next, the methodology used for estimating the model is described. Then the detailed information on the database is presented, together with the characteristics of the variables used in the estimation, estimation results and conclusions.

Rating system

Appropriate risk assessments are one of the most important aspects of the activity of financial institutions. In 1999, the Basel Committee on Banking Supervision published several postulates for changes in the current regulations in terms of the capital adequacy structure of financial institutions, which contributed to the preparation of the New Capital Agreement, known as Basel II. The main modification postulate was the reinforcement of the risk management process in the banking sector. One of the key changes was the introduction of the possibility of internal risk management, and therefore – determination of the minimal capital requirements. In particular, the bank can select among three approaches. The first of them, which is a continuation of the approach presented by the previous regulation, obliges the bank to maintain the ratio between the minimal capital and the sum of risk-weighted assets at the level of 8%, where the weights are determined by the national regulatory body. As part of the second approach, called IRB (*internal rating based*), the bank is obliged to prepare an internal estimation of the likelihood of the obligation not being fulfilled (*probability of default*). The other risk parameters, such as the loss coefficient arising from the failure to fulfil the commitment (*Loss Given Default*) and the exposure at the time of insolvency (*Exposure At Default*) are provided by the regulatory body. The third, and at the same time the broadest approach, known as *Advanced* IRB, enables banks to estimate all risk parameters.

Each bank using one of the IRB approaches is obliged to estimate the likelihood of insolvency for each loan granted. A popular method of achieving that is credit scoring. Financial institutions can use external scoring or rating assessments (external rating approach), however, they are applicable to only several, largest business entities. In the vast majority of cases, an internally developed risk assessment method (*Internal Rating Approach*) is used. The use of the bank's own rating boards, called master scales, is a common practice. Entities with low risk levels are grouped together and assigned to one rating class. Each rating class has a top and bottom threshold expressed by the default probability, as well as an average value. The allocation of a given entity to one of the rating classes automatically determines its default probability, which is equal to the average value for the given class. The number of classes depends on the bank's individual approach, however, at least seven classes are required for solvent entities. Usually, lower probability values are assigned to the "lower" classes, which are denoted by digits or appropriate abbreviations, such as "AAA". It is therefore a process of discretization

of the default probability estimations. On one hand, this causes a certain loss of accuracy, and on the other hand, this approach has several important benefits. Firstly, it facilitates further aggregate analysis, simplifies the reporting and model monitoring process. Secondly, it allows for expert knowledge to be used by way of manual relocation of entities to higher of lower rating classes.

Together, the default probability determination model and the master scale are known as the rating system. It is used to forecast the default probability of each entity, expressed by a rating class. There are two approaches used to establish a rating system. The first, called PIT (*point in time*), assumes maximum adjustment to changes resulting from the business cycle. The default probability estimation includes an individual and macroeconomic component. A high level of migration of units to lower classes is expected in the period of economic growth, and to higher classes at the time of crisis. The second approach, known as TTC (*trough the cycle*), maximally reduces the influence of the macroeconomic component. All changes are only determined by changes in the individual estimation component, while the percentage share of entities should remain relatively unchanged . There is also a broad range of intermediate hybrid approaches, which include individual elements of those two methods.

It is worth referring to the guidelines of the Eurosystem, consisting of the European Central Bank and central banks in countries within the euro zone. In order to ensure financial stability within the European Union, framework principles of debt-or credit assessment were established, known as ECAF (European Credit Assessment Framework), which include instructions in regard to acceptable forms of collateral for lending transactions as part of open market operations. According to the guidelines, quality assessment of a given asset is carried out with the use of credit rating allocated in accordance with the standards that enable a clear and reliable comparison of the creditworthiness of entities. The rating tools currently used for the verification of assets eligible for monetary policy operations are: external entities described as rating agencies (ECAI), internal credit assessment systems (ICAS), internal counterparty rating systems (IRB) and independent external institutions (RT). The abovementioned group, provided for in the Eurosystem, are obliged to comply with the formal terms while carrying out a credit assessment of a given entity, in particular, to apply the definition of default fully consistent with the definition recommended by Basel II. The first of the sources listed, ECAI, refers to agencies whose rating publications can be used – according to the Basel Committee - while calculating risk-weighted assets: DBRS, FitchRatings, Moody's and Standard & Poor's. In this case, the ratings need to be public, and the process of their allocation needs to be objective, independent and clear. The ICAS source includes Deutsche Bundesbank, the Banco de España, the Banque de France, the Oesterreichische Nationalbank, the Banca d'Italia and, since 2013, the Banque Nationale de Belgique and Banka Slovenije. Whereas in terms of the external ratings mentioned above, the counterparty which has received approval from a financial supervisory body to use this tool is obliged to submit appropriate information to the Eurosystem, at least once per year, to enable continued applicability of the system to be determined. The last source of credit assessment, RT, refers to entities whose estimated credit ratings are not announced publicly.

Literature review

Credit Scoring Statistical Techniques

A wide range of statistical techniques are used in building the scoring models (Table 1).

Credit Scoring Statistical Techniques

Source: own calculation — Table 1

Method	Authors
Weight-Of-Evidence measure	Bailey, 2001; Banasik et al., 2003; Siddiqi, 2006; Abdou, 2009
Regression analysis	Lucas, 1992; Henley, 1995; Hand, Henley, 1997; Hand, Jacka, 1998
Discriminant analysis	Altman, 1968; Desai et al., 1996; Hand, Henley, 1997; Caouette et al., 1998; Hand et al., 1998; Sarlija et al., 2004; Abdou, Pointon, 2009; Wiginton, 1980; Crone, Finlay, 2012
Probit analysis	Finney, 1952; Grablowsky, Talley, 1981
Logistic regression	Lenard et al., 1995; Desai et al., 1996; Lee and Jung, 2000; Baesens et al., 2003; Crook et al., 2007; Abdou et al., 2008; Wiginton, 1980; Yap, Ong, Husain, 2011; Kočenda, Vojtek, 2009; Stepanova, Thomas, 2002; Thanh Dinh oraz Kleimer, 2007; Crone, Finlay, 2012
Linear programming	Yang, Wang, Bai, Zhang, 2004
Cox's proportional hazard model	Stepanova, Thomas, 2002
Support Vector Machines	Deschaine, Francone, 2008
Decision trees	Baesens et al., 2003; Stefanowski, Wilk, 2001; Thomas, 2000; Fritz, Hosemann, 2000; Hand, Jacka 1998; Henley, Hand, 1996; Coffman, 1986; Paleologo et al., 2010; Yap, Ong, Husain, 2011; Kočenda, Vojtek, 2009; Frydman, Altman, Kao,1985; Novak, LaDue, 1999; Thomas, Bijak, 2012; Crone, Finlay, 2012
Neural Networks	Amari, 2002; Al Amari, 2002; Gately, 1996; Irwin et al., 1995; Masters, 1995; Palisade Corporation, 2005; Desai, Conwey, Crook, Overstreet,1996; Crone, Finlay, 2012
Genetic algorithms and genetic programming	Goldberg, 1989; Koza, 1992; McKee and Lensberg, 2002; Etemadi et al., 2009; Huang et al., 2006; Huang et al., 2007
Markov switching model and Bayesian estimation	Chuang, Kuan, 2011; Frydman, Schuermann, 2008; Jacobs, Kiefer, 2011; Tasche, 2013

Statistical tools and methods used to establish scoring models correspond to the most popular and effective methods used in statistics for modelling similar phenomena with binary explanatory variables. For that reason, the dominant method for a long time remained the discriminant analysis, described by, among others, Forgy, Myers (1963) and Altman (1968). Over time , the developments in computer technology stirred great interest in logistic regression, which became the most widely used tool for building scoring models, applied, among others, by Wiginton (1980) and Kleimeier, Dinh (2007). An important advantage of the logit model relates to a better adjustment of the logistic distribution to the issue analysed when compared to normal distribution. The discriminant analysis and the probit model, however, require an assumption of a normal distribution of variables. The popularity of logistic regression resulted from, among others, the reliability of estimations based on the available scope of data and the range of probability results contained within the range 0 to 1, which simplifies interpretation of the phenomenon that is being explained. An important factor, while selecting methods,

is also the ease of interpretation of the results. The biggest advantages of the nonparametric model (decision tree), used by Altman, Kao, Frydman (1985), LaDue, Novak (1999), are its intuitive character, ease and effectiveness. The use of the survival analysis (Stepanova, Thomas, 2002; Baesens, Gestel, Poel, 2005; Andreeva, 2006; Pazdera, Rychnovsky, Zahradnik, 2009; Giambony, 2012), in particular, the comparison between Cox proportional and nonproportional models, is justified when the researcher wishes to focus not only on creating a model of default probability, but also on determining the time of its occurrence. Cox proportional model is characterised by the assumption, often not actually met, of the stability of hazard over time; the nonproportional model is free of this flaw. However, the direct benefit of using the duration analysis is determination of the default probability, changeable in time, for the analysed product. Baesens, Gestel and Poel (2005) verified the hypothesis on the superiority of the predictive power of artificial neural network models over logit models and Cox proportional model. According to the authors, Cox model is superior to the logit model due to its universal approximatively characteristics and also as it is not necessary to make assumptions on the baseline hazard function. Whereas its flaw is the inability to cope with nonlinear correlations between variables which need to be indicated by the researcher beforehand.

The neural networks used by, among others, Convay, Crook, Desai, Overstreet (1997), Baesens, Gestel and Poel (2005) generate satisfactory results, which are not inferior to the results obtained using parametric methods, however, the interpretation and understanding of the results obtained is much more complicated. According to Baesens, Gestel and Poel (2005), neural network models lack most limitations imposed by Cox proportional hazards model. In case of the simplest neural network model, censored observations are removed from the dataset, thus determining the estimator load. Whereas Ohno-Machado model uses a diverse neural network, which facilitates implementation of censored observations throughout their existence in the dataset.

In terms of satisfying the requirements relating to the application of a given method, nonparametric methods have an advantage, as they do not make assumptions on the functional form of the correlation. They are more effective in finding interactions between explanatory variables. Every method has its strengths and weaknesses, therefore the selection of the right method should be determined by the type of issue to be analysed.

Is selection of a statistical method really that important? In the study by Yap, Ong, Husain (2011), the discriminative power of a model created with the use of decision trees (CHAID algorithm) was much lower than the results obtained with the use of logistic regression. Kočenda and Vojtek (2009), who used the CART algorithm, were unable to decide which of the models was better, whereas Thomas, Edelman, Crook (2002) compared the percentage of correct classifications for various statistical methods in several studies (Table 2.)

A comparison of percentage correctly classified from publish research

Source: Thomas L., Edelman D., Crook J., Credit scoring and its applications, Philadelphia, 2002, p. 101-103

Table 2

Method/Authors	Henley (1995)	Boyle (1992)	Srinivisan, Chakrin (1987)	Yobas (1997)	Desai (1997)
Linear regression	43,4	77,5	87,5	68,4	66,5
Logistic regression	43,3	-	89,3	-	67,3
Decision tree	43,8	75,0	93,2	62,3	-
Math programming	-	74,7	86,1	-	-
Neutral nets	-	-	-	62,0	66,4
Genetic programming	-	-	-	64,5	-

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. For that reason, it is crucial while selecting the research method to consider all good and bad points and to choose the method that is most suited to the issue at hand.

Calibration and mapping to Ratings

This subsection deals with the issue of rating system calibration, i.e. allocation of rating classes to entities in order to ensure that the calibration power of the division created is as high as possible.

At first, the form of the function depicting the transition of score into default probability is estimated. The methods presented can be divided into two groups. The first contains the methods of approximating conditional score distributions for defaults and entities with a good financial standing to the parametric distribution which can be expressed with the use of a density function and distribution functions (Dey, 2010; Bennett, 2003; Krężołek, 2007; Tasche 2006; Tasche 2008; Tasche 2009). Taking into consideration that those distributions are usually rightward or leftward skewed, only those of the types of distributions are described that allow for a description of both density function asymmetry variants with the use of the appropriate parameters (e.g. asymmetric Gauss distribution, asymmetric Laplace distribution, skew normal distribution and scaled beta distribution). On that basis, with the use of Bayes formula, it is possible to define PD values. Those methods are recommended for the purpose of calibration of a score which is not interpreted as probability (e.g. the score as a result of discriminant analysis).

The second group covers various variants of regression on binary variables denoting the default status of a given company (Tasche, 2009; Neagu, Keenan, 2009; Koenker, Yoon, 2009; Neagu, Keenan, Chalermkraivuth, 2009; Zadrozny, Elkan, 2002; Van der Burgt, 2008). They are universal methods facilitating calibration of a score which can be interpreted as probability (e.g. the score as a result of logistic regression). Firstly, apart from the most popular transition functions (probit and

logit), others have also been suggested: cauchit and the complementary log-log function. Another alternative is the application of Platt adjustment and Box-Cox transformation of the explanatory variable. Apart from that, each regression can also use the polygonal curve model. Another option is the quasi-moment-matching method and isotonic regression. Based on the probability values found, the rating is allocated with the use of the master scale with set threshold values for individual classes.

Data description

The empirical analysis was based on the individual data from different sources (from the years 2007 to 2012), which are:

- Data on banking defaults are drawn from the **Prudential Reporting (NB300)** managed by Narodowy Bank Polski. 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 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 (**AMADEUS (*Bureau van Dijk*); Notoria OnLine**). Amadeus (Bureau van Dijk) 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 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)*.

The following legal forms were analyzed: partnerships (unlimited partnerships, professional partnerships, limited partnerships, joint stock-limited partnerships); capital companies (limited liability companies, joint stock companies); civil law partnership, state owned enterprises, branches of foreign entrepreneurs.

For the definition of the total number of obligors the following selection criteria were used:

- The company is existent (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 1.5 Mio EUR for each reporting date.

The dataset, after its initial preparation and while keeping only the observations on which the model can be based, contained 5091 records. However, the number of observations marked as "bad" was 298 (Table 3). 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 "bad" observations

and adding them to all "good" observations was rejected due to the excessive size of the dataset that would be 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 "bad" observations will be added to a selected part of observation from the other class. The proportions were established at 20:80 for several reasons. A smaller number of "good" observations drawn would cause difficulty with drawing a representative sample. Whereas a higher number would extend the calculation time while improving the quality of the model only slightly.

General statistics for 2012

Source: own calculation Table 3

Number of obligors	Thereof Insolvent	Thereof defaulted	Insolvency rate	Default rate
5091	28	298	0,55%	5,85%

Before estimating the model it was tested whether the constructed sample is representative following the results of the non-parametric Wilcoxon-Mann-Whitney test, Kołogomorow-Smirnow test and the parametric *t-Student* test for equality of averages for the continuous variables and the χ^2 Pearson test and the *Population Stability Index* (PSI) for the discrete variables. The PSI coefficient is applied in order to investigate the differences in distribution of 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.

Forecasting defaults concern only companies that had impaired loans (loans from portfolio B for which objective evidence of impairment and decrease in the value of expected cash flows have been recognised (*in banks applying IFRS*) or loans classified as irregular pursuant to the Regulation of the Minister of Finance regarding principles for creating provisions for the risk of banking activity (*in banks applying the Polish accounting standards*)).

Based on the literature, the potential defaults predictors were chosen with the focus on financial indicators. Signals for deteriorating financial condition of the company are: negative dynamics for revenue, assets and equity, decreasing profits, negative equity, increasing indebtedness, problems with financial liquidity, deteriorating 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 2 years average).

Methodology

Construction of credit scoring model

In order to construct an indicator which would enable assessing the probability of a company to go default, a logistic regression was used. Due to a 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[i], 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*[ii] 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
$p_i^{non-defaults}$	-	the percentage of non-default companies that belong to category *i*
$p_i^{defaults}$	-	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).

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

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.

Calibration & Mapping to rating

In order to perform the calibration, the scores were bucketed with (more or less) same number of defaults in each bucket. After that, Default Rate in each bucket was transformed. Such modified Default Rate was transformed into odds.

PD was calculated using the below formula:

$$PD = e^X / 1 + e^X$$

where:

$$X = \beta_2 * SCORE + \beta_1$$

The theoretical relationship between the **score** and **logarithm of odds** (which from the nature of logistic regression should be linear) was used to obtain estimates of the calibration function. The accuracy of obtained estimated *PD*'s for each calibration function was tested Population Stability Index. According to common usage of the *PSI*, values between 0 and 0,1 mean no significant changes. After obtaining PD values, scores were mapped to ratings according to the master scale.

The calibration of the scoring system which is another important task in scoring model validation.

- **The first group of tests** can only be applied to one single rating grade over a single time period *(binomial test Clopper and Pearson, binomial test Agresti and Coulla, binomial test Wald, corrected binomial test Wald, binomial test Wilson, corrected binomial test Wilson, one-factor-model, moment matching approach and granularity adjustment)*.
- **The second group of tests** provide more advanced methods that can be used to test the adequacy of the default probability prediction over a single time period for several rating grades (*Spiegelhalter test, Hosmer-Lemeshow test, Blöchlinger test*).

Results

The research was performed on the sample included companies observed in 2011. In Model the default probability was predicted for a one year horizon.

After performing the initial data analysis, a dataset was prepared, based on which the model could be built. The models were estimated using logistic regression, preceded by one-dimensional and multidimensional analysis. Many of the 611 explanatory variables available which were not excluded during the previous stages of research are correlated, which negatively affected the estimations of the logistic regression model. Variable correlation partly results from the manner of establishing the dataset, as on the basis of one general variable, multiple detailed variables were established with the use of various aggregates and timelines. Therefore, the selection of variables based on which the logistic regression model is to be estimated will take place in two stages. During the first phase, variables characterised by the lowest predictive power measured with the use of the Information Value statistics will be rejected. Next, cluster analysis will be used to select the best variables from the groups of correlated variables.

The first stage of establishing the model, i.e. the reduction of the number of variables, will be performed with the use of automatic categorisation (predictive power maximising division) in order to calculate the Information Value. Categorisation of all variables was performed in the following variants, by dividing them into 3, 4, 5 or 6 categories. Finally, the division was selected in which the statistical value was the highest, and each category established had at least 2% of the training sample, in order to ensure stable results. In case of variables that did not meet the observation number requirement in any automatic configuration and which were characterised by a high Information Value, manual categorisation was used in order to correct the insufficient number in the categories. The statistical value equal to 0.1 was determined as the minimum value to make variables eligible for further analysis. The selection of this value complies with the value most widely found in literature, especially where there is a sufficient number of explanatory variables of a high predictive power in the dataset. Due to the high number of variables which met the requirement of the minimum Information Value, it was decided that the maximum of three variables with the highest statistical value, originating from the same original class, are to be used for further analysis.

After the conclusion of the one-dimensional analysis, 85 explanatory variables remained in the dataset, which then underwent cluster analysis. The division into groups was carried out until the percentage of explanatory variables in each of them was higher than or equal to 70%. In order to eliminate excessive correlation between variables in the model, only one variable was chosen from each cluster.

Having selected 20 variables, the last stage of the model establishment process was commenced with the use of a stepwise regression algorithm. The limit value, p-value, for adding of removing a variable form the model, was determined at 0.05. The algorithm stopped after 10 iterations (Table 4.), during which variables were only added to the model. The greatest weight was assigned to the indicator of ROA (16%).

In the definition of the regulatory body, validation involves a range of techniques used to verify the model's ability to distinguish between "bad" and "good" entities and the calibration quality of the estimated parameters, facilitating correct quantification of the risk incurred. Although the Basel Committee assigns the obligation of controlling the validation process to the local authorities who establish detailed guidelines, Basel II includes six fundamental principles in regard to this issue:

1. the bank is obliged to validate the scoring model,

2. the validation determines the assessment of the predictive power of the model and ratings used, as part of which four characteristic features of the model are analysed:

- objectivity – the model used ensures a standardised process of allocating risk parameters to borrowers,

- accuracy – achieving small acceptable deviations between the estimated risk parameters and the actual implementations,

- stability – risk parameters are constant for the same risk,

- conservatism – the use of restrictive risk parameters in case of unverified information,

3. the validation process is carried out at least once per year,

4. there is no single common validation method,
5. validation concerns both qualitative and quantitative assessment,
6. the validation results are assessed by an independent entity .

Final scorecard

Source: own calculation

Table 4

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	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	6,87%	-INF	-4.501	0
		-4.501	12.641	59
		12.641	+INF	85
ROA	16,39%	-INF	-10.49	0
		-10.49	1.907	53
		1.907	6.502	91
		6.502	+INF	189
Solvency ratio (Liability based)	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

In accordance with the guidelines of Basel II, the decision to implement the scoring model should be determined by the results of the validation process: the discriminatory power and calibration quality. The validation process used in practice by banks involves filling out a validation report, taking into consideration qualitative and quantitative validation. In regard to the first of them, it should be noted that the data used for the purpose of this paper meet the requirements of Basel II both in terms of the definition of default applied, the timeline and the representativeness

of the selected sample, while the direction of impact of the majority of variables is consistent with business logic.

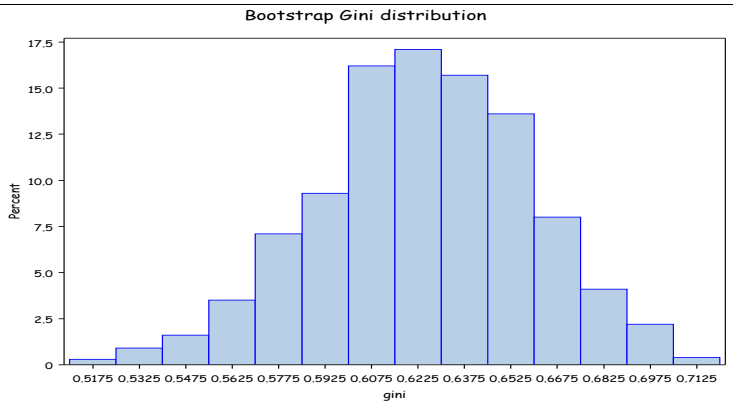
Even though there is no single common method of validating scoring systems, the Basel Committee recommends the use of the Gini coefficient (Accuracy Ratio) and its graphic equivalent - the CAP curve. The Basel Committee also points to techniques popular in theoretical literature and in the banking practice, such as: the ROC curve, AUROC measure, Pietra index, Bayes error rate, measures based on entropy, e.g. CIER, IV coefficient, divergence, Kendall and Somers'D parameter, Brier score.

The GINI and K-S value of the model were equal to, respectively, 62.3 and 51.4, which means satisfactory discrimination. The hypothesis on the combined insignificance of explanatory variables in the model was rejected (p-value = 0.000). While using the Wald method, tests were carried out on the significance of individual variables separately and the p-value for each of them was below the established 5% significance level. There are also no grounds to reject the zero hypothesis on the good adjustment of the model to the data (p-value = 0.19). The VIF (Variance Inflation Factor) value does not indicate any issues of excessive collinearity.

While using the bootstrap method, the stability of the calculated GINI value was verified. To achieve this, a sample was drawn and returned a thousand times, which contained 2/3 observations from the original set.

Results of bootstrap analysis for Gini coefficient

Source: own calculation Graph 1



The operation of the model was verified with the use of a validation set. The hypothesis on the combined insignificance of the parameters was rejected (p-value = 0.000). There are also no grounds to reject the zero hypothesis on the insignificance of individual explanatory variables in the model. The GINI and K-S values were equal to, respectively, 68.5 and 54.9, which confirms the stability of the results of the control set. The hypothesis on the good adjustment of the model was rejected for the validation set (p-value equal to 0.2867).

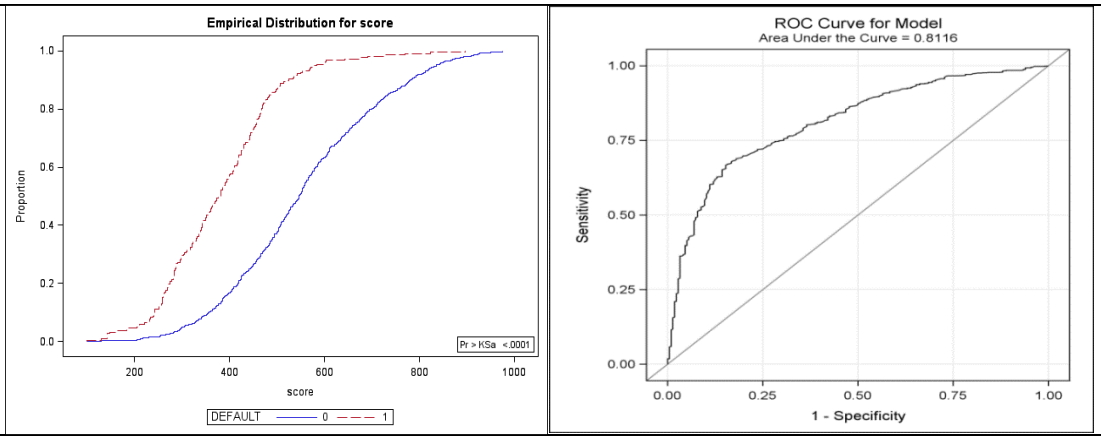
In order to analyse the appropriate cut-off point, Graph 2. is presented of the empirical distribution functions for "good" and "bad" entities, on the basis of which the K-S value is calculated. The maximum difference between those two distribution functions is achieved for 483 scoring points, which means that in order to maximise the percentage of correct classifications, the cut-off point should be established at this value. The ROC curve was prepared for the training sample. The AUROC value

was equal to 0.8116, which means a satisfactory quality model. In terms of the validation sample, the graph of the ROC curve looked very similar, while the AUROC value was equal to 0.80.

Empirical distribution for score (K-S) and ROC curve on development sample

Source: own calculation

Graph 2



In order to evaluate the classification power of the model established, the following were also used: the CAP concentration curve and the directly related Gini coefficient. From the mathematical perspective, the curve presents a correlation between the distribution function of the score obtained for the insolvent group and the distribution curve of the score of the entire sample.

Population Stability Index was used to test for variables' time stability. As suggested by literature the rule of rejecting the hypotheses that default rate distributions are close to each other is when PSI exceeds 0.25. Default rate distributions for the model observation date (2011) were compared to 3 other moments in time (Table 5).

Population Stability Index

Source: own calculation

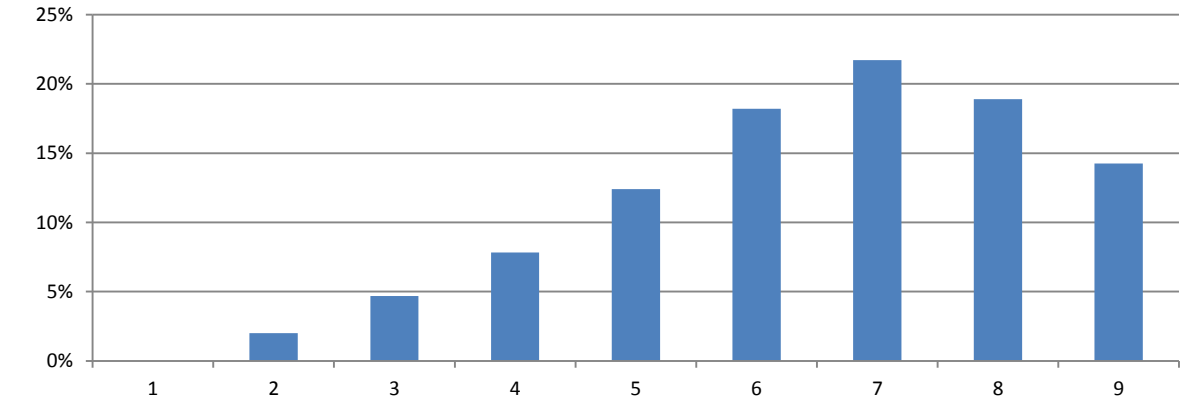
Table 5

Variables	PSI in 2010	PSI in 2009	PSI in 2009
Credit period (Creditors / Operating revenue)*360	0,026	0,021	0,019
Industry sectors	0,018	0,061	0,099
EBIT	0,013	0,015	0,020
Bank-firm relationships	0,009	0,024	0,010
ROCE	0,050	0,064	0,019
ROA	0,024	0,065	0,060
Solvency ratio (Liability based)	0,015	0,033	0,033
(Interest due / Total exposure)*100 (median of 4 q)	0,011	0,022	0,039
(Bank loans denominated in PLN / Total exposure)*100 (median of 6 q)	0,012	0,016	0,007
(Open credit lines / Total exposure)*100 (median of 6 q)	0,009	0,013	0,012

The available values of the PD parameter for individual risk classes determine their calibration.

Rating

Source: own calculation Graph 3



The first method recommended by the Basel Committee is verification (backtesting) which involves comparison of the estimated PD ex ante values assigned to individual risk classes with ex post values. Among the recommended statistical tests that facilitate backtesting, the Basel Committee distinguishes: the binomial test, Hosmer-Lemeshow test, normality test, traffic lights approach. The biggest challenges related to the verification of the PD parameter on the basis of statistical tests are the limitations arising from the required dataset, in particular, the rarity of defaults and their correlation. Due to the existing correlation, the insolvency indicators observed systematically exceed the levels of probability obtained, if they are estimated with the assumption of independency of events. As a consequence, most tests are characterised by conservatism and, based on that, even a well-calibrated model receives a low mark. On the other hand, tests which take correlation into consideration only in extreme circumstances allow for the low-quality calibration of a model to be determined. For that reason, the Basel Committee recommends that banks should also use a complementary tool – benchmarking, which involves the identification of differences between the values of estimated parameters obtained with the use of various statistical techniques or external comparative data.

Score ranges per rating with average PD

Source: own calculation

Table 6

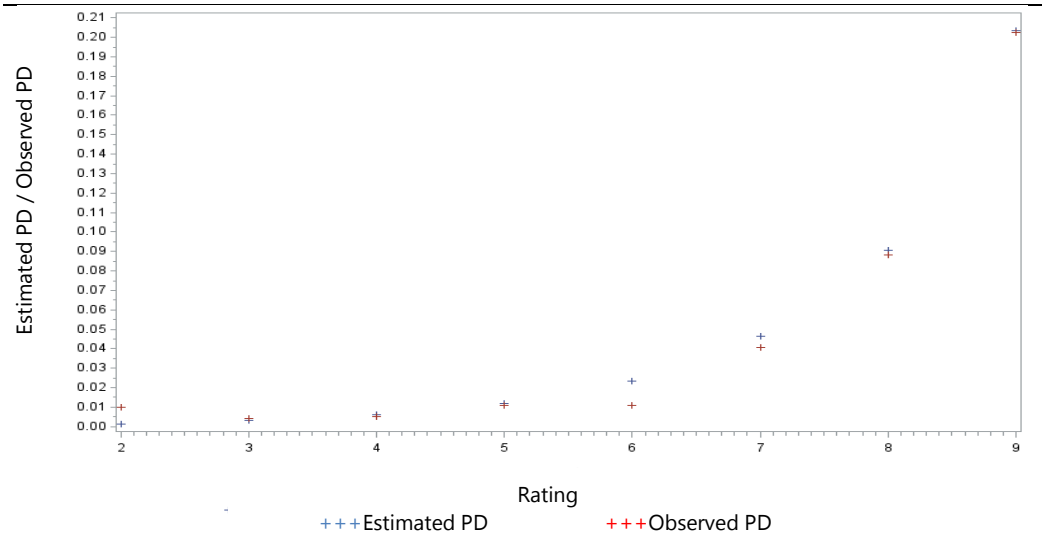
Rating	Min score	Max score	Masterscale average PD	Estimated PD	Observed PD
1	977		0,07%	0%	0%
2	897	977	0,14%	0,15%	0,98%
3	811	896	0,28%	0,31%	0,42%
4	725	810	0,57%	0,59%	0,50%
5	638	724	1,13%	1,17%	1,11%
6	551	637	2,26%	2,36%	1,08%
7	461	550	4,53%	4,66%	4,07%
8	366	460	9,05%	9,08%	8,84%
9		365	18,10%	20,33%	20,25%

Another element of the validation of scoring models is the verification of the correctness of default probability allocation to specific risk classes. The calibration quality is especially important in terms of the calculation of credit reserves. The initial assessment of the element described was carried out on the basis of a graphic summary of the model and the observed PD parameter for individual risk classes. In order to carry out the calibration, the model parameter was determined as the average PD parameter value for each rating. Based on Graph 4., a slight deviation was determined between the estimated and actual PD parameters for specific groups.

Monitor Calibration

Source: own calculation

Graph 4



While validating model calibration, it is worth testing the calibration power of individual classes, as well as the entire rating system. In case of testing individual classes, it mainly involves the binomial test with all its modifications. A crucial aspect here is to take into consideration the default correlation between entities, therefore three additional tests were carried out: one-factor-model, moment matching

approach and granularity adjustment. While assessing the calibration power of the rating system on the basis of multiple tests carried out on individual classes, the error of decreasing the value of the established p-value level is made. One solution to this problem is to use the Bonferroni or Sidak correction. Another method is to follow the Holm, Hochberg or Hommel procedures. The most popular test of the entire rating system is the Hosmer-Lemeshow test, which involves examination of the differences between the observed and the estimated default probability. For the purpose of this research, also the Spiegelhalter and Blöchlinger tests were used, which facilitate verification of the calibration power achieved in a different manner to the Hosmer-Lemeshow test.

The model calibration testing procedure was commenced with the use of the Spiegelhalter, Hosmera-Lemeshow[iii] and Blöchlinger tests. Based on the test statistical values, the models are considered to have a good calibration power. Apart from testing the rating calibration as a whole, there is a range of tests which allow for the calibration to be assessed in terms of individual rating classes. Six variants of the binomial test were used, along three tests which take correlation into consideration. While performing the first of the abovementioned tests, examining each individual risk class separately, the zero hypothesis on the correct allocation of default probability is not to be rejected.

The use of the IRB approach forces banks to perform periodic validation of the quality of rating systems used, with the aim to verify the discrimination and calibration power, as well as system stability. The stability refers to the level of migration between individual rating classes and changes in the values of model parameters over time. Usually, migration matrices are used for the purpose of assessment.

Migration Matrix

Source: own calculation

Table 7

		Rating 31/12/2011									
		1	2	3	4	5	6	7	8	9	
Rating 31/12/2010	1	0	1	0	0	0	0	0	0	0	1
		0%	100%	0%	0%	0%	0%	0%	0%	0%	100%
	2	0	49	19	10	4	3	0	0	0	85
		0%	57,65%	22,35%	11,76%	4,71%	3,53%	0%	0%	0%	100%
	3	0	20	92	52	35	11	3	1	0	214
		0%	9,35%	42,99%	24,30%	16,36%	5,14%	1,40%	0,47%	0%	100%
	4	0	5	46	130	87	69	14	8	1	360
		0%	1,39%	12,78%	36,11%	24,17%	19,17%	3,89%	2,22%	0,28%	100%
	5	0	0	14	73	180	159	68	17	3	514
		0%	0%	2,72%	14,20%	35,02%	30,93%	13,23%	3,31%	0,58%	100%
	6	0	1	6	29	109	265	218	78	24	730
		0%	0,14%	0,82%	3,97%	14,93%	36,30%	29,86%	10,68%	3,29%	100%
	7	0	0	1	3	24	134	394	207	65	828
		0%	0%	0,12%	0,36%	2,90%	16,18%	47,58%	25%	7,85%	100%
	8	0	0	0	1	9	40	114	364	176	704
		0%	0%	0%	0,14%	1,28%	5,68%	16,19%	51,70%	25%	100%
	9	0	0	0	1	1	3	26	90	323	444
		0%	0%	0%	0,23%	0,23%	0,68%	5,86%	20,27%	72,75%	100%
		0	76	178	299	449	684	837	765	592	3880

Migration Matrix

Source: own calculation

Table 8

	+1	+2	+3	+4	+5	+6	+7	+8	+	
	0	26	67	107	143	177	140	90	750	worsened ratings
	0,00%	0,67%	1,73%	2,76%	3,69%	4,56%	3,61%	2,32%	19,33%	
	-1	-2	-3	-4	-5	-6	-7	-8	-	
	1	19	62	126	242	303	311	269	1333	improved ratings
	0,03%	0,49%	1,60%	3,25%	6,24%	7,81%	8,02%	6,93%	34,36%	

While analysing the transition matrix (Table 8) in terms of ratings allocated to companies in 2011 and 2010, it can be seen that in case of 46% of companies the rating remained unchanged. Nearly 57.65% of companies which were allocated the second rating category in 2010, kept this rating, while around 72.75% of companies which scored the worst in 2011 remained in the highest risk group. The rating worsened in case of 19.33% of companies, while a higher rating in 2011 compared to 2010 was assigned to 34.36% of companies.

Conclusions

Appropriate risk assessments are one of the most important aspects of the activity of financial institutions. Assessment of Credit Risk of Companies by Central Banks important for many reasons, a.o. for: Banking Supervision and Evaluation of Financial Stability, Assessment of Credit Quality of Collateral.

This article has the aim of constructing PD Model and Rating System for non-financial corporations in Poland, developed on the basis of individual data from the following databases: (i) balance-sheet and profit and loss account data from Amadeus (Bureau van Dijk) and Notoria, (ii) credit information from the Prudential Reporting (NB300) managed by Narodowy Bank Polski and (iii) insolvency data from The National Court Register (KRS).

The statistical model is built on logistic regression model, and produces an estimate of the annual Probability of Default (PD) of the assessed company. Models were estimated on categorized variables transformed using the weight of evidence (WoE) approach. The outputs of the Scorecard are used in the PD model. Performance measurement purposes and thus model's ability to differentiate between good and bad clients was measured using Gini's coefficient, Kolmogorov-Smirnov statistic, and Information Value.

In order to perform the calibration, the scores were bucketed with (more or less) same number of defaults in each bucket. After that, Default Rate in each bucket was transformed. The theoretical relationship between the score and logarithm of odds was used to obtain estimates of the calibration function. After obtaining PD values, scores were mapped to ratings according to the master scale.

In accordance with the guidelines of Basel II, the decision to implement the scoring model should be determined by the results of the validation process: the discriminatory power and calibration quality.

References

Altman Edward, Financial Ratios, Discriminant analysis and the prediction of the corporate bankruptcy, The Journal of Finance, Vol. 23, No. 4, 1968.

Altman Edward, Duen-Li Kao, Frydman Halina, Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress, Journal of Finance, Vol. 40, No. 1, 1985.

Anderson Raymond, The Credit Scoring Toolkit: Theory and practice for retail credit risk management and decision automation, Oxford University Press, New York, 1999.

Bijak Katarzyna, Thomas Lyn C., Does segmentation always improve model performance in credit scoring?, Expert Systems with Applications, Vol. 39, 2012.

Bo-Wen Chi, Chiun-Cgieh Hsu, A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model, Expert Systems with Applications, Vol. 39, 2012.

Convay Daniel, Crook Jonathan, Desai Vijay, Overstreet George, Credit-scoring models in the credit-union environment using neural networks and genetic algorithms, Journal of Mathematics Applied in Business & Industry, Vol. 8, 1997.

Correa Arnildo, Terra Jaqueline, Neves Myrian, Magalhaes Antonio, Credit Default and Business Cycles: An Empirical Investigation of Brazilian Retail Loans, Working Paper Series no. 260.

Crone Sven, Finlay Steven, Instance sampling in credit scoring: An empirical study of sample size and balancing, International Journal of Forecasting, Vol. 28, 2012.

Ćwik Jan, Koronacki Jacek, Statystyczne systemy uczące się, Wydawnictwo Naukowo-Techniczne, Warszawa, 2005.

Forgy Edward., Myers James, The development of numerical credit evaluation systems, Journal of the American Statistical Association, Vol. 58, Iss. 303, 1963.

Greene William, Econometric Analysis, Upper Saddle River, New Jersey, 2003

Hull John, Zarządzanie ryzykiem instytucji finansowych, Wydawnictwo Naukowe PWN, Warszawa, 2011.

Husain Nor Huselina Mohamed, Ong Seng Huat, Yap Bee Wah, Using data mining to improve assessment of credit worthiness via credit scoring models, Expert Systems with Applications, Vol. 38, 2011.

Jankowitsch Rainer, Pichler Stefan, Schwaiger Walter, Modelling the economic value of credit rating systems, Journal of Banking & Finance, Vol. 31, Iss. 1, 2007.

Karlan Dean, Jonathan Zinman, Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment, Econometrica, Vol. 77, Iss. 6, 2009.

Kleimeier Stefanie, Dinh Thi Huyen Thi, Credit Scoring for Vietnam's Retail Banking Market: Implementation and Implications for Transactional versus Relationship Lending, International Review of Financial Analysis, Vol. 58, Iss. 303, 2007.

Kočenda Evžen, Vojtek Martin, Default predictors and credit scoring models for retail banking, CESIFO WORKING PAPER no. 2862, 2009.

LaDue Eddy, Novak Michael, Application of Recursive Partitioning to agricultural credit scoring, Journal of Agricultural and Applied Economics, Vol. 31, No. 1, 1999.

Lovie Alexander, Lovie Patricia, The flat maximum effect and linear scoring models for prediction, Journal of Forecasting, Vol. 5, Iss. 3, 1986.

Matuszyk Anna, Credit Scoring, CeDeWu, Warszawa, 2012.

Thomas Lyn, Edelman David, Crook Jonathan, Credit scoring and its aplications, SIAM Publishing, Philadelphia, 2002.

Wiginton John, A note on the comparision of logit and discriminant models of consumer credit behavior, The Journal of Financial and Quantitive Analysis, Vol. 15, No. 3, 1980.

[i] The *Gini* coefficient is used for a one dimensional assessment of the discriminating force of a variable. For this purpose a model with only one explanatory variable is estimated and the coefficient measures its predicting force. GINI=$1-\sum_{i=1}^{n}\left(\left(c_i^{defaults}-c_{i-1}^{defaults}\right)\left(c_i^{non-defaults}-c_{i-1}^{non-defaults}\right)\right)$, where $c_i^{defaults}$ is the cumulative share of defaults in the category *i* of the chosen trait. The result is equivalent to the Somer's D statistic.

[ii] The VIF statistic is defined based on the determination coefficient for a regression of a dependent variable X_j in respect to other explanatory variables R_j^2 $\left(VIF=\frac{1}{1-R_j^2}\right)$.

[iii] For example, as a result of the second test, the statistical value of χ_7^2=12.56 was obtained, on the basis of which there are no grounds to reject the zero hypothesis on the consistence of the distributions compared at level of statistical significance 0.05, as a consequence, good quality calibration was determined.