Proceedings of the 7th Symposium FENS, Lublin, May 14–17, 2014

Application of Data Envelopment Analysis to Calculating Probability of Default for High Rated Portfolio

U. Grzybowska* and M. Karwański
Department of Informatics, Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences, Nowoursynowska 159, PL–02776 Warszawa, Poland

The aim of our research is to propose a method of rating companies which is based on efficiency measure given by Data Envelopment Analysis (DEA). Proper rating of borrowers is an essential requirement of PD estimation. The difficulty in DEA application is the selection of input and output from the set of indicators describing evaluated objects, which is usually based on expert knowledge. Therefore we apply random forests and gradient boosting to select financial indicators used by the DEA approach and to obtain a ranking of companies needed for PD estimation.

DOI: 10.12693/APhysPolA.127.A-66
PACS: 89.65.Gh, 88.05.Lg

1. Introduction

According to the Capital Requirements Directive [1–4] banks applying the internal-rating based approach have to estimate probabilities of default (PDs) for their obligors. PDs are a core input to modern credit risk models. In credit risk estimation obligors with the same credit quality are assigned to the same class out of several rating categories. One of the obstacles connected with PD estimation is a low number of defaults, especially in high rating categories that may experience many years without any default. A substantial part of bank assets consists of portfolios with low default rate, especially high rated portfolios are LDP (Low Default Portfolio). In our paper we propose a method of rating that can be used for PD calculation. Our idea was to apply Data Envelopment Analysis (DEA) to obtain division of potential obligors into homogeneous groups. We have also used a new approach to select variables utilized by DEA. Our idea was to support the selection of indicators by ensemble classifiers: random forests and gradient boosting. We illustrate our idea on an example.

2. Methods and models

Low default portfolio (LDP) is a portfolio with only few actual defaults or a portfolio free from any defaults. Usual bank practice for deriving PD values for such exposures is often connected with mapping mechanisms to master scales. Several methods have been proposed to estimation of PD for LDP (see e.g., [5]). In our calculations we have applied the model of K. Pluto and D. Tasche [6].

Assume that there are rating classes $\{c_i\}_{i\in\{1,2,\ldots,n\}}$ numbered according to the decreasing creditworthiness from 1 to n. Then the probability of one[†] default in the class j, PD_j, for $1 \le j \le n$ can be dominated according to the Bernoulli distribution by

$$p_j \le 1 - (1-\alpha)^{1/\sum_{i\ge j} n_i}, \tag{2.1}$$

where n_i is the number of objects in the i-th class, $i = 1, \ldots, n$ and α is a confidence level.

The only key assumption in the method is a correct ordinal rating of the borrowers. Therefore the aim of our research is to propose a method of rating which is based on efficiency measure given by Data Envelopment Analysis (DEA). DEA is a mathematical programming tool for evaluating the performance of a set of peer entities called Decision Making Units (DMU), see for example [7]. The method gives an efficiency rating, i.e., a score θ for each DMU and an efficiency reference set (a peer group of objects that are efficient), which is a target for the inefficient DMUs. In most of many DEA models the DMUs with the efficiency score equal to 1 are called efficient. Calculation of the efficiency can be helpful in improving productivity and performance of an inefficient DMU. For the purpose of our research we have concentrated our efforts not on efficiency measure but on distinguishing groups of homogeneous DMUs.

In order to obtain division into homogeneous groups of companies, we have performed the DEA algorithm to the whole set of DMUs. The efficient units with efficiency score 1 constitute the first homogeneous group [8, 9]. After removing all efficient units we applied DEA algorithm to the remaining set. This resulted in distinguishing the next group of units. The procedure was repeated until the number of DMUs in the remaining group was not sufficient to perform further divisions. In our calculations we have applied input-oriented BCC model. The model can be formulated in the following way.

*corresponding author; e-mail: urszula_grzybowska@sggw.pl

[†]For two or more correlated defaults probit model can be used.

Let us assume that we have n DMUs, denoted by DMU_o, $o = 1, 2, ..., n$. We denote by x_{ij}, $i = 1, 2, ..., m$ the inputs and by y_{rj}, $r = 1, 2, ..., s$ the outputs for $j = 1, 2, ..., n$. For each DMU_o, $o = 1, ..., n$, described by the inputs x_{io}, $i = 1, 2, .., m$ and outputs y_{ro}, $r = 1, 2, ..., s$ the efficiency measure θ_o is the solution of the following problem: $\theta_o^* = \min \theta_o$ subject to

$$\sum_{j=1}^{n} x_{ij}\lambda_{jo} \leq \theta_o x_{io}, \quad i = 1, 2, ..., m, \tag{2.2}$$

$$\sum_{j=1}^{n} y_{rj}\lambda_{jo} \geq y_{ro}, \quad r = 1, 2, ..., s, \tag{2.3}$$

$$\sum_{j=1}^{n} \lambda_{jo} = 1, \quad j = 1, 2, ..., n. \tag{2.4}$$

A very important issue in DEA approach is variable selection that involves also division of variables into inputs and outputs. Variable selection in DEA is usually based on expert knowledge. In our calculations we have decided to follow the choice of financial ratios suggested by experts and compare it with a selection of variables obtained with help of data mining ensemble methods: random forests and gradient boosting [10–13].

Random forests were introduced in 2001 by L. Breiman as a method of classification [13]. In this approach a large number of simple trees is constructed with a random sample of predictors taken before each node is split. The object is classified based on an average vote of the set of de-correlated trees [10]. One can use random forests to rank the importance of variables in a classification problem. The importance of predictors can be measured in terms of a Gini or Breiman importance measures [10, 12]. Random forests and gradient boosting [10–12] are extensions of regression trees, that is simply the partition of the space X, which consists of predictors of target variable y, into disjoint regions R_j.

The relevant algorithms were implemented in R package randomForest and SAS Enterprise Miner. The main advantage of random forests and gradient boosting approach is their high performance on a large set of variables. Their application for economic data does not require examining the structure of financial ratios, their interactions or correlations.

3. Data and results of the research

We have conducted our analysis on a set of companies traded on Warsaw Stock Exchange (WSE). We have used financial indicators published in financial reports to describe the companies under consideration. The sets of financial indicators applied in DEA by various authors differ considerably [14, 15]. In our calculations we have decided to follow the expert knowledge and choose Assets Turnover and Total Liabilities/Total Assets (Debt Ratio) as input indicators and Return on Assets (ROA), Return on Equity (ROE), Current Ratio (CR), Operating profit margin (OPM) as output indicators. Our data for a set of 68 production companies traded on WSE with quarterly financial reports covered two years: 2011 and 2012. The results of our calculations are shown in column DEA1 of Table I. We have distinguished 5 groups of homogeneous objects. The first group consists of the best 10 companies. One can venture an opinion that for these companies the probability of default is very low. We were not interested in examining the ways of improving efficiency of the remaining companies but in division into groups of similar objects. We were also interested in selecting variables that determine obtained DEA classification. In order to select variables that influence division into DEA groups we have applied two ensemble methods: random forests and gradient boosting. The calculations were done both in SAS (ver. 13.2) and R (ver. 3.1.0). We have used 21 financial indicators, which were divided into four groups: profitability ratios, liquidity ratios, activity ratios and debt ratios. The results are shown in Table II.

TABLE I

DEA rating for 68 production companies.

No.	Company	DEA1	DEA2	No.	Company	DEA1	DEA2
1	AC	1	1	35	MIESZKO	5	7
2	ALKAL	3	3	36	MOJ	4	7
3	AMICA	4	7	37	MUZA	4	5
4	APATOR	2	1	38	NOVITA	3	4
5	BERLING	1	3	39	PAMAPOL	5	7
6	BIOMAXIM	3	4	40	PANITERE	1	1
7	BSCDRUK	2	2	41	PATENTUS	4	5
8	BUDVAR	3	4	42	PEPEES	3	4
9	CIGAMES	2	1	43	PGE	1	1
10	CITYINTE	2	2	44	PLASTBOX	5	6
11	DEBICA	3	5	45	POLICE	1	3
12	DUDA	2	2	46	POLNA	3	2
13	EKOEXP	1	1	47	POZBUD	4	4
14	ENERGOIN	5	7	48	PROJPRZM	4	5
15	ERG	5	6	49	PULAWY	1	1
16	ESSYSTEM	2	3	50	RAFAKO	4	7
17	FASING	4	6	51	RAFAMET	5	6
18	FERRO	5	7	52	RELPOL	3	4
19	FERRUM	5	7	53	SNIEZKA	5	1
20	FORTE	4	4	54	SONEL	2	2
21	GRAAL	5	7	55	STALPROD	2	2
22	GROCLIN	5	7	56	STOMIL	3	3
23	HUTMEN	3	5	57	SUWARY	5	7
24	HYDROTOR	2	1	58	SYNEKTIK	4	3
25	INTEGER	4	2	59	TAURON	4	4
26	INTERCAR	3	6	60	VISTULA	5	7
27	INVICO	4	6	61	WAWEL	2	2
28	IZOLJAR	1	5	62	WIELTON	5	7
29	IZOSTAL	3	4	63	WINDMOB	1	2
30	KPPD	2	5	64	WOJAS	5	7
31	LENTEX	5	6	65	ZPCOTM	5	6
32	LOTOS	3	5	66	ZPUE	4	6
33	MEGAR	2	2	67	ZUE	4	5
34	MENNICA	1	3	68	ZUK	3	4

TABLE II

Variables importance in various ensemble methods.

SAS Miner Random forests		SAS Miner Gradient boosting		R-CRAN randomForest	
Variable	Gini's importance	Variable	Variable importance	Variable	Variable importance
ROA	0.047	ROA	1	EBIT	2.99
RC	0.023	RC	0.770	RT	2.59
ROE	0.022	DSR	0.721	DSR	2.50
GPM	0.022	GPM	0.671	ROA	2.47
DSR	0.021	EBIT	0.545	AT	2.45
NPM	0.014	RT	0.531	NPM	2.08
OPM	0.010	ROE	0.525	RC	1.90
CR	0.009	AT	0.514	GPM	1.82
DR	0.009	CR	0.496	ROE	1.79
QR2	0.007	NPM	0.450	QR1	1.64
QR1	0.007	CCC	0.431	OPM	1.48
RT	0.006	GPMoS	0.426	AR	1.39
AR	0.006	DR	0.386	DR	1.33
EBIT	0.005	QR1	0.363	QR2	1.24
AT	0.005	OC	0.353	RA	1.22
OC	0.002	OPM	0.346	CR	1.15
RA	0.002	QR2	0.332	OC	1.08
GPMoS	0.002	AR	0.279	GPMoS	0.94
WC	0.001	IT	0.264	CCC	0.76
CCC	0.001	RA	0.232	WC	0.73
IT	0.000	WC	0.149	IT	0.63

In our further analysis we have decided to use five indicators that were simultaneously distinguished by at least two of applied ensemble methods: Liabilities Turnover (RC), ROA, Debt to EBITDA (EBIT), EBITDA/Financial expenses (DSR) and Gross Profit Margin (GPM). Two ratios can be regarded as input: Debt to EBITDA and Liabilities turnover (RC). The other ratios, Return on Assets (ROA), EBITDA/Financial expenses (DSR), and Gross Profit Margin (GPM), can be regarded as output. The influence of the indicator ROE proved to be unimportant.

TABLE III

Probabilities of default for obtained rating groups for $\alpha = 0.9$.

Group	No. of elements	PD
1	9	0.033
2	10	0.038
3	7	0.046
4	10	0.053
5	9	0.069
6	9	0.095
7	14	0.152

After performing DEA again for selected set of ratios we have obtained 7 groups of companies. The results of the division are shown in column DEA2 of Table I. The first group of efficient objects consists of 9 companies, for which the sufficiency score was equal to 1. The second group consists of 10 companies, etc. The division into 7 DEA groups is more precise but, with minor exceptions, reflects previous ordering. The correlation coefficient between both assignments to DEA groups is quite high. It is equal 0.78. The obtained division into homogeneous groups of objects can be used for calculating PD. The relevant results are shown in Table III.

4. Concluding remarks

In the paper we propose a new approach to classification of companies based on DEA that can e.g., be used in PD calculation especially in a situation where standard scoring methods were not adequate due to a low number of events. The method we propose can be regarded as an alternative approach to classical statistical classification methods although based on a different philosophy. We have shown, on the example, that application of random forests and gradient boosting provides a good tool for variable selection. Both methods, random forests and gradient boosting, are particularly well suited to the search for factors that could be used in DEA because of their response to highly local features of the data and possibility of using in cases with small numbers of observations without risk of over-fitting. The other advantage is that DEA approach can give an insight into the performance of considered object and can provide instructions for increasing the performance of objects that are inefficient. The only obstacle is that DEA can be applied only to the sets of similar objects i.e., described by similar indicators. Random forests and gradient boosting can be expected to improve the automation of procedures to evaluate the status of companies as well as individual borrowers by banks and other financial institutions. In addition, these methods go on under the supervision demands for use in risk calculations the objective, repeatable methodology.

References

[1] Capital requirements regulation and directive, European Commission, 2006, http://ec.europa.eu/finance/bank/regcapital/legislation-in-force/index_en.htm.

[2] Capital requirements regulation and directive, European Commission, 2009, http://ec.europa.eu/finance/bank/regcapital/legislation-in-force/index_en.htm.

[3] Capital requirements regulation and directive, European Commission, 2010, http://ec.europa.eu/finance/bank/regcapital/legislation-in-force/index_en.htm.

[4] Capital requirements regulation and directive, European Commission, 2013, http://ec.europa.eu/finance/bank/regcapital/legislation-in-force/index_en.htm.

[5] L. Dzidzeviciute, *Ekonomika* **91**, 132 (2012).

[6] K. Pluto, D. Tasche, *Estimating Probabilities of Default for Low Default Portfolios* in: *The Basel II Risk Parameters*, Eds. B. Engelmann, R. Rauhmeier et al., Springer, Berlin 2006, p. 79.

[7] W.W. Cooper, L.M. Seiford, K. Tone, *Introduction to Data Envelopment Analysis and Its Uses with DEA-Solver Software and References*, Springer, New York 2006.

[8] B. Kaczmarska, *Econometrica* **20**, 79 (2010).

[9] P. Andersen, N.C. Petersen, *Manage. Sci.* **39**, 1261 (1993).

[10] R.A. Berk, *Statistical learning from a regression perspective*, Springer, New York 2008.

[11] J. Koronacki, J. Ćwik, *Statystyczne systemy uczące się*, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2008.

[12] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning. Data Mining, Inference and Prediction*, Springer, New York 2009.

[13] L. Breiman, *Mach. Learn.* **45**, 5 (2001).

[14] A. Feruś, *Bank i Kredyt* **37**, 44 (2006).

[15] E. Chodakowska, K. Wardzińska, *Quantitative Methods in Economics* **14**, 74 (2013).