QUANTITATIVE METHODS IN ECONOMICS
Vol. XV, No. 2, 2014, pp. 94 –101

FAMILIES OF CLASSIFIERS – APPLICATION IN DATA ENVELOPMENT ANALYSIS

Urszula Grzybowska, Marek Karwański
Department of Informatics
Warsaw University of Life Sciences – SGGW
e-mails: urszula_grzybowska@sggw.pl; marek_karwanski@sggw.pl

Abstract: Economic description of firms and companies is based on a number of indicators. The indicators are related to each other and can be considered only in a specific context. Regression models allow for such approach. Unfortunately, the problems we deal with are usually nonlinear and the choice of relevant information is very difficult. The aim of the paper is to present a method of variable selection based on random forest and gradient boosting approach and its application to companies ranking in DEA method. The results will be compared with the ordering obtained using expert supported approach for variable selection in DEA.

Keywords: random forests, gradient boosting, DEA, rating classes, variable selection, ranking, high rated portfolio

INTRODUCTION

In many economic issues it is essential for a decision maker to obtain a ranking of entities under consideration. So is the application of internal-rating based approach to estimation probabilities of default (PDs) for the bank obligors. One of the obstacles connected with PD estimation is a low number of defaults, especially in high rating grades. High rating categories might experience many years without any default, for example a part of bank assets called Low Default Portfolios (LDP). These portfolios may consist of assets of the same type, e. g. trust funds. Several methods have been proposed to estimation of PD for LDP [Dzidzevičiūtė 2012]. The only key assumption in the method is a correct ordinal rating of borrowers. Therefore we propose a method of rating which is based on efficiency measure given by Data Envelopment Analysis (DEA). We illustrate our research on an example. The example presents rating in a group of companies from the production sector traded

on Warsaw Stock Exchange (WSE). Our approach involves application of financial indicators describing financial standing of considered firms. The most important thing in DEA approach is a proper selection of indicators and their assignment to output and input. In our research we present two approaches. At first we use the set of indicators suggested by experts to obtain DEA rating. Next we apply ensemble classifiers: random forests and gradient boosting, to select indicators that influence the division into classes. We also compare the results and draw conclusions.

METHOD

In our approach we apply Data Envelopment Analysis (DEA), see for example [Cooper et al. 2006], to obtain division of companies into homogeneous groups. DEA is an Operation Research approach for evaluating the performance of a set of peer entities called Decision Making Units (DMU). DEA can be applied to a wide variety of activities. It can be used to evaluate the performance of governmental agencies, hospitals, universities, non-profit organizations, banks, firms. 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. Traditionally, the efficiency is measured as the output to input ratio. In DEA approach the output and input are linear combinations of variables describing performance of the DMU and the efficiency score is obtained by solving linear programming problems in their primal or dual form. The DMUs with the efficiency score equal to 1 are called efficient. The exception is a super-efficiency DEA where the efficiency score can be greater than 1 in input orientation [Andersen et al.1993]. An important advantage of the method is that the inputs and outputs can be measured in various units. Calculation of the efficiency can be helpful in improving productivity and performance of an inefficient DMU. We have however concentrated our efforts not on efficiency measure but on distinguishing groups of similar i. e., 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 group – see for example [Kaczmarska 2010]. 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. The most important obstacle is that the results obtained with DEA refer only to the considered set of DMUs and can be neither generalized nor compared with results concerning even slightly differing sets of objects, not to mention sets of different objects. There are many various DEA models. In our calculations we have applied input-oriented BCC model. The model can be formulated in the following way:

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

$\theta_0^* = min\theta_0$ subject to

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

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

$$\sum_{j=1}^{n} \lambda_{jo} = 1,\ \lambda_{jo} \ge 0 \quad j = 1,2,\dots,n \tag{3}$$

A very important issue in DEA approach is variable selection that involves also division of variables into inputs and outputs. A variable classified as an output should have a positive correlation with efficiency while a variable classified as an input should have a negative correlation with efficiency (see [Demirova 2010]). Variable selection in DEA is usually based on expert knowledge and is subject to many discussions during scientific conferences. 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 ensemble methods: random forests and gradient boosting [Berk 2008, Hastie et al. 2009, Koronacki et al. 2008].

Random forests were introduced in 2001 by L. Breiman as a method ofclassification [Breiman 2001]. In this approach a large number of unpruned trees is constructed with a random sample of predictors taken before each node is split. The object is classified based on a majority vote of the full set of trees [Berk 2008]. 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 index or by Breiman's importance measure [Breiman 2001, Berk 2008].

Random forests and gradient boosting [Berk 2008, Hastie et al. 2009, Koronacki et al. 2008] 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. Let f be the prediction function for regression tree (sometimes simply referred to as a tree):

$$x \in R_j \Rightarrow f(x) = \hat{y}_j \tag{4}$$

Thus regression tree can be represented as

$$T(x;\Theta) = \sum_{j=1}^{J} \hat{y}_j I\left(x \in R_j\right), \tag{5}$$

where $\Theta = \left\{R_j, \hat{y}_j\right\}_{j\in\{1,\dots,J\}}$.

The idea behind random forest is to build a large collection of de-correlated trees and then to average prediction functions. Each tree was constructed based on a random selection of the predictor variables. After B such trees $\{T(x,\Theta_b)\}_{b\in\{1,\dots,B\}}$ are grown the random forest predictor is:

$$\hat{f}^{B}_{random\ forest} \Leftrightarrow \frac{1}{B}\sum_{b=1}^{B} T(x, \Theta_b) \tag{6}$$

Gradient boosting prediction function is yield by formula

$$\hat{f}_{gradient\ boosting} \Leftrightarrow T(x, \Theta_g) \tag{7}$$

where the parameters Θ_g should be found by minimizing the loss function L [Hastie et al. 2009]:

$$\widehat{\Theta} = \arg\min_{\Theta} \sum_{j=1}^{J} \sum_{x_i \in R_j} L(y_j, \hat{y}_j) \tag{8}$$

The solution can be constructed in an iterative way. At m-th iteration it is needed to find:

$$\widehat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i, \Theta_m)) \tag{9}$$

The above equation can be reformulated as numerical optimization task analog to steepest descent method,

$$f_m = f_{m-1} - \rho_m g_m \tag{10}$$

where ρ_m is the *step length* and g_m is gradient vector:

$$\begin{bmatrix} g_{1m} \\ \cdots \\ g_{Nm} \end{bmatrix} = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=f_{m-1}(x_i)} \tag{11}$$

The difference between stochastic gradient boosting and an ordinary steepest descent is at the points x_i. Gradient boosting should be applied to the new points that are not represented in training set X used by optimization procedure. The simple solution is to induce a tree *f* using square error to get the tree as close as possible to the gradient vector

$$\widehat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} w_i \left(-g_{im} - T(x_i, \Theta)\right)^2 \tag{12}$$

In our calculations we used weights w_i derived from multinomial distribution, i.e., we use multinomial deviance as a loss function.

The relevant algorithms were implemented in R package randomForest and SAS 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.

RESULTS OF THE RESEARCH

The sets of financial indicators applied in DEA by various authors differ considerably [Feruś 2006, Demirova 2010, Chodakowska et al. 2013]. 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 76 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 2. We have distinguished 6 groups of homogeneous objects. The first group consists of the best 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 20 financial indicators, which were divided into four groups: profitability ratios, liquidity ratios, activity ratios and debt ratios. The results are shown in Table 1.

Table 1. Variables importance in various ensemble methods

	R-CRAN randomForest		SAS Miner Random forests		SAS Miner Gradient boosting	
	Ratio	Variable importance	Ratio	Gini coefficient	Ratio	Variable importance
1	RC	3.91	ROA	0.043	RC	1
2	ROA	3.39	GPM	0.025	EBIT	0.983
3	RT	2.96	RC	0.022	ROA	0.935
4	EBIT	2.64	DSR	0.018	GPMoS	0.772
5	GPM	2.47	OPM	0.016	QR1	0.696
6	DR	2.39	NPM	0.015	DSR	0.691
7	GPMoS	2.07	CR	0.012	ROE	0.678
8	QR1	2.07	EBIT	0.009	WC	0.625
9	OPM	2.04	DR	0.009	DR	0.616
10	NPM	2.02	QR2	0.008	GPM	0.613
11	DSR	1.98	ROE	0.008	AR	0.580
12	QR2	1.95	QR1	0.008	IT	0.565
13	ROE	1.72	AR	0.005	CR	0.539
14	AR	1.56	RT	0.003	OPM	0.533
15	CR	1.48	OC	0.003	NPM	0.498
16	CCC	1.44	WC	0.002	CCC	0.491
17	RA	1.37	CCC	0.002	RT	0.474
18	IT	1.31	GPMoS	0.002	RA	0.432
19	OC	1.1	IT	0.002	OC	0.306
20	WC	0.86	RA	0.002	QR2	0.294

Source: own calculations

We have decided to use four indictors that were simultaneously distinguished by at least two of applied ensemble methods: Liabilities Turnover (RC), ROA, Debt to EBITDA (EBIT) and Gross Profit Margin. Two ratios can be regarded as input: Debt to EBITDA and Liabilities turnover (RC). The other ratios, Return on Assets (ROA) and Gross Profit Margin (GPM), can be regarded as output.

Table 2. DEA rating for 76 production companies and their efficiency scores

Company	DEA1	DEA2	Eff.	Company	DEA1	DEA2	Eff.
AC	1	1	1.00	DEBICA	4	6	0.48
APATOR	2	1	1.00	IZOSTAL	3	5	0.47
CIGAMES	2	1	1.00	PATENTUS	4	6	0.43
CITYINTE	2	1	1.00	ZPUE	4	7	0.42
EKO_EXP	1	1	1.00	BIOMAXIM	3	4	0.42
HYDROT.	1	1	1.00	ZUE	6	5	0.41
PANITERE	1	1	1.00	WINDMOB	1	2	0.41
PGE	1	1	1.00	MIESZKO	6	8	0.40
PULAWY	2	1	1.00	ZPC_OTM	6	8	0.38
SONEL	2	1	1.00	ZYWIEC	2	4	0.37
WAWEL	2	1	1.00	MOJ	4	8	0.37
ZELMER	1	1	1.00	POLNA	3	2	0.37
BERLING	1	2	0.95	INTERCAR	3	7	0.37
DUDA	3	2	0.91	INVICO	4	7	0.36
RELPOL	3	3	0.85	SUWARY	6	8	0.36
MEGAR	2	2	0.81	PLASTBOX	5	9	0.35
BSCDRUK	2	2	0.78	ENERGOIN	6	8	0.34
STALPROD	2	3	0.77	AMICA	5	8	0.33
SYNEKTIK	5	3	0.75	PAMAPOL	6	10	0.31
ESSYSTEM	2	3	0.73	FERRO	5	8	0.29
MENNICA	1	3	0.73	WIELTON	6	10	0.28
POLICE	2	2	0.68	MUZA	6	5	0.28
NOVITA	3	4	0.67	POZBUD	4	4	0.28
BUDVAR	3	4	0.67	FASING	5	7	0.27
ALKAL	3	3	0.66	BORYSZ.	4	7	0.27
TAURON	4	3	0.63	INTEGER	4	2	0.27
HUTMEN	3	4	0.61	RAFAMET	5	8	0.27
IZOL_JAR	2	5	0.60	SNIEZKA	6	9	0.26
KETY	3	3	0.60	GROCLIN	5	10	0.23
FORTE	4	4	0.59	VISTULA	6	10	0.22
LOTOS	4	5	0.57	GRAAL	6	10	0.20
STOMIL_S	3	4	0.56	FERRUM	6	10	0.20
ZUK	3	5	0.54	RAFAKO	5	9	0.20

Company	DEA1	DEA2	Eff.	Company	DEA1	DEA2	Eff.
LENTEX	5	6	0.54	WOJAS	6	10	0.17
KPPD	3	6	0.54	KOELNER	6	10	0.14
PROJPRZM	4	5	0.53	RAWLPL.	6	10	0.14
PEPEES	3	5	0.53	GRAJEWO	6	10	0.14
ERG	5	7	0.48	ARMATUR.	6	10	0.12

Source: own calculations

After performing DEA again for selected set of ratios we have obtained 10 groups of companies. The results of the division are shown in column DEA2 of Table 2. The column Eff. contains relevant efficiency measure for each DMU. The first group of efficient objects consists of 12 companies. The second group consists of 8 companies, etc. It has to be noticed, that the ordering given by efficiency measure does not reflect the ranking of companies given by DEA groups (compare [Chodakowska et al. 2013]). For example, firms with quite low efficiency score were assigned to the second or third DEA group. The division into 10 DEA groups is more precise but, with minor exceptions, reflects previous ordering. The correlation coefficient between both assignments to DEA groups is high. It is equal 0.87.

CONCLUSIONS

In the paper we propose a new approach to classification of companies based on DEA. The method can be regarded as an alternative approach to classical statistical classification methods. 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 overfitting.

Application of ensemble methods seems to be a promising approach to variable selection for the needs of DEA. Our calculations repeated on the group of 17 construction companies revealed that the ratios distinguished by ensemble methods differ depending on the companies' profile. Moreover, membership into DEA groups will be violated even if the set of considered DMUs will differ by one object only. Nevertheless, DEA seems to be a promising tool, alternative to traditional scoring models. It enables ranking of agents and it can be used for distinguishing classes of homogeneous object, e.g., rating classes. The support of ensemble methods in variable selections makes DEA approach an universal tool.

Random forests and gradient boosting can be expected to improve the automation of procedures to evaluate the status of companies by banks and other financial institutions.

REFERENCES

Andersen P., Petersen N. C. (1993) A Procedure for Ranking Efficient Units in Data Envelopment Analysis, Management Science, Vol. 39, pp.1261-1264.

Berk R. A. (2008) Statistical learning from a regression perspective, Springer, New York.

Breiman L. (2001) Random Forests, Machine Learning, Vol. 45 (1), pp. 5-32.

Chodakowska E., Wardzińska K. (2013) The attempt to create an internal credit risk rating of production companies with the use of Operational Research method, Quantitative Methods in Economics, Vol. XIV, No. 1, pp. 74-83.

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

Demirova M. (2010) An empirical application of data envelopment analysis in credit rating. Theses and dissertations, Paper 981, Ryerson University, Canada.

Dzidzevičiūtė L. (2012) Estimation of default probability for low default portfolios, Ekonomika 2012, Vol. 91 (1), pp.132-156.

Feruś A. (2006) The Application of the DEA Method to Define the Level of Company Credit Risk, Bank i Kredyt, Vol. 37, No. 7, pp. 44-59.

Hastie T., Tibshirani R., Friedman J. (2009) The elements of statistical learning. Data Mining, Inference and Prediction, Second Edition, Springer, New York.

Kaczmarska B. (2010) The Data Envelopment Analysis Method in Benchmarking of Technological Incubators, Operations Research and Decisions, Vol. 20, No. 1, pp. 79-95.

Koronacki J., Ćwik J. (2008) Statystyczne systemy uczące się, Akademicka Oficyna Wydawnicza EXIT, Warszawa.