

A CREDIT SCORE MODEL ON ROMANIAN REAL ESTATE MARKET USING SAS

Ionela-Cătălina Zamfir¹
Ana-Maria Mihaela Iordache²

ABSTRACT

Nowadays, many banks have their own system to identify good clients from bad clients. Even though we talk about persons or companies, many banks rely their decision for giving a loan on the loans history from their databases systems, the financial results of the entity (revenue for person and profit for companies), or the "friendship" between the bank and the client (a bank's friend is an old client). Very few banks develop a market analysis or a credit score model in order to analyze the market on which the company operates. In this article, we propose a credit score model, taking into account 12 activity indicators for 21 companies in the real estate market. Using the principal component analysis, cluster analysis and discriminant analysis, we obtain two score functions that can make the difference between companies with a low risk of bankruptcy and companies with a high risk of not refunding the loan.

Keywords: real estate market, credit score, principal components analysis, cluster analysis, discriminant analysis

1. INTRODUCTION

Starting from the Altman (1968) studies about linear discriminant analysis used to predict bankruptcy the recent studies reveal a much more complicate problem when we talk about bankruptcy and models that can predict this fact. The credit score model is very important for a bank's management system, because, using a model like this, a manager can decide whether he gives the loan or not.

The main goals of this article are to identify the variables that fit to a bankruptcy predictable model in the real estate field and to determine the function (discriminant function) that can "decide" if a new firm (that is not analyzed in the model) may get a loan or not according to the real estate market. Finally, we will provide the answer to the questions: "what is the function (or functions) that can help a bank manager to decide giving the loan to a listed company from the real estate field?", "is this model a good model that a manager can rely on?", and "how precise is this model?".

Section 2 of the article presents a part of the existing literature in this domain and some of the newest articles about bankruptcy's prediction. Section 3 reveals the methodology used to create the discriminant functions, starting from the principal component analysis, in order to extract the unique information from the dataset, continuing with the cluster analysis to classify the companies and ending with the discriminant analysis. Section 4 presents the

¹ Phd Student, Economic Cybernetics and Statistics Doctoral School, Bucharest University of Economic Studies

² Assitant Lecturer, Phd, School of Computer Science for Business Management, Romanian-American University, Bucharest

empirical results of the entire analysis, while section 5 reveals and concludes the new ideas identified in the research.

2. LITERATURE REVIEW AND THE METHODOLOGY USED

Using a set of 66 companies from manufacturing area, Altman (1968) used the discriminant analysis in order to determine functions that can predict a new company's bankruptcy. The author developed a z-score formula with 5 financial indicators that proved to be "72% accurate in predicting bankruptcy two years before the event"[5]. For the next 31 years, tests were made on this z-score function and it proved that the accuracy raised up to 80-90%. In this respect, the Altman's work is considered to be the most important z-score for predicting the bankruptcy.

In another study, using three hundred US commercial banks, P. Gurný and M. Gurný (2013) develop credit scoring models and apply them to the sample, in order to find the most accurate one, with the best results.

From methodological point of view, we firstly use the preliminary data analysis, which consists in:

- - transforming the primary dataset (refining the dataset by eliminating the companies that do not provide all information);
- - eliminating the outliers - values for different variables that are not included in the numeric interval $[\mu - 3 * \sigma, \mu + 3 * \sigma]$. This interval contains 99.98% of total information from a statistical sample. The outliers are values that have a very big influence on the mean value, the standard deviation and can influence (in a bad way) the entire analysis.
- - standardizing the dataset - this process involves transforming each value using the formula: $x_{ijnew} = \frac{x_{ijold} - \bar{X}_i}{\sigma_i}$, where: *new* means the new standardized value, *old* means the old value, $\bar{X}_i$ is the average value of the variable *i*, and σ_i is the standard deviation for variable *i*. This process is used when the variables are expressed in different measurement units, in order to eliminate the false influence of a variable (false influence given by the measurement unit). In this way, we obtain variables which have the mean value equal with zero, and the standard deviation equal with one.
- After we build the dataset we use in the further analysis, we use the following methodologies [Ruxanda, 2009] in order to determine the discriminant function that will tell us if a new company is bankrupt or profitable (or if a bank that loan the company will get its money back or it will lose them):
- - The principal component analysis - is the first analysis used in this model, that extracts the essence of dataset by calculating new variables, named principal components. This analysis is used when the correlation matrix of variables in the original dataset denote a high correlation between variables. The principal components are completely new variables, built with the original variables, using the matrix writing formula: $W = \alpha^{(t)} X$, where *W* is the principal components, $\alpha^{(t)}$ is the transpose eigenvectors of the covariance matrix of the original variables, and *X* is the dataset composed by original variables.
- - The cluster analysis is used to classify the companies into two classes: profitable companies and bankrupt companies. There are two major methods in this respect:

hierarchical methods (such as Ward, nearest neighbor, farthest neighbor or average value method) and algorithmically methods (like K-means). Most of these methods are based on the Euclidian distance between the observations (companies in this case), which is: $d(O_1, O_2) = \sqrt{\sum_{i=1}^n (O_{1i} - O_{2i})^2}$, where O_1 represents an object (company or class), O_2 represents another object, and i from 1 to n refers to the n variables from the dataset.

- - The discriminant analysis is the final analysis used to determine the coefficients of the discriminant function. Based on this function, it is possible to calculate a score for a new company. If the score is negative, the company belongs to a class, and if is positive, the company is classified into the other class.

3. EMPIRICAL RESULTS

The entire model is based on an original data set composed by 12 indicators that reflect the return, the solvability, the liquidity or other financial indicators about 21 companies that are traded on the stock market from Romania sharing the same market (real estate market) and published their balance sheet for 2013 year. The indicators used in this analysis and their calculating formulas are presented in table 1 from below:

Table 1. List of used indicators

Code	Indicator	Formula
CP	market capitalization	
RC	commercial return	(net profit / turnover) * 100
ROE	return on equity	(net profit / equity) * 100
ROA	return on assets	(net profit / total assets) * 100
RLC	current liquidity ratio	(current assets / current liabilities) * 100
SP	patrimonial solvency	equity / total assets
SF	financial stability	long-term debt / (long-term debt + equity)
RAF	financial autonomy ratio	(equity / permanent equity) * 100
RSG	general solvency ratio	(total debt / total assets) * 100
VCA	book value of the share	equity / number of shares issued
RVP	rate market value	price of action / book value of the share
EPS	earnings per share	net profit / No of shares issued

From all 21 companies that match the criterions from above, 5 of them were eliminated from the model, because they were considered to be outliers for one or more variables. With 16 companies left, representing almost half of the real estate market companies that are currently traded on the stock market in Romania, the sample is representative.

Table 2. Descriptive statistics

			R O E	R O A				RA F	RS G	VC A	R V P	E PS
	CP	RC			RLC	SP	SF					

Mean	37024	13.	1.	1.	534.4	0.	0.	97.3	12.	18.4	1.	0.
Standard Error	012.36	67	89	71	4	85	02	2	72	8	22	26
Median	15366	6.2	1.	1.	200.2	0.	0.		3.7		0.	0.
Standard Deviation	568.51	3	20	05	0	04	01	0.91	8	8.96	32	17
Sample Variance	85898	9.3	1.	1.	254.2	0.	0.	99.6	8.1		0.	0.
Kurtosis	85.63	0	34	10	1	90	00	7	5	6.91	73	04
Skewness	61466	24.	4.	4.	800.8	0.	0.		15.	35.8	1.	0.
Range	274.04	92	81	20	2	17	04	3.64	11	5	27	68
Minimum	3.7781	620	.1	.6	6413	0.	0.	13.2	228	128	1.	0.
Maximum	E+15	.94	7	8	12.39	03	00	5	.34	4.93	61	47
Sum	2.0278	-	-	-		-	-	-	9.4	13.0	2.	8.
Count	9601	7	10	18	8.02	24	12	0.44	2	8	78	29
	1.8380	0.1	0.	0.		2.	1.	-	2.8		1.	2.
	3942	6	03	01	2.70	14	25	1.07	2	3.50	77	59
	18449	85.	.7	.1	3139.	0.	0.		62.	147.	4.	3.
	3010.2	53	4	0	88	62	10	9.57	39	27	47	08
	11124	-	-	-								-
	5.79	36.	7.	7.	16.77	0.	0.	90.4	1.3		0.	0.
	18460	49	51	14		36	00	3	9	0.30	13	57
	4256	49.	.2	8.	3156.	0.	0.	100.	63.	147.	4.	2.
	59238	04	4	96	65	99	10	00	78	56	60	51
	4197.7	30	27			13					19	
	59238	218	.1	.4	8550.	.6	0.	155	203	295.	.5	4.
	4197.7	.79	8	0	97	5	39	7.08	.49	60	2	21
Count	16	16	16	16	16	16	16	16	16	16	16	16

Source: Excel output

Table 2 from above presents the descriptive statistics for all 12 variables. The mean value represents the average value of the variable. The kurtosis value denote how flat is the probability distribution of each variable: if kurtosis is higher than 3 (for RLC, SP, RSG, VGA and EPS), than, the probability distribution is leptokurtic, which means that the values of the variable are grouped around the mean value, and, if kurtosis is lower than 3 (for all other variables), the probability distribution is platikurtic, which means that the values are spread around the mean on a larger range. On the other hand, a negative skewness (variables like: RC, ROE, ROA, SP and RAF) denotes a left asymmetry of the probability distribution, while a positive value is a right asymmetry.

After the standardization of the dataset (each difference between the old value and the average value is divided by standard deviation), the first major analysis is the principal component analysis.

Table 3. The correlation matrix

CO R	CP	RC	RO E	RO A	RLC	SP	SF	RAF	RSG	VC A	RVP	EP S
CP	1											
	0.26											
RC	1	1										
RO	0.37	0.74										
E	4	5	1									
RO	0.34	0.76	0.98									
A	2	3	7	1								
	-											
RL	0.15	0.35	0.21	0.25								
C	9	5	8	8	1							
	0.04	0.27	0.13	0.15	0.38							
SP	5	8	3	4	2	1						
					-	-						
SF	0.83	0.17	0.20	0.15	0.33	0.12						
	7	2	0	9	4	7	1					
	-	-	-	-		-						
RA	0.80	0.16	0.19	0.15	0.34	0.10	0.95					
F	2	9	7	5	9	1	8	1				
		-		-	-	-		-				
RS	0.02	0.20	0.01	0.05	0.35	0.92	0.21	0.17				
G	6	4	7	6	9	2	6	7	1			
		-	-	-	-			-	-			
VC	0.55	0.12	0.05	0.05	0.11	0.10	0.40	0.43	0.07			
A	6	0	3	1	6	6	2	2	9	1		
	-					-	-			-		
RV	0.03	0.23	0.31	0.25	0.02	0.41	0.11	0.14	0.54	0.24		
P	0	6	9	4	6	1	0	7	8	8	1	
								-	-		-	
EP	0.14	0.45	0.53	0.58	0.24	0.18	0.04	0.06	0.15	0.40	0.04	
S	8	6	8	6	9	1	4	9	2	9	0	1

Source: Excel output

Table 3 from above is the correlation matrix for all 12 variables. Because there are high correlations between some variables (these correlations come from the fact that some indicators use the same balance sheet indicator in their formula, and because of the balance sheet variable relations), it appears the phenomenon named: "informational redundancy", that means that a part of the information from a variable can be found in another variable. Therefore, the principal component analysis is required.

```
proc princomp data=proiect.std cov out=proiect.scoruri n=12
outstat=proiect.rez_ACP plots=all;
var CP RC ROE ROA RLC SP SF RAF RSG VCA RVP EPS;
title 'Principal component analysis n=12';
run;
```

The SAS code from above is the procedure for the principal component analysis, which has the following SAS output:

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.75638943	0.70313266	0.3130	0.3130
2	3.05325677	0.69394406	0.2544	0.5675
3	2.35931271	1.33539757	0.1966	0.7641
4	1.02391514	0.36836779	0.0853	0.8494
5	0.65554735	0.16177292	0.0546	0.9040
6	0.49377443	0.17488994	0.0411	0.9452
7	0.31888449	0.14983106	0.0266	0.9718
8	0.16905343	0.07141244	0.0141	0.9858
9	0.09764099	0.04793454	0.0081	0.9940
10	0.04970645	0.02730417	0.0041	0.9981
11	0.02240227	0.02228572	0.0019	1.0000
12	0.00011656		0.0000	1.0000

According to the Kaiser³ criterion for choosing the number of principal components in order to develop the credit score model, we consider 4 principal components (because 4 eigenvalues of the covariance matrix are higher than 1). Also, the coverage percentage criterion is accomplished, because these 4 principal components synthesize 84.94% of the information contained in the original set of data.

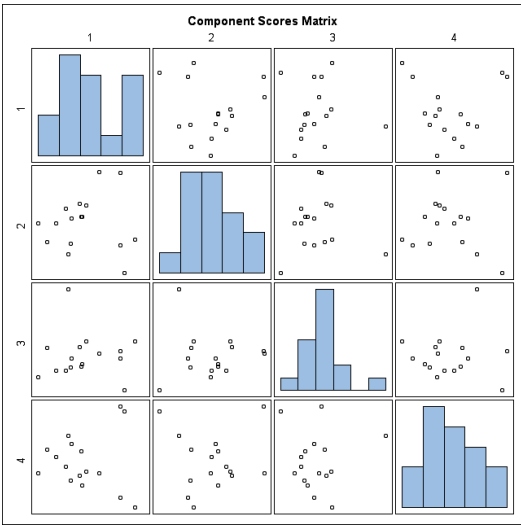


Figure 1. Component Scores Matrix

Source: SAS Output

Figure 1 from above confirms one of the most important characteristics for new variables: the principal components are not correlated one with another. Or, the correlation coefficient

³ Kaiser criterion can be used, because we used the standardized dataset for principal component analysis

between component W_i and component W_j is null, where $i \neq j$. In this way, the redundancy in information is eliminated.

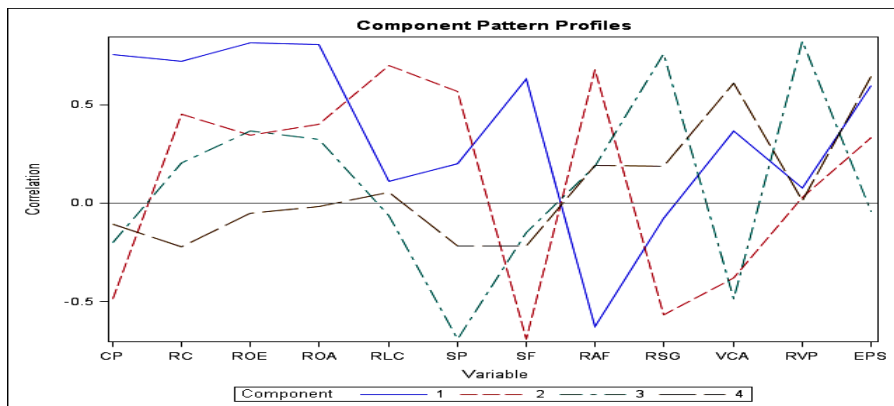


Figure 2. Component Pattern Profiles

Source: SAS Output

Figure 2 presents the component pattern profiles, or the correlation coefficients between the original variables, and the principal components. Like this, we can name the new variables that we work with, according to the economical information from inside. Therefore, we have new variables like:

- - W_1 takes more information from CP, ROA, ROE, RC and SF, and can be named: the performance variable
- - W_2 takes more information from RLC and RAF, and can be named: the independence variable
- - W_3 takes more information from RSG and RVP, and can be named: the market solvency
- - W_4 takes more information from VCA and EPS, and can be named: the value of the share

```
/*the partitioning algorithms -> K-MEANS*/
proc fastclus data=proiect.scoruri out=proiect.w1 maxclusters=2;
var prin1 prin2 prin3 prin4;
title 'K-means algorithm';
run;
```

The SAS code from above represents the second major analysis, the cluster analysis, made in SAS, using the fastclus procedure, and the K-means algorithm. Because of the fact that there are analyzed only 16 companies, we split the sample into 2 major classes: a profitable class, and a bankruptcy class. It is important to notice that, in reality, the companies that are in the bankruptcy class may not bankrupt in the future, but, for a bank that gives a loan to these companies, is a higher risk that these companies to not refund the loan, than for the other companies. The most important SAS output of this procedure is:

The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=2 Maxiter=1

Initial Seeds				
Cluster	Prin1	Prin2	Prin3	Prin4
1	-1.061964328	-2.035614986	4.598516190	0.939077098
2	2.746806583	-3.240778721	-2.369417307	1.834654541

Cluster Means				
Cluster	Prin1	Prin2	Prin3	Prin4
1	-1.091651817	0.362483922	0.369856148	0.033707695
2	1.403552336	-0.466050756	-0.475529334	-0.043338464

In this respect, the two classes are:

- - 7 companies into the first class
- - 9 companies into the second class

Looking back in the original dataset, after classifying the companies, it is easy to notice that the first class represents the profitable companies (or companies with a lower risk for the bank), and the second class has the companies with a higher bankruptcy risk.

```

title 'The discriminant analysis';
proc candisc data=proiect.date2 all out=proiect.dscr;
class cluster;
run;
proc discrim data=proiect.date2 crossvalidate crosslist
testdata=proiect.date2 LIST testout=proiect.dscr2;
class cluster;
run;

```

The SAS code from above represents the final analysis: the discriminant analysis, when the discriminant function can be determined, in order to use it for new companies, that are not included in this study. The main SAS output is:

The discriminant analysis				
The CANDISC Procedure				
Total Sample Size	16	DF Total	15	
Variables	4	DF Within Classes	14	
Classes	2	DF Between Classes	1	
Number of Observations Read			16	
Number of Observations Used			16	
Class Level Information				
Variable				
CLUSTER	Name	Frequency	Weight	Proportion
1	_1	9	9.0000	0.562500
2	_2	7	7.0000	0.437500
Linear Discriminant Function for CLUSTER				
	Variable	1	2	
	Constant	-0.46064	-0.76146	
	Prin1	-0.63841	0.82082	
	Prin2	0.26080	-0.33532	

```

Prin3          0.34438      -0.44277
Prin4          0.07232      -0.09298

The DISCRIM Procedure
Classification Results for Calibration Data: PROJECT.DATE2
Resubstitution Results using Linear Discriminant Function
Posterior Probability of Membership in CLUSTER
Classified
From          into
Obs   CLUSTER CLUSTER      1      2
1      1      1      0.8434  0.1566
2      1      1      0.8663  0.1337
3      1      1      0.8608  0.1392
4      1      1      0.9765  0.0235
5      1      1      0.6884  0.3116
6      1      1      0.8148  0.1852
7      1      1      0.9066  0.0934
8      1      1      0.9880  0.0120
9      1      1      0.9734  0.0266
10     2      2      0.3090  0.6910
11     2      2      0.0080  0.9920
12     2      2      0.0110  0.9890
13     2      1 *      0.5708  0.4292
14     2      1 *      0.5236  0.4764
15     2      1 *      0.5254  0.4746
16     2      2      0.0007  0.9993
* Misclassified observation
Number of Observations and Percent Classified into CLUSTER

From CLUSTER      1      2      Total
1      9      0      9
100.00      0.00      100.00
2      3      4      7
42.86      57.14      100.00
Total      12      4      16
75.00      25.00      100.00
Priors      0.5      0.5
Error Count Estimates for CLUSTER
1      2      Total
Rate      0.0000      0.4286      0.2143

```

According to the output above, the linear discriminant function is:

- - for class 1: $D_1(x) = -0.46 - 0.63*W_1 + 0.26*W_2 + 0.34*W_3 + 0.7*W_4$
- - for class 2: $D_2(x) = -0.76 + 0.82*W_1 - 0.33*W_2 - 0.44*W_3 - 0.09*W_4$

Where: W_1 , W_2 , W_3 , and W_4 are the 4 principal components analyzed. The output also shows all 16 companies reclassified using the discriminant functions identified. There are 3 misclassified observations, fact that diminishes the accuracy of the analysis. Therefore, the probability to classify correctly a new company into one of the two classes, using the functions obtained and taking into account the highest calculating score is: $(13/16)*100=81.25\%$.

4. CONCLUSIONS

Finally, we have discovered the answers for the questions: "what is the function (or functions) that can help a bank manager to decide giving the loan to a listed company from the real estate field?", "is this model a good model that a manager can rely on?", and "how precise is this model?", by making a quantitative research, using three of the most important data analysis methods: principal component analysis (reducing the number of variables, and extracting the essence from the dataset), cluster analysis (classifying the companies) and discriminant analysis (identifying the functions and their accuracy).

The most important conclusion is that a bank manager can use a similar analysis in order to decide if a company "deserves" a loan (the bank can recover the money, plus the interest rate, and this type of loan is considered to be efficient) or not (the bank gives inefficient loans and lose money, by not recovering them). Even thou this type of analysis is not unique (the bank manager may take into account the "friendship" with the company that requests a loan, or the historical data on the loans so far), it provides useful information, and if there are taken into account more companies in a field, the results reflect accurately the economic reality. For further researches, we propose to study more economic areas, using different other methods, in order to identify the most accurate method for analyzing a market, and the company that needs a loan.

5. REFERENCES

1. Altman, Edward I. (September 1968). ""Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy"". *Journal of Finance*: 189–209.
2. Altman, E., Sabato, G. „Modeling Credit Risk for SMEs: Evidence from US Market”, *SSRN working paper*, Decembre, 2005
3. Gurný, P., Gurný, M., (2013), *Comparison of Credit Score Models on Probability of Default Estimation for US Banks*, Prague Economic Papers, no.2, pp. 163-181
4. Ruxanda, Ghe., (2009) *Analiza multidimensională a datelor*, Academia de Studii Economice - Școala Doctorală, București
5. http://en.wikipedia.org/wiki/Altman_Z-score#Precedents
6. <http://www.goldring.ro/gradul-de-indatorare-teorie.html>
7. http://www.macrostandard.ro/indicatori_analiza_financiara.html
8. <http://www.bvb.ro/Companies/ListedCompanies.aspx?t=4>