DATA MINING AND CUSTOMER RELATIONSHIP MANAGEMENT FOR CLIENTS SEGMENTATION

Ionela-Catalina Tudorache (Zamfir)[1], **Radu-Ioan Vija**[2]
[1), 2)]The Bucharest University of Economic Studies, Economic Cybernetics and Statistics Doctoral School
E-mail: tudorachecata@yahoo.com, radu.vija@gmail.com

Abstract
Starting from the idea that, nowadays, data mining techniques are applied in more and more different domains, one of the most important economic domain is Customer Relationship Management. In this respect, many studies were developed, from market research studies, to clients segmentation. We use principal components analysis to extract essential information from our dataset, and to eliminate redundancy, and k-means algorithm to classify clients of an audit company. Finally, we conclude that data mining techniques can be used for clients segmentation and provide useful results for marketing, products and management departments of the company.

Keywords: classification, clients. CRM, data mining, segmentation

JEL Classification: C38, D49

Introduction and literature review

Knowing what customers want is a one of the most important issues for a company, no matter what services/products that company provides. In this respect, many models and techniques were developed: from statistical segmentation of clients, up to data mining techniques (such as classification models) and neural networks.
When we say clients segmentation, we take into account the CRM (Customer Relationship Management) department of a company, as well as the research department. Many of big companies develop and use their own model to detect their customers behavior, characteristics, patterns and, finally, their preferences.
During the fast decades, many articles were wrote and new models were developed in this area. In 2002, Rygielski, Wang and Yen have reviewed several applications of data mining techniques from: discovery, predictive modeling and forensic analysis areas.
They start from the idea that customer relationship management is possible due to data mining techniques that have become tools that answer business questions regarding customers.
Villanueva and Hansseus (2007) believe that the interest of managers is shifted from product management to customer relationship management. There are two ways to accomplish this goal: to focus on what customers value the most, and to identify the methods to maximize customers equity.
Shaving the same idea, the focus on customers value, Verhoef, Doorn and Dorotic (2007) make a wide literature review and reade to the conclusion that customer lifetime value is a research area that has too few studies.

In 2014, a recent literature review, having as authors Janakiraman and Umanmahes reveals the major role of Data Mining techniques used in customer relationship management. In this regard data mining techniques are used more for classification, clustering and prediction. Studying the pros and cons of data mining techniques, the authors suggest that besides the wide range of applications areas, the only argument against them is related to security and privacy. A big issue is how data is taken, what it is used for and if it is used in unethical way.
There are 5 major parts in this article: section 1 represents the introduction and literature review, section 2 presents the methodologies applied, section 3 is about the dataset used, variables and observations, section 4 is the case study, while the final section presents the conclusions and further areas of interest.

Methodology

From methodological point of view, we used data mining techniques after we collected, selected and filtered the data set. The first part in data processing is eliminating the outliers (observations that influence the analysis by having very big or very small values for one or more variables). Data standardization is the next step before running any analysis. This step is important because there are variables that influence the further analysis by having different measurement units (for examples: turnover variables is measured in billions, while the net result is measured in millions).
For further step is the major analysis runned: principal components analysis. Using this method, we reduce the dimensional space from 24 variables to 6 variables (called principal components). The main reason why this analysis is necessary is the redundancy in data set (the information from one variable is partially retrieved in one or more variables). In this respect, principal component analysis "creates" new variables (principal components) that take a reasonable percent of total information (over 80%) and are not correlated (the redundancy is zero). The principal components model is: $W_j = \sum_{i=1}^{n} \alpha_i^{(j)} \cdot X_i$ (1)
where:

- W_j is principal component number *j* ($j = 1,6$ in this case)
- $\alpha_i^{(j)}$ is component *i* of *j* eigenvector of the covariance matrix (the j eigenvector is associated to the *j* eigenvalue of the covariance matrix). In this case: $i = 1,24$
- X_i is the *i* variable (in this case there are 24 original variables, that are financial or economical indicators or ratios of all companies).

Once the principal scores are calculated for each observations, we use all 6 new variables (after we name them, according to their correlation with original variables) to classify companies.
Recent literature review study reveals that data mining techniques are used especially for classification and prediction. Because the main objective of this study is client segmentation, we use supervised recognition techniques is order to classify clients into profitability and algorithmically method that is used to identify the number of classes taken into account. This method is preferred because it accomplishes in a better way the general criterion of classification (high homogeneity within classes and high heterogeneity between them).
After identifying the number of classes that accomplishes the criterion above, K-Means Algorithm is used to classify the companies.

Database, variables and descriptive statistics

The database used in this paper is represented by the main clients of Deloitte company.
We used a sample of a confidential database of Deloitte client database and the variables studied are presented in the below table (data for 2014) : turnover(mld); Net Resultat (mil); Number of employees; LOW 52 Week Low ;HIGH 52 Week High ;Moody Rating ; S&P Rating; Fitch Rating ;Market; ADJ EPS Est Next Yr; Sales LFY; Sales Est Current Yr; Sales Est Next Yr; PE Ratio; Price to Book; Price To Sales; Price to Cash Flow; Price to EBITDA; Return on Assets; Return on Equity; Return on Capital

Table no. 1. Descriptive statistics

	ca	*pr*	*emp*	*p*	*low*	*high*
Mean	839.2	3998.8	49294.7	217.0	175.3	289.7
Std Error	813.8	1169.7	13632.4	98.2	78.7	130.4
Median	6.9	1100.0	13000.0	42.2	37.7	49.6
Std Dev	5578.8	8019.0	93458.7	673.1	539.3	894.1
Kurtosis	47.0	12.4	7.3	23.2	21.2	17.8
Skewness	6.9	3.4	2.7	4.7	4.5	4.3
Range	38270.0	39996.0	433362.0	3985.0	3100.0	4365.9
Minimum	0.0	4.1	0.0	0.0	0.0	0.0
Maximum	38270.0	40000.0	433362.0	3985.0	3100.0	4365.9
Count	47.0	47.0	47.0	47.0	47.0	47.0
	adjcy	*adjny*	*sales*	*salcy*	*salny*	*pe*
Mean	12.0	13.0	112983.7	117557.1	122293.8	35.1
Std Error	4.8	5.1	52700.7	54327.1	56109.6	13.9
Median	2.9	3.1	12998.0	12304.4	14117.5	17.7
Std Dev	32.8	35.1	361298.0	372448.0	384668.2	95.3
Kurtosis	14.7	14.1	37.7	37.4	36.9	45.5
Skewness	3.9	3.8	5.9	5.9	5.8	6.7
Range	169.2	177.7	2415230.8	2486615.8	2561181.3	661.5
Minimum	0.0	0.0	34.8	35.8	38.5	7.9
Maximum	169.2	177.7	2415265.6	2486651.6	2561219.7	669.4
Count	47.0	47.0	47.0	47.0	47.0	47.0
	cap	*val*	*gaap*	*gaapcy*	*gaapny*	*adj*
Mean	179568.0	160233.9	12.5	11.7	12.6	12.7
Std Error	59784.2	52327.3	5.6	4.8	5.1	5.6
Median	25293.2	24585.6	2.4	2.4	2.7	2.4
Std Dev	409859.7	358738.0	38.4	32.9	35.1	38.1
Kurtosis	22.8	16.4	14.5	14.7	14.1	14.4
Skewness	4.4	3.9	3.9	3.9	3.8	3.9

Range	2496488.0	1922452.3	186.4	169.2	177.7	186.3
Minimum	233.5	77.9	0.0	0.0	0.0	0.0
Maximum	2496721.5	1922530.3	186.4	169.2	177.7	186.4
Count	47.0	47.0	47.0	47.0	47.0	47.0
	ptb	*pts*	*ptcf*	*ptebitda*	*roa*	*roe*
Mean	3.7	3.3	10.7	9.4	9.9	19.3
Std Error	0.4	0.5	1.0	0.9	1.1	2.1
Median	2.9	2.4	10.5	9.3	8.1	15.7
Std Dev	2.6	3.3	7.2	6.5	7.3	14.6
Kurtosis	4.2	12.4	-0.4	-0.2	1.6	4.0
Skewness	2.0	3.0	0.4	0.5	1.2	1.9
Range	12.5	19.1	26.6	23.6	32.1	69.4
Minimum	0.9	0.3	0.1	0.0	0.0	0.0
Maximum	13.4	19.4	26.6	23.6	32.1	69.4
Count	47.0	47.0	47.0	47.0	47.0	47.0

source: Excel output

The table from above is the descriptive statistics for the variables. For almost all variables, the average value and the standard deviation is representative from statistical point of view.

Application and results

As we mentioned in the methodology part, the main reason for this type of analysis is information redundancy that is given by correlation matrix. In this respect, a 0.13 correlation is between turnover and net result, 0.29 between number of employees and net result, 0.95 between market capitalisation and enterprise value, 0.74 between enterprise value and estimated sales for next year; 0.9 between price to cash flow and price to EBITDA.

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	11.1225518	7.9586335	0.4634	0.4634
2	3.1639183	0.4814582	0.1318	0.5953
3	2.6824601	1.0950852	0.1118	0.7070
4	1.5873749	0.3609720	0.0661	0.7732
5	1.2264029	0.1699400	0.0511	0.8243
6	1.0564629	0.1813789	0.0440	0.8683
7	0.8750840	0.1036241	0.0365	0.9048
8	0.7714598	0.1086045	0.0321	0.9369

Figure no. 1. Eigenvalues of the covariance matrix
source: SAS Output

The figure from above show the eigenvalues of the covariance matrix, and reveals the number of principal components taken into analysis. According to coverage percentage criterion 6 components are enough (because all of them take 86.83% of total information), and, according to Kaiser's criterion (applied only on standardised variables), 6 components are considered (because there are 6 eigenvalues higher than 1).

Taking into account 6 principal components that take over 85% of total information, each component "takes" more information from several variables, and care be named, like:

- W_1 is highly correlated with: p, low, high, cap, val, gaap, gaapcy, gaapny, adj, adjcy, adjny, sales, salcy, salny. From this point of view, the first principal component can be named prices and commercial component;
- W_2 is highly correlated with ptcf and ptebitda, that means that the second principal component is "money price";
- W_3 is correlated with roa and roe, and it's name is profitability component;
- W_4 is correlated with emp, and is the resource component;
- W_5 is correlated with pts and the name is price to sales;
- W_6 takes most of ca and pr and can be named as income;

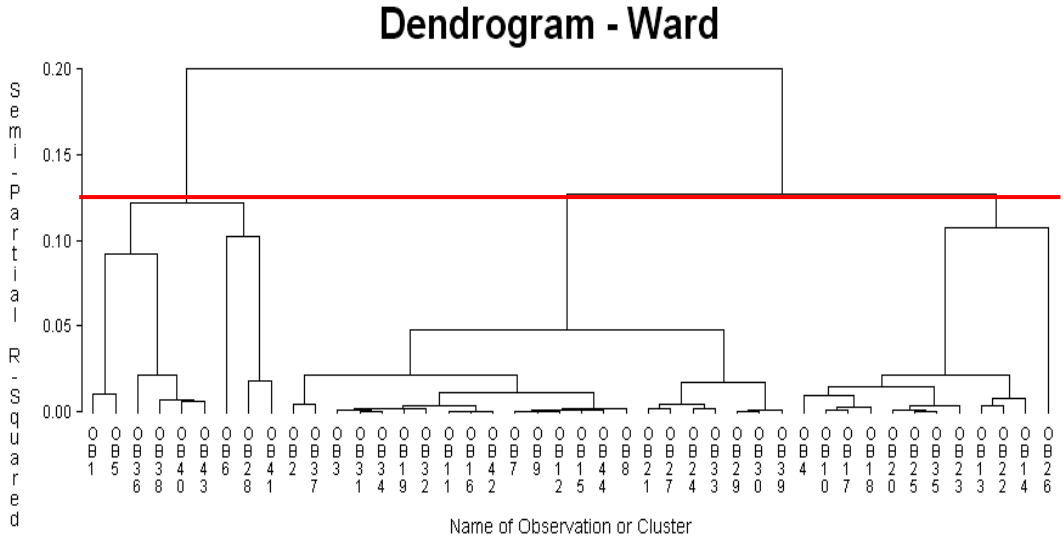

Figure no. 2. Ward's dendrogram
source: SAS output

The figure from above is the hierarchically classification tree (using Ward method). The horizontal line delimitates between three classes chosen to be customers segmentation criterion. The major objective of this analysis is to group all 44 companies into three major clusters, in order to identify: hidden patterns for each class, major characteristics for classes, and use this analysis as decision support for marketing campaigns.

It is important to mention that, because of the fact that three companies were grouped into a single cluster, this type of segmentation is not part of our analysis objectives, and those companies were removed from further analysis.

Table no. 2. Classes centroids

CLUSTER	Avg of roe	Count of nume	Avg of roa	Avg of emp	Avg of pr	Avg of ca	Avg of p	Avg of val
1	35.12	13	16.95	102534.61	9909.93	65.62	95.86	177394.93
2	13.29	18	5.99	15973.48	1684.00	2136.19	32.45	50382.50
3	12.38	13	7.91	51158.31	1895.59	9.48	58.88	43134.85

source: Excel output

The table from above shows all three classes, and a part of the centroids, calculated for original data. The first and the third classes contain 13 companies, while the second one has 18 companies. The first class is represented by the big customers, that bring an annual big profit for the company, either by consulting or audit services. Class 2 has small companies that also bring an important part of total income, because there are many customers that use audit and consulting services, while the third class is represented by middle customers.
Even if we took into account only the first 44 major clients for Deloitte, we have demonstrated that this sample may be split into customers categories and, for each class can be developed special marketing advertisings or promotions.

Conclusions and further research

Applying data mining methods allows as to study different clusters and to construct different categories considering a various set of variables as the turnover and the ratings of agencies specialized in the field.
We can observe that our variables are strongly correlated which lead us to conclude a business maturity and performance. For further research, we propose to extend this research to for a bigger sample, in order to identify more hidden patterns and behaviors for observations.

References

1. Janakiraman, S., Umanmaheswari, K., 2014, A Survey on Data Mining Techniques for Customer Relationship Management*; iJEBEA*; vol. 7(1); pp. 55-61
2. Ruxanda, G., 2009, *Analiza multidimensională a datelor*, Academia de Studii Economice, Ş coala Doctorală, Bucureş ti
3. Rygielski, C., Wang, J-C, Yen D.C., 2002, Data Mining techniques for customer relationship management; *Technology in Society* , (24), pp. 483-502.
4. Scarlat, E., Chirita, N., 2012, *Bazele ciberneticii economice*, Ed. Economica, Bucuresti
5. Villameva, J., Hanssens, D.M., 2007, Customer Equity: Measurement, Management and Research Opportunities*; Foundations an Trends in Marketing*; vol. 1(1), pp. 1-95;
6. Verhoef, P.C., Doorn, J.van, Dorotic, M., 2007, Customes Value Management: An Overview and Research Agenda, *Marketing-JRM*, vol. 2, pp. 51-68
7. Bloomberg.com, [Accessed April 2015]
8. Moodys.com, [Accessed April 2015]
9. Fitch.com, [Accessed April 2015]