BULETINUL Universităţii Petrol – Gaze din Ploieşti	Vol. LXII No. 1/2010	88 - 96	Seria Matematică - Informatică - Fizică

Using Principal Component Analysis in Loan Granting

Irina Ioniţă, Daniela Şchiopu

Petroleum - Gas University of Ploiesti, Informatics Department, Ploieşti, Romania
e-mail: tirinelle@yahoo.com, daniela_schiopu@yahoo.com

Abstract

This paper describes the utility of Principal Component Analysis (PCA) in the banking domain, more exactly in the consumer lending problem. PCA is a powerful tool for analyzing data of high dimension. When an applicant requests a loan for personal needs, a credit officer collects data from him and makes a scoring. The factors analyzed can be significant as well as insignificant. The principal component analysis can help in this case to extract those factors, which produce a better credit scoring model. The data set used for the analysis is provided by a public database containing credit data from a German bank. The results emphasize the utility of PCA in the banking sector to reduce the dimension of data, without much loss of information.

Keywords: *principal component analysis, consumer lending, credit scoring model, banking domain*

Introduction

In banking domain to know what are the best decisions to make is a permanent concern for managers. An active banking area with higher risk is represented by credit department. Here, credits officers analyze the customers application credit forms and calculate a score. The factors considered can influence more or less the credit scoring model. Identifying those factors with higher significance is not a simply task.

Principal Component Analysis (PCA) represents a powerful tool for analyzing data by reducing the number of dimensions, without important loss of information and has been applied on datasets in all scientific domains [16, 17]. On the other hand, PCA is known as an unsupervised dimensionality reduction technique which transfers the data linearly and projects original data to a new set of parameters called the factors (further on, we will use the term “factor” with the meaning of “principal component”), while retaining as much as possible of the variation present in the data set.

In this paper we discuss use of PCA in credit approval problem, considering a set of records provided by a German bank [24]. The results indicate the utility of eliminating variables with a minimum influence in credit scoring model in order to make a better decision for consumer loan granting. The instrument used to apply the PCA technique was SPSS [25]. The paper structure contains a section with theoretical sequences referring to PCA, a section regarding the PCA application in banking domain and a final section presenting a case study.

Principal Component Analysis

PCA is considered the oldest technique in multivariate analysis and was first introduced by Pearson in 1901, and it has been experiencing several modifications until it was generalized by Loeve in 1963 [21].

PCA is a method that reduces the dimensionality of a dataset, by finding a new set of variables, smaller than the original set of variables [15]. This efficient reduction of the number of variables is achieved by obtaining orthogonal linear combinations of the original variables – the so-called Principal Components (PCs) [12]. PCA is useful for the compression of data and to find patterns in high-dimensional data.

PCA and Factor Analysis (FA) are both methods for data reduction. FA analyzes only the variance shared among the variables, while PCA analyzes all of the variance. Concepts such eigenvalues, eigenvectors, loadings and scores are characteristics for these statistical methods.

The main steps of the PCA algorithm are as shown in Figure 1, adapted from [18].

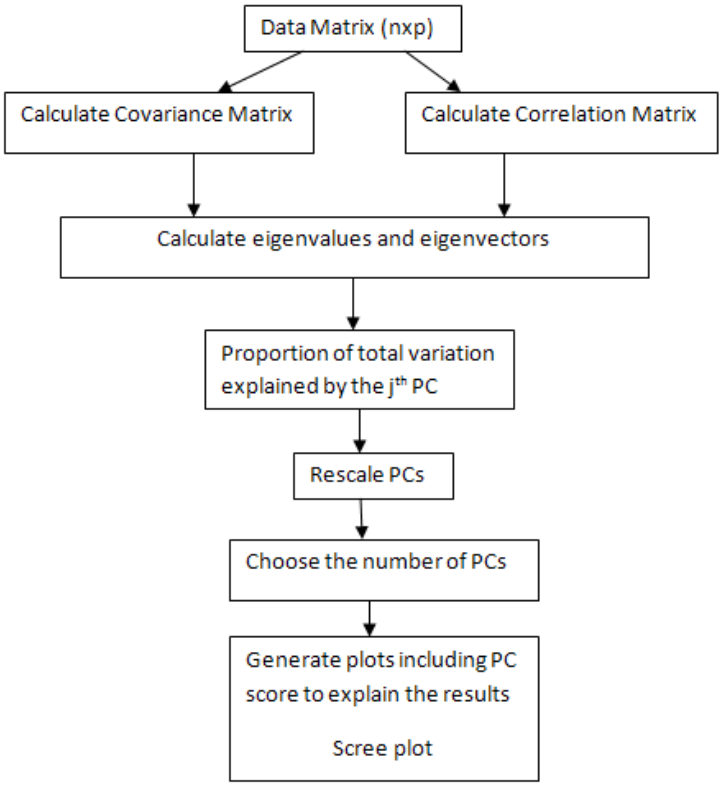

Fig. 1. Principal Components Analysis steps

The mathematical equations for PCA are presented below.

We consider a set of n observations on a vector of p variables organized in a matrix X (n x p):

$$\{x_1, x_2, \cdots, x_n\} \in \Re^p \quad . \tag{1}$$

The PCA method finds p artificial variables (principal components). Each principal component is a "linear combination of X matrix columns, in which the weights are elements of an eigenvector to the data covariance matrix or to the correlation matrix, provided the data are centered and standardized" [7]. The principal components are uncorrelated.

The first principal component of the set by the linear transformation is:

$$z_1 = a_1^T x_j = \sum_{i=1}^{p} a_{i1} x_{ij}, \quad j = 1,..., n \ . \tag{2}$$

In equation (2), the vectors a_1 and x_j are:

$$a_1 = (a_{11}, a_{21},..., a_{p1}) \tag{3}$$

$$x_j = (x_{1j}, x_{2j},..., x_{pj}) \ . \tag{4}$$

One chooses a_1 and x_j such as the variance of z_1 is maximum. All principal components start at the origin of the ordinate axes. First PC is direction of maximum variance from origin, while subsequent PCs are orthogonal to first PC and describe maximum residual variance.

For example, when we work with two dimensions, we have the situation depicted in Figure 2, adapted from [23].

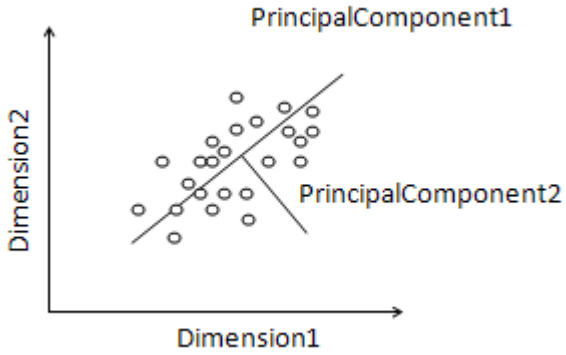

Fig. 2. Principal Components

In the next section, we make a survey of various applications of the PCA method and we consider the use of this data reduction method in banking sector.

PCA and Banking Domain

PCA is applied in various domains such as medicine [11], face detection and recognition [3], signal processing [5], banking [1] etc.

As we discussed earlier in this paper, PCA is an effective transformation method for reduction of a large number of correlated variables in situations in which variable selection is hard to achieve. The result of PCA is a set of new independent variables that can be directly used by credit scoring techniques.

Continuous changes in the banking world produce strategies remodeling, adaptation on new financial trends and manage knowledge. A bank wants to maintain it in balance on the market, to obtain benefits with minimum costs. Managers have to find better solution to make decisions to increase their credibility and to situate their institution on the top.

The basic objectives of bank management have focused on the need to balance between liquidity, assets, credit, interest rate risks, in order to minimize the risks for bankruptcy. Analyzing the factors that may affect these risks is an important job for mangers. An example of factor analysis (similarly with PCA) applied in banking domain to identity the risk exposure is presented in [14]. The results of this analysis indicated that liquidity and interest, domestic market, international market, business operation and credit are the factors affecting banks' risk

exposure. The managers have to consider these factors in formulating the risk management strategy to avoid any situation of bankruptcy.

Credit department confronts with various problems regarding loan granting process. When a customer applies for a loan, the credit officer requests some financial and nonfinancial date and calculates a score. If the customer obtains a good score, the file with necessary documents will be analyze and sent to Central Bank. Other verifying procedures will be applied (for example, analyzing of customer credit history by Credit Bureau). A response affirmative or negative will be send bank to credit officer and the customer will be announced. In the favorable case, after signing the contract, the bank will supply the customer account with the value of loan granted.

The credit score system used today was designed to provide lenders with financial profiles on consumers who wished to borrow money. The lenders' biggest concern was whether or not an individual had the ability to repay a loan on established terms, and find what percentage of risk might be involved [6, 8, 10]. Credit scoring is calculated by a mathematical equation that evaluates many types of information found in a consumer's credit file (duration of loan, loan amount, number of years on service, number of years of residence, marital status, education level etc.) [2, 4, 9]. By comparing this information to the repayment patterns store in hundreds of thousands of consumers' past credit reports, the score identifies the lender's level of future credit risk.

For example, the FICO credit score model takes into consideration five factors to create a model for credit scoring [19]:

- Payment history (35% significance);
- Outstanding credit balances (30% significance);
- Credit history (15% significance);
- Type of credit (10% significance);
- Inquiries (10% significance).

A FICO score can range between 300 and 850 and is a measure of client creditworthiness. Most credit bureau scores used in the U.S. are produced by Fair Isaac and Company, or FICO. FICO scores are provided to lenders by the three major credit reporting agencies: Equifax, Experian, and TransUnion [20]. The FICO score is considered an efficient predictive scoring model designed to evaluate costumer credit risk [19]. This credit scoring model has been widely developed and is used in many credit bureaus around the world, such as [19]: Asia/Pacific Rim, including South Korea, Singapore, India, Taiwan and Thailand, Europe/Middle East, including Ireland, Poland, Sweden, Saudi Arabia and Turkey, Latin America, including Brazil, Mexico, Peru and Panama. The FICO score is presented in Romania, since January 2009.

In order to develop a credit scoring system to assist the credit officers in their decision process, we formulate the hypothesis of using PCA to identify those variables with minimum effect in credit scoring computing and to eliminate them from the scoring model. The next sections present our case study. The software package used in this case is SPSS [25].

Case Study

The dataset that has been used in our case study has been obtained from a public database that contains credit data of a German bank [22]. We have chosen 500 records from this database, data that were organized in a table in SPSS [25]. In Table 1 we present the first five rows. The table also contained information about duration of the credit, credit history, purpose of the loan, credit amount, savings account, years employed, payment rate, personal status, residency, property, age, housing, number of credits at bank, job, dependents and credit approval (target variable). The first 15 variables will be further on denoted with *V1* to *V15* and the target variable with *VT*.

Table 1. Credit Data

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	VT
1	6	2	0	1169	0	4	4	1	4	2	67	1	2	1	1	1
2	48	0	0	5951	1	2	2	0	2	2	22	1	1	1	1	0
3	12	2	1	2096	1	3	2	1	3	2	49	1	1	0	2	1
4	42	0	2	7882	1	3	2	1	4	3	45	0	1	1	2	1
5	24	1	3	4870	1	2	3	1	4	0	53	0	2	1	2	0

The criteria we take into account for the number of the retained factors are: the cumulated percent in variation explained by the retained factors should be higher than 50% and the variance of each retained factor should be higher than 1.

First, PCA summarizes the pattern of intercorrelations between variables. The variables that are highly correlated with one another are grouped together into factors. The correlation matrix for the first 13 variables is shown in Table 2.

Value 0 for correlation coefficient indicates the absence of statistical linkage between variables. We note that only the V5 variable is not well correlated with some others.

Table 2. The Correlation Matrix

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
V1	1												
V2	,103	1											
V3	,108	,149	1										
V4	,611	,115	,201	1									
V5	-,058	-,054	,006	-,084	1								
V6	,011	,122	-,019	-,074	,035	1							
V7	,057	,087	-,063	-,288	,007	,136	1						
V8	-,053	-,049	,060	-,025	,075	,091	,033	1					
V9	,052	,104	,121	,024	,011	,220	,030	-,112	1				
V10	-,188	-,096	-,069	-,247	-,049	-,094	-,024	,002	-,103	1			
V11	-,036	,198	,104	,021	,003	,270	,124	,038	,327	-,130	1		
V12	-,141	-,091	-,067	-,136	,058	-,095	-,079	-,036	-,046	,318	-,342	1	
V13	-,031	,452	,131	,016	,015	,099	,043	-,046	,064	-,022	,149	-,090	1

KMO (Kaiser-Meyer-Olkin) is a statistic test that indicates the degree of association of the variables [16]. In our case, KMO test is 0.555. This value is an argument favorable to the existence of factors, suggesting that factoring is appropriate.

The communality for a given variable can be interpreted as the proportion of variation in that variable that is explained by the analyzed factor. A factor loading represents the correlation between a variable and a factor that has been extracted from the data. The communalities for the V_i variable (i = 1,…,15) are computed by taking the sum of the squared loadings for that variable. Lowest values of communality indicate that the analyzed variable is inadequate represented by the factorial model. Here, the most variable communalities are between 0.5 and 0.9 (see Table 3).

Using the PCA method the fifteen variables are reduced to seven factors as shown in Table 4. These seven factors can be used further as predictors.

The variance in the correlation matrix is reassembled into 15 eigenvalues. Each eigenvalue represents the amount of variance that has been captured by one component.

In general, once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives the components in order of significance and we can decide to ignore the components of lesser significance. This procedure implies to lose some information, but if the eigenvalues are small, the loss of information is minimal.

Table 3. Communalities

	Initial	Extraction
V1	1,000	,837
V2	1,000	,707
V3	1,000	,482
V4	1,000	,819
V5	1,000	,713
V6	1,000	,633
V7	1,000	,691
V8	1,000	,714
V9	1,000	,730
V10	1,000	,458
V11	1,000	,602
V12	1,000	,651
V13	1,000	,707
V14	1,000	,621
V15	1,000	,553

In the first part of Table4 (the first three columns) are presented the eigenvalues for our data and proportions of variance for the fifteen components. Only the first seven components have eigenvalues greater than 1.

Table 4. Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,325	15,500	15,500	2,325	15,500	15,500	1,641	10,942	10,942
2	1,854	12,357	27,858	1,854	12,357	27,858	1,638	10,921	21,863
3	1,330	8,870	36,727	1,330	8,870	36,727	1,609	10,725	32,588
4	1,172	7,815	44,542	1,172	7,815	44,542	1,551	10,339	42,927
5	1,123	7,488	52,030	1,123	7,488	52,030	1,229	8,191	51,118
6	1,077	7,183	59,213	1,077	7,183	59,213	1,147	7,648	58,766
7	1,037	6,910	66,123	1,037	6,910	66,123	1,104	7,358	66,123
8	,933	6,222	72,345						
9	,813	5,422	77,767						
10	,725	4,831	82,598						
11	,693	4,618	87,216						
12	,629	4,195	91,411						
13	,543	3,623	95,034						
14	,474	3,161	98,195						
15	,271	1,805	100,000						

Extraction Method: Principal Component Analysis.

We select *varimax*, the most used method of factors' rotation. Its purpose is to reduce the variance of values not accounted in composition of the factor. *Varimax* minimize the complexity of the components.

The columns Extraction Sums of Squared Loadings contain eigenvalues, explanatory and cumulative version for that three factors, in the context of the initially factorial solution (without rotation). The explanatory version for each factor is distributed as follow: factor I – 15.500%; factor II - 12.357%, factor III – 8.870% and so on. The proportion of the total variation explained by the seven factors is 66.123%. The individual communalities tell how well the model is working for the individual variables, and the total communality gives an overall assessment of performance. In the columns Rotation Sums of Squared Loadings, we have the same values for all the factors, but, as a result of rotation, a new distribution of explanatory variance of each factor can be observed (factor I – 10.942%; factor II - 10.921%, factor III - 10.725% and so on), with the same total variation (66.123%). The remaining variation till 100% is unexplained by this model. With the method of rotation, factor I and factor II loose from the saturated degree in favor of factor III and factor VII.

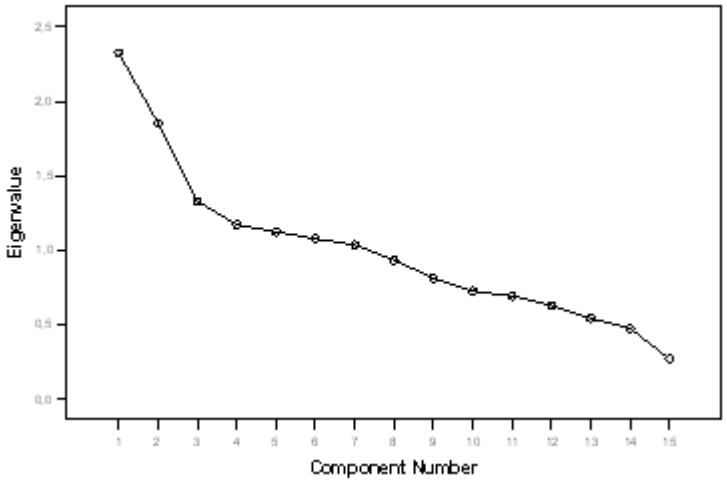

Fig. 3. SPSS Scree Plot

Scree Plot (see Figure 3) presents graphically the eigenvalues for all the principal components obtained, the numeric values being stored in Table 4.

In Table 5 are indicated the data after factors' rotation.

Table 5. Rotated Component Matrix

	Component						
	1	2	3	4	5	6	7
V9	,724	,016	,063	,007	,172	-,391	,136
V6	,695	,085	-,124	,074	-,231	,260	-,038
V11	,619	-,151	,405	,165	,003	1,98E-005	-,071
V1	,013	,904	,091	,034	-,089	-,038	-,001
V4	-,054	,809	,186	,038	,347	-,025	-,066
V12	-,174	-,045	-,711	-,047	,102	-,160	,272
V14	-,248	,052	,687	,029	,138	-,192	,171
V10	-,160	-,290	-,581	,021	,034	-,071	-,062
V13	,028	-,081	,026	,835	,038	-,018	-,001
V2	,116	,142	,012	,815	-,045	-,032	-,078
V7	,100	-,069	,152	,179	-,777	,033	,127
V3	,069	,087	,193	,310	,550	,085	,163
V8	-,025	-,040	,055	-,056	,030	,831	,120
V5	,120	-,063	-,037	-,004	,067	,225	,799
V15	,294	-,040	-,040	,139	,268	,326	-,516

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 14 iterations.

The data in Table 5 lead to the final conclusions on the factorial structure of the analyzed variables, as follows:

- Factor I has a good correlation to variables V9, V6, V11 (and will be labeled *TimeFactor*);
- Factor II has a good correlation to V1, V4 (and will be labeled *CreditFactor*);
- Factor III has a good correlation to V14 (and will be labeled *JobFactor*);
- Factor IV has a good correlation to V13, V2 (and will be labeled *CustomerCreditFactor*);
- Factor V has a good correlation to V3 (and will be labeled *PurposeFactor*);
- Factor VI has a good correlation to V8, V15 (and will be labeled *CustomerFamilyFactor*);

- Factor VII has a good correlation to V5 (and will be labeled *SavingsFactor*).

A related work is presented in [13], where the authors presents an improved credit scoring to the Chinese commercial bank for credit card risk management. Also, they choose PCA to calculate the weights of the original indexes and to obtain a predicting function for computing the score of new applicants. In their case, the KMO value given by SPSS is 0.787 .

Conclusions

PCA is a method for multivariate data analysis and it is used in many fields to extract relevant information from confusing data sets. Also, PCA provides a way of identifying patterns in data, and of expressing the data in a way that highlights their similarities and differences between them. An advantage of using PCA consists in quantifying the importance of each dimension for describing the variability of a data set. This method reduces the number of dimensions, without much loss of information.

In this paper we presented the possible use of PCA in the banking domain to reduce the dimension of data related to the credit problem. In our case study, we used 15 variables describing consumer loan data; by means of PCA, we have got only seven factors that concentrate more than 60% of the information provided by the original 15 variables. Future work will focus on developing a credit scoring system based on use of PCA in the banking domain.

References

1. Abu-Shanab, E., Pearson, M. - Internet Banking in Jordan: An Arabic Instrument Validation Process, *The International Arab Journal of Information Technology*, Vol. 6, No. 3, 2009, pp. 235-245
2. Basno, C., Dardac, N. - *Management bancar*, Editura Economică, Bucureşti, 2002
3. El-Bakry, H.M. - New Fast Principal Component Analysis for Face Detection, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol.11, No.2, 2007, pp.195-201
4. Hand, D.J., Heney, W.E. - Statistical Classification Methods in Consumer Client Scoring: A Review, *J. R. Statist. Soc*., 160, 1997, pp.523-541
5. Helmy, A.K., El-Taweel, GH.S. - Authentication Scheme Based on Principal Component Analysis for Satellite Images*, International Journal of Signal Processing, Image Processing and Pattern Recognition*, Vol. 2, No.3, 2009, pp.1-10
6. Lyn, C.T. - A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers, http://www.sciencedirect.com/science, accessed 2010
7. Marinoiu, C. - Grades-based Characterization of Freshmen Using PCA, *Petroleum-Gas University of Ploiesti Bulletin, Mathematics, Informatics, Physics Series*, vol. LXI, no.2, 2009
8. Mester, L.J. - What's the point of Credit Scoring*, Business review,* September-October 1997
9. Olteanu, A., Olteanu, F.M., Badea, L. - *Management bancar. Caracteristici, strategii, studii de caz*, Editura Dareco, Bucureşti, 2003
10. Rosenbaum, B.Jr. - Your Credit Score. What It Means to You as a Prospective Home Buyer*,* www.mytalentedlender.com/~mytalent/mystuff/File/CreditScoring.pdf , accessed 2010
11. Sinescu, I. et al. - Principal Component Analysis and Classification with Application in Medicine, http://www.adcl-ase.com/wp-content/uploads/2010/04/articcol.pdf, accessed 2010
12. Sustersic, M., Mramor, D., Zupan, J. - Consumer Credit Scoring Models With Limited Data, *EFA 2007 Ljubljana Meetings Paper,* http://ssrn.com/abstract=967384, accessed 2010
13. Xu, W., Liu, J., Li, J. - An Improved Credit Scoring Method for Chinese Commercial Banks, http://web.cenet.org.cn/upfile/52253.pdf , accessed 2010
14. Yap, V.C. et al. - Factor Affecting Banks' Risk Exposure: Evidence from Malaysia, *European Journal of Economics, Finance and Administrative Science*, ISSN 1450-2887 Issue 19, 2010, pp. 121-126

15. Ye, J. - *Principal Component Analysis,* http://www.public.asu.edu/~jye02/CLASSES/Fall-2007/NOTES/PCA.ppt, accessed 2010
16. * * * - *Annotated SPSS Output Principal Components Analysis*, http://www.ats.ucla.edu/stat/SPSS/output/principal_components.htm, accessed 2010
17. * * * - *A Tutorial on Principal Component Analysis*, http://www.cs.cmu.edu/~elaw/papers/pca.pdf, accessed 2010
18. * * * - *Essential Steps of Principal Component Analysis*, http:// www.morris.umn.edu/~sungurea/multivariatestatistics/principalcomponent.html, accessed 2010
19. * * * - *FICO score*, http://www.fico.com/en/FIResourcesLibrary/FICO_Score_1655PS.pdf, accessed 2010
20. * * * - http://www.myfico.com, accessed 2010
21. * * * - *Introduction to Pattern Recognition.* Lecture 5: Dimensionality reduction (PCA), http://courses.cs.tamu.edu/rgutier/cs790_w02/l5.pdf, accessed 2010
22. * * * - *Know the Score. Educate Yourself about Credit Scores and Critical Credit Issues,* www.goodmortgage.com, accessed 2010
23. * * * - *Principal Component Analysis (PCA*), http://www.uga.edu/strata/software/pdf/pcaTutorial.pdf, accessed 2010
24. * * * - *Source data*, http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/ , accessed 2010
25. * * * - *SPSS (Statistical Package for the Social Sciences)*, http://www.spss.com/ , accessed 2010

Utilizarea analizei în componente principale în problema acordării creditelor

Rezumat

Această lucrare descrie utilitatea Analizei în Componente Principale (ACP) în domeniul bancar, mai exact în soluţionarea problemei acordării creditelor de consum. ACP reprezintă un instrument puternic pentru analiza datelor de mari dimensiuni. Atunci când există o cerere pentru un împrumut de nevoi personale, ofiţerul de credite colectează date de la această persoană şi îi calculează un punctaj. Factorii consideraţi pot influenţa diferit analiza de credit realizată. ACP poate ajuta în acest caz la extragerea acelor factori care descriu cel mai bine modelul de credit scoring. Datele folosite în aplicaţie au fost preluate dintr-o bază de date publică ce conţine date de creditare de la o bancă germană. Rezultatele subliniază utilitatea aplicării ACP în sectorul bancar pentru a reduce dimensiunea datelor, fără pierderi mari de informaţii utile.