THE BEHAVIOR OF CREDIT RISK EVALUATION MODELS UNDER RECESSION AND THE INTRODUCTION OF A GENERAL MODEL BASED ON SEMANTIC INTEROPERABILITY AND NOMOGRAMS[1]

Laura POPESCU
PhD Candidate, University of Economics, Bucharest, Romania
Developer, BRD Groupe Societe Generale, Bucharest, Romania

E-mail: laura.popescu@brd.ro

Andreea DIOSTEANU
PhD Candidate, Instructor, University of Economics, Bucharest, Romania
.NET Programmer, Totalsoft, Bucharest, Romania

E-mail: andreea.diosteanu@ie.ase.ro

Anca Nicoleta POPESCU
Master Student, SIMPRE, University of Economics, Bucharest, Romania
IT analyst, Outsource Bazaar, Bucharest, Romania

E-mail: anca.popescu@outsourcebazaar.com

Abstract: *The article analysis the old credit risk evaluation models performance and highlights the failure of complex econometric models to predict recession. Furthermore, this article is intended to propose a software solution for implementing a general, orientative consumption credit risk evaluation scoring model based on semantic web. The use of semantic web enables us to discover the common features from analyzing the evaluation forms used by several banks. These characteristics are taken into account when designing the model.*
The application will have a web interface and the weights and the cut-off value will be represented with the help of nomograms. This software product is intended to offer clients guidance and to provide, with a certain amount of risk, the information related to the chances that client has for obtaining a credit.

Key words: *credit risk; nomogram; semantic interoperability; OWL; recession*

Introduction

The new economic conditions under which we activate prove that we pass through an authentic recession period that has represented, metaphorically speaking, an "earthquake" for the traditional evaluation instruments. The theory has proved its consistency with reality, emphasizing that when environmental conditions change, classical models may lose their efficiency and lead to important losses.

The importance of the decision of granting credits has been confirmed by many specialists and business analysts as being one of the main sources of risk in the banking system. Therefore, it is recommended to create and implement efficient evaluation instruments that overcome the recession particularities and introduce an efficient manner of evaluating and granting credits with a minimized risk.

After presenting the results of a practical study carried out by the authors in order to prove that classical statistical models that are used by the great majority of banks in the financial system are less efficient during *recession*, the paper proposes web semantics and document mining as a solution for identifying a set of common features required by different banks in order to build a general credit scoring model. Furthermore, some theoretical aspects related to the visual instruments called *nomograms* are introduced. The authors present some aspects related to the creation of a general orientative *credit risk* evaluation model that brings the advantage of a visual instrument and of taking into account all the details related to a particular customer.

After building the nomogram – profile for a particular client and discussing his chances of being granted a credit, we emphasize the facilities offered by a visual decision tool. Finally, the authors summarize their main conclusions.

Romanian credit risk evaluation models

An evaluation of the indicators concerning credit institutions computed by the National Bank of Romania (BNR) brings to attention the fact that the percentage of defaulted credits as part of the total amount of existing credits at national level has clearly increased during the last quarter of 2008. If the percentage of "bad" credits was 0.24 in September 2008 (at the end of the third quarter in 2008 when the effects of the economic crisis were not perceived yet), the percentage increased at 0.35 at the end of 2008. This increase of 0.11 percentage points is a big increase if we take into account the relative stability of this indicator's values during the previous 4 quarters (0.22, 0.21, 0.30 and 0.24).

Table 1. The evolution of defaulted credits of credit institutions over 2008

Indicators concerning credit institutions	UM	Dec/2007	Mar/2008	Jun/2008	Sep/2008	Dec/2008
Defaulted credits/Total net value of granted credits	%	0.22	0.21	0.30	0.24	**0.35**

***)** Indicators are computed based on the provisions data reported by credit institutions
****)** Indicators comprise only commercial banks and Creditcoop as foreign banks branches do not report their solvability.
NB: Indicators refer to all the credit institutions in Romania, including commercial banks, foreign banks branches, Creditcoop.Starting from January 1, 2008, indicators are computed based on the financial reports FINREP and COREP of credit institutions. Source: National Bank of Romania (Banca Națională a României)

These results confirm the general expectations that the repayment ability of credit customers would decrease during recession. This feature is one of the first indicators of the economic recession.

Another observation is that the gross value of granted credits is still higher than the total value of deposits as the percentage of credits related to deposits is 122% (we can notice a small decrease in comparison to the third quarter of 2008: 124.71%) (Fugure 1). Many analysts talk about the need of increasing the amount of liquidities of every financial institution during the crisis.

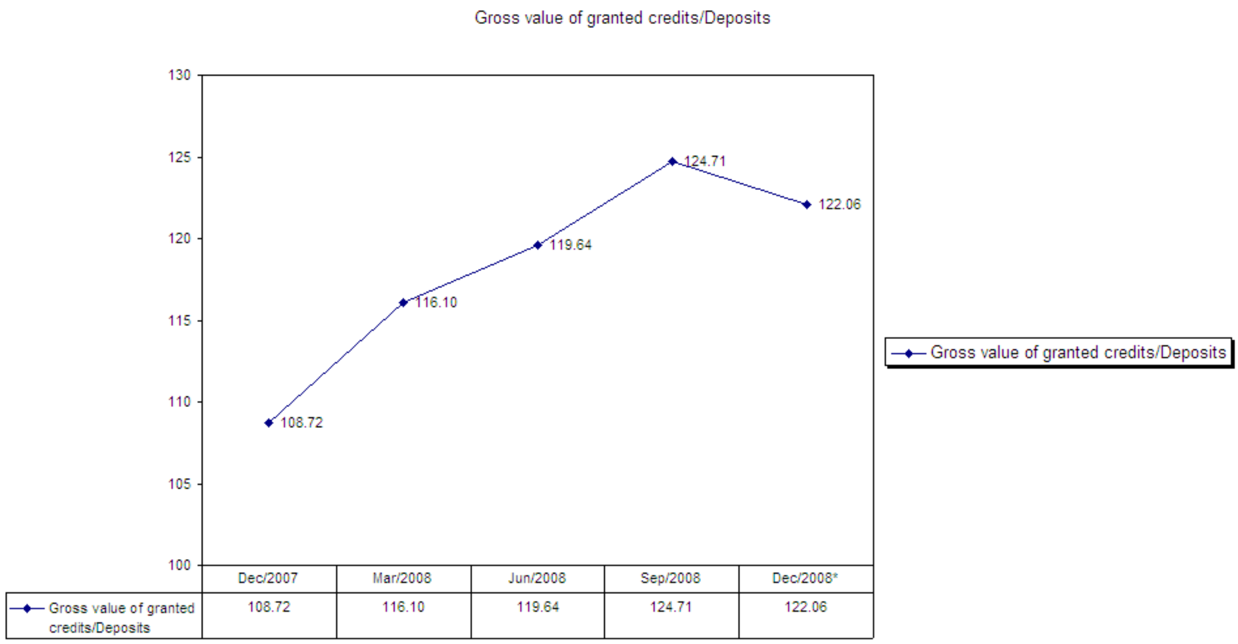

Figure 1. The evolution of granted credits related to deposits over 2008

Data Source: National Bank of Romania (Banca Nationala a Romaniei)

Most of the Romanian financial institutions use very simple methods of computing the credit score by taking into account several features and using weights which are established by experts without making usage of statistical, mathematical or informatic instruments. Consequently, Romanian banks and credit institutions still prefer to use traditional statistical methods (models that are determined based on regressions) as they are less expensive and easy to build and to understand and their coefficients can be rapidly adapted by internal risk analysts to new changes in customers' profiles identified in the annual credit reports.

An experiment carried out by the authors proved that, although very easy to build, use and adjust, those tools may prove less efficient under crisis conditions. The experiment used two data sets that contain information concerning several features recorded for 62 companies. The data source was represented by the financial statements available on Bucharest Stock Exchange or on the KTD Invest and the values reflect the status at the half of 2008 (June 30, 2008) and at the end of the year (December 31, 2008). Same features are taken into account for both periods in order to allow a comparison between the two moments in time. The net income was assumed to estimate a company's financial stability and ability to repay its loans.

The identification of the most significant characteristics is the first step in the process of creating a statistical credit scoring model (Piramuthu, 2006) as the different scores associated to variables in the credit scoring model are derived from the coefficients of the most representative variables estimated through regression.

Different regression models were built through the usage of Eviews 6.0, by applying Ordinary Least Squares and the process was reiterated as several characteristics were repeatedly eliminated and introduced under different scenarios. Finally, the most significant regression model (Dardac, 2005) was obtained when choosing capital, total debts and turn over as independent variables and total debts as dependent variable. The regression model was applied for both data sets and the following results were obtained:

Results for June 30, 2008

NET_INCOME=6,335.501+0.000751*CAPITAL+0.051040*TURNOVER - 0.096196*TOTAL_DEBTS

(0.0000) (0.0157) (0.0010)

Results for December 31, 2008

NET_INCOME=418.9257+0.001783*CAPITAL-0.043378*TURNOVER -0.141287*TOTAL_DEBT

(0.0000) (0.0195) (0.0110)

The fact that the p-values (probabilities that express the significance of the influence of selected variables over the dependent variable) are still quite high is determined by the inclusion of only three independent variables in the regression model. We are aware that there are other characteristics that influence the financial ability of paying the loan and that they may lead to a decrease of the p-values and an increase of the model's efficiency.

The most important conclusion that can be drawn by analyzing the results is that the relevancy of the selected variables decreases during the second half of the year. The capital remains the most significant feature for both data sets, while the p-value for turnover increases from 0.0157 to 0.0195 and the one for total debts increases from 0.0010 (30.06.2008) to 0.0110 (31.12.2008). Therefore, factors tend to lose their significance and the model proves to be less efficient, leading to a need to redesign the credit scoring models.

Semantic Web and customer features' selection

The novelty of the model that we propose is represented by the fact that we implement semantic web and document mining in order to analyze several banks' client profile evaluation forms and to determine a set of common features that are taken into account. The determined features will represent the points for nomogram graphic that is used to calculate the scoring. In the end, we will obtain a general minimalistic and orientative model for clients.

The semantic web is used to process the knowledge that is contained in the evaluation forms and also to facilitate the information integration and long time analysis. The semantic web is based on formal ontologies that structure data in order to represent messages across World Wide Web in a machine processable manner. The ontologies offer the possibility of representing knowledge in an easily interpreted language called *OWL* (Web Ontology Language). OWL also facilitates the implementation of a wide area of classes of semantic applications. The organization that is preoccupied with developing the semantic web proposed the RDF meta language to implement and design semantic messages in an easily machine interpreted language (W3C).

According to Anubhav Madan, the use of semantic web in business applications is very important because it offers *semantic interoperability* possibilities and improves document analysis and research. Business interoperability is one very important factor in assuring economic performance. Furthermore, semantic web is applied in knowledge management.

Nowadays, high quality knowledge management is a key factor in assuring competitive advantage and risk reduction by efficiently exploiting the intellectual assets. Our article introduces the semantic web for knowledge management with impact in credit risk reduction in the actual economic context. Semantics are targeted to improve the traditional knowledge management models of the credit risk evaluation that on one hand were unable to predict the economic crisis and on the other hand proved their inefficiency in evaluating the risk. The consequences were very bad because many banks went bankrupt due to the clients' impossibility of covering their debts.

The model that we propose represents the starting point for a software application that will have two major components: a bank side component and a client side component

The bank side component has banks as main beneficiaries. The application will be based on semantic web services and will be composed of the bellow enumerated modules:

- **a semantic web service** that will implement semantic search for clients' characteristics and ontology development tools for structuring information into machine processable messages
- **a knowledge database** that contains complex credit necessary information, clients' detailed profiles obtained through semantic web search and document mining, and a set of structured rules that will be used by a expert system that will offer support in credit evaluation activity.
- **an expert system** that is designed to offer support for credit officers based on past decisions and a complex set of rules.
- **semantic web service description development tool**

The working scenario for the module will be the following:

Step 1: each bank will expose service that has a specific web service description referring to the functions that access certain databases containing non confidential information and a complete list of clients that were granted credits

Step 2: the semantic web service will provide tools for service discovery and semantic search for both lexical and semantic similarities. With the help of these tools the banks may collaborate in finding the common clients, their profile, and to validate the accuracy of the information the customer provided when applying for a certain amount of money.

Step3: after all necessary information is provided, the expert system will determine a set of structured rules based on previous similar situations and past experiences.

Step 4: in the end a client evaluation scoring model is implemented based on the rules determined. The result will be a percent that measures the client's associated risk level.

Step 5: judging by the determined percentage the credit officers will decide whether the bank will grant a credit for the analyzed client request and the level of warranty the client has to pay or guarantee with.

However, this article will present a detailed analysis of the client side component while the bank's module will be presented in our future articles.

Client semantic component

The client semantic component is designed for clients that intend to apply for a credit and are willing to determine the chances they have to obtain the sum of money they need.

This application proposes a general minimalistic and orientative credit scoring model based on the common features that banks are taking into account when designing the client's profile.

The application will have the following components:

- **a semantic web service** that will perform semantic search and document analysis on the banks evaluation forms and ontologie building tools.
- **a business logic web service** that will contain the general scoring model, based on the features discovered by the semantic web service (the model will be presented in "The usage of nomograms in determining a customer's credit score" subchapter)
- **a user web interface** that will display required information in nomogram graphics, with the help of which the scoring series will be developed, and also client solicited information in a user friendly and easy to understand format

Our model is based on analyzing the evaluation forms from three important Romanian banks: BRD, BCR and Piraeus Bank. For each bank taken into account we identify a set of general terms which represent the first level in the semantic network. Afterwards, we implemented the associated semantic models corresponding to each bank evaluation form.

The words having similar meanings in a standard semantic network, are organized into groups or clusters called synsets. There are two kinds of relations represented by pointers: lexical and semantic. The lexical relations refer to the word forms, while the semantic relations refer to word meaning. This lexical model represents the support for automatic text, document analysis and management. The type of semantic relationships established between terms in a semantic network are the following (grouped by parts of speech) :

Nouns: *hypernym*: B is a hypernym of A if every A is a (kind of) B**,** *hyponyms*: B is a hyponym of A if every B is a (kind of) A**,** *coordinate terms*: B is a coordinate term of A if A and B share a hypernym**,** *holonym*: B is a holonym of A if A is a part of B**,** *meronym*: B is a meronym of A if B is a part of A

Verbs: *hypernym*: verb Y is a hypernym of verb X if activity X is a (kind of) Y**,** *troponym*: verb Y is a troponym of verb X if activity Y is doing X in some manner**,** *entailment*: verb Y is entailed by X if by doing X you must be doing Y, *coordinate terms*: those verbs sharing a common hypernym

Adjectives: related nouns and participle of verb

Adverbs: root adjectives

For each bank we develop the semantic network of the general term "Criterii de evaluare" ("Evaluation Criteria") which represents the root of an arborescent structure. Words having similar meanings are grouped into synsets and correspond to the evaluation criteria in the client evaluation form. Between the synsets we can identify relations such as those mentioned above. However, there might be situations in which terms are expressed differently (for different banks) but have similar meanings.

In order to identify similarities between semantic networks, we take into account some parameters that are related to the type of relationships in the the semantic network (hypernyms, hyponyms, coordinate terms, holonyms, meronyms). For example, "Situatie locuinta"(Client's home situation) (BRD),"Locuinta clientului" (Client's Home) (BCR) and "Tipul locuintei"(Home type) (PIRAEUS) is a meronym of "Criterii de evaluare"(Evaluation criteria) and have similar meanings. We do this for all the elements of the semantic network. Furthermore, we can apply specific algorithms to determine the similarities such as (Frankel, 2007):

- Path Finder
- Depth Finder
  - Wup: Shortest path by scaling sum of values between node and root
  - Lch: (Leacock and Chodrow) Shortest path by scaling the max path
- Path: Inverse of the Shortest Path measures
- Information Content Finder
  - Resnik: Max Distance b/w concepts of both words
  - Jcn (Jiang and Conrath): Inverses the difference between Sum and LCS
  - Lin: Scales LCS IC with the description
- Gloss Finder
  - Lesk (Banerjee and Pederson) finds and scores overlaps between glosses
  - Vector creates a co-occurrence matrix with glosses in vectors
- Hso (Hirst and St-Onge): Specifies Direction between Words

The application based o semantic networks has to have a knowledge base that contains terms and semantic relationships as they are presented in DEX(Explicative Dictionary) for the Romanian language, or an WordNet database for English.

Our pilot solution (the first stage for developing a complex client module) computes semantic models similarities based on LIN algorithm and with a database composed of around 1000 terms and definitions from DEX.

A part of the semantic generated OWL Lite code can be seen below:

```
<rdf:Description rdf:about="VenituriNete">
<rdf:type><rdf:Description rdf:about="http://www.w3.org/2002/07/owl#Class"/></rdf:type>
<rdfs:subClassOf><rdf:Description rdf:about=" CriteriiEval"/>
</rdfs:subClassOf><owl:unionOf rdf:parseType="Collection">
<rdf:Description rdf:about=" VenitNesalarNetLunar"/>
<rdf:Description rdf:about=" VenituriDinDobLaDep"/>
<rdf:Description rdf:about=" VenitDinDreptDeAutor"/>
<rdf:Description rdf:about=" VenituriSocialeLunare"/>
<rdf:Description rdf:about=" SalariuNetLunar"/>
<rdf:Description rdf:about=" ChiriiIncasateLunar"/>
<rdf:Description rdf:about=" SalariuNetLunar"/>
<rdf:Description rdf:about=" PensiiIncasateLunar"/>
<rdf:Description rdf:about=" ChiriiIncasateLunar"/>
</owl:unionOf></rdf:Description>
```

The formula used for computing semantic similarities based on Lin algorithm is the following: $sim(A,B) = \dfrac{\log P(common(A,B))}{\log P(description(A,B))}$

Nomograms – Basic concepts

The nomogram is a statistic tool for concepts and complex data set visual representation. This type of graphic is also known as Kiviat diagram, radar diagram or spider. The great advantage brought by this type of graphic is that of offering a clear representation of the client scoring process with an emphasis on the elements with a higher risk. (Ivan, 2007)

If we consider a P_i client that is part of a set of n persons { P_1, P_2,..., P_n } and we also create the set of scoring features made up of *m* features { C_1, C_2, ..., C_m } for risk analysis represented by the client when granting him a credit.

After specific credit scoring computations, for each feature scoring indicators are being determined { I_1, I_2, ..., I_m }.

The scoring indicators representation by using a nomogram for a certain client (P_i) is developed by following the bellow presented steps:

- the base circle is being drawn;
- the circle is divided into m-1 sectors, each radius representing a certain feature
- for each characteristic C_j, j=1..m a corresponding indicator is defined and calculated I_j, j=1..m. The values are normalized into the [0,1] interval . The 0 value corresponds to the circle's centre and 1 value is represented on the circle's circumference
- each indicator I_j for C_j characteristic is afterwards set on the circle radius.
- the points represented on the radiuses are connected and a specific client's scoring characteristics associated polygon is obtained

In order to simplify the representation, the number of the scoring characteristics that are taken into consideration is m=6, and the selected characteristics are { C_1, C_2, C_3, C_4, C_5, C_6 }. The associated weights for each characteristic are calculated and we consider that each characteristic has the same level of importance. A circle is drawn and the radiuses corresponding to each identified scoring characteristic. After following these steps six equal circle sectors are obtained (the angle for each sector has sixty degrees).

Figure 2. The base circle for a nomogram

Each I_i indicator is represented as a point on the radiuses and finally the points are connected and the client's characteristics' polygon inside of the circle is obtained. Consequently, the nomogram is obtained.

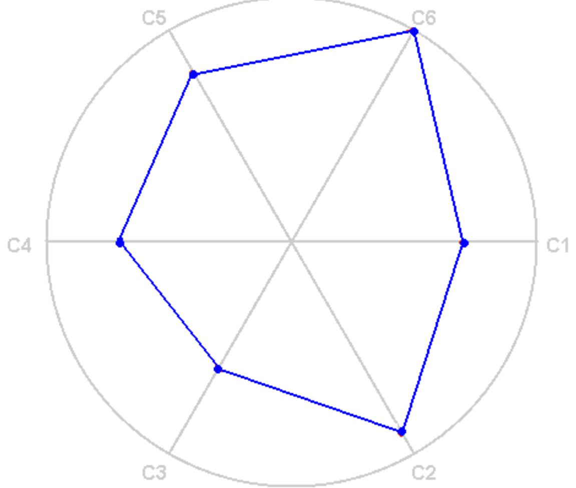

Figure 3. Client specific nomogram

The nomogram is developed so that to cover the general number of scoring characteristics, n. The formula that is used to determine the circle sector's angle is presented below: $teta = \frac{360^{\circ}}{n}$ where:

teta =circle sector angle;
n = scoring characteristics angle.

The n-regular polygon's points inscribed in the base circle is determined by using the following formulas:

$$X_i = \cos(\mathrm{i} * teta) * R + X_{centru}$$
$$Y_i = \sin(\mathrm{i} * teta) * R + Y_{centru}$$

where,

X_i = i point's abscise i=1..n;
Y_i = i point's ordinate , cu i=1..n;
X_{centru}= circle's centre point abscise;
Y_{centru}= circle's centre point ordinate;
R = base circle radius.

The usage of nomograms in determining a customer's credit score

After determining the common characteristics of the credit scoring applications, we obtained the results presented in Table 2. Consequently, we followed several steps in order to represent a particular customer's answers as points on a nomogram. As the goal of the process is to determine the client's risk degree based on the obtained scoring model, aggregated nomograms are used.

The representation of the aggregated nomogram is obtained by overlaying the client's scoring images and the bank's scoring requirements for the characteristics that are taken into account.

Table 2. The characteristics of the credit scoring model

FEATURE	OPTIONS	PTS	Normed indicators	Bank etalon
C1 = AGE	20 – 50 years (men) or 20-45 years (women)	6	1 (6/6)	0.66
	50 – 60 years (men) or 46-55 years (women)	4	0.66 (4/6)	
	Over 60 years (men) or over 55 years (women)	2	0.33 (2/6)	
C2=RESIDENCE TYPE	The person owns the place	8	1	0.5
	The person rents the place	4	0.5	
	The person doesn't have a house	2	0.25	
C3=YEARS AT CURRENT WORKPLACE	Retired	1	0.08	0.5
	Less than 1 year	3	0.25	
	Between 1 and 2 years	6	0.5	
	Between 2 and 5 years	9	0.75	
	More than 5 years	12	1	
C4=FAMILY MEMBERS	Married and both of them have a job	6	1	0.66
	Married, but only one of them has a job	4	0.66	
	Single	2	0.33	
C5=NUMBER OF PERSONS TO FEED	Less than 2	12	1	0.5
	Exactly 2	9	0.75	
	3-4	6	0.5	
	5 or more	3	0.25	
C6=NET INCOME	Less than 500 RON	1	0.11	0.66
	500-1000 RON	3	0.33	
	1000-2500 RON	6	0.66	
	More than 2500 RON	9	1	
C7=INDEBT DEGREE	Less than 10% of his net income	12	1	0.5
	10-20% of his net income	9	0.75	
	20-30% of his net income	6	0.5	
	More than 30% of his net income	3	0.25	

Figure 4 presents the aggregated nomogram for the number of characteristics m=6 and the characteristics **C_1, C_2,.., C_6.** The client's characteristics' polygon is blue shaded, and the banks' polygon is green shaded. By overlaying the two polygons we can highlight the positive and negative differences between client's self calculated scoring model and the one banks use.

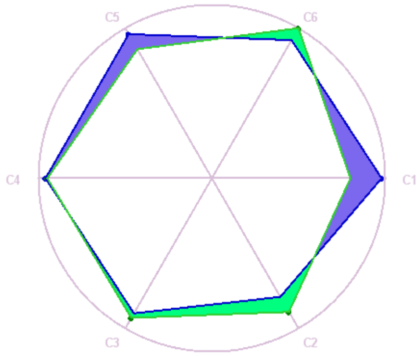

Figure 4. Aggregated nomogram

where,

S_{poz} positive diferences surface area

S_{neg} negative diferences surface area

In order to decide whether the analyzed client respects the imposed scoring legislation imposed by the etalon bank's scoring. The bank defines a quality evaluation indicator IC which is determined by dividing the positive and negative differences that were calculated according to the above presented model of the aggregated nomogram.
The following formula is used:

$$IC = \frac{S_{neg}}{S_{poz}}$$

where,

IC = analyzed client's scoring evaluation indicator;
Sneg = negative diferences surface area ;
Spoz = positive diferences surface area ;

The indicator can be interpreted as it follows: a value equal to 1 indicates a good scoring level (we can state that the client approaches the cut-off value), a value above 1 indicates a bad performance and might lead to the rejection of the customer's loan application, while a value below one indicates a very good performance and the fact that the client is a credible one and can be granted a loan.

For example, let's analyze the case of customer A that is 30 years old and lives in a rented apartment, has been working in the same workplace for less than one year and he is married to another employed person. The customer doesn't have other children and his net income is of 2,700 RON. His indebt degree is between 20-30%. Consequently, customer's A normalized values of the characteristics (as established in Table 2) are the following ones and his nomogram is pictured in Figure 5.

C1 = 1, C2 = 0.5, C3 = 0.25, C4=1, C5=1, C6=1, C7=0.5

It has to be mentioned that the blue line in Figure 5 represents the client's nomogram and the green area represents the characteristics for which the client outperforms the bank's requirement's. the red area represents the risk area and it is determined by customer's A low performance for the third characteristic.

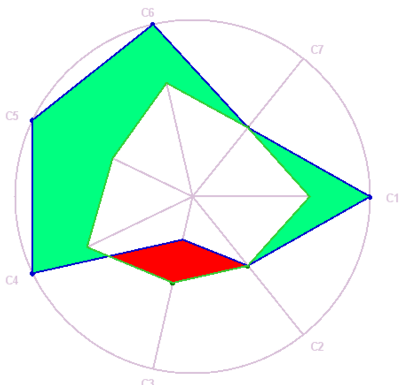

Figure 5. Customer's A nomogram

On the other hand, customer B is 52 years old and he lives in his own apartment. He works in the same place for 20 years and he is married, but has a retired wife. He has two children and his net income is of 5,000 RON, while his indebt degree is less than 10%. Therefore, customer's B characteristics are:

C1=0.66 ; C2=1; C3=1; C4=0.66; C5=0.75; C6=1; C7=1

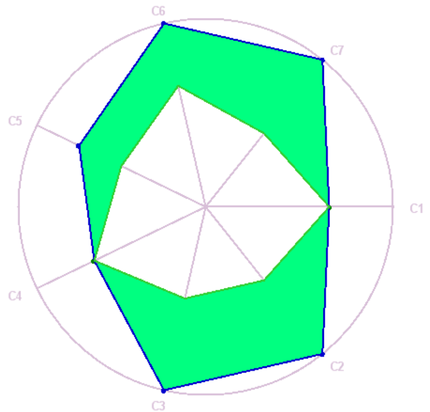

Figure 6. Customer's B nomogram

Conclusions

During economic crisis, the behavior of the credit institutions is different than in a steady or economic growth period. Statistics presented by The National Bank of Romania reveal an increased restrictiveness of credit conditions during the last quarter of 2008. After a short analysis of some general indicators concerning the national status of credit institutions, the authors present the results of a study which proves that, in the case of regression models, coefficients are less significant (higher p-values) during recession (the second half of 2008) when compared with the first half of the year when the economic crisis impact was not perceived in Romania. Therefore, a conclusion is drawn that under economic crisis conditions companies tend to have an abnormal behavior which leads to the necessity of redesigning classical credit scoring models.

The authors propose semantic networks and document mining as a solution for determining common features that can be included in a general, orientative credit scoring model. Finally, some basic aspects related to nomograms are introduced as they can be used as a visual tool for building a credit scoring models. The utility of a user-friendly credit scoring model is proved by taking some particular examples of loan applicants and drawing and discussing their nomograms.

Bibliography

1. Dardac, N. and Moinescu, B. **A quantitative credit risk evaluation from the BASEL II perspective,** Economie teoretica si aplicata, 2005, pp. 41-46
2. Frankel, D. **Business Network Transformation and Semantic Interoperability,** SAP Company, December 2007
3. Ivan, I., Boja, C. and Ciurea, C. **Metrici ale sistemelor colaborative,** Editura ASE, Bucharest, 2007

4. Madan, A. **WordNet: An Overview,** Tutorial, http://www.scribd.com/doc/6890426/tutorialWordNetSimilarity
5. Piramuthu, S. **On preprocessing data for financial credit risk evaluation,** Expert Systems with Applications, 30, 2006, pp. 489–497
6. * * * **An inquiery concerning the creditation of nonfinancial organizations and population,** The National Bank of Romania: www.bnr.ro, November 2008
7. * * * Bucharest Stock Exchange, http://www.bvb.ro
8. * * * KTD Invest, http://www.ktd.ro
9. * * * **W3C Semantic Web Activity,** http://www.w3.org/2001/sw/

[1] **Acknowledgements**
This article is a result of the project „Doctoral Program and PhD Students in the education research and innovation triangle". This project is co funded by European Social Fund through The Sectorial Operational Programme for Human Resources Development 2007-2013, coordinated by The Bucharest Academy of Economic Studies.