SADM – An Automated System Based on Data Mining for Credit Scoring

Irina IONIȚĂ

Petroleum-Gas University of Ploieşti,
39, Blvd. Bucureşti, 100680, Ploieşti, Romania
tirinelle@yahoo.com

Abstract: The credit approval represents a complex banking process which requires a significant involvement of the people and departments responsible with the final decision. The decisions taken by the Central Biro as a result of credit analysis may be alsoaffected by the subjectivity of the working people in charge, causing thedecision-making quality. In order to facilitate the credit approval process, the automation of the credit function made by an automated system may represent a solution in this case. In this paper an automated system based on data mining techniques is proposed and designed to assist the decision credit process. The strong points of this prototype consist in increasing the accuracy of credit decision, increasing the bank's profitability, decreasing the credit risk etc.

Keywords:automated system, data mining, credit scoring

1. Introduction

The banking domain represents a dynamic area with processes that need a significant attention from the managers. Knowledge has become more and more appreciated by organizations because of the value offered to manage the entire activity of them. In a competitive world banks have to apply strategies to assure their continuous functionality and to increase their profit. Most of the knowledge in the banking system is currently generated by daily transactions and operations. The repository of data contains an enormous volume of records than hide valuable information for banks. An important task in this case is to discover that kind of information with significant implication for bank’s management. Data mining as an artificial intelligent technique appears to solve this problem. Mining data is similar to the process of mining gold and supposes a difficult work, strong algorithms and methods to satisfy the quality level of the knowledge discovered.

The data generated by the bank's information systems, manual or automated such as ATM's and credit card processing, were designed to support or track daily transactions, simultaneously satisfying internal and external audit requirements, and meeting government or central bank regulations. Before data mining can proceed to find nontrivial information from large databases a data warehouse will have to be created first. Data warehousing is known as the process of extracting, cleaning, transforming, and standardizing incompatible data from the bank's current transaction systems so that these data can be mined and analyzed for useful patterns, relationships, and associations. Data mining is applied with success to prevent attrition, to cross sell and to do target marketing, to detect and deter fraud, to prevent defaults, bad loans, to increase loyalty and customer retention etc.

A challenge for the authoress of this article was to consider the bank as an automated control system, combining the Automatics theories with banking regulations and artificial intelligence techniques in order to implement a prototype of an automated system based on data mining for credit scoring.

2. Bank System as Automated Control System

As systems, banks have inputs (objectives, desired behaviour), outputs (results), actuating signals and disturbances. A cybernetic approach for banking activities was proposed by authoress, as reflected in the next three graphical representations of automated control systems, adapted from [9]. An economic control process model with a feedback loop was also proposed.

In Figure 1 a control system is illustrated for a bank’s mapping as a physical system. The bank’s management is acting in pursuit of economic and financial targets included in the economic state of reference (SEC). To achieve these objectives, the bank’s management owns financial, human and information resources. These resources are used in the production departments to obtain banking products and services (e.g. credit cards, bank checks etc.). Launch and market capitalization of the

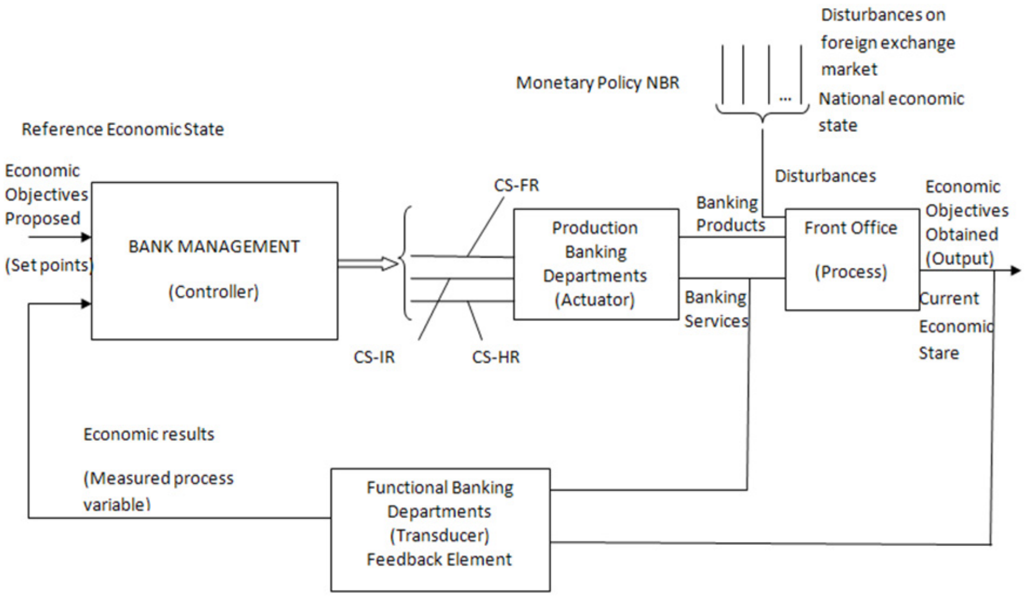
Disturbances on foreign exchange market
Monetary Policy NBR
...
National economic state
Reference Economic State
Economic Objectives Proposed
(Set points)
BANK MANAGEMENT
(Controller)
CS-FR
CS-IR
CS-HR
Production Banking Departments (Actuator)
Banking Products
Banking Services
Disturbances
Front Office
(Process)
Economic Objectives Obtained (Output)
Current Economic Stare
Economic results
(Measured process variable)
Functional Banking Departments (Transducer) Feedback Element

Figure 1. Feedback Control System for banking activities [adapted from 9]:
CS-FR – control signal financial resources; CS-IR – control signal informational resources;
CS-HR – control signal human resources; RES – Reference Economic State;
CES – Current Economic State

mentioned products/services is possible only through front office that establishes communication with bank customers. Functional departments process the data regarding production and market bank activities, pursuing the impact of these outputs of the banking system, and provide information to bank management as economic outcomes (e.g. loan portfolio, volume of deposits etc.).

In case the disturbances (monetary policy of National Bank of Romania, the national economy, foreign exchange fluctuations etc.) deviate the results from the initial goals settled by bank's management, controlling variables are adjusted in order to eliminate deviations.

In case of feedback control system regarding bank's activities, functional banking department evaluates the business objectives achieved. The results obtained are compared and, if there are any deviations from the proposed objectives, new control signals are processed to diminish or to remove the existing error. A common example for a control signal is changing banking strategy by reducing the interest.

An advantage for this feedback control architecture is that this type of control ensures the desired performance (bank's objectives) by altering the inputs, as soon as deviations are identified, regardless of what caused the disturbance. An additional advantage of the proposed feedback control presented in the Figure 1 is that, by analysing the output of a system (economic objectives), unstable processes may be stabilized.

A weakness of the feedback control is the impossibility to recognize and analyse a process deviation that occurred near the beginning of the process until the process output [4]. In this situation, the feedback control will then have to adjust the process inputs in order to correct (remove) the deviation produced.

For the structure proposed in Figure 1, a disadvantage of this control loop is that error cannot be removed immediately, which triggers an imbalance in banking activities during the transient time.

The authoress proposed o second control loop for banking activities represented in Figure 2. In this case, a feedforward control has been chosen. The bank management is informed about the current state and economic trends from the banking market. In these conditions, the control signals are elaborated taking into account both the set points (the banking objectives) and the additional information.

If a disturbance occurs regarding the banking market (i.e. lower demand for certain banking services, the emergence of a new banking market competition, increasing the interest rate

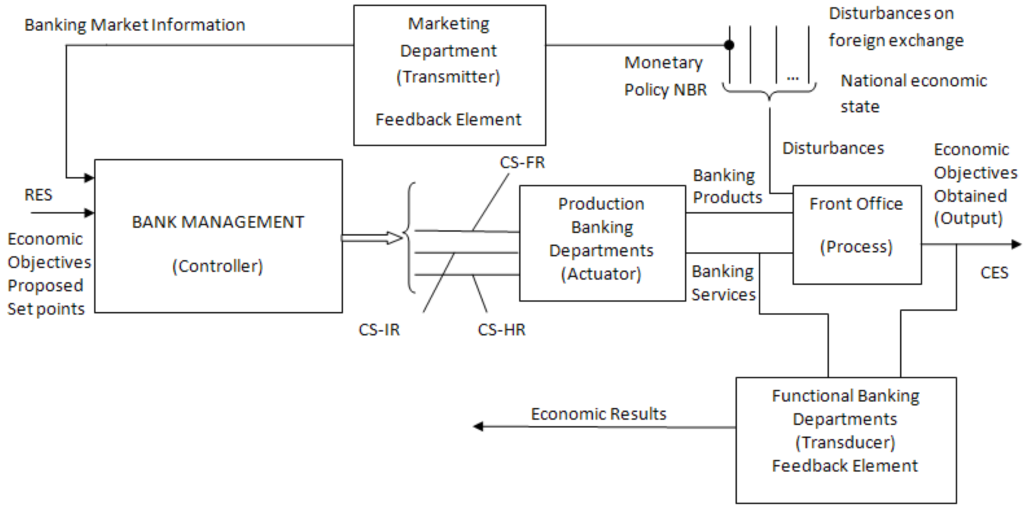
Banking Market Information
Marketing Department (Transmitter)
Feedback Element
Monetary Policy NBR
Disturbances on foreign exchange
...
National economic state
Disturbances
RES
BANK MANAGEMENT
(Controller)
Economic Objectives Proposed Set points
CS-FR
CS-IR
CS-HR
Production Banking Departments (Actuator)
Banking Products
Banking Services
Front Office
(Process)
Economic Objectives Obtained (Output)
CES
Economic Results
Functional Banking Departments (Transducer) Feedback Element

Figure 2. Feedforward Control System for banking activities [adapted from 9]
CS-FR – control signal financial resources; CS-IR – control signal informational resources;
CS-HR – control signal human resources; RES – Reference Economic State; CES – Current Economic State

etc), the bank management will adopt those decisions and will generate those control signals that may prevent the disturbances between economic results. The efficiency of the controller outputs is conditioned by the prediction ability of the bank management. SEC do not change the value for the disturbances taken into account, but when new disturbance factors occur, the current state of the system cannot be equalized with setpoint.

The combination of the two proposed control loops (feedback control and feedforward control) considers both tracing output (controlled variable) relative to the setpoint and the disturbances evolution [8]. In this case, bank management is permanently informed about the changes on the banking market. The resulted automated control system that contains the feedback control and the feedforward control is proposed in Figure 3.

The advantages of both control loops (feedback and feedforward) are cumulated to maintain the outputs of the bank system in concordance with the objectives aimed at and taking into account the disturbances traced.

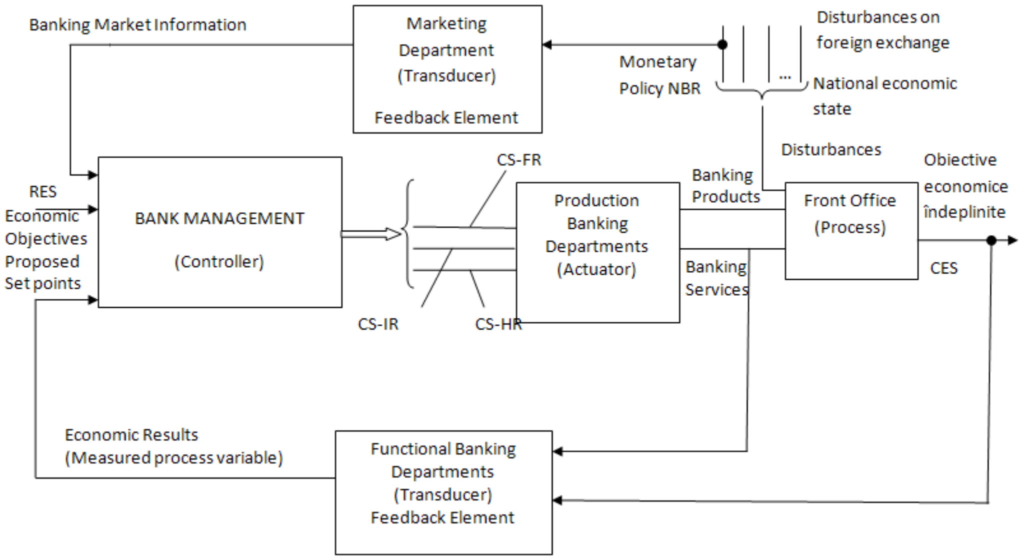
Banking Market Information
Marketing Department (Transducer)
Feedback Element
Monetary Policy NBR
Disturbances on foreign exchange
...
National economic state
Disturbances
RES
Economic Objectives Proposed Set points
BANK MANAGEMENT
(Controller)
CS-FR
CS-IR
CS-HR
Production Banking Departments (Actuator)
Banking Products
Banking Services
Front Office
(Process)
Obiective economice îndeplinite
CES
Economic Results (Measured process variable)
Functional Banking Departments (Transducer) Feedback Element

Figure 3. Feedforward Control System for banking activities [adapted from 9];
CS-FR – control signal financial resources; CS-IR – control signal informational resources;
CS-HR – control signal human resources; RES – Reference Economic State;
CES – Current Economic State

2. Data Mining and Credit Scoring

Data mining represents a domain with a wide range of application in finance, medicine, education, engineering, telecommunication etc. Mining data is a sensible task that requires experience and adequate methods for practitioners who want to discover new meanings and connections between data from large repositories [1].

Data mining techniques (clustering, decision trees, association rules, neural networks etc.) can be successfully applied if the following conditions are met:

- large quantities of data are provided;
- variety of data is analyzed;
- database techniques are used;
- information obtained from past data is used;
- real time analyses are made.

Data mining is the analysis of historical business activities, stored as static data in data warehouses, to reveal hidden patterns and trends [2]. Examples of what businesses use data mining for include performing market analysis to identify new products and services, finding the root cause of manufacturing problems, to prevent customer attrition and to attract new customers, cross-sell to existing customers, and profile customers with higher accuracy [2, 7].

Data mining can contribute to solving business problems in banking and finance by finding patterns and associations in business information and market prices that are not immediately apparent to managers. Data mining techniques can be used to discover hidden knowledge, unknown patterns and new rules from large data sets, which may be useful for a variety of decision-making activity [1, 7].

For bank managers, credit risk assessment is the key component in the process of lending. The objective of credit scoring is to help credit providers to quantify and manage the financial risk involved in providing credit, so that they can make good lending decisions in a short time and more objectively and to predict the probability that a loan applicant or an existing borrower may default or became delinquent.

As a definition, credit scoring represents a numerical expression that indicates the creditworthiness of a person who applies for a loan [5]. To determine whether credit applicants will fraud, it is necessary to use a credit scoring model described by variables indicating the risk of fraud. Identifying the most significant variables for these models is a difficult task, given the large number of data sets used in the designing the credit model phase. In practice, there have been mentioned several methods for determining the characteristics of a credit scoring model, as follows [6, 10]:

- use of the experts' knowledge in lending practices;
- use of the statistic procedure stepwise;
- selection of individual features, by using a distance measure between the feature distributions in case of occurrence/not occurrence of an event.

Variables that usually describe a credit scoring model are classified in four categories as follows: demographic indicators, financial indicators, employment status of the applicant indicators, behavioural indicators [5, 6, 11].

Demographic indicators focus on differences between categories of applicants regarding age, sex, marital status etc. During a numerous of analysed cases it was found that an elder person has a low risk of fraud, as well as a married person who is supposed to be more responsible than a person without such obligations. Also, a person living in a house which is a personal property can guarantee with that building and therefore, has a lower risk of fraud, unlike another person who lives in a rented house.

Financial indicators are significant because these indicators provide the opportunity to identify financial resources of repayment and the gearing ratio for applicant.

Indicators describing the employment status of an applicant provide information about the existence of a salary (if the person is employed for an indefinite period) or on the contrary, the lack of this source of income (for an unemployed person).

Behavioural indicators offer important information about the relations between the applicant and the bank (the history of banking activities regarding the applicant). Guarantees play a major role among the characteristics analysed when applying a credit scoring model, in that they indicate the possibility to recover the borrowed amount and also provide a financial safety of the applicant.

In the next section, the authoress presents a solution for credit scoring, namely an automated system based on data mining techniques used as a decision system.

3. SADM Architecture and Functionality

The proposed system consists of eleven modules (Figure 4), each module having a particular task.

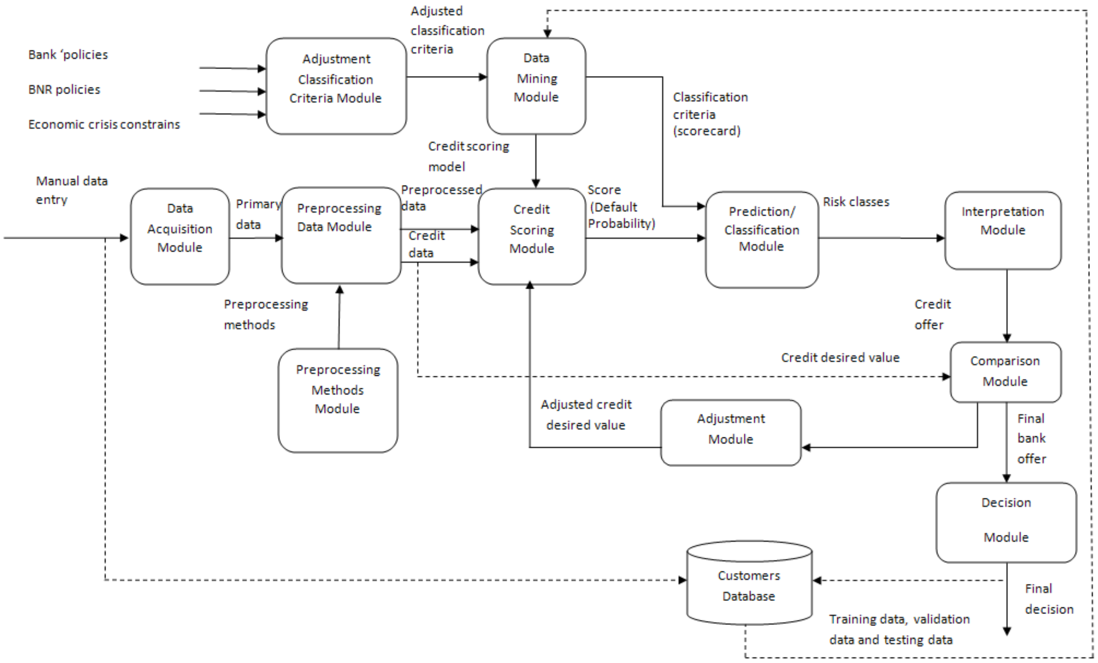

Figure 4. The SADM architecture [3]

Data Acquisition Module (DAM) represents the input block and allows storing data entered manually by the customer. Raw data are preprocessed by Data Preprocessing Module (DPP) based specific methods provided by the Preprocessing Methods Module (PMM). The "cleaning" task refers to the frequent operations applied to raw data, operations such as: eliminating extreme values (outliers), filling the missing values (missing values), removing incorrectly entered data, data sampling, variable selection etc. The core module of the automated system is represented by Credit Scoring Module (CSM), which provides the output as default probability of borrower. The role of Data Mining Module (DMM) is to automate the discovery of patterns from massive customer data. The feedforward control behaviour of the system gives to this module the ability to adapt, in the sense that tracking disturbances (central bank policy, bank policy, economic crisis, etc.) may modify the credit scoring model by adjusting the criteria for the classification of customers. Adjusting Module Classification Criteria (AMCC) is the module responsible for making adjustments to the classification criteria imposed by the bank management. The input variable of Prediction/Classification Module (PCM) is represented by the obtained score or default probability associated to customer creditworthiness. Decision module (DM) produces the final decision (that can have the value approved/rejected) and sends the results to the customers' data base to update the stored data. The loan officer is informed about the final decision, in order to transmit it to the borrower. Interpretation Module (IM) has the role to "translate" the input variable (risk classes) into a credit offer which is compared by the Comparison Module (CM). The final decision takes into account the credit desired value (setpoint for the CM block). Adjustments Module (AM) modifies the customer credit option (setpoint for CSM block) in order to obtain a score which will permit a better credit offer.

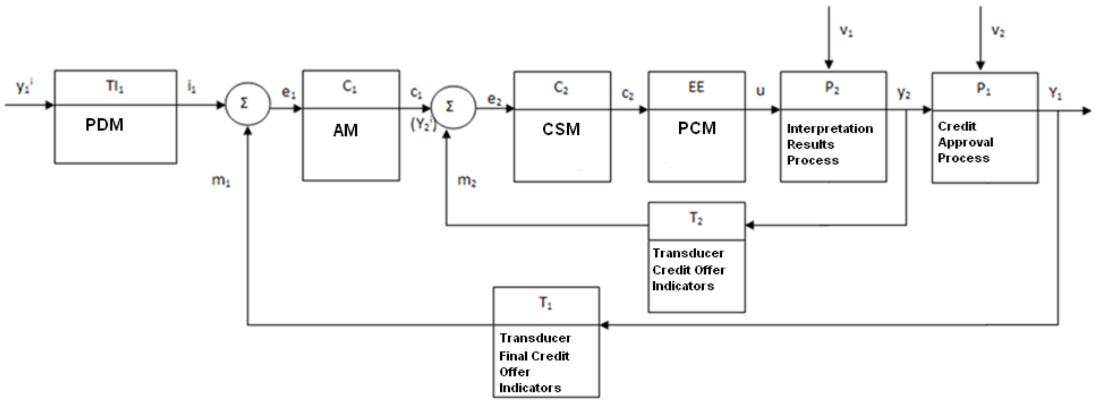
v₁
v₂
y₁ⁱ
TI₁
PDM
i₁
Σ
e₁
C₁
AM
c₁
(Y₂ⁱ)
Σ
e₂
C₂
CSM
c₂
EE
PCM
u
P₂
Interpretation Results Process
y₂
P₁
Credit Approval Process
Y₁
m₁
m₂
T₂
Transducer Credit Offer Indicators
T₁
Transducer Final Credit Offer Indicators

Figure 5. SADM – the proposed cascade control scheme; PDM – Preprocessing Data Module; AM – Adjustment Module (Setpoints); CSM – Credit Scoring Module; PCM – Prediction/Classification Module; TI1 – Transducer Input; T1, T2 – Transducers; C1, C2 – Controllers; EE – Execution Element (Actuator); P1, P2 - Processes

The comparative study made by the authoress regarding control loops, constituted a reference support for building cascade control structure that can be associated to the automatic system based on data mining proposed as decision support for credit approval process. Figure 5 represents the obtained control structure where:

- y_1 represents the final credit offer of bank (process variable);
- y_1^i – customer initial credit application (request);
- i_1 – credit request indicators (credit amount, credit period etc);
- m_1 – final credit offer indicators;
- e_1 – error between credit offer (measure process variable) and request offer (setpoint);
- $c_1 = y_2^i$ - intern control loop setpoint (control signal produced by controller C1);
- y_2^i – adjusted indicators for credit setpoint;
- c_2 - score (probability of default) (control signal produced by controller C2);
- y_2- calculated credit offer (process variable);
- m_2 – calculated credit offer indicators (measured process variable);
- u – risk classes;
- e_2 – intern control loop error;
- v_1, v_2 – disturbances;
- $\sum -$ evaluation elements (comparison elements) of errors.

Generally, the output of the primary controller determines the set point for the secondary controller and the output of the secondary controller is used to adjust the control variable [8, 12].

Intern control loop acts more quickly than extern control loop. Once the cascade structure is implemented, disturbances from rapid changes of the secondary controller will not affect the primary controller [8, 13].

In figure 5 the role of secondary controller is taken by the Credit Scoring Module (CSM), while primary controller is represented by the Adjustment Module (AM). In order to have a constant flow of information throughout the control system some conditions need to be applied:

- there must be an evident relationship between the measured variables of the primary and secondary loops (Credit Offer Indicators and Final Credit Offer Indicators);
- the secondary loop must influence the primary loop (the value of credit scoring model influences the final bank offer);
- the major disturbance to the system should act in the primary loop .

4. Credit Scoring Model for SADM

SADM uses a hybrid credit scoring model which is a combination of four credit scoring models: two regression models (modelRegLog1and modelRegLog2), a neural network model (modelRNA) and a decision tree model (modelDT).

The process model tested for SADM has the following expression:

$$\begin{aligned} modelSADM = 0{,}15 \times modelRegLog1 \\ + 0{,}25 \times modelReglog2 \\ + 0{,}50 \times modelRNA \\ + 0{,}10 \times modelAD \end{aligned} \quad (1)$$

The coefficients of the SADM process model were empirically determined after several experiments and tests.

Each credit scoring model produces as output a value (approved/rejected) that is binary encoded (1/0). The final decision regarding credit approval is taken relative to an established threshold (0.60).

$$BankDecision = \begin{cases} approved, modelSADM > 0{,}60 \\ rejected, modelSADM \leq 0{,}60 \end{cases} \quad (2)$$

According to the control structure associated to the SADM proposed in Figure 5, the process variable (output) is represented by BankDecision, and the cause-and-effect relation has the mathematical expression presented in (2). Considering the binary codification, the cause-and-effect function may be described as follows:

$$f: [0,1] \rightarrow \{0,1\}, y = f(u), \quad (3),$$

where: u is modelSADM and y is BankDecision.

The cause-and-effect relation associated to the actuator (EE)(Prediction/Classification Module /PCM) has the mathematical expression:

$$f: \mathbb{R} \rightarrow \{0,1,2,3,4\}, u = f(c_2), \quad (4)$$

where u is the risk class and c_2 represents the score obtained for an applicant. As a result of a significant number of tests realized by the authoress, five risk classes were identified, as it is shown in Table 1.

Table 1. Risk classes

Score	Risk classes
score > 500	No risk (0)
300 ≤ score ≤ 500	Minimum risk (1)
200 ≤ score ≤ 300	Medium risk (2)
100 ≤ score ≤ 200	Significant risk (3)
score < 100	Maximum risk (4)

5. Results and Conclusions

The current paper is based on the research work of the authoress regarding the interpretation of the banking system as an automated system and the modelling of credit scoring process. As a result of the systemic approach of the bank management, the authoress proposed three control structures (feedback control, feedforward control and a combination of the two control structures). By using this framework, an automated control system was designed for mapping the credit scoring process. SADM (automated system based on data mining) contains eleven modules, each module having an associated function. A significant contribution of the authoress is the proposed cascade control structure corresponding to the SADM architecture. The credit scoring model is a linear expression with four components: two regression models (modelRegLog1and modelRegLog2), a neural network model (modelRNA) and a decision tree model (modelDT), the coefficients of the model being obtained after a significant number of experiments. The cause-and-effect relation describing the bank decision uses a binary codification for the response variable (0-rejected, 1- approved) and takes into consideration a threshold as 60%. Also, five risk classes were identified for credit applicants.

Future work will focus on developing the credit scoring model proposed in this paper and in increasing the accuracy of the results.

REFERENCES

1. GORUNESCU, F., **Data Mining. Concepte, Modele şi Tehnici**, EdituraAlbastră, Cluj-Napoca, 2006.
2. HAN, J., M. KAMBER, **Data Mining Concepts and Techniques**, San Francisco, Academic Press, 2001.
3. IONITA, I., IONITA, L., **A Decision Support based on Data Mining in e-Banking**, Roedunet International Conference (RoEduNet) 2011 10th, 2011, pp. 1-5.
4. ILAS, C., **Teoria sistemelor de reglare automată**, Bucureşti, Matrix Rom, 2006.
5. LIU, Y., **A Framework of Data Mining Application Process for Credit Scoring**,

http://www.econbiz.de/archiv1/2010/102367_datamining_creditscoring_framework.pdf

6. KISS, F., **Credit Scoring Processes from a Knowledge Management Perspective,** Periodica Polytehnica Services, Social, and Management Sciences, vol. 11, no. 1, 2003, pp. 95-110.

7. MEGGS, G., M. FAGAN, **Knowledge Discovery – Practical Methodology and Case Studies,** Tutorial Notes in PADD98 – 2nd International Conference on Practical Applications of Knowledge Discovery and Data Mining, London, 1998.

8. MIHALACHE, S., **Elemente de ingineria reglării automate**, Bucureşti, MAtrix Rom, 2008.

9. PARASCHIV, N., G. RĂDULESCU, **Introducere în ştiința sistemelor şi a calculatoarelor**, Editura Matrix Rom, Bucureşti, 2007.

10. SCHOMAKER, C. A. M., **Credit Scoring. An Overview of Traditional Methods and Recent Improvements**, BMI Paper, December, 2006.

11. SIDDIQI, N., **Credit Risk Scorecards. Developing and Implementing Intelligent Credit Scoring**, John Wiley &Sons Inc., New Jersey, 2006.

12. WANG, X. Z., **Data Mining and Knowledge Discovery for Process Monitoring and Control,** Advances in Industrial Control, Springer, 1999.

13. WEISS, S. M., N. INDURKHYA, **Predictive Data Mining: a Practical guide**, Morgan Kaufmann Publishers, San Francisco, California, 1998.