Selection of Variables for Credit Risk Data Mining Models: Preliminary research

M. Pejić Bach*, J. Zoroja*, B. Jaković*, N. Šarlija**
* University of Zagreb, Faculty of Economics & Business, Zagreb, Croatia
** University of Osijek, Faculty of Economics, Osijek, Croatia
mpejic@efzg.hr, jzoroja@efzg.hr, bjakovic@efzg.hr, natasa@efos.hr

Abstract - Credit risk is related to the risk of the borrower that the lender will not be able to return their debt including interest. Numerous researches have been conducted in the area of credit risk, both using classical models such as Altman Z-score and using machine learning methodology. However, the research using the data from Croatian financial institutions is scarce, especially research focused on the selection of the demographic and/or behavior variables. In addition, it is important to develop robust models that estimate credit risk as accurately as possible. The goal of this research is to develop a data mining model for prediction of credit risk, using the data from Croatian financial institutions on defaulted clients (demographic and behavior data). Decision tree models are constructed for the prediction of credit risk. Different algorithms for the variable selection are evaluated based on the classification accuracy of the decision trees developed based on the selected variables. This work has been fully supported by the Croatian Science Foundation under the project "Process and Business Intelligence for Business Performance" - PROSPER (IP-2014-09-3729).

I. Introduction

Data mining methods are used for finding undiscovered valuable information from large databases. In other words, the main goal of data mining techniques is to extract knowledge in order to make successful management decisions [1]. Applications of data mining methods are used in almost every industry: banking, marketing, finance, manufacturing, medicine, education, trade, supply [2, 3, 4]. Each industry has its own characteristics, which implies usage of different data mining methodologies. Therefore, in the banking industry, characterized with a high level of fraud and risks, which requires successful prediction of credit default, scoring and applicants, usage of data mining techniques is very common [5]. Data mining usage is one of the most common techniques used in the financial analysis, especially in the banking industry. Prediction of credit risk, mostly prediction of credit default, presents an important activity of the banking industry [6]. There are several different data mining techniques that can be used for financial data analysis because of their high level of success. However, their success also depends on on the data available, its cleaning, and transformation. Therefore, decision trees are one of the most commonly used methods [7, 8]. Decision trees are one of the classification methods which group variables into one or more categories of the target variables [9]. When using the decision trees process it is important to follow three main steps: (i) determine the sample, (ii) choose variables, and (iii) select an appropriate algorithm. In this paper we have used the algorithm C4.5, as one of the ten most popular variables. The goal of the paper is to compare the classification of banking clients according to the credit default, with the C4.5 decision tree algorithm, using different sets of the variables: Entrepreneurial idea; Growth plan; Marketing plan, Personal characteristics of entrepreneurs, Characteristics of SME, Characteristics of credit program, and Relationship between the entrepreneur and a financial institution. The variables are selected by the usage of three different algorithms, provided in the Weka software: Class CfsSubsetEva algorithm, ChiSquaredVariableEval algorithm, and ConsistencySubsetEval algorithm. Previous research that tested the efficiency of algorithms for the selection of variables was mostly conducted on the retail credit risk datasets, e.g. Oreski et al. [10]. The scientific contribution of our paper is that the algorithms for the selection of variables are tested on the real-world dataset of credit risk of Croatian financial institution's business clients (entrepreneurs from Eastern Croatia). The paper consists of six sections. After the Introduction, as the first section, there is Literature review. In Literature review, data mining methodology and its usage for predicting credit default are presented. Decision trees, as one of the data mining methods, are described as well as variables and techniques selection approaches used in this research. In the third section named Methodology, data, decision trees techniques and the variable selection process are discussed. Research results are provided in the fourth and the fifth section. The fourth section elaborates on results of the different variable selection strategies, while the fifth section of the paper discusses results regarding classification efficiency measures, classification matrices and falsely predicted good and bad debtors with different variable selection approaches. The last section is Conclusion.

II. LITERATURE REVIEW

A. *Data mining methodology*

The amount of data has been constantly increasing, which creates difficulties for managers and successful decision making. A high growth of valuable as well as invaluable data in databases has created a need for the use of different methodologies which can help finding, extracting and analyzing data important for decision makers [11].

Data mining technology combines different approaches, e.g. machine learning, statistics and database management, which are used for finding valuable patterns in data for further prediction and decision making. In addition, data mining techniques can also be used for determining relationships among data in order to create knowledge [12]. The main purpose of data mining is to find and analyse disorganized information with the goal of improving business knowledge and activities.

The most commonly used data mining methods are: classification, regression, clustering, visualization, decision trees, association rules, neural networks, support vector machine [5, 13]. In our research we conducted a decision tree analysis, which is mainly used for classification, in order to predict the credit default.

The main purpose of the decision tree analysis is to predict behavior of the target variable using different algorithms to get the best outcome [14]. Some of possible algorithms are: C4.5, ID3, CART, CHAID, and MARS. One of the disadvantages of the decision tree analysis is that, in the process, analysts should pay more attention to a high variance. Other disadvantage of the decision tree is the overfitting, which occurs when model is excessively complex, and it has a poor predictive efficiency. However, the most important advantage of the decision trees is the possibility of interpretation, which together with the simple usage and implementation make the decision tree method appropriate for a wide range of research.

B. Predicting credit default with data mining approach

Countries, especially their economies and financial institutions, have been facing a strong financial crisis in the last years. Therefore, in many countries, governments have brought many saving measures in order to decrease costs and to restart economy development. In addition, credit default has increased and nowadays banks pay much more attention to credit risk assessment and to prediction of credit default with the goal to reduce risk [15].

Financial institutions and banks are using different intelligent techniques, e.g. mathematical models, statistics analysis, data mining methods with the goal to make efficient credit decisions. A detailed analysis of data on the characteristics of current and previous credit users plays an important factor in forecasting the future credit default of new clients [16].

C. Variables and techniques selection approaches

In order to predict credit default, financial institutions and banks mostly use behavioral and demographic variables of previous and current clients, e. g. monthly income, marital status, real-estate owner, employment, age, gender [17].

The main purpose of our research is to classify banking clients regarding credit default with a decision tree analysis, using different variables related to entrepreneurship activity. In addition, financial institutions and banks, when approving credits to clients, strive to select those clients who will be able to repay it in the given period of time [18]. In other words, they are focused on good clients.

There are also studies about methods used in credit scoring. One of the examples is a research which used demographic and behavioral data and three data mining methods: credit scorecard, logistic regression and a decision tree model [9]. The results of the research showed that all three methods are appropriate for use but the scorecards method is the easiest to apply.

There is also a study in which authors have investigated recent researches conducted in the field of credit risk assessment regarding clients and their ability to repay credits [19]. Research results showed that logistic regression is the most commonly used method to group clients into good or bad debtors.

Recent studies showed that intelligence methods used for discovering credit scoring are mostly non-parametric methods and computational intelligence techniques, e.g. decision trees, artificial neural networks, support vector machines and evolutionary algorithms [20, 21, 17].

III. METHODOLOGY

A. Data

Data used in this research were collected from an entrepreneurship credit dataset. Data were collected randomly from the database of clients (entrepreneurs from Eastern Croatia) of the financial institution, that is focused on financing small and medium enterprises, mostly start-ups. There are 200 applicants in the sample.

There are two main reasons for a small dataset: (i) a quite low level of business activity regarding entrepreneurship in Croatia (Total Early State Entrepreneurial Activity (TEA) index for the year 2015 = 7.69; TEA index for the year 2004 = 7.97) which means that a low number of people is taking a credit to start a business, and (ii) financial institutions are rejecting too risky start-ups applications for a credit. Therefore, collecting a larger sample will be possible in the next few years, when the entrepreneurial climate and perception of entrepreneurship activity improves.

The following variables were used for the development of the credit scoring model: Entrepreneurial idea; Growth plan; Marketing plan, Personal characteristics of entrepreneurs, Characteristics of SME, Characteristics of credit program, Relationship between the entrepreneur and a financial institution, and Creditworthiness. Most of the variables are nominal, while two of them are numeric (Entrepreneurs' age, Number of employees, and Credit amount). Variables related to the future plans for the SME were estimated by a banking clerk (Table I). First, it was estimated whether the entrepreneur has a clear vision of the business development (for newly established SMEs), or it is already an established business (Variable Vision).

Second, the variable Better estimated what the main competitive advantage of the SME (better quality, technology, price, or expertise of employees) is.

Third, it was estimated what the main market for SME’s products/services is: local, national, wider region, or narrow targeted customers (Variable Market). Entrepreneurs stated which percentage of the profit is planned to be reinvested in the business operations (Variable Reinvest), and what the plans for the promotion of products/services are (Variable Ad). Also, it was estimated whether the entrepreneur can identify who SME's main competitors are.

TABLE I VARIABLES RELATED TO THE FUTURE PLANS FOR THE SME

Variable	Question asked	Answers
I Entrepreneurial idea		
Vision	Does the entrepreneur have a clear vision of the business?	1 – no clear vision (for newly established SMEs) 2 – clear vision (for newly established SMEs) 3– established business
Better	Advantages of products/services	1 – better quality 2 - better technology 3 – good price 4 – expertise of employees 100 – no answer
Market	Market for products/services	1 – local 2 – narrow targeted customers 3 – wider region 4 – Croatia 100 – no answer
II Growth plan		
Reinvest	Projected percentage of the invested profit (reinvested profit/profit*100)	1 - 0 to 30% 2 - 30.01 to 50% 3 - 50.01 to 70% 4 –70.01 to 100% 100 – missing value
III Marketing plan		
Ad	Promotion of products/services	1 – without promotion 2 – no need for promotion 3 – all media 5 – personal selling, presentation 6 – posters, leaflets, internet 100 – missing value
Compet	Can the entrepreneur identify competition?	1 – no competition 2 – not defined 3 – defined competition 100 – no answer

Source: Authors, using Entrepreneurship credit dataset

Table II presents the variables related to the characteristics of the entrepreneur and SME. Entrepreneurs' occupations are grouped into 5 main groups: 1 - farmer, veterinarian; 2 - trader, restaurateur; 3 - construction worker; 4 - engineer, physician, and pharmacist, 5 - Technologist, chemist. Entrepreneurs' ages are expressed as a numeric variable. Entrepreneurs' locations refer to 4 geographic areas in Croatia.

Table III represents the variables related to the credit program and the bank: interest repayment frequency (monthly, quarterly, half-yearly), grace period, principal repayment, repayment period (expressed in months), interest rates, and amount of credit (expressed in local currency). Also, the variable Client measures whether an entrepreneur has applied for a credit before.

Table IV represents the classification variable that was used for the credit scoring (variable Default), and groups clients as "bad" or "good" based on the regularity of their payment.

TABLE II VARIABLES RELATED TO THE CHARACTERISTICS OF ENTREPRENEUR AND SME

Variable	Question asked	Answers
IV Personal characteristics of entrepreneurs		
Occup	Entrepreneurs' occupations	1 - farmer, veterinarian 2 - trader, restaurateur 3 - construction worker 4 - engineer, physician, pharmacist 5 - technologist, chemist
Age	Entrepreneurs' ages	numeric
Location	Entrepreneurs' locations	1 – Baranja, Osijek 2 - S.Brod, Požega 3 - Đakovo, Našice 4 - Vinkovci, Vukovar
V Characteristics of SME		
Ind	Industry	1 - plastics, textiles 2 - car service 3 - food production 4 - health and intellectual services 5 - agriculture 6 - construction 13 - tourism
Start	Is this a new business venture	1 – yes 2 – no
Equip	Does the entrepreneur have some equipment?	1 – yes 0 – no
Emp	Number of employees	numeric

Source: Authors, using Entrepreneurship credit dataset

TABLE III VARIABLES RELATED TO THE CREDIT PROGRAM AND THE BANK

Variable	Question asked	Answers
VI Characteristics of credit program		
Int	Interest repayment frequency	1- monthly 2 - quarterly 4 - half-yearly
Grace	Grace period	1 – yes 0 – no
Prin	Principal repayment	1- monthly 5 – yearly
Period	Repayment period (months)	numeric
I_rate	Interest rate	4,9% 6,9% 8,9%
Amount	Credit amount (local currency)	numeric
VII Relationship between the entrepreneur and a financial institution		
Client	Is this the first time the entrepreneur is applying for a credit?	1 – yes 2 – no

Source: Authors, using Entrepreneurship credit dataset

TABLE IV GOAL VARIABLE USED FOR THE CREDIT SCORING

Variable	Question asked	Answers
Default	"bad" clients are defined as those who have been late with their payments for more than 45 days at least once. Other clients who have not been late for more than 45 days are labeled as "good".	1-bad 0-good

Source: Authors, using Entrepreneurship credit dataset

B. Decision trees

Table V represents the Weka Description of the used algorithm. The C4.5 (weka.classifiers.trees.J48-C0.25-M5) algorithm was used for developing models for classification of debtors as good or bad. There are 200 instances in the model and the used test mode is 10-fold cross-validation.

TABLE V WEKA DESCRIPTION OF THE USED ALGORITHM

Weka feature	Used feature
Scheme:	weka.classifiers.trees.J48 -C 0.25 -M 5
Relation:	credit_scoring.arff
Instances:	200
Test mode:	10-fold cross-validation

Source: Authors, using Entrepreneurship credit dataset

C. Variable selection

Three approaches to the variable selection were applied: (1) selection of the variables using the Class CfsSubsetEval algorithm (searching approach BestFirst), (2) selection of the variables using the ChiSquaredVariableEval algorithm (searching approach Ranker), and (3) selection of the variables using the ConsistencySubsetEval (searching approach Greedy Stepwise).

There are differences in definition and usage of the three mentioned approaches to the variable selection which were applied. The Class CfsSubsetEval algorithm is based on the individual estimation of the variables that are highly correlated with the class variables but are not highly mutually correlated. The ChiSquaredVariableEval calculates the value of a variable regarding the value of the chi-squared statistic with respect to the class. The ConsistencySubsetEval calculates the value of a subset of variables by the level of reliability in the class values [22].

Table VI presents the variables used for a different approach to the variable selection. The Class CfsSubsetEval algorithm selected only three variables: Variable Vision - Does the entrepreneur have a clear vision of the business?; Variable Ad - Promotion of products/services, and Variable Reinvest - Projected percentage of the invested profit.

The ChiSquaredVariableEval algorithm selected all of the variables except the Variable Location. The ConsistencySubsetEval selected all of the variables related to the Entrepreneurial idea, Growth plan, and Marketing plan. However, only a few algorithms were selected that were related to the Personal characteristics of entrepreneurs, Characteristics of SME, Characteristics of a credit program, and Relationship between the entrepreneur and a financial institution.

TABLE VI VARIABLES SELECTED BY DIFFERENT ALGORITHMS

Variable	Class CfsSubsetEval	ChiSquaredVariableEval	ConsistencySubsetEval
I Entrepreneurial idea			
Vision	✓	✓	✓
Better	∅	✓	✓
Market	∅	✓	✓
II Growth plan			
Reinvest	✓	✓	✓
III Marketing plan			
Ad	✓	✓	✓
Compet	∅	✓	✓
IV Personal characteristics of entrepreneurs			
Occup	∅	✓	✓
Age	∅	✓	∅
Location	∅	∅	∅
V Characteristics of SME			
Ind	∅	✓	✓
Start	∅	✓	∅
Equip	∅	✓	∅
Emp	∅	✓	∅
VI Characteristics of credit program			
Int	∅	✓	✓
Grace	∅	✓	✓
Prin	∅	✓	∅
Period	∅	✓	∅
I_rate	∅	✓	∅
Amount	∅	✓	∅
VII Relationship between the entrepreneur and a financial institution			
Client	∅	✓	∅

Source: Authors, using Entrepreneurship credit dataset

IV. RESULTS

Table VII presents the results of different strategies for variable selection. The largest tree is produced by the ChiSquaredVariableEval (34 leaves and 48 nodes), which could be expected since this algorithm generated the largest number of independent variables. The next is the ConsistencySubsetEval algorithm with 24 leaves and 32 nodes. The smallest tree is produced using the Class CfsSubsetEval algorithm (3 leaves and 4 nodes).

TABLE VII CHARACTERITICS OF THE TREES DEVELOPED WITH DIFFERENT VARIABLE SELECTION APPROACHES

Variable selection	Number of Leaves	Size of the tree (number of nodes)
Class CfsSubsetEval	3	4
ChiSquaredVariableEval	34	48
ConsistencySubsetEval	24	32

Source: Authors, using Entrepreneurship credit dataset

Figure 1 represents the complete decision tree generated with the C4.5 algorithm, using the algorithms selected by the Class CfsSubsetEval algorithm. However, only one variable was selected for the tree development. The first number in the bracket is the total number of cases classified in the leaf. The second number is the number of those cases that are misclassified. For example, among 153 entrepreneurs that had a vision of established business, 104 were correctly classified as good, while 49 were incorrectly classified as bad.

```
vision = established_business: good (153.0/49.0)
vision = no_clear_vision: bad (14.0/2.0)
vision = clear_vision: good (28.0/5.0)
```

Figure 1 Decision tree developed using variables selected by the Class CfsSubsetEval algorithm (Source: Authors, using Entrepreneurship credit dataset)

Figure 2 represents a part of the decision tree generated with the C4.5 algorithm, using the algorithms selected by the ChiSquaredVariableEval algorithm.

```
vision = established_business
|  ad = missing_value
|  |  grace = yes
|  |  |  period = 12_months
|  |  |  |  occup = farmer_veterinarian: good (17.0/2.0)
|  |  |  |  occup = construction_worker: bad (1.0)
|  |  |  |  occup = trader_restaurateur: bad (2.0)
|  |  |  |  occup = engineer_physician_pharmacist: bad
(1.0)
|  |  |  |  occup = technologist_chemist: good (0.0)
|  |  |  period = 24_months
|  |  |  |  occup = farmer_veterinarian
|  |  |  |  |  reinvest = missing_value
|  |  |  |  |  |  age <= 38: bad (9.0/1.0)
|  |  |  |  |  |  age > 38
|  |  |  |  |  |  |  amount <= 22227.6: bad (2.0)
|  |  |  |  |  |  |  amount > 22227.6: good (6.0)
|  |  |  |  |  reinvest = 30.1.1950: bad (4.0/1.0)
|  |  |  |  |  reinvest = 70.01-100: good (1.0)
|  |  |  |  |  reinvest = 0-30: bad (2.0)
|  |  |  |  occup = construction_worker: good (1.0)
|  |  |  |  occup = trader_restaurateur: bad (0.0)
|  |  |  |  occup = engineer_physician_pharmacist: good
(4.0/1.0)
```

Figure 2 Decision tree developed using variables selected by the ChiSquaredVariableEval algorithm (Source: Authors, using Entrepreneurship credit dataset)

Figure 3 represents a part of the decision tree generated with the C4.5 algorithm, using the algorithms selected by the ConsistencySubsetEval algorithm.

```
vision = established_business
|  ad = missing_value
|  |  int = quarterly
|  |  |  occup = farmer_veterinarian: good (15.0/1.0)
|  |  |  occup = construction_worker: bad (1.0)
|  |  |  occup = trader_restaurateur: bad (2.0)
|  |  |  occup = engineer_physician_pharmacist: good (0.0)
|  |  |  occup = technologist_chemist: good (0.0)
|  |  int = monthly: good (36.0/10.0)
|  |  int = half-yearly: bad (14.0/4.0)
|  ad = personal_selling_presentation: good (20.0/1.0)
|  ad = all_media
|  |  compet = not_defined
|  |  |  reinvest = missing_value
|  |  |  |  grace = yes: good (3.0)
|  |  |  |  grace = no: bad (7.0/1.0)
|  |  |  reinvest = 30.1.1950: bad (2.0)
|  |  |  reinvest = 70.01-100: good (6.0/1.0)
|  |  |  reinvest = 0-30: good (0.0)
```

Figure 3 Decision tree developed using variables selected by the ConsistencySubsetEval algorithm (Source: Authors, using Entrepreneurship credit dataset)

V. DISCUSSION

The following section elaborates on research results regarding classification efficiency measures, classification matrices and falsely predicted good and bad debtors with different variable selection approaches.

Table VI presents classification efficiency measures. According to the percentage of correctly classified instances and the root mean squared error the best approach was to use the ConsistencySubsetEval algorithm. However, according to the Kappa statistics, the best approach was to use the Class CfsSubsetEval algorithm.

Table VIII Classification Efficiency Measures

Variable selection	Correctly classified instances	Root mean squared error	Kappa statistic
Class CfsSubsetEval	71,28 % (2)	0.4556 (2)	0.2059 (1)
ChiSquaredVariableEval	65.64 % (3)	0.5188 (3)	0.1747 (3)
ConsistencySubsetEval	70.77 % (1)	0.4548 (1)	0.1954 (2)

Note: Rank of the measure in parenthesis

Source: Authors, using Entrepreneurship credit dataset

Table VII presents classification matrices for the decision trees generated by the different sets of variables selected by the three algorithms. The decision tree generated using the variables selected by the Class CfsSubsetEval falsely predicted 27% of good clients as bad clients, but it falsely predicted only 1% of bad clients as good clients. The decision tree generated using the variables selected by the ChiSquaredVariableEval falsely predicted 22% of good clients as bad clients, but it falsely predicted 12% of bad clients as good clients. Finally, the decision tree generated using the variables selected by the ConsistencySubsetEval falsely predicted 27% of good clients as bad clients, but it falsely predicted only 2% of bad clients as good clients, thus producing the similar results as the Class CfsSubsetEval.

TABLE VII CLASSIFICATION MATRICES

	Predicted - good	Predicted - bad
Class CfsSubsetEval		
Observed - good	127 (64%)	2 (1%)
Observed - bad	54 (27%)	12 (6%)
ChiSquaredVariableEval		
Observed - good	105 (53%)	24 (12%)
Observed - bad	43 (22%)	23 (12%)
ConsistencySubsetEval		
Observed - good	126 (63%)	3 (2%)
Observed - bad	54 (27%)	12 (6%)

Source: Authors, using Entrepreneurship credit dataset

Figure 4 presents falsely predicted good and bad debtors with different variable selection approaches.

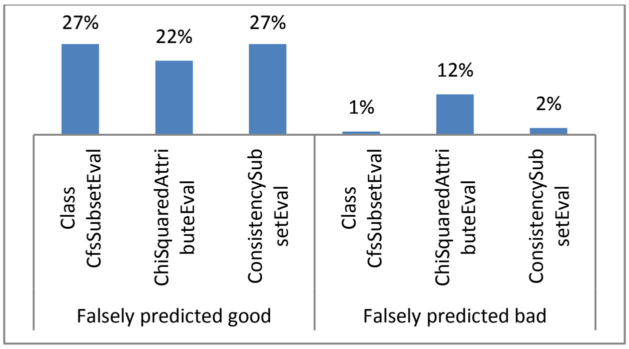

Figure 4 Falsely predicted good and bad debtors with different variable selection approaches (Source: Authors, using Entrepreneurship credit dataset)

VI. CONCLUSION

The main purpose of the paper is to compare the classification of banking clients regarding credit default through the analysis of different entrepreneurial variables using decision tree algorithms. Research results showed that variables selected by the algorithm Class CfsSubsetEval have the best results regarding the percentage of correctly classified instances. On the other hand, according to the percentage of bad debtors falsely predicted as the good ones, the decision tree generated using the variables selected by the ChiSquaredVariableEval is the worse. According to the criteria of the minimal percentage of falsely predicted bad debtors as good, the best approach was to use the decision tree generated using the variables selected by the Class CfsSubsetEval or the decision tree generated using the variables selected by the ConsistencySubsetEval. In addition, for financial institutions, especially for banks, the most valuable data are the data on prediction of bad debtors, and in our case two mentioned algorithms should be used. However, since the Class CfsSubsetEval generates a decision tree that is based only on the Variable Vision, it is prone to subjective mistakes, since this variable was estimated by a banking clerk. The ConsistencySubsetEval could be considered as more reliable, since it produces similar results as the Class CfsSubsetEval, and it is based on a larger number of variables. Most of the variables are related to ‘what has been done’ instead of ‘who is doing it’. In other words, variables related to Entrepreneurial idea, Growth plan, and Marketing plan were more relevant than variables related to Personal characteristics of entrepreneurs, Characteristics of SME, Characteristics of credit program, and Relationship between the entrepreneur and a financial institution.

REFERENCES

[1] R.-S. Wu, C.S. Ou, H.-Y. Lin, S.-I. Chang, D.C. Yen, “Using Data Mining Technique to Enhance Tax Evasion Detection Performance,” Expert Systems with Applications, vol. 39, no. 10, pp. 8769-8777, 2012.

[2] J.-T. Wei, M.-C. Lee, H.-K. Chen, H.-H. Wu, „Customer Relationship Management in the Hairdressing Industry: An Application of Data Mining Techniques,” Expert Systems with Applications, vol. 40, no. 18, pp. 7513-7518, 2013.

[3] M.A.P.M. Lejeune, “Measuring the Impact of Data Mining on Churn Management,” Internet Research: Electronic Networking Applications and Policy, vol. 11, no. 5, pp. 375-387, 2001.

[4] A.K. Choudhary, J.A. Harding, M.K. Tiwari, “Data Mining in Manufacturing: A Review Based on the Kind of Knowledge,” Journal of Intelligent Manufacturing, vol. 20, no. 5, pp. 501-521, 2008.

[5] E.W.T. Ngai, Y. Hu, Y.H. Wong, Y. Chen, X. Sun, “The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature,” Decision Support Systems, vol. 50, no. 3, pp. 559-569, 2011.

[6] L.C. Thomas, R.W. Oliver, D.J. Hand, “A Survey of the Issues in Consumer Credit Modelling Research“, Journal of the Operational Research Society, vol. 56, no. 9, pp. 1006-1015, 2005.

[7] J. Quinlan, „C4.5: Programs for Machine Learning,“ San Francisco, Calif.: Morgan Kaufmann, 1992.

[8] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, „Classification and Regression Trees,“ Belmont, Calif.: Wadsworth, 1984.

[9] B.W. Yap, S.H. Ong, N.H.M. Husain, “Using Data Mining to Improve Assessment of Credit Worthiness via Credit Scoring Models”, Expert Systems with Applications, vol. 38, no. 10, pp. 13274-13283, 2011.

[10] S. Oreski, D. Oreski, G., Oreski, “Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment”, Expert systems with applications, vol. 39, no. 16, pp. 12605-12617.

[11] P. I. Priya, D. K. Ghosh, ”A Survey on Different Clustering Algorithms in Data Mining Technique” International Journal of Modern Engineering Research (IJMER), vol. 3, no. 1, pp. 267-274, 2013.

[12] E.W.T. Ngai, L. Xiu, D.C.K. Chau, “Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification,” Expert Systems with Applications, vol. 36, no. 10, pp. 2592-2602, 2009.

[13] S. Strohmeier, F. Piazza, “Domain Driven Data Mining in Human Resources Management: A Review of Current Research,” Expert Systems with Applications, vol. 40, no. 7, pp. 2410-2420, 2013.

[14] H.G. Patel, K. Sarvakar, “Research Challenges and Comparative Study of Various Classification Technique Using Data Mining”, International Journal of Latest Technology in Engineering, Management & Applied Science, vol. 3, no. 9, pp. 170-176, 2014.

[15] Y. Marinakis, M. Marinaki, M. Doumpos, N. Matsatsinis, C. Zopounidis, “Optimization of nearest neighbor classifiers via metaheuristic algorithms for credit risk assessment,” Journal of Global Optimization, vol. 42, no. 2, pp. 279–293, 2008.

[16] L.C. Thomas, “A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Consumers,” International Journal of Forecasting, vol. 16, no. 2, pp. 149-172, 2000.

[17] A. Lucas, “Statistical challenges in credit card issuing, ”Applied Stochastic Models in Business and Industry, vol. 17, no. 1, pp. 83–92, 2001.

[18] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A.F.M. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, “Top 10 Algorithms in Data Mining”, Knowledge and Information Systems, vol. 14, no. 1, pp. 1-37, 2008.

[19] J.N. Crook, D.B. Edelman, L.C. Thomas, ”Recent Developments in Consumer Credit Risk Assessment”, European Journal of Operational Research, vol. 183, no. 3, pp. 1447-1465, 2007.

[20] D.F. Zhang, S. Leung, Z.M. Ye, “A decision tree scoring model based on genetic algorithm and K-means algorithm,” In: Proceedings of the 3rd International Conference on Convergence and Hybrid Information Technology, Daejeon, Korea, pp. 1043–1047, 2008.

[21] A. Mahmoud, M. Pourzandi, K. Babaei, “Using genetic algorithm in optimizing decision trees for credit scoring of banks customers,” Journal of Information Technology Management, vol. 2, no. 4, pp. 23–38, 2010.

[22] M.A. Hall, “Correlation-Based Feature Subset Selection for Machine Learning”, Hamilton, New Zealand, 1998.