MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
SUMY STATE UNIVERSITY
UKRAINIAN FEDERATION OF INFORMATICS

PROCEEDINGS

OF THE VI INTERNATIONAL SCIENTIFIC CONFERENCE

ADVANCED INFORMATION SYSTEMS AND TECHNOLOGIES

AIST-2018

(Sumy, May 16–18, 2018)

SUMY
SUMY STATE UNIVERSITY
2018

UDC 004(063)
A

A Advanced Information Systems and Technologies: proceedings of the VI international scientific conference, Sumy, May 16–18 2018 / Edited by S. I. Protsenko, V. V. Shendryk – Sumy: Sumy State University, 2018 – 145 p.

This book comprises the proceedings of the VI International Scientific Conference “Advanced Information Systems and Technologies, AIST-2018”. The proceeding papers cover issues related to system analysis and modeling, project management, information system engineering, intelligent data processing, computer networking and telecomunications, modern methods and information technologies of sustainable development. They will be useful for students, graduate students, researchers who interested in computer science.

UDC 004(063)

ISSN 2311-8504

International Scientific Committee:

A.N. Chornous, Sc.D (Ukraine)
A.S. Dovbysh, Sc.D (Ukraine)
O.A. Borisenko, Sc.D (Ukraine)
E.A. Lavrov, Sc.D (Ukraine)
V.O. Lyubchak, PhD (Ukraine)
S.I. Protsenko, Sc.D (Ukraine)
A.M. Kulish, Sc.D (Ukraine)
M.M. Glybovets, Sc.D (Ukraine)
Yu. I. Grytsyuk , Sc.D (Ukraine)
I.V. Grebennik, Sc.D (Ukraine)
O.O. Yemets, Sc.D (Ukraine)
D.D. Peleshko, Sc.D (Ukraine
V.I. Lytvynenko, Sc.D (Ukraine)
N. B. Shakhovska, Sc.D (Ukraine)
E.A. Druzynin, Sc.D (Ukraine)
S.I. Dotsenko, Sc.D (Ukrain)
T.V. Kovalyuk, PhD (Ukraine)
A. Pakštas, PhD (United Kingdom)
O.Romanko, PhD (Canada)
I. Polik, PhD (USA)
V.Kalashnikov, Sc.D (Mexico)
S. Berezyuk, PhD (Canada)
P. Davidsson, PhD (Sweden)
M. Biagi, PhD (Italy)

Organizing Committee:

**Protsenko S. I., Sc.D, chairman (Ukraine);
Shendryk V. V., PhD, co-chairman(Ukraine);
Vaschenko S. M., PhD, co-chairman (Ukraine);
Parfenenko Y. V., PhD (Ukraine), Nahornyi V. V. , PhD (Ukraine),
Zakharchenko V. P. (Ukraine), Shendryk S.O. (Ukraine),
Boiko O. V., PhD, executive secretary (Ukraine).**

Contacts:

Address: AIST conference, Sumy State University
2 Rimsky-Korsakov Str., Sumy, 40000, Ukraine
website: www.aist.sumdu.edu.ua
e-mail: aist@sumdu.edu.ua.

SESSION 5 INTELLIGENT DATA PROCESSIN

SESSION 6 COMPUTER NETWORKING AND TELECOMMUNICATIONS

SESSION 7 MODERN METHODS AND INFORMATION TECHNOLOGIES OF SUSTAINABLE DEVELOPMENT

The VIth International Conference « Advanced Information Systems and Technologies, AIST 2018»
16-18 May 2018, Sumy, Ukraine

Simulation of Scoring of the Bank's Borrowers Creditworthiness

Konstantin Gritsenko
Sumy State University, Ukraine, k.hrytsenko@uabs.sumdu.edu.ua

***Abstract* – The advantages of credit scoring are considered. The characteristics of bank borrowers used in the construction of logit model, decision tree and neural network are described. The choice of the best model is made.**

***Keywords* – scoring model, credit risk, borrower creditworthiness, intellectual data analysis, logit model, neural network, decision tree.**

I. INTRODUCTION

The current economic situation in Ukraine has led to an increase in credit risks associated with non-repayment of loans. One of the ways to reduce credit risks is the use of scoring technologies that allow you to rapidly assess the creditworthiness of potential borrowers based on questionnaire data. The use of credit scoring facilitates the speed of decision-making on granting loans and conducting an express analysis of creditworthiness in the presence of the borrower. Most banking professionals consider that credit scoring is the most suitable technology for consumer lending. For example, see [1]. The introduction of credit scoring system in the practice of banks is necessary both for the banks themselves to assure the return of the loan as well as for borrowers because the scoring system significantly reduces the time taken by the bank to decide on a loan.

It is important to note that for a scoring model it is typical to use a certain set of variables (characteristics of the borrower) that reflects the credit risk associated with the borrower. The construction of scoring models is based on statistical methods in which the qualitative and quantitative characteristics of a potential borrower are compared with the level of credit risk, which is determined on the basis of retrospective credit histories.

For each variable of the scoring model cut-offs are determined, according to which the scoring model divides borrowers into "bad" and "good". For each scoring model the own cut-off determined, which reflects the boundary of vulnerability in relation to the bank's credit policy and external factors. The purpose of credit scoring is to calculate the level of credit risk inherent in one or another borrower, in other words, his credit rating. Comparison of the obtained results with the limit value allows providing borrowed funds to one or another borrower.

The result of the credit scoring is, as a rule, a certain integral indicator, which is proportional to the borrower's creditworthiness. Based on the received credit scoring estimates, the bank has the opportunity to classify borrowers by their level of creditworthiness. For example, see [2]. Unfortunately, there is currently no qualitative statistical database on borrowers in Ukraine, and credit bureaus are not yet operational. Ukrainian banks have to rely on their own methods of assessing credit risk and take the full burden of credit risk.

II. RESEARCH RESULTS

The source of information used for the practical implementation of scoring models is the characteristics of borrowers of one of the Ukrainian banks: borrower's age, gender, work experience, annual income, annual household income, the existence of borrower's own real estate, the purpose of the loan agreement (consumer credit, mortgage, credit for study fees, business development loan, car loan, credit card, building real estate lending), the average amount of payment for a loan, the amount of the last payment for a loan, the amount of the next payment for a loan, the frequency of payments on the loan, the discipline of payments, the amount of past due payments in the last year, the total amount of past due payments in credit history, the result of the loan agreement (loan was not repaid, loan was repaid). The SAS Enterprise Miner software product was selected for scoring models development. It is an integrated component of the SAS system of intellectual data analysis and designed to detect information in large amounts of data which is necessary for management decisions to be made. For example, see [3].

To construct scoring models, we used the SAS Enterprise Miner tools of regression analysis, decision trees and neural networks. The entire incoming data set (*TRAIN*) of 14559 borrowers was randomly divided into two parts (80% – training data and 20% – validation data) with the preservation of the distribution of the positive response (loan was repaid) and the negative response (loan was not repaid) of the target variable (the result of the loan agreement). Thus, all models were constructed and tested on equivalent data sets (Fig.1).

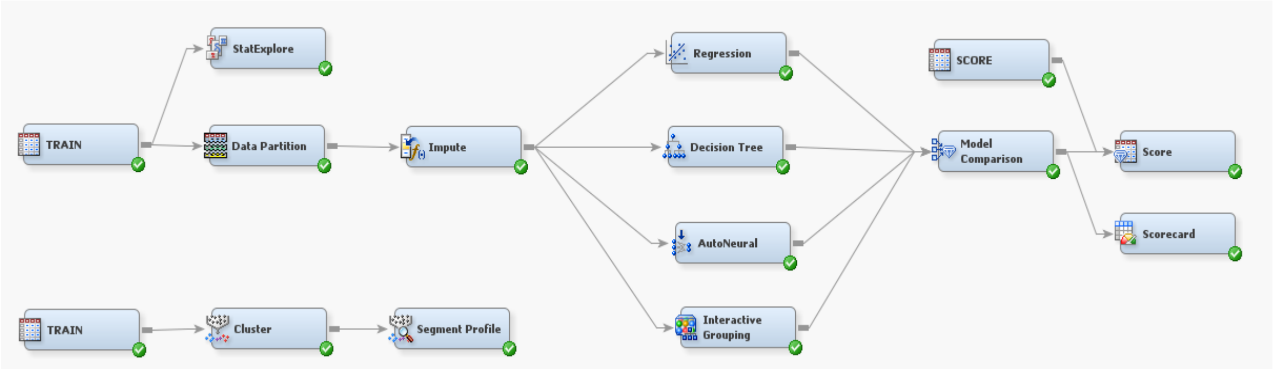


Figure 1. ETL-diagram of the simulation process of scoring of the bank's borrowers creditworthiness

Incoming data was divided into equal parts in percentage terms (*Data Partition* tool). They are given in table 1.

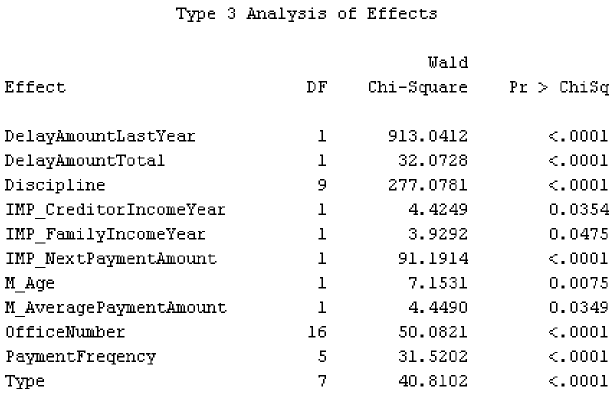
TABLE I. BREAKDOWN OF INCOMING DATA SET

	Number of credit order with positive response		Number of credit order with negative response	
	units	%	units	%
Incoming data	9 716	66,74 %	4 843	33,26 %
Training data	6 800	66,74 %	3 389	33,26 %
Validation data	2 916	66,73 %	1 454	33,27 %

As a result of the primary analysis (*StatExplore* tool), the basic statistical characteristics of the input data were obtained, the role of variables in the simulation was determined, and it was found that the input data array has missing values in the interval variables that were filled (*Impute* tool). Missed data was filled in this way: for interval variables all missed values were replaced by the average for all available values, for categorical input variables, all missed values were replaced by the most frequently encountered category

As a result of the cluster analysis carried out by the *k*-means method (*Cluster* tool), four clusters were obtained. The first cluster (19,7%) has been formed by senior citizens with significant (higher than the average) work experience (*Segment Profile* tool). The second cluster (34,9%) has been formed by borrowers without own real estate, with an average work experience, middle age and average income. The third cluster (39,1%) has been formed by borrowers with own real estate, with an average work experience, middle age and middle income. The fourth cluster (6,3%) has been formed mainly by borrowers with an income higher than the average.

To optimize the complexity of the logit model, the method of stepwise exclusion of non-significant variables was selected (*Regression* tool). The significance of the latters were determined by us according to the statistical Wald chi-squared test. Result is given in figure 2.



```
                    Type 3 Analysis of Effects

                                           Wald
Effect                         DF    Chi-Square    Pr > ChiSq

DelayAmountLastYear             1      913.0412        <.0001
DelayAmountTotal                1       32.0728        <.0001
Discipline                      9      277.0781        <.0001
IMP_CreditorIncomeYear          1        4.4249        0.0354
IMP_FamilyIncomeYear            1        3.9292        0.0475
IMP_NextPaymentAmount           1       91.1914        <.0001
M_Age                           1        7.1531        0.0075
M_AveragePaymentAmount          1        4.4490        0.0349
OfficeNumber                   16       50.0821        <.0001
PaymentFreqency                 5       31.5202        <.0001
Type                            7       40.8102        <.0001
```

Figure 2. Significance of the logit model variables

It turned out that the following variables have high statistical significance: the purpose of the loan agreement (*Type*), the frequency of payments on the loan (*PaymentFrequency*), the amount of the next payment for a loan (*IMP_NextPaymentAmount*), the discipline of payments (*Discipline*), the amount of past due payments in the last year (*DelayAmountLastYear*), the total amount of past due payments in credit history (*DelayAmountTotal*).

By evaluating the odds ratio, we have investigated how selected variables of the logit model affect the target variable. According to the results of a constructed logit model, a borrower with a minimum amount of past due payments in the last year, a senior age and with a maximum average amount of payment for a loan, who has paid payments every two months and received a credit card, is most likely a creditworthy borrower.

The decision tree was built in an automatic mode (*Decision Tree* tool). Optimization of the complexity of the decision tree was carried out using average squared error. Result is given in figure 3.

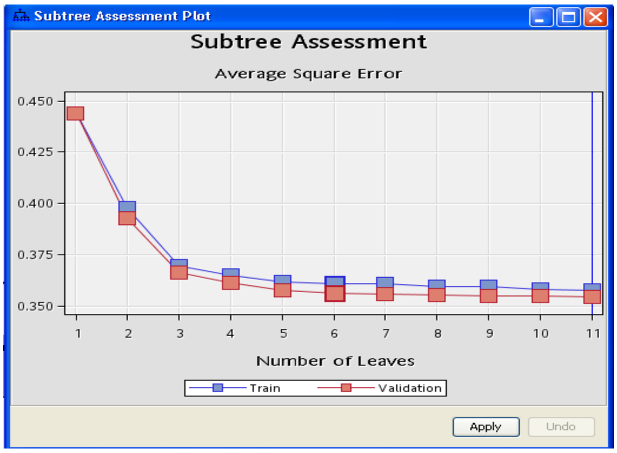


Figure 3. Graph of changes of average squared error

For the training and validation data sets, the graph has a declining trend. At the 6th step, the value of the average squared error starts to decrease less rapidly, and therefore further increase in the number of branches is not needed. Thus, as the optimal option, a tree with 6 branches of branching was selected. Classification rules for this model are given in table 2.

According to the results of the decision tree, the most likely creditworthy borrower is a client who had overdue payment to four months, paid without past due in the last year or repay the loan using a pledge.

The neural network was built in an automatic mode (*AutoNeural* tool), the hyperbolic tangent was chosen as an activation function. Optimization of the complexity of the neural network has been made on the basis of minimizing the proportion of incorrectly classified borrowers. The result is a neural network, which consists of one hidden layer with two neurons. The constructed neural network is characterized by low values of the average squared error (0,189 and 0,191 for training and validation data respectively). The misclassification rate of the model data also has low values, namely: 0,296 and 0,295, respectively. Thus, we can conclude that the built neural network is qualitative and adequate.

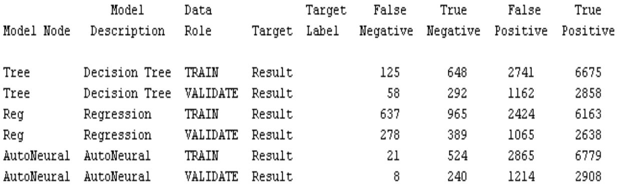
The choice of the best model was performed (*Model Comparison* tool) on the basis of the misclassification rate (*MISC*), the average squared error (*ASE*) and the Gini coefficient (*G*). Result is given in table 3.

The lowest values of the misclassification rate, the average squared error and the highest values of the Gini coefficient has a decision tree. In the second place – the neural network, the last place occupies the logit model.

Classification capabilities of the built models were also studied using the *Classification Table* of *Model Comparison* tool. Result is given in figure 4.

TABLE II. RULES FOR CLASSIFYING A DECISION TREE

Classification rule	Training data	Validation data
The amount of past due payments in the last year (*DelayAmountLastYear*):		
more than 5	100% borrowers not repayed loan	100% borrowers not repayed loan
less or equals 5	70 % borrowers repayed loan, 30% – no	70 % borrowers repayed loan, 30% – no
more than 1	64% borrowers repayed loan, 36% – no	64% borrowers repayed loan, 36% – no
less or equals 1	94% borrowers repayed loan, a 6% – no	93% borrowers repayed loan, 7% – no
The discipline of payments (*Discipline*):		
overdue payment more than 120 days, deferment of loan payment	64% borrowers repayed loan, 36% – no	66% borrowers repayed loan, 34% – no
overdue payment to 119 days, payment without past due last year, repayment on a loan using a pledge	98% borrowers repayed loan, 2% – no	98% borrowers repayed loan, 2% – no
The amount of the last payment for a loan (*LastPaymentAmount*):		
more than 1 403 hrn	70% borrowers repayed loan, 30% – no	71% borrowers repayed loan, 29% – no
less or equals 1 403 hrn	61% borrowers repayed loan, 39% – no	61% borrowers repayed loan, 39% – no
The total amount of past due payments in credit history (*DelayAmountTotal*):		
more than 11	59% borrowers repayed loan, 41% – no	59% borrowers repayed loan, 41% – no
less or equals 11	66% borrowers repayed loan, 34% – no	69% borrowers repayed loan, 31% – no



```
                   Model          Data                Target    False      True      False     True
Model Node      Description      Role       Target   Label    Negative  Negative  Positive  Positive

Tree            Decision Tree    TRAIN      Result                  125       648      2741      6675
Tree            Decision Tree    VALIDATE   Result                   58       292      1162      2858
Reg             Regression       TRAIN      Result                  637       965      2424      6163
Reg             Regression       VALIDATE   Result                  278       389      1065      2638
AutoNeural      AutoNeural       TRAIN      Result                   21       524      2865      6779
AutoNeural      AutoNeural       VALIDATE   Result                    8       240      1214      2908
```

Figure 4. Classification table

TABLE III. COMPARATIVE CHARACTERISTICS OF THE QUALITY OF THE LOGIT MODEL, DECISION TREE AND NEURAL NETWORK

Data	Coefficient		
	MISC	*ASE*	*G*
Logit model			
Training	0,30	0,19	0,41
Validation	0,31	0,20	0,37
Decision tree			
Training	0,28	0,18	0,46
Validation	0,28	0,18	0,46
Neural network			
Training	0,28	0,18	0,44
Validation	0,28	0,18	0,43

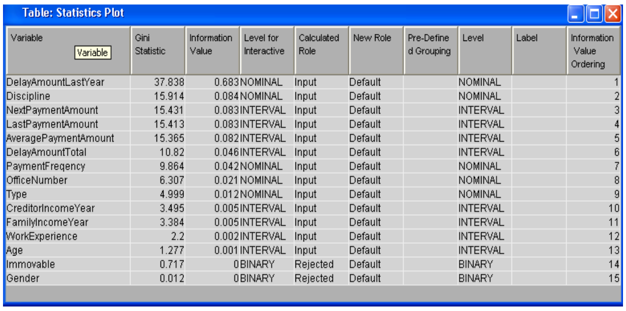
At the last stage of the study we attach the output of the *Model Comparison* tool and a set of score data (*SCORE*) into the *Score* tool (Fig.1). This tool generates forecasts using the model that was selected as the best by *Model Comparison* tool. In our case, this is a decision tree.

The SAS Enterprise Miner package includes specialized *Credit Scoring for Banking* solution that address specific credit scoring tasks. For example, see [4]. The component *Scorecard* automatically calculates scorecards based on the results of a model built on the training data. The *Interactive Grouping* component automatically selects the most significant input variables and interactively generates groups of values of input variables with continuous values (Fig.1).

The *Gini coefficient* and *Information Value criteria* were used to automatically select the most significant input variables. For the automatic formation of groups of values as criteria for the breakdown of the range of values into groups, the *Weight of Evidence* was used.

As a result of automatic implementation of the *Interactive Grouping* tool on the basis of the *Gini coefficient* and *Information Value criteria* from the input data set, it was suggested that 13 variables be used to form a scorecard. Each variable based on the weight of the *Weight of Evidence* has been broken down to the levels. Binary variables were automatically rejected (the role of *Rejected* is set) due to the fact that the *Information Value criteria* equals 0. Result is given in figure 5.

The final stage is the development of a scorecard using the *Scorecard* tool based on the results obtained at *Interactive Grouping* and *Model Comparison* tools (Fig.1). The main coefficients characterizing the quality of the built scorecard are given in the table 4.



Variable	Gini Statistic	Information Value	Level for Interactive	Calculated Role	New Role	Pre-Defined Grouping	Level	Label	Information Value Ordering
DelayAmountLastYear	37.838	0.683	NOMINAL	Input	Default		NOMINAL		1
Discipline	15.914	0.084	NOMINAL	Input	Default		NOMINAL		2
NextPaymentAmount	15.431	0.083	INTERVAL	Input	Default		INTERVAL		3
LastPaymentAmount	15.413	0.083	INTERVAL	Input	Default		INTERVAL		4
AveragePaymentAmount	15.365	0.082	INTERVAL	Input	Default		INTERVAL		5
DelayAmountTotal	10.82	0.046	INTERVAL	Input	Default		INTERVAL		6
PaymentFreqency	9.864	0.042	NOMINAL	Input	Default		NOMINAL		7
OfficeNumber	6.307	0.021	NOMINAL	Input	Default		NOMINAL		8
Type	4.999	0.012	NOMINAL	Input	Default		NOMINAL		9
CreditorIncomeYear	3.495	0.005	INTERVAL	Input	Default		INTERVAL		10
FamilyIncomeYear	3.384	0.005	INTERVAL	Input	Default		INTERVAL		11
WorkExperience	2.2	0.002	INTERVAL	Input	Default		INTERVAL		12
Age	1.277	0.001	INTERVAL	Input	Default		INTERVAL		13
Immovable	0.717	0	BINARY	Rejected	Default		BINARY		14
Gender	0.012	0	BINARY	Rejected	Default		BINARY		15

Figure 5. Selection of the most significant input variables

TABLE IV. COEFFICIENTS OF SCORECARD QUALITY

Coefficient	Sample	
	Training	Validation
Misclassification Rate	0,299	0,289
Average Squared Error	0,192	0,190
Kolmogorov-Smirnov criterion	0,269	0,289

According to the results of built scorecard, the most likely creditworthy borrower is a middle-aged borrower who has taken a consumer loan and has a maximum amount of the next payment for a loan, made a payment delay to the month during the past year or paid off loan using a pledge.

CONCLUSIONS

The developed scoring models are primarily aimed at reducing the number of irreversible and “problem” loans, that is, the fulfillment by borrowers of the terms of the loan agreement. Using a scoring model for assessing a bank borrower’s creditworthiness based on a decision tree will provide an opportunity to achieve a range of effects: reducing the role of the subjective component when deciding on lending, significant acceleration of the decision-making process, improving the quality of the loan portfolio as a result of minimizing the share of problem loans. Thus, the simulation of credit scoring is one of the most effective ways to increase the efficiency of lending to bank customers.

REFERENCES:

[1] A. Kaminsky, “The expert model of credit scoring of the bank borrower”, *Banking*, no.1, pp.75-81, 2006.

[2] G. Velikoivanenko, L. Trocoz, “Modeling of internal credit ratings of commercial bank borrowers”, *Economic analysis*, vol.11, no.1, pp.313-319, 2012.

[3] Georges J., *Applied Analytics Using SAS Entepise Miner Course Notes.* Cary, NC: SAS Institute Inc., 2010.

[4] SAS Institute Inc. *Developing Credit Scorecards Using Credit Scoring for SAS Enterprise Miner 12.1*, Cary, NC: SAS Institute Inc., 2012.

Наукове видання

СУЧАСНІ ІНФОРМАЦІЙНІ СИСТЕМИ І ТЕХНОЛОГІЇ

Матеріали
Шостої міжнародної науково-практичної конференції
(Суми, 16 – 18 травня 2018 року)

Відповідальний за випуск В. В. Шендрик
Комп’ютерне верстання: О. В. Бойко

Формат 60×84/16. Ум. друк. арк. . Обл.-вид. арк. .

Видавець і виготовлювач
Сумський державний університет
вул. Римського-Корсакова,2, м.Суми, 4007
Свідоцтво суб’єкта видавничої справи ДК № 3062 від 17.12.2007