Vol. 33, No. 02, pp. 608-615/Diciembre 2020

ISSN-E 1995-9516
Universidad Nacional de Ingeniería

http://revistas.uni.edu.ni/index.php/Nexo
https://doi.org/10.5377/nexo.v33i02.10796

# Collection Scoring Models Development and Research Based on the Deductor Analytical Platform

# Desarrollo e investigación de modelos de puntuación de colecciones basados en la plataforma analítica Deductor

Ilyas Idrisovich Ismagilov[1], Ajgul Ilshatovna Sabirova[2], Dina Vladimirovna Kataseva[3], Alexey Sergeevich Katasev[4]

[1] Doctor of Technical Sciences, Professor of the Department of Economic Theory and Econometrics, Institute of Management, Economics and Finance, Kazan (Volga Region) Federal University
[2] Ph.D. in Economics, Senior Lecturer at the Department of Accounting, Analysis and Audit of the Institute of Management, Economics and Finance, Kazan (Volga Region) Federal University.
[3] Senior Lecturer, Department of Information Security Systems, Institute of Computer Technologies and Information Security, Kazan National Research Technical University named after A.N. Tupolev-KAI.
[4] Doctor of Technical Sciences, Professor of the Department of Information Security Systems of the Institute of Computer Technologies and Information Security, Kazan National Research Technical University named after A.N. Tupolev-KAI.

Corresponding author email: iiismag@mail.ru

*(recibido/received: 16-July-2020; aceptado/accepted: 19-September-2020)*

## ABSTRACT

This article solves the problem of collection scoring models constructing and researching. The relevance of solving this problem on the intelligent modeling technologies basis: decision trees, logistic regression and neural networks is noted. The initial data for the models was a set of 14 columns and 5779 rows. The models construction was performed in Deductor platform. Each model was tested on the set of 462 records. For all models, the corresponding classification matrix were constructed and the1st and 2nd kind errors were calculated, as well as the general error of the models. In terms of minimizing these errors, logistic regression showed the worst results, and the neural network showed the best. In addition, the constructed models effectiveness was evaluated according to «income» and «time» criteria. By the time costs the logistic regression model exceeds other models. However, in terms of income the neural network model was the best. Thus, the results showed that in order to minimize the time spent on work with debtors it is advisable to use a logistic model. However, to maximize profits and minimize classification errors, it is appropriate to use a neural network model. This indicates its effectiveness and practical use possibility in intelligent scoring systems.

**Keywords:** overdue credit debt, collection scoring, decision tree, logistic regression, neural network, data mining.

## RESUMEN

Este artículo resuelve el problema de la construcción e investigación de modelos de puntuación de colecciones. Se destaca la relevancia de resolver este problema sobre la base de las tecnologías de modelado inteligente: árboles de decisión, regresión logística y redes neuronales. Los datos iniciales de los modelos fueron un conjunto de 14 columnas y 5779 filas. La construcción de los modelos se realizó en plataforma Deductor. Cada modelo fue probado en el conjunto de 462 registros. Para todos los modelos se construyó la correspondiente matriz de clasificación y se calcularon los errores de 1º y 2º tipo, así como el error general de los modelos. En términos de minimizar estos errores, la regresión logística mostró los peores resultados y la red neuronal mostró los mejores. Además, se evaluó la efectividad de los modelos construidos según criterios de «ingresos» y «tiempo». Por el tiempo que cuesta el modelo de regresión logística supera a otros modelos. Sin embargo, en términos de ingresos, el modelo de red neuronal fue el mejor. Así, los resultados mostraron que para minimizar el tiempo dedicado al trabajo con los deudores es recomendable utilizar un modelo logístico. Sin embargo, para maximizar las ganancias y minimizar los errores de clasificación, es apropiado utilizar un modelo de red neuronal. Esto indica su eficacia y posibilidad de uso práctico en sistemas de puntuación inteligentes.

**Palabras clave:** deuda crediticia vencida, puntaje de cobranza, árbol de decisión, regresión logística, red neuronal y minería de datos.

## 1. INTRODUCTION

Currently, in Russia there is an increase of the number of loans issued to individuals by various banks and financial organizations (Shikimi, 2020; Gemzik-Salwach, 2020). There is also an overdue credit debt increase in all types of lending (Xie & Hansen, 2020; Du & Palia, 2016). This leads to increase the load on various collection agencies and banks collection departments. To reduce the increasing load, it is necessary to raise the collection departments activity efficiency.

Efficiency in this case is time and material costs reduction for working with borrowers who have overdue credit debt. The collection activity optimization is possible due to the modern methods of intellectual analysis of accumulated data use (Katasev et al., 2016; Dela Cruz Galapon, 2020) and the effective collection scoring models construction (Shen et al., 2020; Terko et al., 2019). Such models are able to minimize the time spent working with clients to collect overdue debts, and the total time spent on activities, and maximize profits from collection activities. In addition, the collection scoring models should be able to assess the opportunities of obtaining cash from borrowers, to predict the outcome of measures taken to collect overdue debts, to segment debtors into groups with varying degrees of debt repayment probability, and also to develop a strategy for dealing with overdue credit debt. Therefore, an actual task is the collection scoring models construction and study for collection services activities optimization.

## 2. METHODS

There are many data mining methods that can be used for credit debt overdue estimation models construction. This work is focused on three most effective and frequently used methods: a decision tree (Nagra et al., 2020; Ke et al., 2017), a logistic regression (Ansori et al., 2019; Meier et al., 2008; Asar & Wu, 2020) and a neural network (Ismagilov et al., 2018; Mustafin et al., 2018; Swiderski et al., 2012; Katasev & Kataseva, 2016; Akhmetvaleev & Katasev, 2018).

When using decision trees to classify loan applications, a set of rules is applied, which is formed when constructing a tree based on a training set (Alqam & Zaro, 2019). A decision tree example is shown in Figure 1.

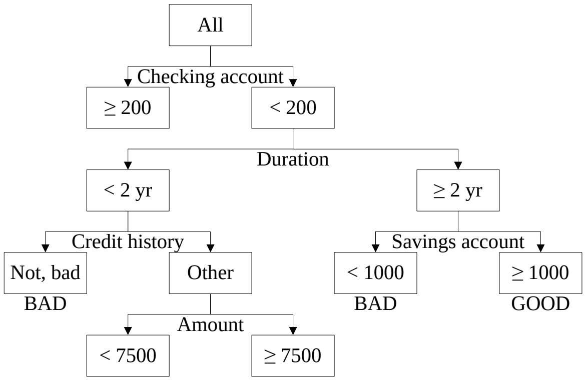

Figure 1: Decision tree example

The tree includes interconnected initial (root), intermediate and final nodes (leaves of the tree). Connections between nodes are called branches. Each node corresponds to a condition (rule) for classifying objects. In the initial and intermediate nodes, in accordance with this condition, the tree branches, and the leaves define a class of objects whose attributes correspond to the conditions that determine the way leading to this leaves.

In the case of a complete decision tree constructing, it accurately describes the X vector realizations set classification, from which the tree was built, into subsets of credits whose results are considered «good» or «bad». However, in most cases, a truncated tree is built, acting by analogy with other classification problems, in which this approach can be justified by various factors (for example, errors in the initial data). The result of truncation in a scoring application is usually a decrease of classification accuracy. In addition, due to truncation, those implementations of the vector X for which there are no data on the results of lending can be included in the tree. These are such data sets, applications with which either did not arrive at the bank or were not satisfied. The results obtained for loans with partially matching attributes corresponding to uncut tree nodes extend to these implementations of the decision tree.

In the regression model, the scoring function is approximated relative to the X vector components by a linear function of the following form (Aaserud et al., 2013):

$$p = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n, \qquad (1)$$

where $a_0$ – free coefficient, $a_i$ , ($i$=1…$n$) – weight coefficients of the request, $x_i$ – application signs, i.e. X vector components.

The $a_i$ coefficients are determined by one of the statistical estimation methods, for example, by the maximum likelihood estimation method (Chen et al., 2019). If the proportion of «good» or «bad» loans is used as the scoring function $p$, then it should be in the interval from 0 to 1. However, the value of the right side of equation (1) can exceed this range. This fact indicates a weak adequacy of the model. To overcome this drawback, the «bad» credit outcome chance logarithm is used as a scoring function:

$$p = ln\ (q/(1\text{-}q)), \qquad (2)$$

where $q$ – the probability of «bad» credit outcome.
This approach is called logistic regression (Shen et al., 2020). The $p$ function varies in the interval from - ∞ to + ∞ (Fig. 2).

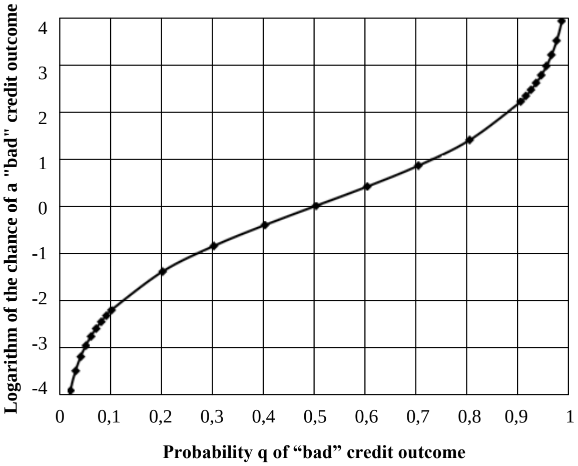

Figure 2: Example of scoring function using logistic regression

The frequencies of the «bad» credit outcome are used as estimates of $q(X)$ values for each implementation of the $X$ vector (application features), for which the bank has data on loans issued. Based on these data, the $a_i$ coefficients are estimated, which calculate the scoring function approximated value by the formula (2) for any feasible implementation of the $X$ vector.
The neural networks can also be used to approximate the scoring function (Swiderski et al., 2012). A neural network is a mathematical model whose parameters for a specific task are formed by training the model on a training data set. For scoring, such a sample may be a data set of previously issued loans or part of this set. As a result, a piecewise linear approximation of the $p(X)$ function is formed, which is specified algorithmically, and can be calculated using a neural network for any feasible implementation of the $X$ vector.

## 3. RESULTS AND DISCUSSION

Let's consider the intelligent collection scoring models construction and research on the base of Deductor analytical platform (Lomakin et al., 2019). The initial data for models constructing is a sample consisting of 14 columns and 5779 rows. The sample fields structure is presented in table 1.

Table 1: File fields structure of the initial data of debtors

| № | Field name | Description | Field type |
|---|---|---|---|
| 1 | Credit amount | Credit amount in rubles | Real |
| 2 | Credit term | Credit term in months | Integer |
| 3 | Monthly payment | Monthly payment amount | Real |
| 4 | Age | Client age in years | Integer |
| 5 | Gender | Client gender | String |
| 6 | Overdue period | Overdue in days | Integer |

| | | | |
|---|---|---|---|
| 7 | Number of payments before delay | The number of payments before the first arrears | Integer |
| 8 | Availability of the writ | Is there a writ of execution for the borrower (court decision) | Logical |
| 9 | Principal balance | The balance amount of the main debt | Real |
| 10 | Interest debt balance | The balance amount of interest | Real |
| 11 | Fines | Accrued fines and penalties | Real |
| 12 | Delay / debt | The ratio calculated on the three previous fields basis by the formula: 11 / (9+10) | Real |
| 13 | Payments resumption | Did the borrower begin to pay the loan again after the collection department specialist work | Integer |
| 14 | Testing set | A sign of the example participation in a testing set | Logical |

As you can see from the table, information about borrowers is stored in such diverse fields as «Credit amount», «Credit term», etc.

Based on the described initial data, the collection scoring models were constructed in the Deductor: a decision tree, a logistic regression, and a neural network in the form of a two-layer perceptron. Each model is tested on the data set marked in the initial sample as «Testing set».

The testing data set consisted of 462 records, that is about 8% of the initial data volume.

Let’s consider the testing results of the constructed collection scoring models, evaluate these models, compare the models with each other according to various criteria, and choose the most effective. Table 2 presents the constructed models test results (Sulewski, 2019).

Table 2: Classification matrix when testing models

| Actual values | Classified by the model | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0 | | | 1 | | | Total |
| | LR | DT | NN | LR | DT | NN | |
| 0 | 57 | 61 | 64 | 11 | 7 | 4 | 68 |
| 1 | 49 | 37 | 16 | 345 | 357 | 378 | 394 |
| Total | 106 | 98 | 80 | 356 | 364 | 382 | 462 |

The following notation is used in the table: LR – logistic regression, DT – decision tree, NN – neural network. Number «0» means that work with the client was carried out, but the resumption of payments did not follow. The number «1» means that work with the client was carried out, and resumption of payments followed.

Based on the data presented in table 2, we calculated 1st, 2nd kind errors (Zhang et al., 2017), and the general models errors (see Table 3).

Table 3: Model testing errors

| Model | Test results | | |
|---|---|---|---|
| | 1st kind error, % | 2nd kind error, % | Total error, % |
| Logistic regression | 12,44 | 16,18 | 28,61 |
| Decision tree | 9,39 | 10,29 | 19,68 |
| Neural network | 4,06 | 5,88 | 9,94 |

As it can be seen from the table, from the point of view of errors of the 1st and 2nd kind, and the general error of the models, logistic regression showed the worst results. At the same time, the

neural network showed the best results in terms of minimizing these errors. The effectiveness of collection scoring models practical use of is determined not only by the 1st and 2nd kind errors value, but also by the time spent on collecting overdue debts, as well as by the income received from the result of collection activities. We introduce the following destinations:
*t* – time for collection one debt (working time with one debtor), amounting to 1 conventional unit of time;
*z* – the costs of collecting one debt, amounting to 1 cu;
*d* – the average income from one client (if the client has resumed payments), amounting to 1000 cu;
*TP* – the number of truly positive outcomes when working with debtors (the number of positive modeling results that match the actual values);
*FP* – the number of false positive outcomes when working with debtors (the number of positive modeling results that do not match the actual values);
$D=TP(d\text{-}z)\text{-}FP^*z$ – net income from debt collection activities.
Then, based on the data in table 2 and introduced designations, we calculate the constructed models effectiveness according to the criteria of «income» and «time» (see table 4).

Table 4: Comparison of collection scoring models

| Model | *TP* | *FP* | Time, c.u. | Income, c.u. |
|---|---|---|---|---|
| Logistic regression | 345 | 11 | 356 | 344 644 |
| Decision tree | 357 | 7 | 364 | 356 636 |
| Neural network | 378 | 4 | 382 | 377 618 |

As it can be seen from the table, the logistic regression model outperforms other models in terms of time spent on collecting overdue debts. However, in terms of net income from the debt collection measures implementation, the neural network model is the best.

## 4. CONCLUSIONS

Thus, the problem of intelligent collection scoring models effectiveness constructing and evaluating has been solved in this study. The results showed that to minimize the time spent on work with debtors, it is advisable to use the logistic regression model. However, to maximize profits, it is advisable to use a model based on the multilayer neural network training. In addition, this model showed the greatest accuracy in terms of 1st and 2nd kind errors minimizing. This indicates its effectiveness and practical use possibility in intelligent scoring systems (Mehdi et al., 2019).

## 5. ACKNOWLEDGMENTS

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

## REFERENCES

Aaserud, S., Kvaløy, J. T., & Lindqvist, B. H. (2013). Residuals and functional form in accelerated life regression models. In Risk Assessment and Evaluation of Predictions (pp. 61-65). Springer, New York, NY.

Akhmetvaleev, A. M., & Katasev, A. S. (2018). Neural network model of human intoxication functional state determining in some problems of transport safety solution. Computer research and modeling, 10(3), 285-293

Alqam, S. J., & Zaro, F. R. (2019). Power Quality Detection and Classification Using S-Transform and Rule-Based Decision Tree. Int. J. Electr. Electron. Eng. Telecommun, *8*, 45-50

Ansori, M. F., Sidarto, K. A., & Sumarti, N. (2019, December). Logistic models of deposit and loan between two banks with saving and debt transfer factors. In AIP Conference Proceedings (Vol. 2192, No. 1, p. 060002). AIP Publishing LLC.

Asar, Y., & Wu, J. (2020). An improved and efficient biased estimation technique in logistic regression model. Communications in Statistics-Theory and Methods, *49*(9), 2237-2252.DOI:

Chen, J., Ding, F., Zhu, Q., & Liu, Y. (2019). Maximum likelihood based identification methods for rational models. International Journal of Systems Science, *50*(14), 2579-2591.

Dela Cruz Galapon, A. (2020). An assessment: Respiratory analysis using data mining method - A decision support system. Test Engineering and Management, 83, 4824-4829.

Du, B., & Palia, D. (2016). Short-term debt and bank risk. Journal of Financial and Quantitative Analysis (JFQA), Forthcoming.

Gemzik-Salwach, A. (2020). Institutional Analysis of Banks and Personal Loan Companies: Lesson from Poland. Journal of Economic Issues, *54*(1), 142-163.

Ismagilov, I. I., Khasanova, S. F., Katasev, A. S., & Kataseva, D. V. (2018). Neural network method of dynamic biometrics for detecting the substitution of computer. Journal of Advanced Research in Dynamical and Control Systems, 10(10 Special Issue), 1723-1728)

Katasev, A. S., & Kataseva, D. V. (2016, November). Neural network diagnosis of anomalous network activity in telecommunication systems. In 2016 Dynamics of Systems, Mechanisms and Machines (Dynamics) (pp. 1-4). IEEE

Katasev, A. S., Kataseva, D. V., & Emaletdinova, L. Y. (2016, May). Neuro-fuzzy model of complex objects approximation with discrete output. In 2016 2nd International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM) (pp. 1-5). IEEE.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems (pp. 3146-3154).

Lomakin, N., Shokhnekh, A., Sazonov, S., Polianskaia, A., Lukyanov, G., & Gorbunova, A. (2019, October). Hadoop and Deductor Based Digital Ai System for Predicting Cost of Innovative Products in Conditions of Digitalization of Economy. In Proceedings of the 2019 International SPBPU Scientific Conference on Innovations in Digital Economy (pp.1-8).

Mehdi, B., Hasna, C., & Tayeb, O. (2019). Intelligent credit scoring system using knowledge management. IAES International Journal of Artificial Intelligence, *8*(4), 391.

Meier, L., van de Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression Series B Statistical methodology.

Mustafin, A. N., Katasev, A. S., Akhmetvaleev, A. M., & Petrosyants, D. G. (2018). Using Models of Collective Neural Networks for Classification of the Input Data Applying Simple Voting. The Journal of Social Sciences Research, 333-339.

Nagra, A. A., Han, F., Ling, Q. H., Abubaker, M., Ahmad, F., Mehta, S., & Apasiba, A. T. (2020). Hybrid self-inertia weight adaptive particle swarm optimisation with local search using C4. 5 decision tree classifier for feature selection problems. Connection Science, *32*(1), 16-36.

Shen, F., Wang, R., & Shen, Y. (2020). A cost-sensitive logistic regression credit scoring model based on multi-objective optimization approach. Technological and Economic Development of Economy, *26*(2), 405-429.

Shikimi, M. (2020). Bank loan supply shocks and leverage adjustment. Economic Modelling, 87, 447-460.

Sulewski, P. (2019). Some contributions to practice of 2× 2 contingency tables. Journal of Applied Statistics, *46*(8), 1438-1455.

Swiderski, B., Kurek, J., & Osowski, S. (2012). Multistage classification by using logistic regression and neural networks for assessment of financial condition of company. Decision Support Systems, *52*(2), 539-547.DOI: https://doi.org/10.1016/j.dss.2011.10.018

Terko, A., Žunić, E., Đonko, D., & Dželihodžić, A. (2019, October). Credit Scoring Model Implementation in a Microfinance Context. In 2019 XXVII International Conference on Information, Communication and Automation Technologies (ICAT) (pp. 1-6). IEEE

Xie, D., & Hansen, M. E. (2020). Supply of bank loans and business debts: A view from historical bankruptcy cases. Review of Financial Economics,The university of new Orleans, *38*, 170-187.

Zhang, Q., Xia, D., & Wang, G. (2017). Three-way decision model with two types of classification errors. Information Sciences, 420, 431-453.