ASUE

ARMENIAN STATE UNIVERSITY OF ECONOMICS

ISSN 1829-0280

MESSENGER

OF ARMENIAN STATE UNIVERSITY OF ECONOMICS

ВЕСТНИК

АРМЯНСКОГО ГОСУДАРСТВЕННОГО ЭКОНОМИЧЕСКОГО УНИВЕРСИТЕТА

MESSENGER

OF ARMENIAN STATE UNIVERSITY OF ECONOMICS

2022 [5]
YEREVAN

MATHEMATICAL ECONOMICS

ARMEN GHAZARYAN
Head of the Chair of Economic Computer Science and Information Systems of Armenian State University of Economics, PhD, Associate Professor
https://orcid.org/0000-0001-6083-5489

LIANA GRIGORYAN
Head of the Department of Management Accounting and Audit of ASUE, Ph.D., professor,
https://orcid.org/0000-0002-9991-8377

GARNIK ARAKELYAN
Lecturer of the Chair of Economic Computer Science and Information Systems of Armenian State University of Economics,
https://orcid.org/0000-0001-6217-9681

IMPLEMENTATION OF MACHINE LEARNING IN THE CREDIT RISK MANAGEMENT SYSTEM OF INDIVIDUALS

There are many problems in each credit institution. The most important of them is the risk of possible losses in lending. Within the framework of the topic, the studies conducted by other researchers were investigated, from which it was concluded that machine learning tools are often used to optimally solve the above-mentioned problem. Real data on credits were used as a basis for modeling in the work. In this work, based on the available data, several machine learning models were developed, from which the best one was selected, which can contribute to the improvement of the credit risk management process. During the work, the logical connections between data and their interaction with each other were revealed. Then, based on the work done, the appropriate models were built, the quality of which was checked using various tools. The obtained models were compared and the best one was selected. The obtained results are practically applicable and show that each bank and credit organization can develop a better solution based on the large databases they have, which will contribute to curbing credit risk and reducing costs.

Keywords: *Credit risk, logistic regression, decision trees, random forests*
JEL: C81, C88
DOI: 10.52174/1829-0280_2022.5-123

Introduction. The economic situation of recent years has shown that some of the organizations included in the financial and banking system of the RA are quite vulnerable from the point of view of credit risk management, due to which these structures suffer significant losses. In addition, it is necessary to note that credit operations are one of the prerequisites for ensuring the normal functioning of the economy. History shows that credit risk management occupies a unique place in the financial and banking system. To assess and control risk when lending to individuals, credit institutions use various developed models. In this case, the FICO Scoring statistical model is being widely used.

Depending on the borrower's score, credit organizations make a decision on granting the loan. However, this decision often is made incorrectly, due to which the quality of the credit portfolio deteriorates and significant losses occur. That is why it is necessary to introduce a reasonable decision-making system, and by this, the topicality of the research is conditioned.

Currently, banks in several developed countries are solving the above-mentioned problem using machine learning tools.

The purpose of the work is to research the customers' behavior based on certain examples, to develop several decision-making models, and to choose the best of these models.

The implementation of the research objective can be applied in credit organizations, which will contribute to the optimal management of credit risk and cost reduction in these organizations.

Literature review. Within the framework of this research, various articles, theses, books, etc. Related to the topic were studied. Let us discuss some of them.

In the article "Development of the Scoring Map Using Logistic Regression" written by Sorokin A. S. econometric modeling of credit default probability based on logistic regression is considered. In this work, particular attention is paid to the methodology of building the model. The transformation of the obtained coefficients of the logistic regression model into scoring maps is also presented. An example of creating a scoring map is presented.[1]

The article "Implementation of the Credit Scoring System in the Bank" written by Stroev A. A. is devoted to practical issues of scoring algorithm construction. This article considers the process of data preparation for modeling and calculations. The paper also provides a comparative analysis between three modeling algorithms; logistic regression, decision trees, and neuronal networks.[2]

[1] A. S. Sorokin, "Construction of scoring maps using a logistic regression model", Internet journal "SCIENCE", No. 2 (21) 2014 Date of last access: 17.10.22թ. 22:58:00 (In Russian)

[2] A.A. Stroev, SAS company, "Implementation of a credit scoring system in a bank", Methodical journal Calculations and operational work in a commercial bank, (In Russian) number 6/2004, Moscow Date of last access: 17.10.22 թ. 22:58:00

In the article "Credit Scoring with Boosted Decision Trees" written by Bastos J. a credit rating model based on augmented decision trees, which is a robust machine learning method, is considered. It is a classifier built from several decision trees. Each tree makes a prediction, after which the answer with the most votes is accepted as the answer of the classifier. Obtained simulation results show that augmented decision trees consider being a competitive method for building a credit scoring model.[3]

According to the study by G. Paleolog, based on a sample of Italian IBM customers, the main goal is to create and test reliable models that can develop new predictions even in the case of missing information in the scoring system. The essence of the method is to replace the missing data with the so-called "shadow data" and to implement group classification methods. This method is suitable for highly unbalanced data such as credit data. The shadowing stages are structured with an individual cross-check cycle that creates dependencies between different credit inquiries. The methodology was implemented using several classifier methods, support vector machines (Support Vector Machines-SVM), nearest neighbors (K-nearest neighbor) and decision trees.[4] From the study of the above-mentioned articles, as well as other theses and books, it can be concluded that the logistic regression model is the method most often used in the banking system as a method for developing a model for assessing the creditworthiness of customers.

Research methodology Data processing methods were used during the research. To identify patterns in the data, the method of correlation analysis was used.

Correlation is the relationship between two variables. This coefficient shows the linear relationship between the variables. Given the fact that the quality of the machine learning model can be improved by pre-processing the data, data clustering mechanisms have been applied (Data binning).[5] Data were then standardized for modeling.

The main purpose of standardization is to bring all variables into one common form. In practice, there are various methods of standardization. The Weights of Evidence (WOE) method is used in this paper because it is widely used and understood.[6] For each group of grouped factors, WOE is calculated using the following formula.

$$WOE = \ln\frac{Percentage\ of\ good\ in\ the\ class}{percentage\ of\ bad\ in\ the\ class},$$

where ***ln*** is the natural logarithm.

Logistic regression, decision tree, and random forest models were used to build the machine learning models in the research.

[3] J. Bastos, «Credit scoring with boosted decision trees», CEMAPRE, School of Economics and Management (ISEG), Technical University of Lisbon, 2008.

[4] Paleologo, G., Elisseeff, A., & Antonini, G., 2010. Subagging for credit scoring models. European Journal of Operational Research, 201, 490-499. Date of last access: 16.10.22 թ. 22:25:00

[5] https://www.geeksforgeeks.org/binning-in-data-mining/. Date of last access: 17.10.22 թ. 22:58:00

[6] https://medium.com/mlearning-ai/weight-of-evidence-woe-and-information-value-iv-how-to-use-it-in-eda-and-model-building-3b3b98efe0e8. Date of last access: 16.10.22 թ. 23:15:00

Logistic regression (Logit model) is a statistical model used to predict the probability of an event occurring by comparing it to a logistic curve. This model returns a binary event probability (between 0 and 1) as an answer.[7]

A decision tree is a model that uses a tree-like data structure that represents several possible ways to solve a problem and the final result for each of them.[8] Graphically, it can be presented in the form of a tree structure, where decision-making moments correspond to so-called decision nodes, in which branching of the process takes place, dividing it into so-called branches depending on the choice made. The final nodes are called leaves (leaf nodes), each of which represents the final result of the decision made[9].

A random forest model is a set of pre-selected number of decision trees, where all trees have the same parameters. Each tree in the forest is allowed to view not all the data, but only a sub-sample, which is formed by both rows and columns (randomly selected). Based on factors and samples, cross-validation training is performed for each tree. In these models, the quality of each individual tree is generally poor, but the quality of the overall forest model is significantly higher through a group of these trees.

To avoid the problem of overtraining in machine learning models, the cross-validation training method was used. To check the qualities of the obtained models, different types of methods were used and various indicators were calculated, for example, recall score, precision score, F1 score ROC curve, etc.[10]. From the received models, the best model was selected through comparative analysis.

Analysis. The work is based on real lending data obtained by Unibank OJSC, which operates on the territory of the Republic of Armenia. There are missing values in the data for only one variable (Number of delays) that has been processed. Some variables were originally coded. Based on the information received from the bank, such variables were translated into understandable language for further analysis. Such data cleaning was carried out using the Excel program. There are 13 variables in the given dataset:

- **Marital Status**. In the original data, this variable had the values M1, M2, M3, and M4. Each value has been changed to “Married”, “Widow”, “Divorced”, and “Single”, respectively;
- **Education**. In the original data, this variable had the values E1, E2, E3, and E4. Each value has been changed to “Academic degree”, “Higher education”, “Medium professional”, and “High school”, respectively;
- **Availability of property**. In the original data, this variable had the values P1, P2, P3, and P4. Each value has been changed to “Availability of real estate”, “Availability of movable property”, “Availability of real estate and movable property”, and “Absence”, respectively;

[7] https://www.ibm.com/topics/logistic-regression. Date of last access: 17.10.22 թ. 23:08:00

[8] https://www.ibm.com/topics/decision-trees. Date of last access: 17.10.22թ. 23:30:00

[9] https://www.ibm.com/cloud/learn/random-forest. Date of last access: 17.10.22թ. 23:22:00

[10] https://www.pycodemates.com/2022/05/precision-and-recall-in-classification.html. Date of last access: 17.10.22 թ. 20:22:00

- **Sex**. In the original data, this variable had the values S1 and S2. Each value has been changed to “Female”, and “Male”, respectively;
- **Borrower’s Age**. The variable is numeric and contains values from 20 to 65;
- **The number of days past due in the last 12 months**. The variable is numeric and contains values from 0 to 1600;
- **The number of delays**. The variable is numeric and contains values from 1 to 27. This variable contains missing data. Based on the information received from the bank, it became clear that these clients do not have overdue debts. Based on this, the missing values were changed to 0;
- **The number of changes in risk classes**. The variable is numeric and contains values from 0 to 28;
- **Credit load**. The variable is numeric and contains values from 0 to 2,984,303. Data are presented in AMD;
- **Credit history length**. The variable is numeric and contains values from 0 to 488. Data is presented in days;
- **Maximum repaid loans**. The variable is numeric and contains values from 0 to 25,131,048. Data are presented in AMD;
- **Contract sum**. The variable is numeric and contains values from 30,000 to 2,300,000. Data are presented in AMD. There are some outliers in the data. These are the clients to whom the bank has issued a loan in the amount of more than AMD 1,000,000. Similar observations 7. It was calculated that the removal of observational data does not significantly affect the quality of the models. Therefore, when creating models, these observations were not removed from the dataset;
- **Default**. Binary data is presented. 1 – default, 0 – no default. Based on the information received from the bank, a default is considered the presence of overdue obligations for 90 days or more at least once in the last year.

There are 4603 observations in the dataset.

For modeling, the “Default” variable was selected as a predicted (dependent) variable.

Both the analysis and the construction of the models were carried out using the Excel program, the Python programming language, and its corresponding libraries. These are Pandas, Matplotlib, Seaborn, Scikit-learn libraries.

The data file with csv extension was then imported and a preliminary examination was performed.

Descriptive statistics of numerical data are shown below (Table 1). The table shows that there are no missing data since the number of values in all variables is 4,603. The table shows the arithmetic mean, minimum and maximum values for all indicators, as well as percentiles of 25%, 50%, and 75%.

The table below shows that the standard deviation of the variables Credit load, Credit history length, Maximum repaid loans, and Contract sum is high. This is due to the fact that the data contains customers with different credit histories. For example, one client may have no credit history, while another client may have a large debt burden. This is due to the credit policy of the bank.

Given the characteristics of the data, the standard deviation of other parameters is within acceptable ranges.

Table 1

Descriptive statistics

Indicator	Age	Number of days past due in the last 12 months	Number of delays	Number of changes in risk classes	Credit load	Credit history length	Maximum repaid loans	Contract sum	Default
count	4603	4603	4603	4603	4603	4603	4603	4603	4603
mean	37,54	4,03	2,71	0,55	471551,5	59,39	332959,8464	254229,2	0,426027
std	12,09	38,54	4,28	1,46	537158,022	61,27	920340,3925	144770,9	0,494551
min	20	0	0	0	0	0	0	30000	0
25%	27	0	0	0	0	13	0	181900	0
50%	36	0	1	0	300000	40	150000	200000	0
75%	48	0	4	0	825099,5	86	376000	377050	1
max	65	1600	67	28	2984303	488	25131048	2300000	1

A correlation matrix was constructed (Table 2), from which it can be seen that there is a stronger relationship between credit burden and length of credit history, default and credit burden, number of delinquencies, and number of changes in risk class.

Table 2

Correlation analysis

	Age	*Number of days past due in the last 12 months*	*Number of delays*	*Number of changes in risk classes*	*Credit load*	*Credit history length*	*Maximum repaid loans*	*Contract sum*	*Default*
Age	1								
Number of days past due in the last 12 months	0,01	1							
Number of delays	0,05	0,34	1						
Number of changes in risk classes	0,08	0,38	0,57	1					
Credit load	0,14	-0,01	0,05	0	1				
Credit history length	0,21	0,09	0,35	0,22	0,5	1			
Maximum repaid loans	0,12	0,14	0,18	0,12	0,17	0,26	1		
Contract sum	0,07	0,04	0,01	0,01	0,07	0,1	0,13	1	
Default	-0,08	0,08	0,04	-0,05	0,15	-0,05	0	0,1	1

In the framework of the work, an analysis of various variables was carried out, from which it can be seen that defaulted clients are quantitatively concentrated in the age group from 20 to 25 (50.49%, Table 3, Figure 1), in the group of clients with secondary professional education (42.83%, Table 4, Figure 2), as well as in the group of clients whose husband died (54.14%, Table 5, Figure 3). From the analysis, it can be seen that the most overdue clients are male representatives (46.83%, Table 6, Figure 4).

Table 3

Analysis of age groups

Age range	Default		Total	Default %
	No	Yes		
1. 20-25	457	466	923	50,49%
2. 26-35	760	589	1349	43,66%
3. 36-50	880	558	1438	38,80%
4. 51+	545	348	893	38,97%

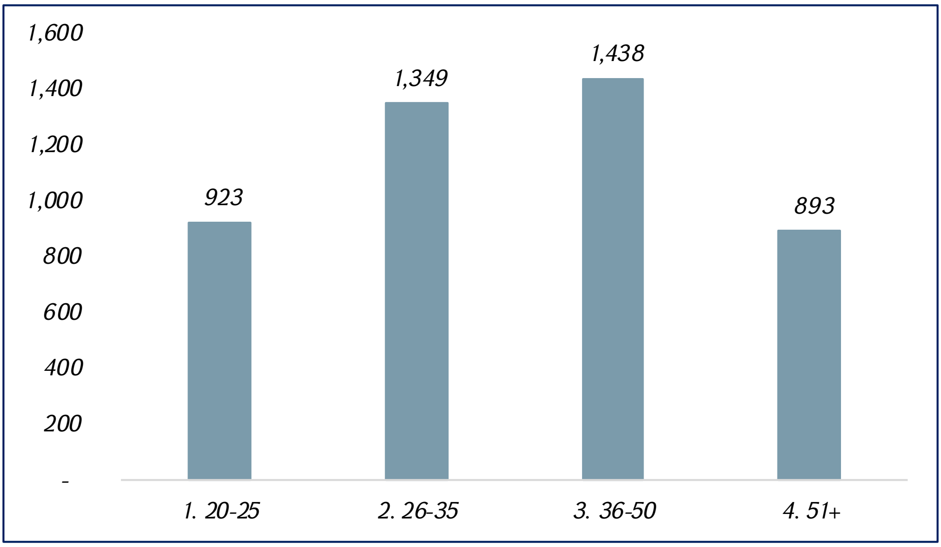

Figure 1. ***Analysis of age groups***

Table 4

Analysis of customers depending on the level of education

Education	*Default*		*Total*	*Default %*
	No	*Yes*		
Academic degree	343	255	598	42,64%
High school	3	2	5	40,00%
Higher education	626	453	1079	41,98%
Medium professional	1670	1251	2921	42,83%

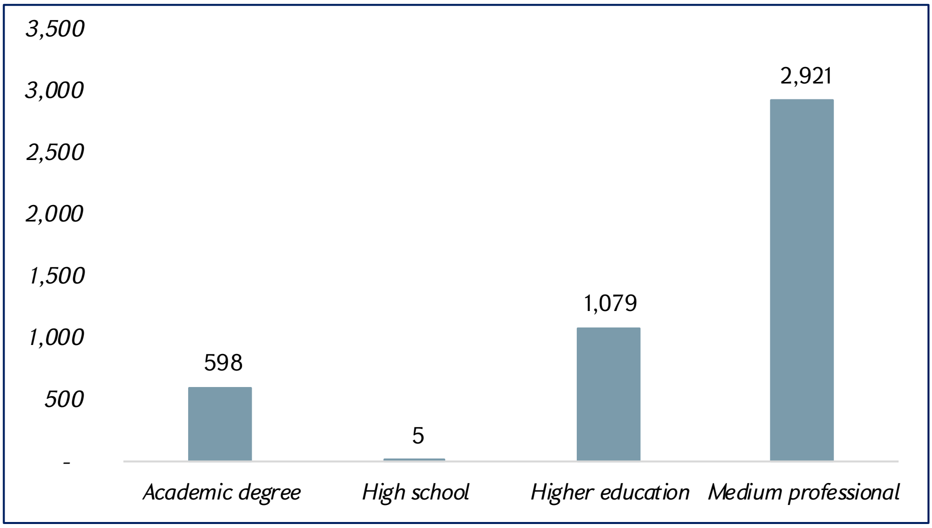

Figure 2. ***Analysis of customers depending on the level of education***

Table 5

Analysis of clients by marital status

Marital Status	*Default*		*Total*	*Default %*
	No	*Yes*		
Divorced	46	13	59	22,03%
Married	2139	1428	3567	40,03%
Single	42	30	72	41,67%
Widow	415	490	905	54,14%

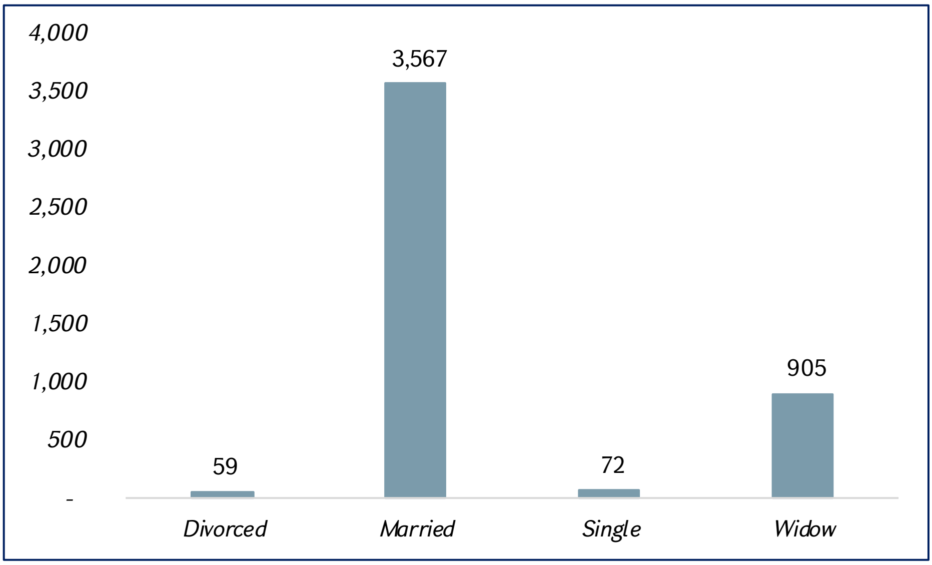

Figure 3. ***Analysis of clients by marital status***

Table 6

Analysis depending on the customer's gender

Sex	Default		Total	Default %
	No	Yes		
Female	1 492	948	2 440	38,85%
Male	1 150	1 013	2 163	46,83%

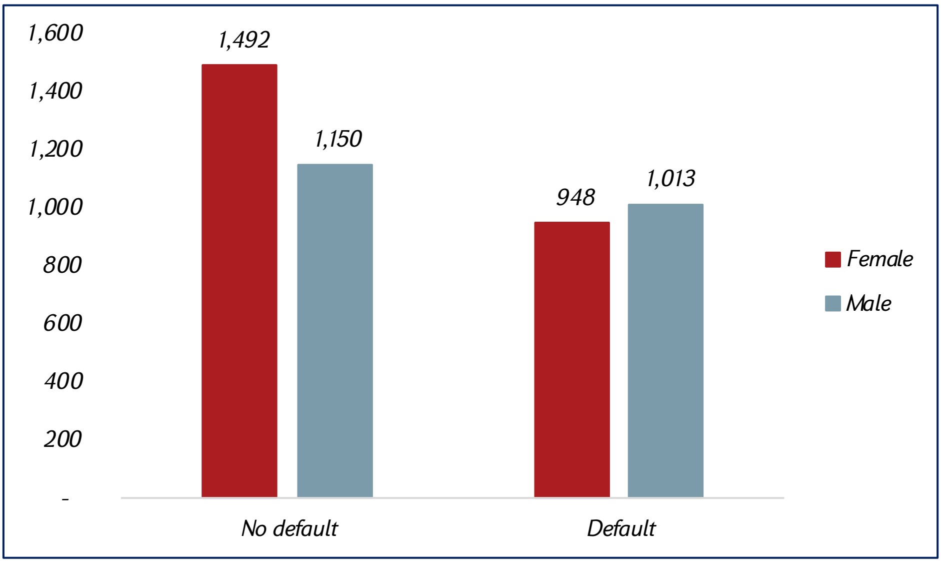

Figure 4. ***Analysis depending on the customer's gender***

After completing the analysis, it is necessary to develop the machine learning models (logistic regression, decision tree, and random forest), evaluate their quality indicators, and perform a comparative analysis to choose the best of these models.

For modeling purposes, it is necessary to group the available data and perform standardization for all groups using the above-mentioned WOE method.

Table 7 shows the grouping of data in terms of available observations and the values obtained as a result of the standardization of each group.

Table 7

Data regrouping, WOE calculation

Group	Sum	Bad	Good	The Share of the Bad	WOE
			Sex		
Female	612,677,502	252,268,372	360,409,130	41.2%	0.168654027
Male	557,539,675	277,974,655	279,565,020	49.9%	-0.182387387
			Education		
Academic degree	155,561,450	67,882,200	87,679,250	43.6%	0.067819077
Higher education	291,127,220	128,092,010	163,035,210	44.0%	0.053125014
Medium professional	722,189,907	333,875,217	388,314,690	46.2%	-0.037043596
High school	1,338,600	393,600	945,000	29.4%	0.68775742
			Availability of property		
Availability of real estate	70,759,900	31,236,900	39,523,000	44.1%	0.047190351

Availability of movable property	280,834,370	123,675,310	157,159,060	44.0%	0.051506408
Availability of real estate and movable property	59,682,480	24,920,490	34,761,990	41.8%	0.14474185
Absence	758,940,427	350,410,327	408,530,100	46.2%	-0.034631577
Marital status					
Married	913,540,317	392,974,777	520,565,540	43.0%	0.093078026
Widow	225,660,060	125,451,550	100,208,510	55.6%	-0.412758855
Divorced	13,085,400	3,069,200	10,016,200	23.5%	0.994694499
Single	17,931,400	8,747,500	9,183,900	48.8%	-0.139408336
Borrower’s age					
20-25	210,717,210	113,159,800	97,557,410	53.7%	-0.336452296
26-35	350,813,150	166,827,850	183,985,300	47.6%	-0.090198921
36-50	370,737,745	153,233,915	217,503,830	41.3%	0.162158508
51+	237,949,072	97,021,462	140,927,610	40.8%	0.185221801
Number of days past due in the last 12 months					
0-30	1,155,315,600	516,660,850	638,654,750	44.7%	0.023885006
31+	14,901,577	13,582,177	1,319,400	91.1%	-2.519673673
Number of delays					
0	540,572,830	219,287,600	321,285,230	40.6%	0.193852843
1	629,644,347	310,955,427	318,688,920	49.4%	-0.163526468
Number of changes in risk classes					
0-1	1,038,196,360	481,878,700	556,317,660	46.4%	-0.044445303
2+	132,020,817	48,364,327	83,656,490	36.6%	0.359864171
Credit load					
0	317,685,120	90,703,290	226,981,830	28.6%	0.729183998
1-300,000	207,519,120	94,002,700	113,516,420	45.3%	0.000531649
300,001+	645,012,937	345,537,037	299,475,900	53.6%	-0.331158231
Credit history length					
0-270	1,156,562,627	524,693,077	631,869,550	45.4%	-0.002222854
271-365	10,142,950	4,176,150	5,966,800	41.2%	0.168728659
366+	3,511,600	1,373,800	2,137,800	39.1%	0.254104298
Maximum repaid loans					
0-350,000	810,314,677	365,608,277	444,706,400	45.1%	0.007759468
350,001+	359,902,500	164,634,750	195,267,750	45.7%	-0.017450032
Contract sum					
Up to 200,000	520,208,537	212,833,517	307,375,020	40.9%	0.179465974
200,001-400,000	418,168,380	176,410,450	241,757,930	42.2%	0.127031212
400,001+	231,840,260	140,999,060	90,841,200	60.8%	-0.627732638

In this work for machine learning, the calculated WOE values for each group are taken as independent variables, and as a dependent variable that needs to be predicted, the attribute “Default”, which can be 0 (not default) or 1 (default).

The above WOE variable values and the predicted variable were compiled into a csv file using the Excel program. Based on these data, it is necessary to create appropriate models. Then, to get the best result, the appropriate parameters should be searched through cross-validation. For training, the data were divided into two groups; training (x_train, y_train) and testing (x_test, y_test) with a ratio of 65% and 35%. Taking into account that the logistic regression model is often used in scoring maps, it was decided not to choose the best parameters for that model.

After performing the above-mentioned actions, the models were trained with x_train and y_train data. The training was carried out based on the identified best parameters for the models (except for the logistic model). This is done by the GridSearchCV function of the Sklearn library. So, based on the available data for decision trees, the best parameters were to calculate using the Gini coefficient and a tree depth of 9. For random forests, the best parameters found are a depth of 10 for each tree and a number of 150 trees.

Models created with those parameters are saved, and regarding those models, a quality index was calculated using the sklearn library's scor method. This is the simplest metric for evaluating the quality of a model and shows the percentage of correctly predicted outcomes. So for logistic regression, this indicator calculated on the training base was 62.0%, and for the test base, it was 58.9%. For decision tree models, this indicator calculated on the training base was 71.4%, and for the test base it was 64.9%, and for random forest models 74.4% and 64.1%, respectively.

Based on the test data, the Precession and Recall coefficients were also calculated, indicating the quality of the models. Then, F1 score was calculated for each model, which ideologically combines the information of the above-mentioned indicators (Table 8).[11] Below are the formulas for calculating these indicators (1), (2), (3).

$$Precesion = \frac{TP}{TP + FP}, \qquad (1)$$

$$Recall = \frac{TP}{TP + FN}, \qquad (2)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}, \qquad (3)$$

where TP (True positive) is the number of correctly predicted observations; FP (False positive) is the number of incorrectly positively predicted observations; FN (False negative) is the number of mis-negatively predicted observations.

Table 8

Qualitative indicators of models based on test data

Models	*Test data*			
	Precession	*Recall*	*F1 Score*	*AUC*
Logistic regression	0.5549	0.2675	0.3610	0.6275
Decision tree	0.6338	0.4506	0.5268	0.6661
Random forest	0.6364	0.4006	0.4917	0.6801

After all this, the ROC curve for each model was constructed. It is a graph that allows you to evaluate the quality of the classifier (model), showing the ratio between correctly classified and incorrectly classified objects (figure 5). Quantitative interpretation of the ROC provides the AUC (Area under Curve) indicator, the surface bounded by the ROC curve, and the axis of the proportion

[11] N. Shakla, "Machine Learning & TensorFlow", – Peter, 2019, p. 118 Date of last access: 10.10.22 թ. 21:25:00 (in Russian)

of false positive classifications. The higher the AUC index, the better the classifier is, while a value of 0.5 indicates the inappropriateness of the chosen classification method. A value less than 0.5 indicate that the classifier is doing just the opposite.[12] Below is the formula for calculating AUC (4).

Table 9 below quantitatively presents the prediction results of all models based on test data, and in Table 10, based on the test data, the results of the predictions of defaulted customers only.

Table 11 presents the results of forecasts of defaulted customers based on test data in quantitative terms, from which it can be seen that in the case of implementing decision trees, the considered financial organization can reduce its expenses by 94.9 million AMD.

Based on the work done, it can be noted that the best of the developed models is the decision tree, because according to Table 8, this model can work more effectively to identify defaulted customers.

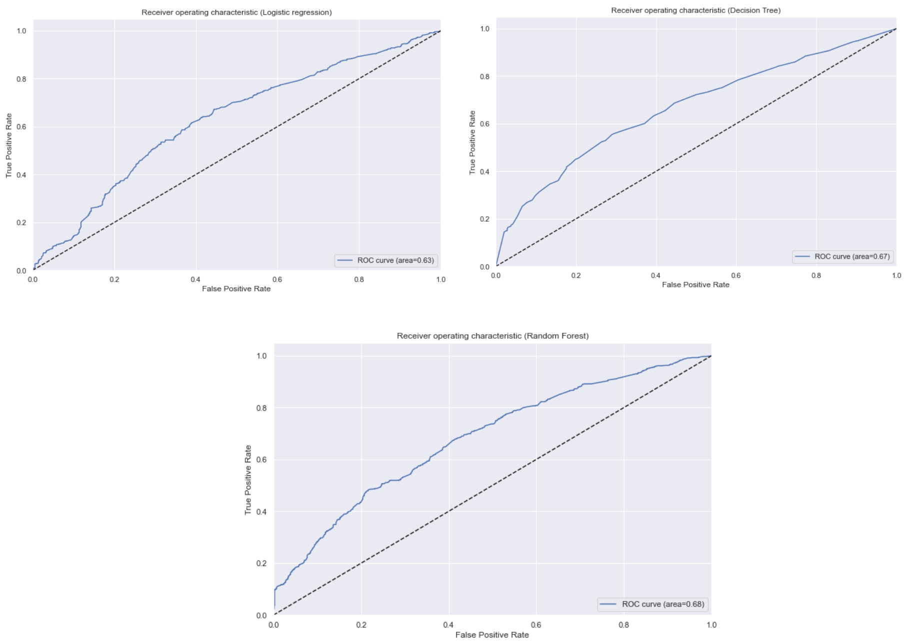

Figure 5. ***ROC curves of the models***

Table 9

Prediction results based on test data (quantitative data are presented)

Prediction status	*Logistic regression*	*Decision tree*	*Random forest*
Right	950	1046	1033
Wrong	662	566	579
Sum	**1612**	**1612**	**1612**

[12] https://www.ibm.com/docs/ru/spss-statistics/beta?topic=features-roc-analysis. Date of last access: 17.10.22 թ. 21:25:00

Table 10

Prediction results only for defaulted customers based on test data (quantitative data are presented)

Prediction status	Logistic regression	Decision tree	Random forest
Right	187	315	280
Wrong	512	384	419
Sum	**699**	**699**	**699**

Table 11

Prediction results only for defaulted customers based on test data (dimensional data are presented)

Model type	Predicted wrong	Predicted right	Sum	The proportion of right predictions
Logistic regression	122,573,150	62,366,862	184,940,012	33.7%
Decision tree	90,006,150	94,933,862	184,940,012	51.3%
Random forest	102,357,550	82,582,462	184,940,012	44.7%

Conclusions. It is important for credit organizations to have a clear understanding of the need to identify, measure, monitor, and manage credit risk. To curb this risk it is necessary to introduce a reasonable decision-making system, and by this, the modernity of the research is conditioned.

The purpose of the work was to research customers' behavior based on certain examples, to develop several decision-making models, and to choose the best of these models. One of the options for the optimal solution to this problem is the use of machine learning methods because currently, credit organizations have generated a fairly large amount of data that needs to be processed, researched, and then build a decision-making model based on the obtained results.

Taking into account that the problem of making a decision on providing a loan during lending is a classification problem, a logistic regression model, a decision tree model, and a random forest model of machine learning were used in this work.

For the work in this research, it was necessary to solve the following problems;

1. Examine existing data to identify relationships between indicators;
2. Develop decision-making models, and test them.

As a result of the analysis of the available data, three models were developed, of which the decision tree model was considered the best according to the obtained results.

Given the peculiarities of risk management and regulatory legal acts, when issuing loans, banks are required to form reserves for possible losses on loans. As the customer's loan service worsens, banks are forced to increase the amount of these reserves, which leads to a decrease in their profits.

Loans from defaulted customers in most cases cannot be recovered, which causes financial institutions to suffer losses in the amount of the entire loan.

As part of this study, using test data that was not used to create machine learning models, we estimated the losses that a bank can reduce when using a decision tree model. This is the sum of loans of defaulted borrowers (from a test dataset that the model has not yet seen) that the bank has issued a loan without using machine learning models and that the decision tree has classified as bad borrowers. Table 11 shows that the amount of these loans is 94.9 million AMD. It can also be seen from this table that the amount of loans issued, for which the decision tree during predictions made a mistake, is the smallest compared to other models (90.0 million AMD).

It is necessary to note that the quality of this or that machine model also depends on the qualitative and quantitative characteristics of the data used for modeling. Therefore, in the presence of other data, another type of model can be recognized as the best.

From the results of the work, it can be concluded that the application of machine learning models in the financial and banking system can reduce the impact of existing credit risks and the costs of organizations.

In addition, the use of machine learning implies the complete automation of the decision-making process, which will lead to a reduction in the number of errors made during manual calculations. As a result of this, the time for issuing a loan will also be reduced, which will lead to the optimization of processes and an improvement in the quality of loan servicing.

References

1. N. Shakla, “Machine Learning & TensorFlow”, - Peter, 2019, p. 336.
2. A. S. Sorokin, “Construction of scoring maps using the logistic regression model”, Internet journal “NAUKOVEDENIE”, No. 2 (21) 2014.
3. A.A. Stroev, SAS company, «Implementation of a credit scoring system in a bank», Methodical journal Calculations and operational work in a commercial bank, number 6/2004, Moscow
4. J. Bastos, «Credit scoring with boosted decision trees», CEMAPRE, School of Economics and Management (ISEG), Technical University of Lisbon, 2008.
5. https://corporatefinanceinstitute.com/resources/knowledge/finance /correlation/. Date of last access: 16.10.22 թ. 22:25:00
6. https://www.geeksforgeeks.org/binning-in-data-mining/. Date of last access: 17.10.22 թ. 22:58:00
7. https://medium.com/mlearning-ai/weight-of-evidence-woe-and-information-value-iv-how-to-use-it-in-eda-and-model-building-3b3b98efe0e8. Date of last access: 16.10.22 թ. 23:15:00
8. https://www.ibm.com/topics/logistic-regression. Date of last access: 17.10.22 թ. 23:08:00
9. https://www.ibm.com/topics/decision-trees. Date of last access: 17.10.22 թ. 23:30:00
10. https://www.ibm.com/cloud/learn/random-forest. Date of last access: 17.10.22 թ. 23:22:00
11. https://www.ibm.com/docs/ru/spss-statistics/beta?topic=features-roc-analysis. Date of last access: 17.10.22 թ. 21:25:00
12. https://www.pycodemates.com/2022/05/precision-and-recall-in-classification.html. Date of last access: 17.10.22 թ. 20:22:00

ԱՐՄԵՆ ՂԱԶԱՐՅԱՆ
ՀՊՏՀ տնտեսական ինֆորմատիկայի և
տեղեկատվական համակարգերի ամբիոնի վարիչ,
տնտեսագիտության թեկնածու, դոցենտ

ԼԻԱՆԱ ԳՐԻԳՈՐՅԱՆ
ՀՊՏՀ կառավարչական հաշվառման և աուդիտի ամբիոնի վարիչ,
տնտեսագիտության դոկտոր, պրոֆեսոր

ԳԱՌՆԻԿ ԱՌԱՔԵԼՅԱՆ
ՀՊՏՀ տնտեսական ինֆորմատիկայի և
տեղեկատվական համակարգերի ամբիոնի դասախոս

Մեքենայական ուսուցման կիրառումը ֆիզիկական անձանց վարկային ռիսկի կառավարման համակարգում.– Յուրաքանչյուր վարկատու կազմակերպությունում առկա են բազմաթիվ խնդիրներ: Դրանցից կարևորագույնը վարկավորման ժամանակ հնարավոր կորուստներ կրելու ռիսկն է: Թեմայի շրջանակում ուսումնասիրվել են այլ հետազոտողների կողմից իրականացված աշխատանքները, եզրակացվել է, որ վերոնշյալ խնդրի օպտիմալ լուծման համար հաճախ են կիրառվում մեքենայական ուսուցման միջոցներ: Աշխատանքում մոդելավորման համար որպես հիմք են ընդունվել վարկերի վերաբերյալ իրական տվյալները: Սույն աշխատանքում առկա տվյալների հիման վրա մշակվել է մեքենայական ուսուցման մի քանի մոդել, որոնցից ընտրվել է լավագույնը, որը կարող է նպաստել վարկային ռիսկի կառավարման գործընթացի բարելավմանը: Աշխատանքի ընթացքում բացահայտվել են տվյալների միջև առկա տրամաբանական կապերը և դրանց փոխազդեցությունը: Այնուհետև կատարված աշխատանքի հիման վրա կառուցվել են համապատասխան մոդելներ, որոնց որակը ստուգվել է տարբեր գործիքների միջոցով: Կատարվել է ստացված մոդելների համեմատություն, ընտրվել է լավագույնը: Ստացված արդյունքները գործնականում կիրառելի են և ցույց են տալիս, որ յուրաքանչյուր բանկ և վարկային կազմակերպություն իր մոտ առկա տվյալների խոշոր բազաների հիման վրա կարող է մշակել ավելի որակյալ լուծում, ինչը կնպաստի վարկային ռիսկի զսպմանը և ծախսերի կրճատմանը:

Հիմնաբառեր. *վարկային ռիսկ, լոգիստիկ ռեգրեսիա, որոշումների ծառեր, պատահական անտառներ*
JEL: C81, C88
DOI: 10.52174/1829-0280_2022_5_123

АРМЕН КАЗАРЯН
Заведующий кафедрой экономической информатики и информационных систем АГЭУ, кандидат экономических наук, доцент

ЛИАНА ГРИГОРЯН
Заведующая кафедрой управленческого учета и аудита АГЭУ, кандидат экономических наук, профессор

ГАРНИК АРАКЕЛЯН
Преподаватель кафедры экономической информатики и информационных систем АГЭУ

Применение машинного обучения в системе управления кредитными рисками физических лиц.– В каждом кредитном учреждении есть множество проблем. Важнейшим из них является риск возможных потерь при кредитовании. В рамках исследования были изучены труды ряда исследователей, на основе которых был сделан вывод о том, что для оптимального решения вышеуказанной задачи часто используются средства машинного обучения. В качестве основы для моделирования в работе использовались реальные данные по кредитам. В данной работе на основе имеющихся данных было разработано несколько моделей машинного обучения, из которых была выбрана лучшая, способная помочь совершенствованию процесса управления кредитным риском. В ходе работы были выявлены логические связи между данными и их взаимодействие друг с другом. Затем на основе проделанной работы были разработаны соответствующие модели, качество которых проверялось с помощью различных инструментов. Проведено сравнение полученных моделей и выбрана лучшая. Полученные результаты применимы на практике и показывают, что каждый банк и кредитная организация может разработать лучшее решение на основе имеющихся у них больших баз данных, что будет способствовать сдерживанию кредитного риска и снижению затрат.

Ключевые слова: *кредитный риск, логистическая регрессия, деревья решений, случайные леса*
JEL: C81, C88
DOI: 10.52174/1829-0280_2022_5_123