ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse52942020.pdf
https://doi.org/10.30534/ijatcse/2020/52942020

Construction of models for conversion of mortgage applications by the method of multiple regression and Neural Networks

Aleksey V. Burkov[1],Ruslan V. Pshenichnov[2], Tatiana V. Yalyalieva[3]
[1]Mari State University, Yoshkar-Ola, Russia, alexey.burkov@gmail.com
[2]Mari State University, Yoshkar-Ola, Russia, psheni4nov@yandex.ru
[3]Volga State University of Technology, Yoshkar-Ola Russia, yal05@mail.ru

ABSTRACT

The article is devoted to the problem of modeling the results of work in the field of mortgage lending. Currently, the relevance of the problems and prospects for the development of the mortgage market both in Russia as a whole and its subjects, as well as its impact on the economy and welfare of the country, is the subject of research by many scientists and researchers. The solution to the problem includes work on several blocks of tasks at once, namely: clarification of the conceptual apparatus; analysis of factors affecting the assessment of mortgage risk; improving statistical models to identify unreliable borrowers; modeling of mortgage lending processes considering the peculiarities of regional economic development.Based on the methods of regression analysis and neural networks, a new technique has been developed that allows you to simulate the conversion of mortgage applications considering the current situation in the housing market and mortgages. Thus, the obtained models allow us to determine the dependence of the number of mortgage loans provided on the factors received. The constructed models are of practical importance, as well as the use of factors that play a key role in the formation of the mortgage market.

Key words: Mortgage, mortgage lending market, modeling, neural networks

1. INTRODUCTION

It is not by chance that the housing sector has become one of the keys in the government's strategy to overcome the crisis. Comfort and quality of living conditions is an indicator of the standard of living of citizens, which is a key issue of public policy. But in recent years, the Russian economy has repeatedly experienced a crisis, including in the field of housing and mortgages. In this regard, it is important to develop statistical analysis tools aimed at modeling the mortgage lending market. It is most efficient to build models at the very early stages of the application life cycle, because more time is left for corrective action. If the flow of applications is large, then to attract additional employees to issue loans; if there are few applications, then pay more attention to attracting applications, strengthen advertising or apply motivational measures for employees receiving applications. In accordance with this goal, the article is supposed to solve the following problems:

a) clarify the concepts of the apparatus used for the purposes of state statistical monitoring in the field of mortgage lending;
b) to analyze the state, degree of development and main trends in the housing market and mortgages at the level of regions and the country as a whole;
c) propose a data clustering model in order to increase the accuracy of forecast values;
d) to generalize and systematize the results of the study, in order to determine the optimal tools for modeling the conversion of mortgage applications, and their implementation in the banking sector.

2. REVIEW OF LITERATURE

Currently, the relevance of the problems and prospects for the development of the mortgage market both in Russia as a whole and its subjects, as well as its impact on the economy and welfare of the country, is the subject of research by many scientists and researchers. The solution to the problem includes work on several blocks of tasks at once, namely: clarification of the conceptual apparatus; analysis of factors affecting the assessment of mortgage risk; improving statistical models to identify unreliable borrowers; modeling of mortgage lending processes considering the peculiarities of regional economic development. Each of these aspects has received coverage in the scientific literature. Most authors note a sharp change in the situation in the mortgage market. So, Lapteva E.V. analyzes the dynamics of average prices in the mortgage lending market of the Russian Federation [1]. Vilchinskaya E.K. He focuses on finding ways to solve the problems of housing mortgages, and considers ways to increase its accessibility [2]. Simakova E.K. focuses on the state-legal regime for supporting mortgages as a factor in infrastructure development in a crisis [3]. Studying the problems of mortgage lending is impossible without using the modeling method described in E. Murzina and A. Burkov [4], [5]. These studies reflect the relevance of issues related to the planning of mortgage activities, and the topic requires further analysis and research.

Despite a deep study of the topic under study and the high professionalism of the authors, the question of a comprehensive analysis of indicators of the mortgage lending market remains open. It is necessary to combine the approaches of

the authors outlined in each of the works for a comprehensive assessment of the state of the mortgage lending market.

3. MATERIALS AND METHODS

The research information base is made up of official data of the Federal State Statistics Service, reporting data of the Central Bank of the Russian Federation (CBR), JSC "DOM.RF", the Analytical Center for Mortgage Lending and Securitization "Rusipoteka", open data of the Public Joint Stock Company Sberbank of Russia, as well as materials of scientific publications, periodicals, official Internet sites and electronic media on the subject under study.

To process the initial information and solve the tasks, the software packages “Microsoft Excel 2010”, “Statistica 12” and “SPSS 10” were used.The theoretical and methodological basis of the study was the work of domestic and foreign authors in the field of statistics, economics, econometrics, as well as work on the modeling of mortgage lending processes. As research tools used multidimensional statistical methods of correlation, regression, cluster analysis, methods of neural networks, tabular and graphical methods for presenting the results of the study.

4. RESULTS

We will build models for converting mortgage applications by multiple regression at the stage of receiving applications. To search for the most accurate model, modeling was performed using linear and ridge regression methods. The initial stage for modeling was the acceptance of applications. In order to exclude insignificant factors from the model, the methods of direct selection, sequential inclusion and reverse exclusion were used.
As a result of parameter estimation by linear regression method, by the method of reverse exclusion of insignificant variables, it was possible to build the most accurate model - as a result, an equation of the form was obtained:

$$\bar{y} = 24{,}79 + 0{,}43x_1 - 2{,}63x_2 + 0{,}10x_3,$$

t-values:(2,73) (13,31) (-2,72) (2,45)

where$\bar{y}$ – number of loans granted, x_1 – number of accepted applications, x_2 – weighted average interest rate, x_3 – dollar rate.

Darbin-Watson coefficient is 2,048, DW>du ((4-2,048)>1,799). The hypothesis of independence of random deviations is not rejected, autocorrelation is absent, at a significance level of 0.05. The determination coefficient is 0.78, which indicates the good quality of the constructed model. The obtained parameters of the regression model are shown in table 1.

Based on the t-test, we can conclude that the null hypothesis is rejected ($|t_{obs.}| > t_{\text{cr.}}(1{,}98)$) at a significance level of 0.05. Based on the p-level, it is clear that all factors of the constructed model are significant (p-level <0.05). To test the model for the adequacy of the distribution of residues and accuracy, Figures 1-2 show a histogram of the distribution of residues and a dispersion diagram.

Table 1: Parameters of the multiple regression model

Independent variables	Standardized Regression Coefficients	Standard error	Unstandardized regression coefficients	Standard error	Probability of the null hypothesis for the free term of the equation
Freemember	-	-	49,13	18,00	0,01
Numberofaccepteda pplications	0,84	0,06	0,44	0,03	0,01
Weightedaverageint erestrate	-0,12	0,07	-2,63	1,53	0,04
Dollarrate	0,09	0,06	0,10	0,07	0,04

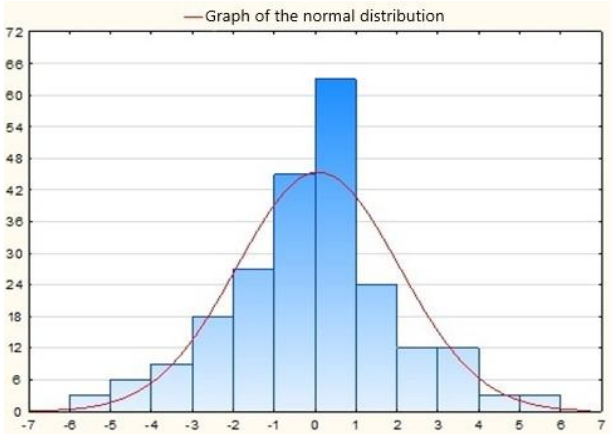

Figure 1: Histogram of the distribution of residues with a graph of normal distribution

The histogram of the distribution of residues does not have critical deviations from the normal distribution graph, the model can be considered adequate. From figure 2. it can be seen that the line fits well on the data. The scatter of the observed values is small, from which it was concluded that the dependence is close to linear. When substituting sample data into the equation, the selected method showed the best results among the models constructed by the linear regression method. The average approximation error of this model was 12.14%.

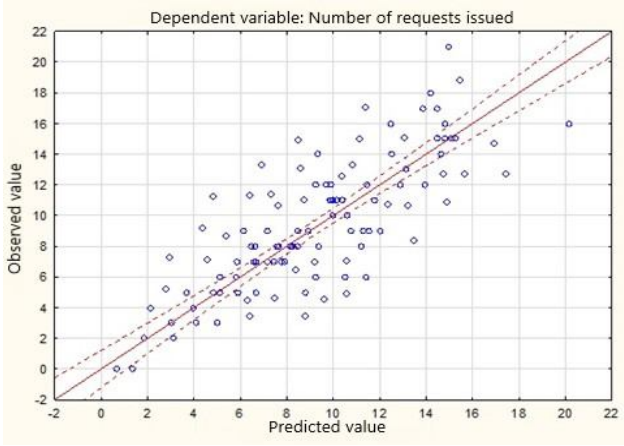

Figure 2: Scatter plot

The next stage in the life cycle of mortgage applications is the stage of operation of the suspensive condition. Modeling at the stage of operation of the suspensive condition is different in that the risks related to underwriting applications are already excluded from the models. Therefore, the factor of the number of approved applications in the data will be of greatest importance.

In the most accurate model constructed by the ridge regression method, the use of sequential inclusion of significant variables allowed us to obtain the greatest accuracy. As a result, an equation of the form was obtained:

$$\bar{y} = 25{,}81 + 0{,}59x_1 + 2{,}05x_2,$$

t-values: (2,61) (14,20) (-2,59)

where$\bar{y}$ – number of loans granted, x_1 – number of approved applications, x_2 – Weighted average interest rate in the Republic of Mari El. The Darbin-Watson coefficient is 1.836, DW> du (1.836> 1.789). The hypothesis of independence of random deviations is not rejected, autocorrelation is absent, with a significance level of 0.05. The coefficient of determination is 0.79. Table 2 shows the parameters of the regression model.

Table 2: Parameters of the regression model

Independent variables	Standardized Regression Coefficients	Standard error	Unstandardized regression coefficients	Standard error	Probability of the null hypothesis for the free term of the equation
Free member	-	-	25,81	16,02	0,01
Numberof approved applications	0,81	0,06	0,59	0,04	0,01
Weighted average interestrate	-0,09	0,06	-2,05	1,29	0,02

Based on the t-criterion, we can conclude that the null hypothesis is rejected ($|t_{obs.}| > t_{cr.}(1{,}98)$) at a significance level of 0.05. Based on the p-level, it is clear that all factors of the constructed model are significant (p-level <0.05). To test the model for adequacy and compliance with real data, a histogram of the distribution of residues and a scatter plot was constructed (Figures 3-4).

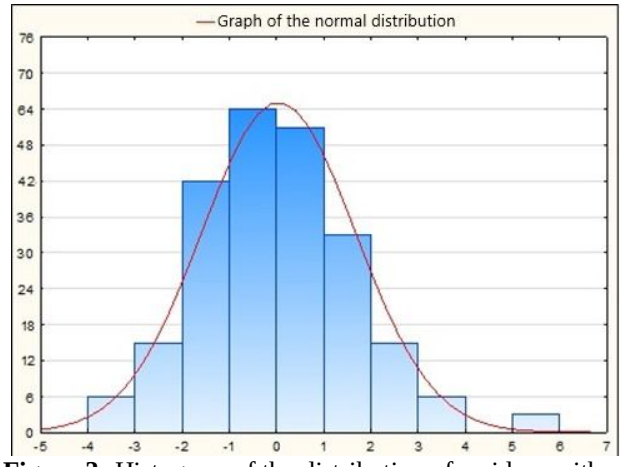

Figure 3: Histogram of the distribution of residues with a graph of the normal distribution

The distribution schedule of the residues does not have critical deviations from the normal distribution schedule. Figure 4 shows that the straight line lays well on the data - this indicates that the dependence is close to linear. To determine the possibility of using the model in practice, the average approximation error was calculated - 9.12%.

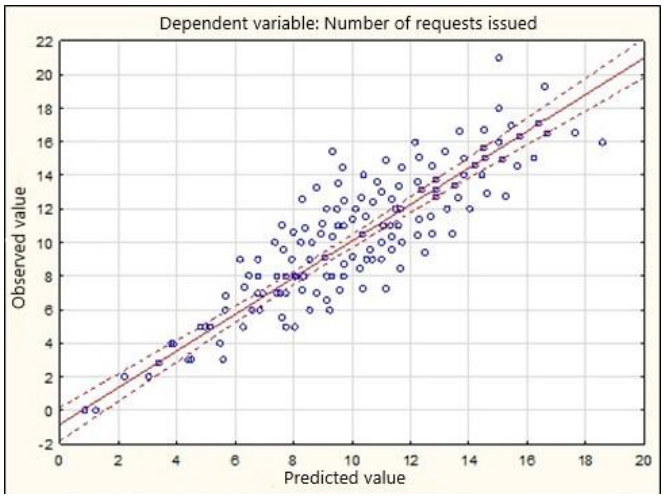

Figure 4: Scatter plot

Table 3: Parameters of neural networks

Networksettings	1. Multilayerperceptron 3-4-1	2. Multilayerperceptron 3-3-1	3. Multilayerperceptron 3-3-1
Quality on the training sample	0,86	0,92	0,87
Quality on the control sample	0,9	0,94	0,9
Quality on the test sample	0,99	0,97	0,98
Error on the training sample	2,7	1,6	2,6
Error on the control sample	1,6	1,34	2,1
Error on the test sample	0,75	0,47	0,32
Learningalgorithm	Broyden-Fletcher-Goldfarb-Shanno Algorithm 6	Broyden-Fletcher-Goldfarb-Shanno Algorithm 128	Broyden-Fletcher-Goldfarb-Shanno Algorithm 6
Errorfunction	Sumofsquares	Sumofsquares	Sumofsquares
HiddenNeuronActivationFunction	Logistic	Hyperbolic	Sinusoidal
Functionofinputneurons	Exponential	Hyperbolic	Exponential

In order to increase the accuracy of the forecast values of the conversion models of mortgage applications, we will construct similar models using the neural network method. To compare the accuracy of the models, we will use the indicator of the average approximation error. Based on the results, it will be possible to conclude which methods are best used

in practice when modeling the conversion of mortgage applications in the Republic of Mari El.
We will build models using the neural network method at the stage of receiving applications. As a result of the calculations, the 3 most optimal neural networks were selected. Network parameters are presented in table 3.

All constructed models have a multilayer perceptron structure. The difference between the first model is that in the second layer the model has 4 neurons, while the rest of the models have 3. The second model has a learning algorithm, but, in general, the networks are very similar in many ways, and this proves insignificant differences in performance indicators. In terms of error, the second network is best. The histogram of the remnants of the network No. 2 multilayer perceptron 3-3-1 is presented in Figure 5.

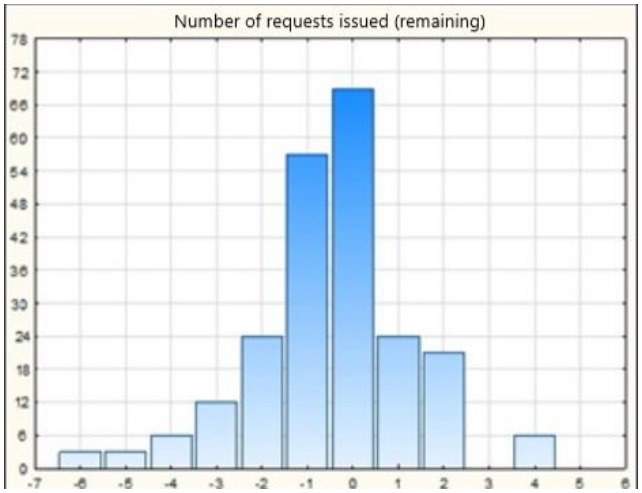

Figure 5: Histogram of the distribution of residues network No. 2 Multilayer perceptron 3-3-1

Of the 3 networks that showed the best results, only network number 2 can be used in practice. The approximation error for this network was 9.79%, which is 2.14% lower than the average approximation error of the best model constructed by the multiple regression method for the total sample at the application acceptance stage.
We construct the model by the method of neural networks at the stage of operation of the suspensive condition. The parameters of the neural networks that showed the best results are shown in table 4.
The table shows that neural networks based on the radial basis function showed worse results compared to networks with a multilayer perceptron structure. The greatest efficiency and the smallest error can be observed in the network No. 3 Multilayer perceptron 2-3-1.
As a result of the analysis of the constructed networks, data were obtained that the network No. 3 (Multilayer perceptron 2-8-1) has the highest accuracy - 9.02%. For comparison, we construct a graph of the average error of approximation of the best models obtained by multiple regression methods and neural networks from a common sample.

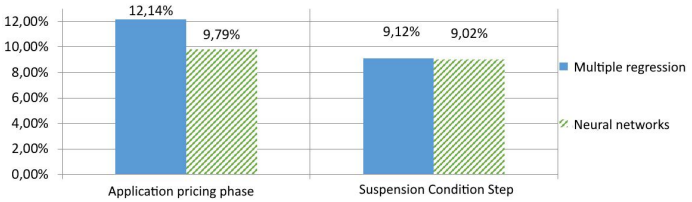

Figure 6: Diagram of the average error of approximation of the most accurate models constructed by the method of neural networks and multiple regression for the total sample

Table 4: Parameters of neural networks

Networks ettings	1. Multilayerperceptron 2-8-1	2. Multilayerperceptron 2-6-1	3. Multilayerperceptron 2-8-1	4. Multilayerperceptron 2-8-1	5. Multilayerperceptron 2-3-1	6. Multilayerperceptron 2-9-1	7. Multilayerperceptron 2-7-1
Quality on the training sample	0,93	0,88	0,93	0,91	0,93	0,93	0,93
Quality on the control sample	0,91	0,93	0,91	0,93	0,91	0,91	0,91
Quality on the test sample	0,98	0,99	0,98	0,98	0,98	0,97	0,98
Error on the training sample	1,49	2,48	1,49	1,9	1,5	1,4	1,49
Error on the control sample	1,43	1,27	1,43	1,11	1,42	1,41	1,38
Error on the test sample	0,2	0,24	0,2	0,28	0,19	0,33	0,33
Learningalgorithm	Broyden-Fletcher-Gold-farb-Shanno Algorithm 7	Broyden-Fletcher-Gold-farb-Shanno Algorithm 5	Broyden-Fletcher-Gold-farb-Shanno Algorithm 7	Broyden-Fletcher-Gold-farb-Shanno Algorithm 6	Broyden-Fletcher-Gold-farb-Shanno Algorithm 7	Broyden-Fletcher-Gold-farb-Shanno Algorithm 19	Broyden-Fletcher-Gold-farb-Shanno Algorithm 8
Errorfunction	Sumofsquares	Sumofsquares	Sumofsquares	Sumofsquares	Sumofsquares	Sumofsquares	Sumofsquares
HiddenNeuronActivationFunction	Identical	Identical	Identical	Exponential	Identical	Exponential	Identical
Functionofinputneurons	Identical	Exponential	Identical	Identical	Identical	Hyperbolic	Logistic

5. CONCLUSION

Based on the indicator of the average approximation error, the greatest accuracy is shown by models constructed by the neural network method. It should be noted that at the stage of receiving applications, the accuracy of the best model obtained by the neural network method is 2.35% higher than the accuracy of the best model constructed by the multiple

regression method and is 9.79%. At the stage of operation of the suspensive condition, the accuracy of the model constructed by the neural network method in terms of the average approximation error is higher by 0.1% and amounts to 9.02%.

6. ACKNOWLEDGMENTS

The reported study was funded by RFBR, project number №20-010-00472 A

REFERENCES

1. E.V. Lapteva (2016) **Analysis of the dynamics of average prices in the mortgage market of the Russian Federation,** Symbol of science. 2016. 5-1 (17). 148-152.
2. E.K. Vilchinskaya (2014) **Accessibility of Housing Mortgages: Problems and Ways of Their Resolution,** Scientific Bulletin of the Omsk Academy of the Ministry of Internal Affairs of Russia. 2014. 1 (52). 71-73.
3. E.K. Simakova**The state legal regime for supporting mortgages as a factor in infrastructure development in a crisis,** Bulletin of the St. Petersburg Law Academy. - SPb. 2016. 1 (30). 76-81.
4. A.V.Burkov, E.A. Murzina (2016). **Analysis method of structural equation modeling,** Advances in Systems Science and Applications, 2016. 4 (16).
5. A.V.Burkov, E.A. Murzina (2019). **The use of logit and probit regression models in the process of graduates' employment**, International journal of scientific & technology research, 2019. 8 (11).
6. Gunawan Wang, LarasAnggitaDestofia, FakhriNurullah and Devi YuriscaBernanda (2019) The Influence of Social Media and Knowledge Management to Improve Employees Creativity. **International Journal of Advanced Trends in Computer Science and Engineering**, (2019), 8 (5), 1927-1936. https://doi.org/10.30534/ijatcse/2019/17852019.
7. T.V. Zavgorodnaya(2010)**Mortgage lending (for example, JSC AKB "Rosbank", Omsk branch)**. 2010.
8. T.S. Korosteleva (2016) **Mortgage lending as a factor in intensifying the growth of regional ecosystems (based on materials from the Samara region)**. Housing strategies. 2016. 4(3). 279-298.
9. A.A. Shumeyko (2017) **Statistical analysis of the development of mortgage lending in Russia,** Actual problems of economics and management: materials of the V International. scientific conf. 2017. 68-73.
10. P.Santhi, N.Deeban, N.Jeyapunitha, B.Muthukumaran, R. Ravikumar (2020) **Prediction of Diabetes using Neural Networks**, International Journal of Advanced Trends in Computer Science and Engineering, (2020), 9 (2), 985-990. https://doi.org/10.30534/ijatcse/2020/13922020
11. The Central Bank of the Russian Federation [Electronic resource]. Access mode: http: //www.cbr.ru. (Date of access: 05.05.2019)
12. Tuga Mauritsius, Annisa Safira Braza and Fransisca (2019) Bank Marketing Data Mining using CRISP-DM Approach. **International Journal of Advanced Trends in Computer Science and Engineering**, (2019), 8 (5), 2322-2329. https://doi.org/10.30534/ijatcse/2019/71852019