Model for Credit Scoring on a Base of Behavioural and Macroeconomic Predictors

Azat R. Abubakirov, Nikolay O. Nikitin, Anna V. Kalyuzhnaya
ITMO University
Saint-Petersburg, Russia
mr.azat.abubakirov@gmail.com, nikolay.o.nikitin@gmail.com, kalyuzhnaya.ann@gmail.com

***Abstract*— This paper is devoted to the research questions to the prediction of defaults on loans, taking into account behavioural predictors and introducing additional macroeconomic predictors. Where behavioural predictors are the variables which characterise the financial behaviour of the borrower, and that could be identified on a base of their transactional history analysis. As the method underlying the scoring model, logistic regression is used. In the experimental studies, the effect of taking into account macroeconomic predictors is investigated. Also, the results of the study of the dependence of the quality and stability of the prediction result are presented when selecting a significant set of variables and changing the size of the training and test samples.**

Keywords*— *behavioural scoring; macroeconomic factors; logistic regression; default on a loan

I. Introduction

Classical applicational scoring is based on data that describes the client of the bank and the financial product that he/she was interested in. The disadvantage of models that use such data for training is that they are static: they do not take into account the dynamics of the client's financial conditions and, accordingly, the creditworthiness. To the dynamic characteristics of the client can be included, for example, information that describes the monetary assets flow of borrower for the last year or the number of purchases paid by credit card. The described problem is solved within the framework of the task of behavioural (transactional) scoring, which involves the use of aggregated data on customer transactions.

Using such approaches, banks assume that the external economic situation does not affect the solvency of borrowers. The current study aims to investigate the possibility of improvement of prediction quality by adding macroeconomic variables.

II. Related work

The use of macroeconomic factors can be implemented in different ways. For example, they can be added to the model by logistic regression [1]. Another solution is the model based on the survival analysis [2], [3]. This approach, which was originally used in medicine to assess the effectiveness of treatment, naturally describes the process of borrower default. Crook and Bellotti added time-dependent economic factors to the model: the unemployment index, the interest rate, the purchasing power index, and others. This data made it possible to improve the accuracy of forecasts slightly, as well as to show the dependence of the borrower's default on the economic situation.

Also, data based on customer transactions history can be added to the survival model [4]. The introduction of such data allowed to improve the predictive ability of the model: the best result was obtained using behavioural and macroeconomic factors for the last 12 and three months, respectively.

The classical survival model assumes [5] that the event of interest eventually will come. Models, called *Mixture cure models*, with macroeconomic parameters [6] allow us first determine what category the client belongs to (by static data), and after to introduce the dynamic data to calculate the probability of survival.

The presence of foreign economic factors in the model allows not only to increase the accuracy of forecasting, but also to model risks in different economic conditions [4], [7].

III. Data

A. Data description

To construct the credit scoring model on a base of behavioural predictors we use anonymised transactional data (for 2014-2016) of the clients of one of the major banks in Russia. Data were provided in aggregate form and contained the following indicators: the date of the loan agreement, the number of transactions, the average amount that accounted for one transaction, the average amount that was cashed, the share of transactions that relate to health, food purchases, the purchase of items wardrobe and finance, and many others. An indicator of delayed payment for more than 30 days was used as an objective variable.

As macroeconomic parameters, the following indicators were taken:

- the dollar exchange rate against the ruble (usd);
- the price of a Brent oil barrel (brent);
- the average amount of the application for consumer credit in Russia (avg. crd);
- number of applications for consumer credit per month in Russia (avg. aps);

978-1-5386-5824-6/18/$31.00 ©2018 IEEE

- the average size of the new deposit in Russia (cnt. crd);
- the number of new deposits per month in Russia (cnt. dps).

The values of the last four parameters were obtained from the information resource "Open data of Sberbank" [9]. The dynamics of changes in the macroeconomic indicators are presented in Fig. 1–3.

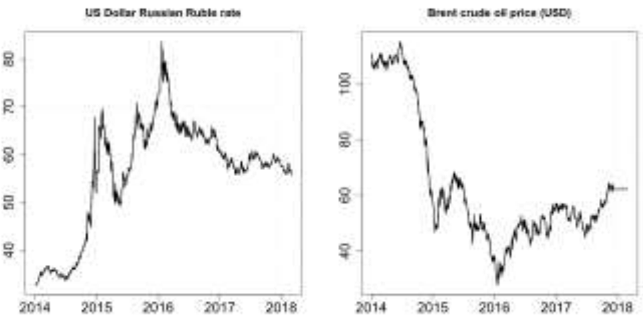

Fig. 1. Dynamics of oil price and the dollar exchange rate over 2014 – 2018 years

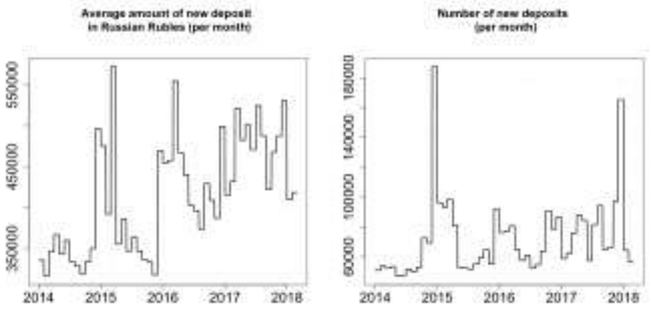

Fig. 2. Dynamics of new deposits in Russia

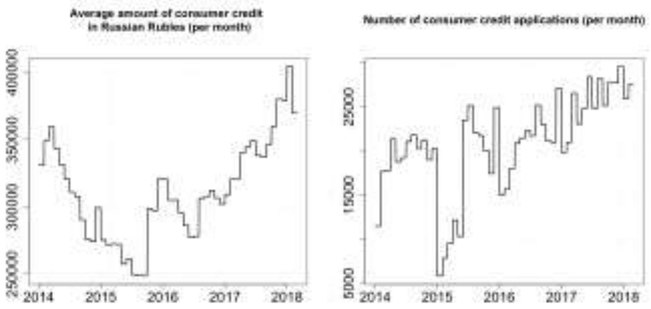

Fig. 3. Dynamics of consumer credits applications in Russia

A portion of borrowers with defaults is substantially less than a portion of responsible borrowers. In the sample, there were about 2% of irresponsible clients.

B. Data preprocessing

In the time series describing such macroeconomic parameters as the dollar rate and the cost of oil, there were missing values. For example, on weekends and holidays, the dollar rate is not updated, so the corresponding values were omitted in the sample used. Such omissions were filled with the last non-empty values.

Data that relate to the banking operations describe the average or total values for the month: on the 15th of each month there corresponds one entry in the sample. These values were duplicated for all other days of the month.

IV. Prediction Model

The prediction model is constructed on the basis of logistic regression since this approach is classical in the problem of credit scoring. The logistic regression is based on the following function:

$$f(z) = \frac{1}{1+e^{-z}},\ где$$

$$z = a_0 + a_1 \times x_1 + \ldots + a_n \times x_n\ ,$$

$$a_i - regression\ parameters,\ i = \overline{1,n},$$

$$x_j - values\ of\ independent\ variables,\ j = \overline{1,n}.$$

A. Selection of informative predictors

The selection of the most informative features was carried out using the *Information Value* metric:

$$IV = \sum_{i=1}^{n} (DistrGood_i - DistrBad_i) \cdot \ln(\frac{DistrGood_i}{DistrBad_i}),$$

$$feature - independent\ variable,$$

$$value_i - unique\ values\ of\ \ feature, i = \overline{1,n}\ ,$$

$$target - dependent\ variable,$$

$$DistrGood_i = P(\,feature = value_i \mid target = 1),$$

$$DistrBad_i = P(\,feature = value_i \mid target = 0).$$

The threshold values for the classification of the informativeness of the characteristics are presented in Table I.

TABLE I. Threshold Values for Information Value metric

Information Value	*Description*
$0 \le$ ***IV*** < 0.02	Non-informative
$0.02 \le$ ***IV*** < 0.1	Weakly informative
$0.1 \le$ ***IV*** < 0.3	Informative
$0.3 \le$ ***IV*** $<+\infty$	Strongly informative

To calculate the *Information Value*, it is necessary to carry out the interval discretisation of the continuous variables that prevail in the sample. The breakdown into intervals was made by 0.25, 0.5 and 0.75 quantiles. Informative features are presented in Table II.

TABLE II. Informative Features

Feature	*Information Value*
Loan rate	0.268
Age	0.247
Mean_mon_pay	0.073
Mean_mon_trans	0.063
Mean_trans	0.049
Mean_day_trans	0.038

Feature	*Information Value*
Mean_pay	0.036
Mean_day_pay	0.034
Sum of loan	0.033

B. *Model quality metric*

Since the positive and negative classes are unbalanced in sample, the area under receiver operating characteristic curve (AUC ROC) was used as the quality metric.

V. Results

The data was divided into train and test samples. To find the optimal train-test size ratio, the set of experiments was conducted. The value of relative test sample size varied from 0.05 to 0.95 with 0.1 step. For each of these values, the process model training and validation on randomly selected subsets was repeated 100 times. The average results are shown in Fig. 4.

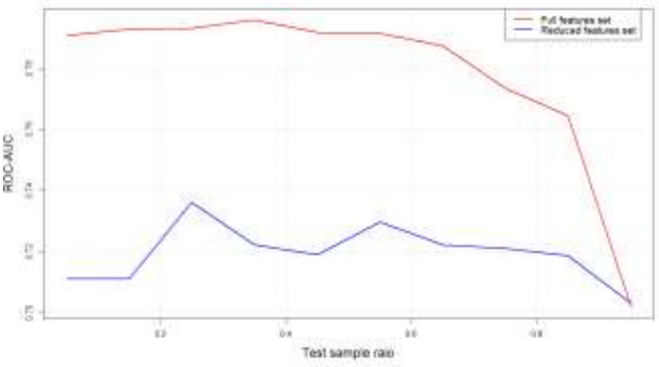

Fig. 4. The results of optimal train-test size ratio search. The 0.7 value was choosed as optimal

The set of numerical experiments with default forecasting macroeconomic variables was conducted with different values of time lag since the financial market is not immediately reflected the economy state. The lags values 3, 6 and 12 months were used. The experiments were conducted with additional macroeconomic variables (Table IV) and without it (Table III); with full variables set and reduced set with most informative features.

TABLE III. THE ERROR METRICS FOR MODEL WITHOUT MACROECONOMIC FEATURES INCLUDED

Features	**Information Value**
All	0.79
Informative	0.73

TABLE IV. THE CHARATERISTICS AND ERROR METRICS FOR MODEL WITH MACROECONOMIC FEATURES INCLUDED

Features	**Macroeconomic features lag[a], month.**						***ROC-AUC***
	usd	***brent***	***avg crd***	***avg aps***	***cnt crd***	***cnt dps***	
All	12	3	6	6	12	12	0.81
Informative	3	3	12	12	3	12	0.76

[a] The best results of lags obtained from full search are presented in table.

VI. Conclusions

The addition of macroeconomic parameters to scoring model variable set allows increasing the quality of default prediction. For the full variable set test sample ROC-AUC metric is increased by 0.019, and for the variable set selected by Information Value metric, the 0.021 quality growth was achieved.

However, the selection of informative features isn't an optimal solution of the variable set dimension reducing problem, because in the calculation of the Information Value metric for real variables the selected sampling method is important.

References

[1] Shen S.-W., Nguyen T.-D., Ojiako U. Modelling the predictive performance of credit scoring. Acta Commercii. 2013. 13(1). 12 p. https://doi.org/10.4102/ac.v13i1.189

[2] Bellotti T., Crook J. Credit scoring with macroeconomic variables using survival analysis. Journal of the Operational Research Society. 2009. 60(12). Pp. 1699–1707. https://doi.org/10.1057/jors.2008.130

[3] Im J. K., Apley D. W., Qi C., Shan X. A time-dependent proportional hazards survival model for credit risk analysis. Journal of the Operational Research Society. 2012. 63(3). Pp. 306–321. https://doi.org/10.1057/jors.2011.34

[4] Bellotti T., Crook J. Forecasting and stress testing credit card default using dynamic models. International Journal of Forecasting. 2013. 29(4). Pp. 563–574. https://doi.org/10.1016/j.ijforecast.2013.04.003

[5] Clark T.G., Bradburn M.J., Love S.B., Altman D.G. Survival Analysis Part I: Basic concepts and first analyses. British Journal of Cancer. 2003. 89(2). Pp. 232–238. https://doi.org/10.1038/sj.bjc.6601118

[6] Dirick L., Bellotti T., Claeskens G., Baesens B. Macro-Economic Factors in Credit Risk Calculations: Including Time-Varying Covariates in Mixture Cure Models. Journal of Business and Economic Statistics. 2017. Pp. 1–14. https://doi.org/10.1080/07350015.2016.1260471

[7] Simons D., Rolwes F. Macroeconomic default modelling and stress testing. 2008. Pp. 1–31.

[8] Dirick L., Claeskens G., Baesens B. An Akaike information criterion for multiple event mixture cure models. European Journal of Operational Research. 2015. 241(2). Pp. 449–457. https://doi.org/10.1016/j.ejor.2014.08.