UDC336.77

SCORING MODELLING OF COLLECTION FINANCIAL FLOWS

СКОРИНГОВОЕ МОДЕЛИРОВАНИЕ ФИНАНСОВЫХ ПОТОКОВ ОТ ВЗЫСКАНИЯ

T.I. Grigorchuk, Z.V. Maksimenko, L.F. Rozanova, G.R. Bikbulatova

FSBEI HPE "Ufa State Petroleum Technological University",
FSBEI HPE "Ufa State Aviation Technical University",
Ufa, the Russian Federation
Ministry of Finance of the Republic Bashkortostan, Ufa,
the Russian Federation

Григорчук Т.И., Максименко З.В., Розанова Л.Ф., Бикбулатова Г.Р.

ФГБОУ ВПО «Уфимский государственный нефтяной технический университет»,
ФГБОУ ВПО «Уфимский государственный авиационный технический университет», г. Уфа, Российская Федерация
Министерство финансов Республики Башкортостан, г. Уфа, Российская Федерация

e-mail: tgrigor@yandex.ru

Abstract. Nowadays, an expansive growth of past due debts in retail crediting is one of the most important problems of financial-credit system of Russia as well as of other countries. Collection activity may be optimized by implementing collection scoring systems. The paper regards one of the main problems of collection scoring – evaluating of collection financial flows and collection probability. The solution with the use of Tobit Type II model is offered, which allows determining the probability that debtor will make payments, dividing debtors to the groups of payers and defaulters and estimating

the probable sum of payment for each debtor. Factors that have the most influence on the probability of payment (surety amount, loan duration, residence region etc.) are detected. As a classification result, the category of debtors who will not return the debt is defined. For other debtors on the basis of linear model the probable monthly sum of payment is calculated. The results of this research may serve as a foundation for planning work with the debtor, will allow identifying hopeless debts and in accordance with that effectively allocating collection resources and reducing transaction costs.

Аннотация. Резкий рост просроченной задолженности по розничному кредитованию – в настоящее время одна из важнейших проблем финансово-кредитной системы, как России, так и западных стран. Оптимизировать деятельность по взысканию долгов возможно с помощью систем коллекторского скоринга. В статье рассматривается одна из основных задач коллекторского скоринга – оценка финансовых потоков от взыскания задолженности и вероятности ее взыскания. Предложено решение с применением модели Тобит-II, позволяющее определить вероятность того, что заемщик будет производить выплаты по долгу, разделить должников на группы плательщиков/не плательщиков и оценить вероятную сумму взыскания с каждого должника. Выявлены показатели, оказывающие наибольшее влияние на вероятность возврата долга (количество поручителей, срок кредита, регион проживания, пол должника и др.). В результате классификации выделена категория клиентов, которые не будут возвращать долг. Для остальных должников на основе линейной модели рассчитывается ежемесячная вероятная сумма взыскания. Результаты исследования могут служить обоснованием для планирования работы с должником, позволят выявить безнадежные долги, и в соответствии с этим эффективно перераспределить коллекторские ресурсы и снизить операционные издержки по взысканию.

Key words: collection, scoring, tobit II model, Heckit method, collection financial flows.

Ключевые слова: взыскание, скоринг, модель Тобит II, Хэкит метод, финансовые потоки от взыскания.

Introduction

Today, the appearance of past due debts is one of the most prevalent problems in banking sphere in Russia. The reason of not returning the loan by debtor may be a hard vital situation, for example, loss of income source or job, as well as an intentional evasion of making payments. One of the most effective collection instruments is collection scoring system. Collection scoring is a system for determining foreground working trends for returning past due debts based on debts' portfolio analysis. The collection scoring practical result will be development of optimal working strategy with every debtor [5].

The paper regards the solution of collection scoring primary problem – debtors' classification and evaluating their possible payments. As the result of classification the class of hopeless clients, who cannot or do not want to return the debt, is distinguished. This doesn't mean that collectors shouldn't have to work with them – in practice even for those kind of debts minimal possible level of payment is inherent. Scoring results allow choosing effective collection methods for each debtor to choose and make collector's agency actions more justified. This paper is a logical continuation of works [2,4].

1 Research description

The research goal is estimation of each debtor's payment amount and collection probability. The objects of research are individual borrowers - consumers of retail loans service (debtors of banks that are operating within the

territory of the Russian Federation), that have past due credit debts, which are in hand of the bank collection department.

Payment probability and amount depend on many factors, some of them are known and may be statically estimated. The following factors that characterize the loan given, the past due debt and the debtor as well - in total, 26 initial indicators, which include financial information and depersonalized social-demographic data, were chosen for the research. In addition, two target variables were given for research. The first variable takes the value of 1, if debtor pays his debt, and 0, if debtor doesn't pay. Second variable represents the average monthly payment of every debtor and is equal to 0 for the debtors who doesn't pay.

On basis of the given data it is required to develop an analytic model, which allows predicting the probability that the debtor will pay his debt and evaluating the most likely sum of payment.

The traditional approach to determination of the probable payment assume applying censored linear regression model or Tobit model (partly observed variable is the probable payment), where debtor's decision-making on making or not making a payment is determined by amount of payment itself. However considering another model, where decision-making process about the sum of payment is separated from making decision "to pay/not to pay", would be more correct. So for this purpose Tobit Type II model will be used in this research [6].

It is possible to obtain consistent and asymptotically effective parameters of Tobit Type II model by using maximum likelihood estimation method where corresponding likelihood function is maximized on all possible model parameters values. But more often this model is estimated by using the two-stage Heckman correction (Heckit method) [6]. So on the first stage of the research the binary classification model of getting into the class of debtors that are not making payments and on the second stage, the model, which allows determining the probable sum of payment for debtors that are making payments, will be developed.

2 Mathematical description of Tobit Type II model

Tobit Type II model has two latent variables, which answer the following models:

$$z^* = z^T c + u, \quad (1)$$

$$y^* = x^T b + \varepsilon, \quad (2)$$

where x^T, z^T – are the model parameters, b,c – vectors of the parameters' coefficients, ν u – error vectors.

1 stage. Binary classification model for determining that the debtor get into the group of debtors that are making payments. z^* defines "will/will not pay".

2 stage. The model for determining the probable sum of payment in conditions that the debtor is in the group of "making payments". If he doesn't pay, then y is not observed (is equal to zero), since no payments are made:

$$y = \begin{cases} y^*, & g = 1 \\ 0, & g = 0 \end{cases}, \quad (3)$$

$$g = \begin{cases} 1, & z^* > 0 \\ 0, & z^* \le 0 \end{cases}. \quad (4)$$

Assuming that the random errors latent variables models are correlated and related by ratio: $\varepsilon = \sigma_{\varepsilon u} u + v$, $\sigma_{\varepsilon u}$ – is the correlation coefficient between the values of ν and u,

$$\Rightarrow E(y \mid g = 1) = x^T b + \sigma_{\varepsilon u} E(u \mid u > -z^T c) = x^T b + \sigma_{\varepsilon u} \frac{f(z^T c)}{F(z^T c)} = x^T b + \sigma_{\varepsilon u} \lambda(z^T c) \quad (5)$$

where F и f – are distribution and probability density functions of a standard normal ($F(x) \equiv \Phi(x)$), or logistic ($F(x) \equiv \Lambda(x)$), or extreme distribution ($F(x) \equiv E(x)$), $\lambda(z^T c)$ – «Heckman's lambda » [1].

Thus, on the second stage, linear model for every debtor is estimated:

$$y_i = x_i^T b + \sigma_{\varepsilon u}\lambda_i + \eta_i \quad (6)$$

3 Estimation and analysis of the binary classification model

A set of initial variables for model estimation and their expected impact on the probability of payment are presented in Table 1. The total amount of cases is 3320 after correction. Traditionally, 70% of cases were used for training and 30 % – for testing the model. At the stage of preliminary analysis initial data the following variables “Date of default” and “Principal balance of debt, in rubles” have been excluded because of significant correlation with other variables. At the model estimation stage variables “R3”, “R8” and “State Tax in rubles” have been excluded as uniquely definable in the model (all observations have the same value of the target variable).

The results of modelling (approx. 50 models were estimated and analyzed) showed that logistic regression describes the model in the best way. The standard logistic distribution function:

$$F(u) = \Lambda(u) = \frac{e^u}{1+e^u},\ u=c_0+c_1z_1+c_2z_2+\ldots+c_nz, \quad (7)$$

where c_0 – is the model’s independent constant, c_j – model’s parameters, z_j – the value of j^{th} independent variable, $j = \overline{1,n}$.

Model parameters were estimated by maximum likelihood method in statistical package EViews. The final model is shown in Figure 1.

Table 1. Variables description

Variable description	Variable type	Designation	The expected impact on the payment probability
Target variable – debt payment probability	Target	TARGET	-
Date of loan issue (date of signing loan agreement)	Quantitative	DATE_OF_LOAN_ISSUE	Unclear
Loansize,in rubles	Quantitative	LOAN_SIZE	Negative
Amount of monthly payment, in rubles	Quantitative	MONTHLY_PAYMEN	Negative
Day of month when the monthly payment is made	Quantitative	MONTHLY_PAY_DATE	Unclear (probably insignificant)
Loan duration in months according to agreement	Quantitative	LOAN_DURATION	Positive
Loan expiry date according to agreement	Quantitative	LOAN_EXPIRY_DATE	Unclear
Annual interest rate according to agreement	Quantitative	INTEREST_RATE	Negative
Pledge availability	Binary	PLEDGE_TYPE	Positive
Surety amount according to agreement	Quantitative	SURETY_AMOUNT	Positive
Type of debt (loans for small business, express credit, consumer credit, credit card, car loan)	Categorial	CREDIT_MSB CREDIT_EXPRESS CREDIT_POTREB CREDIT_CARD CREDIT_AVTO	Unclear
Date of default	Quantitative	DATE_OF_DEFAULT	Positive
Number of days of delay from the date of default	Quantitative	DAYS_DELAY	Negative
Total amount of debt, in rubles	Quantitative	TOTAL_DEBT	Negative
Principal balance of debt, in rubles	Quantitative	PRINCIP_BALANCE	Unclear
Sum of penalties, in rubles	Quantitative	SUM_OF_PENALTIE	Negative
Sum of commission, in rubles	Quantitative	SUM_COMMISSION	Unclear (probably insignificant)
Sum of state tax, in rubles	Quantitative	S_TAX_AMOUNT	Unclear (probably insignificant)
Accrued percents, in rubles	Quantitative	PERCENTS	Unclear
Sum of other charges, in rubles	Quantitative	SUM_OF_OTHER	Unclear (probably insignificant)
Debtor's date of birth	Quantitative	DOB	Unclear
Debtor's gender (0- female, 1- male)	Binary	GENDER	Unclear
Debtor's residence region (Central, Southern, Northwestern, Far-Eastern, Siberian, Ural, Volga, North Caucasian)	Categorial	R1 R2 R3 R4 R5 R6 R7 R8	Unclear
Debtor's residence city (type of settlement in terms of population)	Categorial	G1 G2 G3 G4 G5 G6 G7	Unclear
Phone availability indicator for communication with the debtor	Binary	TEL_FLG	Positive
Last payment date	Quantitative	LAST_PAYMENT_DATE	Negative
Type of hopeless	Binary	TYPE_OF_HOPELESS	Negative

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: TARGET
Method: ML - Binary Logit (Quadratic hill climbing)
Date: 02/23/15 Time: 01:31
Sample (adjusted): 1 2320
Included observations: 2320 after adjustments
Convergence achieved after 12 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
DATE_OF_LOAN_ISSUE	0.013326	0.005942	2.242746	0.0249
LOAN_EXPIRY_DATE	-0.014586	0.005962	-2.446605	0.0144
LAST_PAYMENT_DATE	-0.000138	3.99E-05	-3.446795	0.0006
SURETY_AMOUNT	0.940263	0.210032	4.476755	0.0000
LOAN_DURATION	0.430274	0.177042	2.430356	0.0151
SUM_OF_PENALTIE	0.000132	5.05E-05	2.617602	0.0089
PERCENTS	-2.42E-05	4.75E-06	-5.099013	0.0000
MONTHLY_PAYMEN	0.000133	2.75E-05	4.828407	0.0000
TYPE_OF_HOPELESS	-1.429203	0.568776	-2.512770	0.0120
GENDER	-0.229621	0.100561	-2.283386	0.0224
DOB	-3.85E-05	1.33E-05	-2.898206	0.0038
G2	-0.446686	0.181294	-2.463878	0.0137
G4	0.350077	0.120850	2.896793	0.0038
G7	-0.326921	0.125274	-2.609651	0.0091
R2	-0.713590	0.344304	-2.072558	0.0382
R6	-1.468401	0.611864	-2.399883	0.0164
C	57.52592	7.081312	8.123625	0.0000

McFadden R-squared	0.177360	Mean dependent var	0.374569
S.D. dependent var	0.484116	S.E. of regression	0.430652
Akaike info criterion	1.102750	Sum squared resid	427.1172
Schwarz criterion	1.144878	Log likelihood	-1262.190
Hannan-Quinn criter.	1.118102	Restr. log likelihood	-1534.315
LR statistic	544.2509	Avg. log likelihood	-0.544047
Prob(LR statistic)	0.000000		

Obs with Dep=0	1451	Total obs	2320
Obs with Dep=1	869		

Figure 1. The results of binary classification model estimation

The hypothesis about the model's coefficients significance was tested using likelihood ratio test (LR) at a significance level of 0.05. The goodness of fit for the model has been verified by the Hosmer-Lemeshow test [2] at a significance level of 0.05. The test showed that model is adequate and may be used in analytical purposes (Figure 2).

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Goodness-of-Fit Evaluation for Binary Specification
Andrews and Hosmer-Lemeshow Tests
Grouping based upon predicted risk (randomize ties)

	Quantile of Risk Low	High	Dep=0 Actual	Expect	Dep=1 Actual	Expect	Total Obs	H-L Value
1	0.0224	0.1558	202	206.525	30	25.4747	232	0.90303
2	0.1563	0.2076	193	189.591	39	42.4089	232	0.33531
3	0.2076	0.2426	179	179.764	53	52.2356	232	0.01444
4	0.2430	0.2771	170	171.902	62	60.0985	232	0.08120
5	0.2771	0.3156	162	163.063	70	68.9370	232	0.02332
6	0.3158	0.3615	163	153.718	69	78.2816	232	1.66092
7	0.3615	0.4165	145	142.271	87	89.7288	232	0.13533
8	0.4167	0.5049	118	125.807	114	106.193	232	1.05845
9	0.5049	0.7055	97	94.3181	135	137.682	232	0.12850
10	0.7068	1.0000	22	24.0400	210	207.960	232	0.19312
		Total	1451	1451.00	869	869.000	2320	4.53362

H-L Statistic	4.5336	Prob. Chi-Sq(8)	0.8061
Andrews Statistic	4.5938	Prob. Chi-Sq(10)	0.9166

Figure 2. The Hosmer-Lemeshow test results

The results interpretation due to the model nonlinearity is based on the marginal effects (Table 2).

Table 2. Marginal effects for model's parameters

Parameter	Logit-model coefficient	Marginal effect
DATE_OF_LOAN_ISSUE	0.01333	0.2453%
DOB	-0.00004	-0.0010%
GENDER	-0.22962	-4.2260%
LAST_PAYMENT_DATE	-0.00014	-0.0030%
LOAN_DURATION	0.43027	7.9190%
LOAN_EXPIRY_DATE	-0.01459	-0.2680%
MONTHLY_PAYMENT	0.00013	0.0025%
PERCENTS	-0.00002	-0.0004%
SUM_OF_PENALTIES	0.00013	0.0024%
SURETY_AMOUNT	0.94026	17.3050%
TYPE_OF_HOPELESS	-1.4292	-26.3040%
R2 Southern Federal District	-0.71359	-13.1330%
R6 Ural Federal District	-1.4684	-25.0250%
G2 Million-plus cities	-0.44669	-8.2210%
G4 Cities with population 100-450 ths. people	0.35008	6.4430%
G7 Small settlements (village, country, island)	-0.32692	-6.0170%

The marginal coefficient for every parameter z_j, $j = 1,..,k$ is continuous and depends on the other factors and is determined by the following formula:

$$\frac{\partial P(z_j^* = 1)}{\partial z_j} = c_j \cdot F'(z_j^T c) = c_j \cdot f(z_j^T c), \quad (8)$$

where f – is probability density function.

For logit-model:

$$\frac{\partial P(z_j^* = 1)}{\partial z_j} = c_j \cdot \Lambda'(z_j^T c) = c_j \cdot \lambda(z_j^T c), \quad (9)$$

where $\lambda(u) = \frac{e^u}{(1+e^u)^2}$.

According to calculations, we can make the following conclusions:

1) The following indicators “Surety amount”, “Loan duration in months”, “Debtor’s residence city with population of 100-450 ths. people” have the most positive influence on the probability of debt payment. The surety presence increases the probability of making payments by 17.3%, what is quite predictable result. Every extra month of loan duration increases the probability by 7.91% – this can be explained both by availability of extra time for making payment and smaller monthly payment, that in general reduces the debtor’s household budget load.

For people living in cities with population of 100-450 ths. people probability of making payment increases by 6.44%. Probably, in this kind of cities small amount of banks is represented, therefore, fewer opportunities to take a loan are available, especially with bad credit history, so the payment discipline is higher. Also, cities of this scale tend to maintain enterprises in industry of extraction and processing of commercial minerals and are rich enough.

2) The following indicators significantly reduces the probability of paying the debt: “Type of hopeless” (debtor’s death) by 26.3%, “Debtor’s residence region” – Ural and Southern Federal Districts by 25.02% and 13.13% respectively. The probability of making payment also reduces by 8.22% if a debtor lives in million-plus city and by 6.01% if one lives in a small settlement (village, country, island). For big cities, this may be explained by more expensive level of life along with high salary and high indebtedness level. In case that debtor’s financial creditworthiness changes or some unforeseen circumstances emerge, a debtor turns out to be unable to pay the debt.

For small settlements low financial and legal literacy and also the difficulties of accessing the bank offices for making payment in time are typical, as a result, this entails low level of payment discipline.

3) The gender of a debtor also have a significant impact. It is revealed that men repay the debts not as good as women do. So if a debtor is male, the probability of making payments reduces by 4.22%.

Estimation of classification quality of model was based on ROC-curve analysis and its derivatives calculation: AUC index (Area Under Curve) and Gini coefficient [3]. ROC-curves plots and the values of model's quality indices are showed in Figure 3 and Table 3.

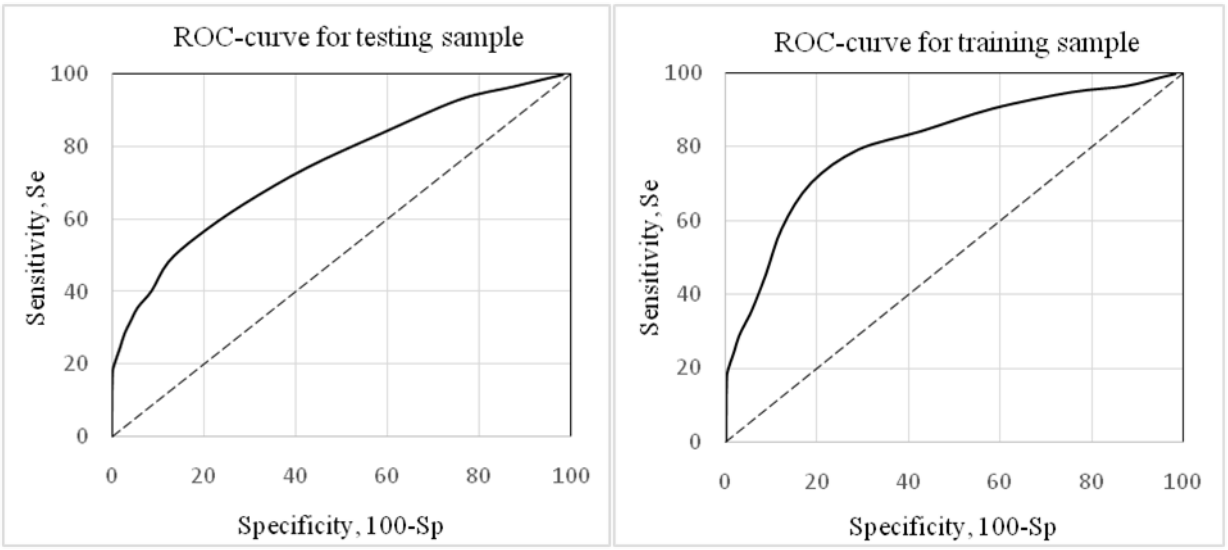

Figure 3. ROC-curve analysis

The closer the ROC-curve is to the upper left corner, the higher is the model's quality. If the curve coincides with diagonal the model is useless. The value of area under curve AUC = 1 corresponds to ideal and AUC = 0.5 corresponds to useless qualifier. Values from 0.7 to 0.9 shows good classification performance of the model. The Gini coefficient has the same interpretation - the closer the value is to 1, the better is the predictive ability of the model [3].

Table 3. The results of the model quality estimation

Coefficients	For training sample	For testing sample
AUC	0.8074	0.7461
Gini coefficient = $2 \cdot AUC\ (AUC\text{-}0.5)$	0.6148	0.4922

ROC-analysis also allows selecting an optimal division probability threshold of debtors to those who will and will not pay the debt. In table 4 there is a fragment of points array "Sensitivity - Specificity".

Table 4. Definition of division threshold

For training sample										
Division threshold	…	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	...
Se+Sp	…	132.24	141.82	150.52	***151.74***	145.95	136.71	130.14	127.81	….
|Se-Sp|	…	48.85	26.86	***8.13***	11.04	29.52	45.93	58.56	64.06	…
For testing sample										
Se+Sp	…	125.24	131.82	135.52	***136.74***	135.95	131.71	130.14	127.81	…
|Se-Sp|	…	41.85	16.86	***6.87***	26.04	39.52	50.93	58.56	64.06	…

For definition of optimal threshold, it is required to set a definition criterion, because for different tasks there are specific optimal strategies. The definition criterion for cut-off threshold may be [3]: maximum of sensitivity and specificity (Se+Sp), balance between sensitivity and specificity (|Se-Sp|) and other approaches. By results of analysis, the value 0.4 was chosen as the debtors' division threshold.

The results obtained with the use of this binary model further are applied for linear model estimation in order to determine the probable sum of payment for each debtor.

The research pursued has an applied character and practical importance because allows estimating of collection financial flows, optimizing work with every debtor (based on calculated probability of positive result), maximizing collection income from debts' portfolio by choosing the effective collection strategy for each debtor.

Conclusions

In this paper, the problem of estimating collection financial flows is considered. The solution with the use of Tobit Type II model is offered, which allows determining the probability that debtor will make payments, dividing debtors to the groups of payers and defaulters and estimating the probable sum of payment for each debtor. The results of this research may serve as a foundation for planning work with the debtor, will allow identifying hopeless debts and in accordance with that effectively allocating collection resources and reducing transaction costs.

Введение

Возникновение просроченной задолженности – на сегодняшний день одна из самых распространенных проблем в банковской сфере в России. Причинами невозврата кредита заемщиком может быть как сложившаяся тяжелая жизненная ситуация, например, потеря источника дохода, или работы, так и умышленное уклонение от уплаты кредита. Одним из наиболее эффективных инструментов взыскания долгов является система коллекторского скоринга. Коллекторский скоринг – это система определения приоритетных направлений работы по возврату просроченной задолженности на основе анализа проблемного портфеля, практическим результатом чего будет построение оптимальной стратегии работы с каждым должником [5].

Данная статья посвящена решению первоочередной задачи коллекторского скоринга – классификации должников и определению вероятной суммы взыскания с них. В результате классификации выделяется категория безнадежных клиентов, которые не могут или не хотят возвращать долг. Это не значит, что коллекторы не должны с ними работать – как показывает практика, даже для таких долгов характерен минимально возможный уровень взыскания. Результаты скоринга позволяют выбрать наилучший способ взыскания в отношении каждого должника и делают действия коллекторского подразделения/агентства более оправданными.

Данная статья является логическим продолжением работ [2,4].

1 Описание исследования

Целью исследования является оценка размера суммы возврата с каждого должника и вероятности взыскания задолженности. Объектом исследования выступают физические лица – потребители услуг розничного кредитования (заемщики банков, действующих на территории Российской Федерации), имеющие просроченную задолженность по кредитам, находящихся в работе коллекторского подразделения банка.

Вероятность и сумма возврата долга зависит от многих факторов, часть из которых известна и может быть оценена статистически. Для исследования были отобраны следующие факторы, характеризующие выданный кредит, просрочку по кредиту, а также личность заемщика – всего 26 исходных показателей, включающих финансовую информацию и обезличенные социально-демографические данные. Также для исследования предоставлены значения двух целевых переменных. Первая принимает значение 1, если заемщик производит выплаты по долгу, и 0, если заемщик не производит выплаты по долгу. Вторая представляет среднюю ежемесячную сумму взыскания с каждого должника, причем для неплательщиков она равна 0.

По заданному набору данных необходимо разработать аналитическую модель, которая для каждого займа позволит спрогнозировать вероятность того, что заемщик будет производить выплаты по долгу и определить вероятную сумму взыскания.

Традиционный метод нахождения вероятной суммы взыскания в данной задаче предполагает построение цензурированной линейной модели регрессии или Тобит-I (частично наблюдаемая переменная – вероятная сумма взыскания), где принятие должником решения платить или не платить определяется самой суммой долга, которую он намеревается выплатить. Однако правильнее бы было рассмотреть другую модель, в которой процесс принятия решения о сумме выплаты отделен от процесса принятия решения «платить/не платить». Для этого в данном исследовании будет применяться модель Тобит-II [6].

Получить состоятельные и асимптотически эффективные оценки параметров модели Тобит-II можно, используя метод максимального правдоподобия, при котором соответствующая функция правдоподобия максимизируется по всем возможным значениям параметров модели. Однако чаще такую модель оценивают, используя простую в вычислительном отношении двухшаговую процедуру Хекмана [6]. В данном исследовании на первом этапе будет строиться модель бинарного

выбора попадания в круг неплательщиков, а на втором этапе, модель, позволяющая определить для должников, делающих взносы, вероятную ежемесячную сумму погашения задолженности.

2 Математическое описание модели Тобит-II

В модели Тобит-II имеются две латентные переменные, удовлетворяющие следующим моделям:

$$z^* = z^T c + u, \quad (1)$$

$$y^* = x^T b + \varepsilon, \quad (2)$$

где x^T, z^T – факторы модели, b,c – векторы коэффициентов при факторах, ε, u – векторы ошибок.

1 этап. Модель бинарного выбора для определения «попадания» клиента в группу осуществляющих платежи, z^* определяет «будет/не будет платить».

2 этап. Модель для определения суммы взыскания при условии попадания должника в группу осуществляющих платежи. Если выбирается «не попадание», то y не наблюдается (равна нулю), так как платежей нет.

$$y = \begin{cases} y^*, & g = 1 \\ 0, & g = 0 \end{cases}, \quad (3)$$

$$g = \begin{cases} 1, & z^* > 0 \\ 0, & z^* \le 0 \end{cases}. \quad (4)$$

Предполагая, что случайные ошибки моделей латентных переменных коррелированы и связаны соотношением: $\varepsilon = \sigma_{\varepsilon u} u + v$, $\sigma_{\varepsilon u}$ – коэффициент корреляции между величинами ν и u,

$$\Rightarrow E(y|g=1) = x^T b + \sigma_{\varepsilon u} E(u|u > -z^T c) = x^T b + \sigma_{\varepsilon u} \frac{f(z^T c)}{F(z^T c)} = x^T b + \sigma_{\varepsilon u} \lambda(z^T c) \quad (5)$$

где F и f – соответственно функция распределения и плотность либо стандартного нормального ($F(x)\equiv\Phi(x)$), либо логистического ($F(x)\equiv\Lambda(x)$), либо экстремального распределения ($F(x)\equiv E(x)$), $\lambda(z^T c)$ - «лямбда Хекмана» [1].

Таким образом, на втором этапе для каждого должника оценивается линейная модель:

$$y_i = x_i^T b + \sigma_{\varepsilon u}\lambda_i + \eta_i \qquad (6)$$

3 Построение и анализ модели бинарного выбора

Набор переменных, использующийся для построения модели, и их ожидаемое влияние на вероятность погашения долга представлены в таблице 1. Общее количество исходных данных после корректировки составило 2320 наблюдений. Традиционно 70% наблюдений выборки применялись для обучения, 30% – для тестирования.

На этапе предварительного анализа исходных данных исключены переменные: Дата выхода в просрочку и Остаток основного долга в рублях, как значительно коррелирующие с другими переменными. На этапе построения моделей исключены переменные R3, R8 и Сумма начисленной госпошлины в рублях, как однозначно определяемые моделью (все наблюдения имеют одинаковое значение целевой переменной).

В результате моделирования (было построено и проанализировано около 50 моделей) выявлено, что наилучшим образом модель описывает логистическая регрессия. Функция стандартного логистического распределения:

$$F(u) = \Lambda(u) = \frac{e^u}{1+e^u},\ u=c_0+c_1z_1+c_2z_2+\ldots+c_nz, \qquad (7)$$

где c_0 – независимая константа модели, c_j – параметры модели, z_j – значение j-й независимой переменной, $j=\overline{1,n}$.

Таблица 1. Описание переменных

Описание переменной	Тип переменной	Обозначение	Ожидаемый эффект на вероятность возврата долга
Целевая переменная - вероятность возврата долга	Целевая	TARGET	-
Дата выдачи кредита (дата заключения договора на выдачу)	Количественная	DATE_OF_LOAN_ISSUE	Неясный
Размер кредита, выраженный в рублях	Количественная	LOAN_SIZE	Отрицательный
Размер ежемесячного платежа в рублях	Количественная	MONTHLY_PAYMEN	Отрицательный
Число месяца, в которое производится ежемесячная выплата	Количественная	MONTHLY_PAY_DATE	Неясный (возможно незначимый)
Срок кредита в месяцах по договору	Количественная	LOAN_DURATION	Положительный
Дата окончания кредита по договору	Количественная	LOAN_EXPIRY_DATE	Неясный
Годовая процентная ставка по договору	Количественная	INTEREST_RATE	Отрицательный
Тип залога	Бинарная	PLEDGE_TYPE	Положительный
Количество поручителей	Количественная	SURETY_AMOUNT	Положительный
Тип долга (кредиты МСБ, экспресс-кредит, потребительский, кредитная карта, автокредит)	Категориальная	CREDIT_MSB CREDIT_EXPRESS CREDIT_POTREB CREDIT_CARD CREDIT_AVTO	Неясный
Дата выхода в просрочку	Количественная	DATE_OF_DEFAULT	Положительный
Количество дней в просрочке с даты выхода в просрочку	Количественная	DAYS_DELAY	Отрицательный
Общая сумма долга в рублях	Количественная	TOTAL_DEBT	Отрицательный
Остаток основного долга в рублях	Количественная	PRINCIP_BALANCE	Неясный
Сумма начисленных штрафов в рублях	Количественная	SUM_OF_PENALTIE	Отрицательный
Сумма начисленной комиссии в рублях	Количественная	SUM_COMMISSION	Неясный (возможно незначимый)
Сумма начисленной госпошлины в рублях	Количественная	S_TAX_AMOUNT	Неясный (возможно незначимый)
Начисленные проценты в рублях	Количественная	PERCENTS	Неясный
Сумма прочих начислений в рублях	Количественная	SUM_OF_OTHER	Неясный (возможно незначимый)
Дата рождения заемщика	Количественная	DOB	Неясный
Пол заемщика (0- женщина, 1- мужчина)	Бинарная	GENDER	Неясный
Регион фактического места проживания должника (центральный, южный, северо-западный, дальневосточный, сибирский, уральский, приволжский, северо-кавказский)	Категориальная	R1 R2 R3 R4 R5 R6 R7 R8	Неясный
Город фактического проживания должника (тип населенного пункта по количеству населения)	Категориальная	G1 G2 G3 G4 G5 G6 G7	Неясный
Индикатор наличия указанного телефона для связи с должником	Бинарная	TEL_FLG	Положительный
Дата последнего платежа	Количественная	LAST_PAYMENT_DATE	Отрицательный
Тип безнадежности	Бинарная	TYPE_OF_HOPELESS	Отрицательный

Оценивание параметров модели производилось с помощью метода максимального правдоподобия в ППП EViews. Построенная модель приведена на рисунке 1.

Гипотеза о значимости коэффициентов модели проверялась с помощью теста отношения правдоподобия (LR) на уровне значимости 0,05. Адекватность подобранной модели реальному процессу была проверена с помощью теста Хосмера-Лемешоу [2] при уровне значимости 0,05. Тест показал, что модель адекватна, и может быть использована в аналитических целях (рисунок 2).

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: TARGET
Method: ML - Binary Logit (Quadratic hill climbing)
Date: 02/23/15 Time: 01:31
Sample (adjusted): 1 2320
Included observations: 2320 after adjustments
Convergence achieved after 12 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
DATE_OF_LOAN_ISSUE	0.013326	0.005942	2.242746	0.0249
LOAN_EXPIRY_DATE	-0.014586	0.005962	-2.446605	0.0144
LAST_PAYMENT_DATE	-0.000138	3.99E-05	-3.446795	0.0006
SURETY_AMOUNT	0.940263	0.210032	4.476755	0.0000
LOAN_DURATION	0.430274	0.177042	2.430356	0.0151
SUM_OF_PENALTIE	0.000132	5.05E-05	2.617602	0.0089
PERCENTS	-2.42E-05	4.75E-06	-5.099013	0.0000
MONTHLY_PAYMEN	0.000133	2.75E-05	4.828407	0.0000
TYPE_OF_HOPELESS	-1.429203	0.568776	-2.512770	0.0120
GENDER	-0.229621	0.100561	-2.283386	0.0224
DOB	-3.85E-05	1.33E-05	-2.898206	0.0038
G2	-0.446686	0.181294	-2.463878	0.0137
G4	0.350077	0.120850	2.896793	0.0038
G7	-0.326921	0.125274	-2.609651	0.0091
R2	-0.713590	0.344304	-2.072558	0.0382
R6	-1.468401	0.611864	-2.399883	0.0164
C	57.52592	7.081312	8.123625	0.0000

McFadden R-squared	0.177360	Mean dependent var	0.374569
S.D. dependent var	0.484116	S.E. of regression	0.430652
Akaike info criterion	1.102750	Sum squared resid	427.1172
Schwarz criterion	1.144878	Log likelihood	-1262.190
Hannan-Quinn criter.	1.118102	Restr. log likelihood	-1534.315
LR statistic	544.2509	Avg. log likelihood	-0.544047
Prob(LR statistic)	0.000000		

Obs with Dep=0	1451	Total obs	2320
Obs with Dep=1	869		

Рисунок 1. Результаты построения модели бинарного выбора

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Goodness-of-Fit Evaluation for Binary Specification
Andrews and Hosmer-Lemeshow Tests
Grouping based upon predicted risk (randomize ties)

	Quantile of Risk		Dep=0		Dep=1		Total	H-L
	Low	High	Actual	Expect	Actual	Expect	Obs	Value
1	0.0224	0.1558	202	206.525	30	25.4747	232	0.90303
2	0.1563	0.2076	193	189.591	39	42.4089	232	0.33531
3	0.2076	0.2426	179	179.764	53	52.2356	232	0.01444
4	0.2430	0.2771	170	171.902	62	60.0985	232	0.08120
5	0.2771	0.3156	162	163.063	70	68.9370	232	0.02332
6	0.3158	0.3615	163	153.718	69	78.2816	232	1.66092
7	0.3615	0.4165	145	142.271	87	89.7288	232	0.13533
8	0.4167	0.5049	118	125.807	114	106.193	232	1.05845
9	0.5049	0.7055	97	94.3181	135	137.682	232	0.12850
10	0.7068	1.0000	22	24.0400	210	207.960	232	0.19312
		Total	1451	1451.00	869	869.000	2320	4.53362

H-L Statistic	4.5336	Prob. Chi-Sq(8)	0.8061
Andrews Statistic	4.5938	Prob. Chi-Sq(10)	0.9166

Рисунок 2. Результаты теста Хосмера-Лемешоу

Интерпретация результатов моделирования в силу нелинейности модели проводится на основе предельных (маржинальных) эффектов (таблица 2). Предельный коэффициент каждого объясняющего фактора z_j, j=1,..,k является непрерывным и зависит от значения остальных факторов и определяется по формуле:

$$\frac{\partial P(z_j^* = 1)}{\partial z_j} = c_j \cdot F'(z_j{}^T c) = c_j \cdot f(z_j{}^T c), \qquad (8)$$

где f – плотность вероятности.

Для логит-модели:

$$\frac{\partial P(z_j^* = 1)}{\partial z_j} = c_j \cdot \Lambda'(z_j{}^T c) = c_j \cdot \lambda(z_j{}^T c), \qquad (9)$$

где $\lambda(u) = \frac{e^u}{(1+e^u)^2}$.

По выполненным расчетам можно сделать следующие выводы:

Наибольшее положительные влияние на вероятность возврата долга оказывают показатели: количество поручителей, срок кредита в месяцах по

договору, город фактического проживания должника с населением 100-450 тыс. чел. Наличие поручителя по кредиту увеличивает вероятность возврата долга на 17,3%, что является вполне предсказуемым результатом. Каждый дополнительный месяц срока кредита повышает вероятность возврата на 7,91% – это может объясняться как наличием дополнительного времени для возврата, так и меньшим среднемесячным платежом, что в целом снижает нагрузку на бюджет должника. Для проживающих в городах с численностью населения 100-450 тыс. чел. вероятность возврата повышается на 6,44%. Вероятно, в таких городах представлено малое количество банков, соответственно меньше возможностей взять кредит, особенно с плохой кредитной историей, поэтому дисциплина исполнения кредитных обязательств более высока. Также города такого масштаба, как правило, обслуживают предприятия сферы добычи и переработки полезных ископаемых и являются достаточно богатыми.

Существенно снижают вероятность возврата долга показатели: тип безнадежности (смерть заемщика) на 26,3%, регион фактического места проживания должника – уральский и южный на 25,02% и 13,13% соответственно. Также уменьшается вероятность возврата долга, если заемщик проживает в городе-миллионнике на 8,22% или в малых поселениях (село, деревня, остров) на 6,01%. Для больших городов это может быть связано с более дорогим уровнем жизни наряду с высокими зарплатами, с большей закредитованностью населения. В случае изменения финансовой кредитоспособности или каких-либо непредвиденных обстоятельств заемщик оказывается не в состоянии выплатить очередной долг. Для малых же поселений характерна низкая финансовая и юридическая грамотность, а также трудности доступа к отделениям банка для своевременной оплаты долга, что влечет низкий уровень кредитной дисциплины.

Также существенное влияние оказывает пол должника. Выявлено, что мужчины хуже возвращают долги. Так если заемщик мужского пола, вероятность возврата снижается на 4,22%.

Таблица 2. Маржинальные эффекты для параметров модели

Параметр	Коэффициент logit-модели	Маржинальный эффект
DATE_OF_LOAN_ISSUE	0,01333	0,2453%
DOB	-0,00004	-0,0010%
GENDER	-0,22962	-4,2260%
LAST_PAYMENT_DATE	-0,00014	-0,0030%
LOAN_DURATION	0,43027	7,9190%
LOAN_EXPIRY_DATE	-0,01459	-0,2680%
MONTHLY_PAYMENT	0,00013	0,0025%
PERCENTS	-0,00002	-0,0004%
SUM_OF_PENALTIES	0,00013	0,0024%
SURETY_AMOUNT	0,94026	17,3050%
TYPE_OF_HOPELESS	-1,4292	-26,3040%
R2 Южный ФО	-0,71359	-13,1330%
R6 Уральский ФО	-1,4684	-25,0250%
G2 Города-миллионники	-0,44669	-8,2210%
G4 Города с нас. 100-450 тыс. чел.	0,35008	6,4430%
G7 Малые поселения (село, деревня, остров)	-0,32692	-6,0170%

Оценка качества классификации проводилась на основе анализа ROC-кривых, а также расчете производных от нее: показателя AUC (площади под кривой) и коэффициента Джини [3]. Графики ROC-кривых приведены и значения показателей качества модели на рисунке 3 и в таблице 3.

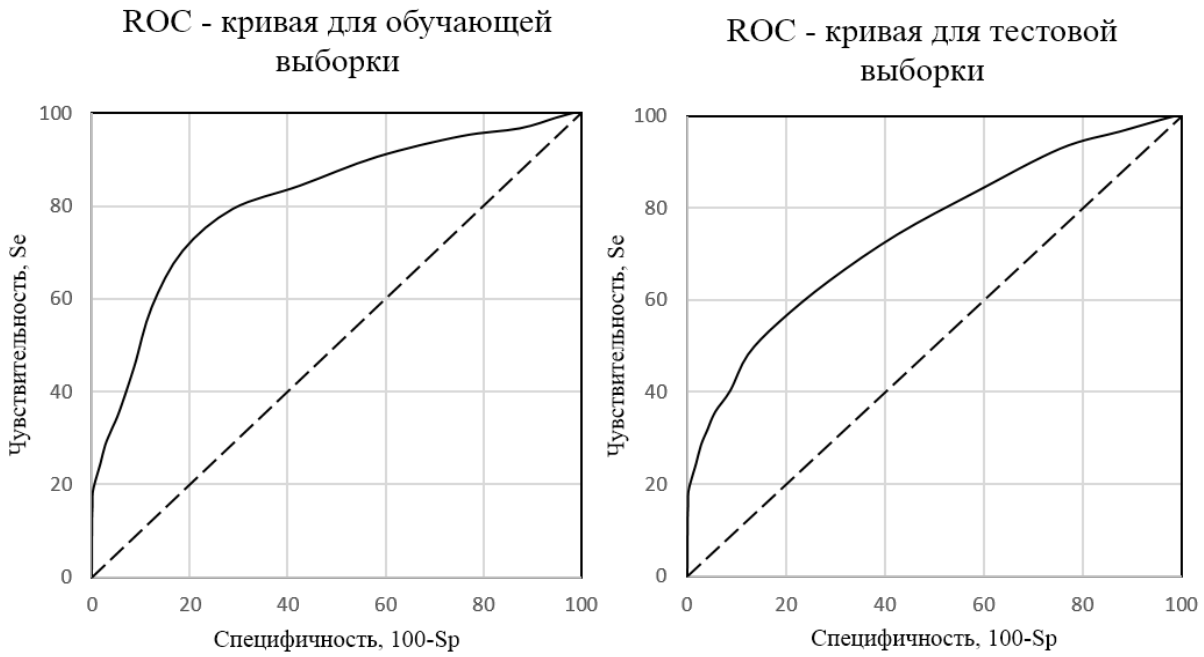

Рисунок 3. Анализ ROC-кривых

Таблица 3. Результаты оценки качества модели

Коэффициенты	Для обучающей выборки	Для тестовой выборки
AUC	0,8074	0,7461
Коэф. Джинни= 2·*AUC* (*AUC*-0,5)	0,6148	0,4922

Чем ближе график ROC-кривой проходит к левому верхнему углу, тем качественнее модель. Если кривая совпадает с диагональю, то модель бесполезна. Значение площади под кривой *AUC*=1 соответствует идеальному, а *AUC*=0,5 – бесполезному классификатору. Значения от 0,7 до 0,9 показывают хорошую классификационную способность модели. Также интерпретируется и коэффициент Джинни – чем ближе его значение к 1, тем выше прогностическая способность модели [3].

ROC-анализ также позволяет выбрать оптимальный порог вероятности разделения должников на тех, кто будет и не будет возвращать долг. В таблице 4 приведен фрагмент массива точек «Чувствительность-Специфичность».

Таблица 4. Определение порога отсечения

Для обучающей выборки										
Порог отсечения	…	*0,25*	*0,3*	*0,35*	*0,4*	*0,45*	*0,5*	*0,55*	*0,6*	...
Se+Sp	…	132,24	141,82	150,52	*151,74*	145,95	136,71	130,14	127,81	….
|Se-Sp|	…	48,85	26,86	*8,13*	11,04	29,52	45,93	58,56	64,06	…
Для тестовой выборки										
Se+Sp	…	125,24	131,82	135,52	*136,74*	135,95	131,71	130,14	127,81	…
|Se-Sp|	…	41,85	16,86	*6,87*	26,04	39,52	50,93	58,56	64,06	…

Для определения оптимального порога необходимо задать критерий его определения, т.к. в разных задачах существует своя оптимальная стратегия. Критериями выбора порога отсечения могут выступать [3]: максимум чувствительности и специфичности (Se+Sp), баланс между чувствительностью и специфичностью (|Se-Sp|) и другие подходы. По результатам анализа в качестве порога разделения должников выбрано значение 0,4.

Полученные на базе данной бинарной модели результаты в дальнейшем применяются для оценки линейной модели для определения суммы взыскания с каждого должника.

Проводимое исследование носит прикладной характер и имеет практическую ценность, т.к. позволяет провести оценку финансовых потоков от взыскания, оптимизацию работы с должником (исходя из рассчитываемой вероятности положительного исхода), максимизировать прибыль от взысканий по портфелю за счет правильного выбора стратегии работы с отдельными должниками.

Заключение

В статье рассмотрена задача оценки финансовых потоков от взыскания просроченной задолженности. Предложено решение с применением модели Тобит-II, позволяющее определить вероятность того, что заемщик

будет производить выплаты по долгу, разделить должников на группы плательщиков/неплательщиков и оценить вероятную сумму взыскания с каждого должника. Результаты исследования могут служить обоснованием для планирования работы с должником, позволят выявить безнадежные долги и, в соответствии с этим, эффективно перераспределить коллекторские ресурсы и снизить операционные издержки по взысканию.

References

1 James J. Heckman. Sample selection bias as a specification error // Applied Econometrics. 2013. №31(3). pp. 129-137.

2 Optimization of work with past due debts for collection department/agency management / Lackman I.A. [at al.] // Eurasian Law Journal. 2015. №4(83). pp. 139-142. [in Russian].

3 Logistic regression and ROC-analysis – mathematical tools. URL: http://www.basegroup.ru/library/ analysis/regression/ logistic. [in Russian].

4 Maksimenko Z.V., Lackman I.A., Rozanova L.F. Information support strategy formation debt collection based on the collection of scoring // Information Technologies and Systems (ITIS' 2015):Proc. of 4th scientific conference. Chelyabinsk: CSU, 2015. pp. 139-140. [in Russian].

5 Mikhmel P.S., Dovgii N.V. Collection scoring // Bankovskoe Delo. 2013. №2. pp. 65-71. [in Russian].

6 Nosko V.P. Econometrics for beginners (additional paragraphs). M: IEPP, 2005.379 p. [in Russian].

Список используемых источников

1 Джеймс Дж. Хекман. Смещение селективной выборки как ошибка спецификации // Прикладная эконометрика. 2013. №31(3). С. 129-137.

2 Оптимизация работ по взысканию проблемной задолженности для управления деятельностью коллекторского подразделения/агентства / Лакман И.А. [и др.] // Евразийский юридический журнал. 2015. № 4 (83). С. 139-142.

3 Логистическая регрессия и ROC-анализ - математический аппарат. URL: http://www.basegroup.ru/library/analysis/regression/logistic/

4 Максименко З.В., Лакман И.А., Розанова Л.Ф. Информационная поддержка формирования стратегии взыскания просроченной задолженности на основе коллекторского скоринга // Информационные технологии и системы: Тр. /Четвертой Междунар. науч. конф./ Отв. ред. Ю. С. Попков, А.В. Мельников. Челябинск, 2015. С. 139-140.

5 Михмель П.С., Довгий Н.В. Коллекторский скоринг // Банковское дело. 2013. №2. С.65-71.

6 Носко В.П. Эконометрика для начинающих (Дополнительные главы). М.: ИЭПП, 2005. 379 с.

About the authors

Сведения об авторах

T.I. Grigorchuk, Candidate Engineering Sciences, Associate Professor of FSBEI HPE “Ufa State Petroleum Technological University”, Ufa, the Russian Federation

Григорчук Т.И., канд. техн. наук, доцент ФГБОУ ВПО УГНТУ, г. Уфа, Российская Федерация

e-mail: tgrigor@yandex.ru

Z.V. Maksimenko, Candidate of Engineering Sciences, Associate Professor of FSBEI HPE “Ufa State Aviation Technical University”, Ufa, the Russian Federation

Максименко З.В., канд. техн. наук, доцент ФГБОУ ВПО УГАТУ, г. Уфа, Российская Федерация

e-mail: maximenkozv@gmail.com

L.F. Rozanova, Candidate of Engineering Sciences, Associate Professor of FSBEI HPE “Ufa State Aviation Technical University”, Ufa, the Russian Federation

Розанова Л.Ф., канд. техн. наук, доцент ФГБОУ ВПО УГАТУ, г. Уфа, Российская Федерация

e-mail: rozanova_lara@mail.ru

G.R. Bikbulatova, Candidate of Engineering Sciences, Chief Specialist of Ministry of Finance of the Republic Bashkortostan, Ufa, the Russian Federation

Бикбулатова Г.Р., канд. техн. наук, главный специалист Министерства финансов Республики Башкортостан, г. Уфа, Российская Федерация

e-mail: guzel-sabiryanova@yandex.ru