Alternative Ways for Loss-Given-Default Estimation in Retail Banking

Alexey Masyutin(✉)

National Research University Higher School of Economics, Moscow, Russia
alexey.masyutin@gmail.com

Abstract. The cornerstone of retail banking risk management is the estimation of the expected losses when granting a loan to the borrower. The expected losses are determined by three parameters. The first is the probability of default (PD) of the borrower. The methods of PD estimation were studied in detail by previous authors, and the most common method is credit scorecard development. The second parameter is exposure at default (EAD). Except for revolving loans, it is known in advance, it is the current balance (principal amount plus accrued interests) of the loan. Finally, there is a third parameter that defines the expected losses. This is the so-called loss given default (LGD) which is in effect the share of EAD, which is irretrievably lost in the event of default. This paper discusses several econometric techniques which allow one to obtain estimates of the LGD parameter.

Keywords: LGD · Survival analysis · Kaplan-Meier estimator · Cox regression · Beta-regression · Recovery rate

1 Introduction

Banks and financial institutions are taking credit risks in order to make profit. Despite the rigorous risk management strategies banks face considerable portion of defaulting borrowers. When the arrears occur it is necessary to carry out activities aimed to recover the significant part of the defaulted loan.

In order to build a system of effective bad debt collection banks have to discriminate among the debtors and detect the groups of those who tend to return the bigger part of the overdue amount and those who will likely give no repay at all. The problem of finding such groups is reduced to identification of significant factors influencing recoveries. Having determined which kind of borrowers pose the greatest risk of no recovery, and (what is more difficult) having obtained quantitative estimates of this risk, banking analysts are able to advise on the collection department resources allocation. If expected recovery rate is too low then bad debts can be sold to third parties. The problem is particularly relevant at the stage of the so-called late recovery (hard collection), when the bank is not limited to auto-dialing systems and instant messaging, but is forced to assign call-centers staff for direct communication with customers,

D.I. Ignatov et al. (Eds.): AIST 2014, CCIS 436, pp. 152–162, 2014.
DOI: 10.1007/978-3-319-12580-0_15

employees for personal contact with the client. In addition, the bank always faces a dilemma: to continue the collection process internally or to give it to outsourcing. If the bank assesses the likelihood of repayment of arrears for some pool of customers as relatively high, then it is reasonable not to conclude agency agreements with collectors, because the commission for services of debt collection can reach more than 30 %. The research of overdue debts can be carried out in two directions: (1) assess the likelihood of the transition from delinquent state to healthy state, (2) evaluate the expected amount of income as a percentage of arrears (recovery rate).

2 Methods Discussion

Traditional and, apparently, the most common way to solve the first problem is to build a scorecard (collection scoring). Like in the case of PD estimation (application, behavioral scoring) the scorecard assesses the likelihood of loan full repayment. The only difference here is target variable definition. The techniques remain the same: those are relevant variables selection, variables transformation (e.g. WOE transformation) maximizing the specified criteria, and, finally, binary logit model calibration. Two important and significant limitations of the method should be mentioned. First, in order to build and validate the model there must be the evidence that an event has occurred or not (in this case, the return from the delinquent status). Therefore analyst must have a dataset containing a sample of closed loans (full repayment took place) and a sample of written-off loans (loan did not return to a healthy state). Meanwhile the active loans with current delinquent status fall out of the sample, since the event of recovery (as well as no-recovery) are not yet determined. This case is highly undesirable, as soon as excluded observations also carry information. Such situation applies to the known problem of *right-censored* data. In other words, the logit model is not designed to work with censored data. This limitation is less painful for large banks, because the history of the portfolio has more than enough observations. But for small and medium-sized banks which find themselves in process of growing its loan portfolio, when the maximum age of the loans reaches only 12–18 months, reduction in the sample size is impossible. Processing the censored data in this case is an inevitable necessity.

Second, collection-scoring answers the question, *how likely* this loan is going to return from the state of delinquency. However it does not aim to answer *when* it is going to recover. Therefore, time aspect remains out of focus. Indeed, scoring can answer the question, whether there will be a return within a specified period of time (usually 12 months), but at what point the transition takes place is unknown. So, collection-scoring does not assess the density distribution function of the moments of recoveries. In this paper, we use survival models (or time-to-event models) as a tool to analyze the repayment of arrears. This branch of statistics was developed in the second half of the XX century. The milestone works are the work of Edward Kaplan and Paul Meier (1958, [1]), Weibull Vallodi (1961, [2]), and David Cox (1972, [3]). Survival analysis is used primarily in the

medical and sociological research. For example, one can verify the effect of certain therapy when patients are divided into those who receive treatment and those who receive placebo. In sociology this tool is used to investigate what factors influence the duration of staying unemployed. The duration of life is understood as time spent in the unemployed state, and "death" is defined as getting the job [4]. However, in credit risk management survival analysis techniques are also applicable. First of all, this is alternative way to estimate PD, when the lifetime is treated as duration of a loan without overdue (or without overdue more than X days), while the "death" refers to falling into default (Lyn C. Thomas et al. [5]). We can find other applications of survival analysis that are not connected with credit insolvency. In Stepanova et al. [8] they consider the fact of the loan early repayment. This phenomena is less painful for banks in comparison with default but still not desirable. Our approach is close to Jiri Witzany et al. [9]. However, we do not equate the LGD parameter with the survival function of the loan. We would rather consider survival function as a way to assess the proportion of the defaulted loans which are not fully recovered up to the point in time. LGD modeling is also used within small and medium enterprise (SME) segment. The relevant work is Sudheer Chava et al. [10] where LGD is estimated using survival analysis techniques. In contrast, we focus on retail segment loans, and the set of predictive variables is primarily socio-demographic characteristics of the borrower.

3 Concepts of Survival Analysis Models

Suppose we have a sample of n objects, each is defined as a random variable, i.e. lifespan: $T_1, T_2, ..., T_n$. The object is called right-censored, if its real lifespan is yet not known. For example, the object has been living for 2 years, and the researcher does not know how long it will live more. So, the observed lifespan is less of equal to the real lifespan. Thus, due to the right censoring, the researcher does not observe real lifespans $T_1, T_2, ..., T_n$, instead he observes the minimum of the real life and observed $(X_i, D_i), i = 1, 2, ..., n$:

$$X_i = min(T_i, C_i) \tag{1}$$

where C_i is a real lifespan

$$D_i = \begin{cases} 0 & \text{the subject is censored, i.e. } C_i < T_i \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

But we are interested in characteristics of the initial series distribution. Namely, $f(t)$ is a lifetime density function, $F(t)$ is a lifetime distribution function, $S(t) = 1 - F(t)$ is a survival function, $h(t)$ is a hazard function. Survival function answers the question about the probability of lifetime greater than t. i.e.:

$$S(t) = Pr(T \geq t) \tag{3}$$

Hazard function (force of mortality, or the failure rate in engineering) is by definition:

$$h(t) = \lim_{dt \to 0} \frac{Pr(t \leq T < t + dt | T \geq t)}{dt} \tag{4}$$

where T is duration of life, in our case, this is time spent in the delinquency state. It is assumed that the fact of censoring does not depend on the lifetime and performance of any subject. This is a realistic assumption, since censored subjects are in the banking collection are active loans. Under this assumption one can show how lifetime functions are related:

$$h(t) = \frac{f(t)}{S(t)} \tag{5}$$

In its turn, the survival function can be expressed in terms of the hazard function:

$$S(t) = \exp^{-\int_0^t h(s)ds} \tag{6}$$

Survival analysis models are aimed to obtain estimates of the last two functions.

4 Description of Data and Analysis

In our case, the object of the study is the delinquency higher than 30 days. The loan is considered to be in a state of delinquency until the borrower repays (a) the overdue principal amount, (b) overdue interests, (c) penalty for each day of delay in installments. Under the lifetime of the subject we mean duration of the delinquent status. Under the "death" of the subject we will understand full recovery, i.e. loan returning from delinquent to healthy status. It is still a question about the interdependency between consequent delinquencies within a single loan. For example, it is possible that if the loan falls into arrears for the third time, then it is highly unlikely to get any recovery. In this paper, this dependence is not modeled. Thus, if the loan had several cases of delinquency, their durations were considered to be independent random variables. In this paper we consider the auto-loan portfolio of one of the top-20 Russian banks. The loans were issued within 2010–2012 years. Due to the non-disclosure agreement the details are confidential. The sample consists of 1370 cases of delinquency. So, written-off, closed and still active loans are all presented within the dataset. If the delinquent loan is active then the observation is right-censored. Analysis of repayment of arrears was conducted within the software package STATA[1]. Despite the SAS high prevalence in the banking sector, the choice was made in favor of the package STATA due to availability of datamarts (no need to manipulate the data), as well as easier syntax of the STATA package[2].

[1] Seminar on time-to-event analysis is available at http://www.ats.ucla.edu/stat/stata/seminars/stata_survival/.

[2] The similar seminar but within SAS framework can be found at http://www.ats.ucla.edu/stat/sas/seminars/sas_survival/.

The variables that were significantly affecting the repayment of arrears, are given in the appendix [Table 2].

The first thing to do before going on to multivariate regressions is to check the relationship between the delinquency duration and variables separately. Non-parametric survival function estimation (Kaplan-Meier estimator) suits well for this task. The evaluation does not require any assumptions about the distribution function of the delinquency duration. The estimator assesses the probability that the duration of the installment delay exceeds t days:

$$\widehat{S(t)} = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i} \tag{7}$$

where t_i is a moment in time at which the full recoveries were observed, n_i - number of loans, which preserve the delinquent status at time t_i less censored observations at this point, and d_i - the number of loans, returned into a healthy state at time t_i. Kaplan-Meier estimates can be constructed within different groups, by maturity, loan amount, and other characteristics of the borrower. Survival functions graphs can be found in appendix [Figs. 2, 3 and 4]:

Long flat tail of survival function shows that after being one year in the state of delinquency the probability of returning to a healthy state is almost zero. Further, note that the probability of exit from the delinquency state is 5–6% lower for men than for women. The same effect is observed for unmarried versus married borrowers. Parallelism of survival curves show the proportionality of risk that allows one to build multiple regression model in which the effects will be evaluated simultaneously. For this we use the semiparametric Cox model. Cox regression estimates the hazard function, suggesting that it depends on factors as follows:

$$h(t|X) = h_0(t) \exp^{X\beta} \tag{8}$$

where $h_0(t)$ is an arbitrary function (baseline hazard), X are factors and β is the vector of coefficients. The model is semi-parametric, since there is a function that is not a priori given, but on the other hand still there are parameters to be estimated. After evaluation of the extended model some insignificant variables were excluded, and eventually the model took the following form (Fig. 1):

The coefficients in column *Haz.Ratio* show how many times the hazard function will increase when the regressor in its turn increases by one. Since many variables (such as education, type of loan, sex) are discrete, the coefficients indicate how the hazard functions differ between the groups. So, if the loan is of the third type, at any point in time "risk" of exiting the delinquency state is up to 2.53 times higher compared to the first type of loan. If the borrower has more than one child, the "risk" to leave the state of delinquency increases in 1.14 times. The difference between the borrower with higher education and complete secondary is 1.3 times. Alternative way of lifetime estimation is presented by the parametric methods. The most popular is the use of the lognormal distribution function, as well as the Weibull distribution. Parameters distributions are estimated within likelihood maximization method. For brevity, we give a report on a model constructed for the lognormal distribution:

```
. stcox age sex_enc i.education_enc credit_sum credit_period num_depend i.type

        failure _d:  vyshel
  analysis time _t:  prosrochka_1

Iteration 0:   log likelihood = -4057.4778
Iteration 1:   log likelihood = -4024.0101
Iteration 2:   log likelihood = -4023.1607
Iteration 3:   log likelihood = -4023.1553
Iteration 4:   log likelihood = -4023.1553
Refining estimates:
Iteration 0:   log likelihood = -4023.1553

Cox regression -- Breslow method for ties

No. of subjects =         1237                     Number of obs   =      1237
No. of failures =          615
Time at risk    =       223963
                                                   LR chi2(11)     =     68.64
Log likelihood  =   -4023.1553                     Prob > chi2     =    0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.9879588	.004548	-2.63	0.008	.9790849	.9969132
sex_enc	.768267	.0681181	-2.97	0.003	.6457148	.9140788
education_~c						
2	.97061	.1416051	-0.20	0.838	.7292232	1.2919
3	1.3043	.1245625	2.78	0.005	1.081649	1.572781
4	1.106318	.1964495	0.57	0.569	.7811452	1.566853
5	2.839574	1.293924	2.29	0.022	1.16247	6.936249
credit_sum	.9999993	1.65e-07	-3.99	0.000	.999999	.9999997
credit_per~d	1.00508	.0027373	1.86	0.063	.9997293	1.010459
num_depend	1.138421	.0523611	2.82	0.005	1.040285	1.245815
type						
2	.4741113	.0854275	-4.14	0.000	.3330481	.6749221
3	.3952999	.0819458	-4.48	0.000	.2633131	.5934457

Fig. 1. Cox regression model output in STATA

The distribution of the delinquency duration is defined as follows:

$$\ln T \sim N(X\beta, \sigma) \tag{9}$$

Coefficient sign shows the direction of delinquency duration change, due to the change in corresponding factor per unit. Coefficients themselves can be used to construct the probability of loan remaining in delinquency state in the next predetermined time interval:

$$Pr(T > t_0 + \delta t | T > t_0) = \frac{N(\frac{X\beta}{\sigma} - \frac{1}{\sigma}\ln t_0 + \delta t)}{N(\frac{X\beta}{\sigma} - \frac{1}{\sigma}\ln t_0)} \tag{10}$$

t_0 is time (in days) since loan has fallen into delinquency. Using this formula, for instance, one can calculate the probability of remaining in a delinquent state within the next 30 days for specific borrowers.

5 Recovery Rate Estimation

To solve the second problem, set at the beginning of the work, namely recovery rate estimation, we will use the beta-regression (Silvia et al. [6]). However, there is a sufficient restriction of the beta regression: censored data cannot be

processed. The density function of the random variable having a beta distribution is defined as follows:

$$f(z,\beta,\gamma) = \frac{\Gamma(\beta+\gamma)}{\Gamma(\beta)+\Gamma(\gamma)} z^{\beta-1}(1-z)^{\gamma-1} \tag{11}$$

And in case of regressors:

$$f(z,\beta,\gamma,X) = \frac{\Gamma(X\beta+X\gamma)}{\Gamma(X\beta)+\Gamma(X\gamma)} z^{X\beta-1}(1-z)^{X\gamma-1} \tag{12}$$

In this formula β and γ are now coefficient vectors. The distribution is used in the analysis of a continuous variable strictly limited by 0 and 1. Since the problem of estimating the probability of full recovery from delinquent to healthy state was considered in the first part, we will focus on those loans, which showed only partial recovery. Its distribution is bimodal around zero and unity: that means that in most cases the debtor either completely pays back, or nothing at all (Table 1).

Table 1. Beta regression output in STATA

betafit rr, alphavar(perc_sum credit_period) betavar(age credit_period type1)						
Iteration 0: log likelihood = 111.88367						
Iteration 1: log likelihood = 159.03024						
Iteration 2: log likelihood = 175.59134						
Iteration 3: log likelihood = 177.08649						
Iteration 4: log likelihood = 177.1083						
Iteration 5: log likelihood = 177.10831						
ML fit of beta (alpha,beta)			Number of obs = 531			
			Wald chi2(2) = 19.38			
Log likelihood = 177.10831			Prob >chi2 = 0.0001			
rr	**Coef.**	**Std. Err.**	**z**	**P>|z|**	**[95% Conf.**	**Interval]**
alpha						
perc_sum	.4156041	.131536	3.16	0.002	.1577983	.6734098
credit_period	-.0122402	.0028874	-4.24	0.000	-.0178994	-.006581
_cons	1.013151	.1503694	6.74	0.000	.7184326	1.30787
beta						
age	-.018639	.0062506	-2.98	0.003	-.0308899	-.0063881
perc_sum	.9988718	.2704875	3.69	0.000	.468726	1.529018
credit_period	.0120325	.0054593	2.20	0.028	.0013324	.0227326
type1	-.0161823	.1245794	-0.13	0.897	-.2603534	.2279889
_cons	.9595355	.3603647	2.66	0.008	.2532337	1.665837

The report shows that in our sample, there are 531 observations with recovery rate strictly within the limits of zero and one. If full and zero recovery cases are added then the sample makes a total of 900 observations. It is less than 1370

delinquency cases in the original sample because beta regression does not handle censored data. When estimating the parameters of these observations have to be ruled out. Only significant variables were left in the regression model. Coefficients are used to build up the distribution of recovery rate (rr), conditional on the characteristics of the loan:

$$Pr(rr \leq z|X) = \int_0^z f(s, \hat{\beta}, \hat{\gamma}, X)ds \tag{13}$$

From this the expected recovery rate for the loan can be determined as:

$$E(rr|X) = \int_0^1 sf(s, \hat{\beta}, \hat{\gamma}, X)ds = \frac{X\hat{\beta}}{X(\hat{\beta} + \hat{\gamma})} \tag{14}$$

For example, one can analyze the differences in the context of the loan types. Indeed the first type loans pose a greater risk of low recovery, rather than the third [Table 3]:

6 Conclusion

We provided the analysis of bad debts in terms of the temporal structure of recoveries. The probabilities of full repayment can be assessed within different groups of borrowers with the help of non-parametric methods such as Kaplan-Meier estimators. They not only help to avoid any assumptions about the density functions but also visualize the result in a comprehensive way. More detailed interconnections between repayments and borrowers characteristics are revealed by Cox proportional hazards model. Finally, the loans which showed only partial recovery were analyzed with beta regression. The described techniques allow one to discriminate bad debtors into groups with high versus low recovery rates. This provides the instrument for efficient debt collection process, when bank focuses primarily on the defaulted borrowers who are likely to pay back. Meanwhile, the portfolio of bad loans with low expected recovery rate can be sold to third parties and collection agencies. In our further research we would like to analyze factors that determine recoveries using concept-based learning [7].

Acknowledgments. Author would like to express his gratitude to Ivan Medvedev, Head of Retail Risks at RN Bank (former RCI Banque representative office) for being a guide in the world of banking risk management.

Appendix

Table 2. Variables influencing the recovery rate

Variable	Block	Description	Possible values	Transcript
Age	Borrower	Age	N	In full years
sex_enc		Sex	1	Male
			0	Female
marital_status_enc		Marital Status	1	Single/Married
			0	Other
education_enc		Education	1 (B)	Complete secondary education
			2 (C)	Incomplete higher
			3 (D)	Higher
			4 (E)	Two or more higher
			5 (F)	Academic degree
num_depend		Number of dependents	N	
is_estate_enc		Availability of real estate owned	1	Yes
			0	No
company_type_enc	The employer	Type of company	0	Without state participation
			1	With public participation
company_age		Age of	N	
company_count_staff		Number of employees	N	
credit_sum	Loan	Amount of credit	N	In the rub
credit_period		Term	N	In months
perc_sum		Principal debt / original loan amount	0 to 1	
Type		Specific loan classification adopted in the bank	1	Not associated with the loan price parameters
			2	
			3	

Table 3. Differences in recovery rate between two types of loan

	Client X	Client Y
age	25	25
credit_period	24	24
perc_sum	20 %	20 %
Type	1	3
Expected rr	**22.6 %**	**44.4 %**

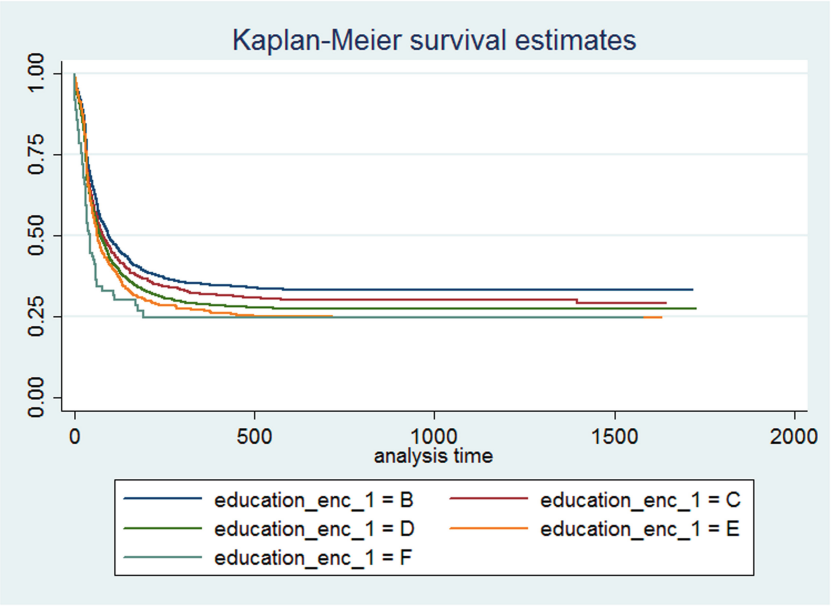

Fig. 2. Survival function within groups by education

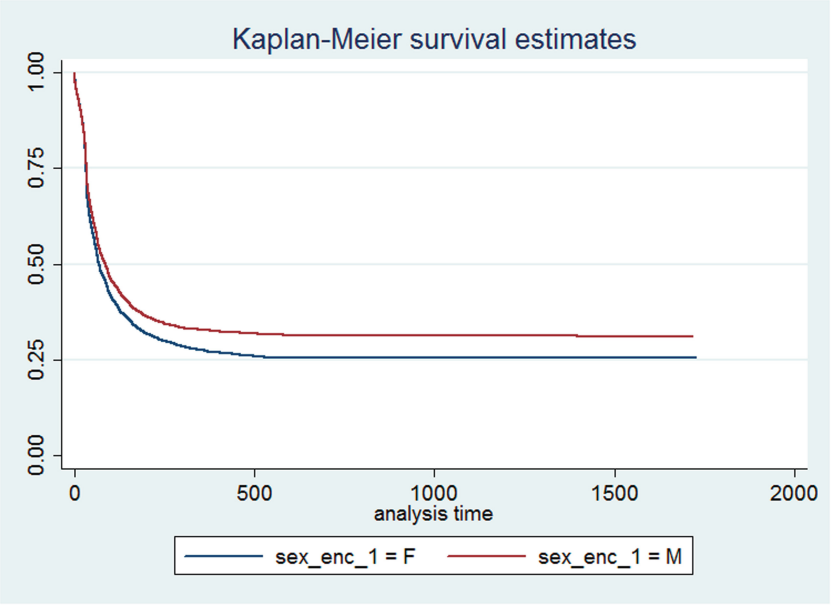

Fig. 3. Survival function within groups by sex

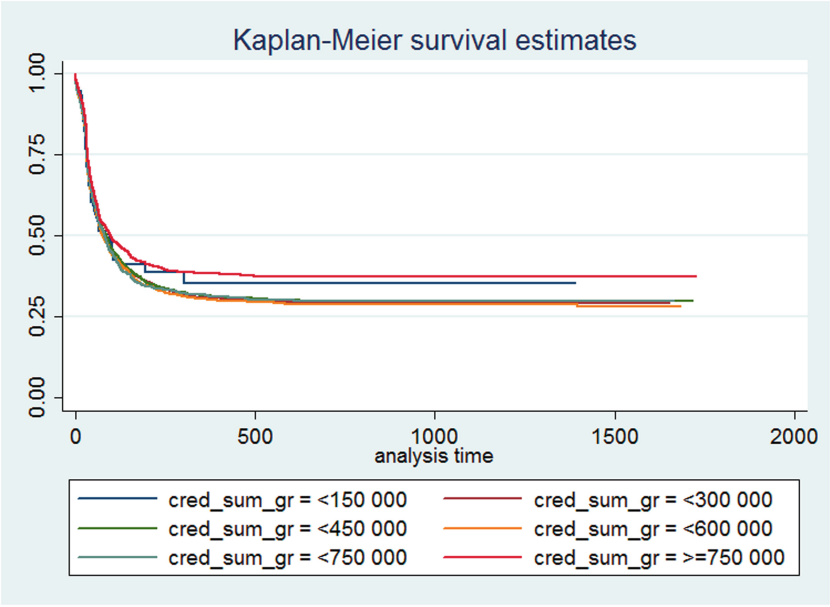

Fig. 4. Survival function within groups by loan amount

References

1. Kaplan, E.L.: Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. **53**, 457–481 (1958)
2. Weibull, W.: Fatigue Testing and Analysis of Results, p. 250. Pergamon Press, New York (1961)
3. Cox, D.R.: Regression models and life-tables. J. R. Stat. Soc. **34**(2), 187–220 (1972)
4. Ratnikova, T. A., Furmanov, K. K. Factual unemployment duration determinants in Russia (in Russian) In: Proceedings of <<Statistical Analysis Implementation in Economics and Quality Estimation>>, pp. 202–206, NRU-HSE – Moscow (2010)
5. Thomas, L.C., Edelman, D.B., Crook, J.N.: Credit Scoring and Its Applications, p. 250. SIAM, Pliladelphia (2002)
6. Silvia, F., Cribari-Neto, F.: Beta regression for modelling rates and proportions. J. Appl. Stat. **31**(7), 799–815 (2004)
7. Ganter, B., Kuznetsov, S.O.: Hypotheses and version spaces. In: Ganter, Bernhard, de Moor, Aldo, Lex, Wilfried (eds.) ICCS 2003. LNCS, vol. 2746. Springer, Heidelberg (2003)
8. Stepanova, M., Thomas, L.: Survival analysis methods for personal loan data. J. Oper. Res. **50**(2), 277–289 (2002)
9. Witzany, J., Rychnovsky, M., Charamza, P.: Survival analysis in LGD modeling. Eur. Financ. Account. J. **7**(1), 6–27 (2012)
10. Chava, S., Stefanescu, C., Turnbull, S.: Modeling the loss distribution. J. Manage. Sci. **57**(7), 1267–1287 (2011)