Paper 11845-2016

Credit Card Interest and Transactional Income Prediction with an Account-Level Transition Panel Data Model Using SAS/STAT® and SAS/ETS® Software

Denys Osipenko, the University of Edinburgh Business School

ABSTRACT

Credit card profitability prediction is a complex problem because of the variety of card holders' behavior patterns and the different sources of interest and transactional income. Each consumer account can move to a number of states such as inactive, transactor, revolver, delinquent, or defaulted. This paper i) describes an approach to credit card account-level profitability estimation based on the multistate and multistage conditional probabilities models and different types of income estimation, and ii) compares methods for the most efficient and accurate estimation. We use application, behavioral, card state dynamics, and macroeconomic characteristics, and their combinations as predictors. We use different types of logistic regression such as multinomial logistic regression, ordered logistic regression, and multistage conditional binary logistic regression with the LOGISTIC procedure for states transition probability estimation. The state transition probabilities are used as weights for interest rate and non-interest income models (which one is applied depends on the account state). Thus, the scoring model is split according to the customer behavior segment and the source of generated income. The total income consists of interest and non-interest income. Interest income is estimated with the credit limit utilization rate models. We test and compare five proportion models with the NLMIXED, LOGISTIC, and REG procedures in SAS/STAT® software. Non-interest income depends on the probability of being in a particular state, the two-stage model of conditional probability to make a point-of-sales transaction (POS) or cash withdrawal (ATM), and the amount of income generated by this transaction. We use the LOGISTIC procedure for conditional probability prediction and PANEL procedures for direct amount estimation with pooled and random-effect panel data. The validation results confirm that traditional techniques can be effectively applied to complex tasks with many parameters and multilevel business logic. The model is used in credit limit management, risk prediction, and client behavior analytics.

INTRODUCTION

Credit card profitability prediction is a complex problem because of a variety of the card holders' behaviour patterns and different sources of the interest and transactional income. Each consumer account can take a number of states such as inactive, transactor, revolver, delinquent, and defaulted. Because of different behavioral types and account income sources for a bank in each state it is required to use an individual model for generated income prediction. Credit cards modelling needs to take into account revolving products dual nature both as standard loan and payment tool. The state of a credit card depends on a type of card usage and payments delinquency. The estimation of status transition probability on account level helps to avoid the memorylessness property of Markov Chains approach which is used for the pool level prediction of income and losses.

General credit cards profit prediction model consists of five stages: i) account or consumer status prediction with conditional transition probabilities, ii) outstanding balance and interest income estimation, iii) non-interest income estimation, iv) expected losses estimation, and v) profit estimation. In current paper the first item the account or consumer status prediction with conditional transition probabilities is discussed.

The paper discovers two problems. Firstly, it describes an approach to credit cards income estimation at the account level based on multistates conditional probabilities model. Secondly, it provides with an

empirical investigation and a comparative analysis of multinomial logistic regression and multistage conditional logistic regression with binary target approaches for transition probabilities estimation model. This model is a part of credit card holders behavioural modeling and can be used for risk management and marketing strategies purposes in retail banking.

GENERAL MODEL SETUP

At the high level credit card holder can be non-active, active, delinquent and defaulted. Active and non-delinquent credit cards holders are split up into two groups: revolvers and transactors. Revolver is user who carries a positive credit card balance and not pay off the balance in full each month – roll over. Transactor is user who pays in full on or before the due date of the interest-free credit period. Competent user does not incur any interest payments or finance charges.

At the highest level the methodology of the credit cards profit prediction model consists of five stages: account (consumer) status prediction with conditional transition probabilities, outstanding balance and interest income estimation, non-interest income estimation, expected losses estimation, and profit estimation.

The model input consists of two types of factors: characteristics (or predictors) and constants. The characteristics are originated from three sources: i) loan application form in the bank's application processing system on account level; ii) core or accounting banking system, aggregated into the data warehouse on account level; iii) credit bureau. Because of the topic of the current research is the credit card (credit line) the loan amount is equal to the credit limit and can change values in time. Thus credit limit value is not constant, but is predictor in the forecasting equations.

The full system of credit card account statuses can be described by the next set: inactive, transactor, revolver, delinquent and default. The account's status is predicted for the next period of time t+1. Each account can transfer into the limited number specific statuses only depending on the current status (see Figure 1. Transition between states). Inactive status account in the next period can be transactor or revolver. Transactor can be revolver or inactive. Revolver can be delinquent or transactor or inactive. Delinquent is unique status which can transit to any possible status, including default. Default status is absorbing status but expected losses estimation is corrected with loss given default estimation. And also each status can be stable without transition to another status for the unlimited period of time.

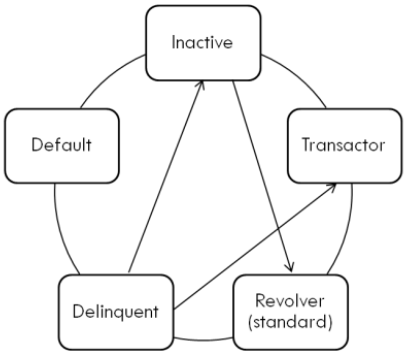


Figure 1. Transition between states

The applications of multinomial regression for credit cards usage states modelling has been proposed by Volker [1]. He defined four type of card usage (hold bankcard, use credit, use regularly, and use moderately) and compared how the same set of predictors (age, professional skills, marital status, region of residence etc.) impact on the customer probability to obtain one of the mentioned statuses. Previous investigations used splitting of customers for users and non-users [2] with discriminate analysis.

Multistage models are widely used, for example, for Loss Given Default estimation in credit risk modelling [3]. We apply methods from [1] and [3] to predict the transition probabilities and compare them.

DATA SAMPLE

The data set for the current research contains information about credit card portfolio dynamics at the account level and cardholders applications. Totally data sample contain information about 150 000 accounts. The data sample is uploaded from the data warehouse of a European commercial bank. The account level customer data sample consist of three parts: i) application form data such as customer socio-demographic, financial, registration parameters , ii) credit product characteristics such as time-dependent credit limit and interest rate, and iii) behavioural characteristics on the monthly basis such as the outstanding balance, days past due, arrears amount, number and types of transactions, purchase and payment turnovers. The macroeconomic data is collected from open sources and contains the main macroindicators such as GDP, CPI, unemployment rate, and foreign to local currency exchange rate. The data sample is available for period Jan 2010 – Dec 2012. The total number of accounts available for the analysis for the whole lending period is 85000.

Behavioural characteristics were created from the original raw data. Index definition in characteristic formulas is the following. Month numeration is calculated backward. For example, Month 1 is the current month at the observation point, Month 2 is previous month (or -1 month). Thus, AvgBalance (1-6) is an average balance for Jan-Jun, AvgBalance (1-3) is an average balance for Apr-Jun.Month numeration is calculated in backward order. For example, Month 1 – current month, observation and calculation point in time, Month 2 – previous month (or -1 month).

Month name	Jan	Feb	Mar	Apr	May	Jun
Month Num	6	5	4	3	2	1

June is current month, month of characteristics calculation and prediction. Thus, AvgBalance (1-6) is average balance for Jan-Jun, AvgBalance (1-3) is average balance for Apr-Jun. The characteristics are presented in the **Ошибка! Источник ссылки не найден.**. The dictionary is not full.

Characteristic	**Description**
Behavioural characteristics (transactional) – Time Random	
Beop1_to_minB2_5	Balance EOP 1 to minimum balance for month 2-5
Beop1_to_maxB2_5	Balance EOP 1 to maximum balance for month 2-5
Beop1_to_avgB2_5	Balance EOP 1 to average balance for month 2-5
DeltaAvgB_1to2	Ratio of Average Balance in month 1 to Average Balance in previous month
DeltaAvgB_1to26	Ratio of Average Balance in month 1 to Average Balance in 5 months
DeltaAvgB_13to46	Ratio of Average Balance in the last 3 month to Average Balance in months 4-6
AvgBeop13_to_AvgBeop46	Average Balance EOP in the last 3 month to Average Balance in months 4-6
maxdpd16	Maximum days past due in the last 6 months
CountDPD1_30	Number of times customer has got to days past due from 1 to 30 for lifetime
CountMonthInDPD	Number of months in any delinquency
Tr_Sum_deb_to_Crd_16	Sum of Debit transactions amounts to Credit transactions amounts for months 1-6
Tr_Avg_deb_to_Crd_16	Average Debit transactions amounts to Average Credit transactions amounts for months 1-6
TR_AvgNum_deb_16	Average monthly number of debit transactions for months 1-6
TR_AvgNum_Crd_16	Average monthly number of credit transactions for months 1-6
TR_MaxNum_deb_16	Maximum monthly number of debit transactions for months 1-6
TR_MaxNum_Crd_16	Maximum monthly number of credit transactions for months 1-6
TR_max_deb_to_Limit16	Amount of maximum debit transaction to limit for months 1-6
TR_avg_deb_to_Beop16	Average of debit transaction to balance EOP 1-6

Characteristic	Description
NoAction_NumM_16	Number of month with no actions for period 1-6
Application characteristics – Time fixed	
Age	As of the date of application
Gender	Assumption that status constant in time
Education	Assumption that status constant in time
Marital status	Assumption that status constant in time
Region	Assumption that is not changed. In case of change – it will be new account
Work at last place	As of the date of application
Position	The position occupied by an applicant
Income	As of the date of application
Spouse income	As of the date of application
Additional income	As of the date of application
Macroeconomic characteristics – Time random	
Unemployment Rate ln lag3	Log of unemployment rate with 3 month lag
GDPCum_ln yoy	Log of cumulative GDP year to year to the same month
UAH-EURRate_ln yoy	Log of exchange rate of local currency to Euro in compare with the same period of the previous year
CPIYear_ln yoy	Log of the ratio of the current Consumer Price Index to the previous year the same period CPI

Table 1. List of the original data, behavioural, application and macroeconomic characteristics

INCOME MODEL BUILDING

Generally (Thomas et. al., 2001; So and Thomas, 2011) risk management approaches define delinquent and non-delinquent account buckets as following: current, day past due (DPD) 1-30 (Bucket 1), DPD 31-60 (Bucket 2), DPD 61-90 (Bucket 3), and default. Current state may differentiate by the level of risk, or score. As the aim of our investigation is the profit prediction from a credit card usage, we propose to define the credit card statuses subject to the revenue source and the revenue availability. The risk assessment is accompanied to the main revenue-based state definition. For profit prediction the risk is estimated as Expected Losses. Expected Losses is a product of the probability of default, loss given default and exposure at default. The probability of default is actually a transition probability to default state. Loss given default in the current research is taken as a constant. Exposure at default is depending on the expected outstanding balance at the point of default, and we use the credit limit utilization rate models in the current investigation for the outstanding balance prediction.

Clients are split up two group: revolvers and transactors. Revolver – user, who carry a positive credit card balance and not pay off the balance in full each month – roll over. Transactor – user, who pay in full on or before the due date of the interest-free credit period. Competent user do not incur any interest payments or finance charges. We propose to define the credit card statuses subject to the revenue source and the revenue availability.

An account in each status exception inactive and defaulted can generate an income. However, the sources of income are different. This point is often not considered by researchers. For instance, delinquent account can generate non-interest income due to interchange fees from merchants and penalty, but does not generate interest income because of non-paid debt. However, delinquent account is not losses like defaulted one.

The number of transition probabilities is N-1, where N is the number of states. For common scoring model such as the probability of default estimation we need the model for only one probability. For example, the

probability of moving to default state is p. Then the probability to stay in non-default state is 1-p. However, in our model of the credit card holder's behaviour the number of states, which account can move in, is more than two, for example, a revolver can move to transactor, delinquent, and inactive states, or stay a revolver. Thus it is necessary to estimate the set of Ns,t+1-1 transition probabilities pj, where j is the transition index, and $\sum_{j=1}^{N_{s,t+1}-1} p_j = 1$,

where Ns,t+1 is the number of states available for moving from the state s at time t+1.

Account status	Symbol	Definition	Risk assessment	Revenue assessment	Note
closed	C	Account is closed or inactive more than 6 months	No	No	Excluded from the analysis
inactive	NA	Average OB = 0 and Debit Turnover Amount = 0	No	No	Expected Loss (EL) can be estimated with state transitions
transactor	TR	OB_eop = 0 and Debit Turnover Amount > 0	No	Debit Transactions Amount x Transaction Profit Rate	TR Profit Rate = (avg interchange rate + fees rate) EL – see inactive note
revolver (current)	RE	Average OB > 0 and DPD = 0	Behavioural (transition) score for current	Limit x Utilization Rate x Interest Rate + Debit Transactions Amount x Transaction Profit Rate	-
delinquent	Dl	Average OB > 0 and (DPD > 0 and DPD <=90)	Behavioural (transition) score for delinquent	No	If credit card is not blocked, the transaction revenue exists
defaulted	D	Average OB > 0 and DPD > 90	LGD	-	Recovery is not revenue. It's EL reduction

Table 2. Account state definition and related assessments

Depending on the status an account has an individual set of the models: probability of transition to another state, probability of action and income estimation for each possible action. Thus, the total income prediction model is presented as a sum of results of three-level conditional models:

i) probability to be in status s,

ii) probability of action,

iii) income estimation after action for specific status.

Expected income is equal to the product of two functions: the probability that customer will use cards for certain transaction (for example, pos transaction, atm cash withdrawal) and the estimation of income from this transaction. The final model in general format sum of the products of three estimations such as the probability to be in status S, the probability to do action POS/ATM and income estimation for each status.

Transactor income:

$$\mathrm{I}(i,t+1 \mid s_i = T) = \Pr(s_{i,t+1} = T \mid s_{i,t} \neq D) \times \\ \times \begin{pmatrix} \Pr(a_{i,t+1} = POS \mid s_{i,t+1} = T) \cdot \mathrm{R}(\mathbf{x_i} \mid a = POS) + \\ + \Pr(a_{imt+1} = ATM \mid s_{i,t+1} = T) \cdot \mathrm{R}(\mathbf{x_i} \mid a = ATM) \end{pmatrix} \quad (1)$$

where R(.) is the revenue function.

Revolver income:

$$\mathrm{I}(i,t+1 \mid s_i = R) = \Pr(s_{i,t+1} = R \mid s_{i,t} \neq D) \times \\ \times \begin{pmatrix} \mathrm{Ut}(\mathbf{x_i} \mid s_{i,t+1} = R) \times IR \times Limit \times \mathrm{P}(s_{t+1} = R \mid s_t = R) + \\ + \Pr(a_{i,t+1} = POS \mid s_{i,t+1} = R) \cdot \mathrm{R}(\mathbf{x_i} \mid a = POS) + \\ + \Pr(a_{imt+1} = ATM \mid s_{i,t+1} = R) \cdot \mathrm{R}(\mathbf{x_i} \mid a = ATM) \end{pmatrix} \quad (2)$$

where Ut(.) is the utilization rate function.

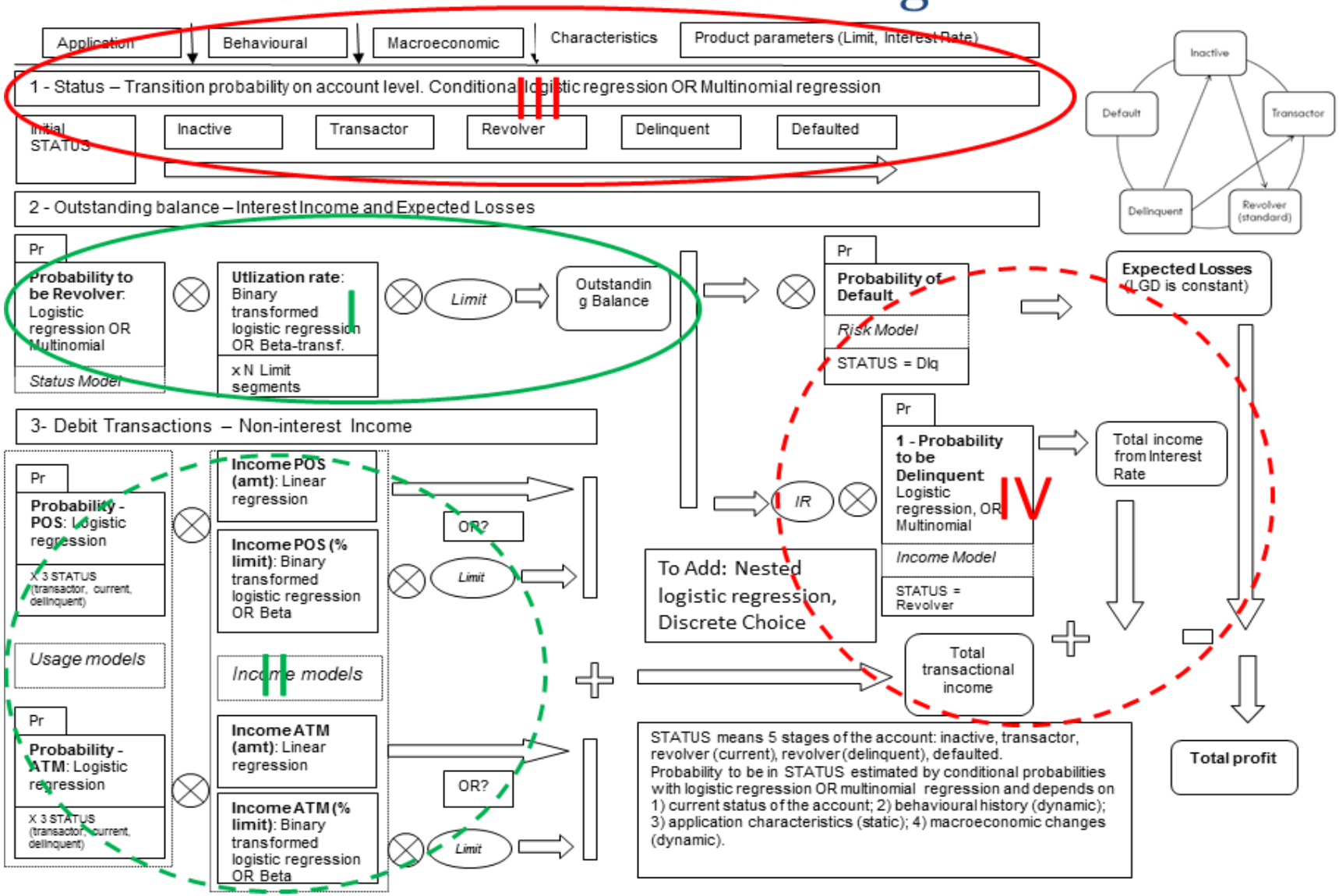
Delinquent income:

$$\mathrm{I}(i,t+1 \mid s_i = Dlq) = \Pr(s_{i,t+1} = Dlq \mid s_{i,t} = Dlq) \times \\ \times \begin{pmatrix} \Pr(a_{i,t+1} = POS \mid s_{i,t+1} = Dlq) \cdot \mathrm{R}(\mathbf{x_i} \mid a = POS) + \\ + \Pr(a_{imt+1} = ATM \mid s_{i,t+1} = Dlq) \cdot \mathrm{R}(\mathbf{x_i} \mid a = ATM) \\ + Penalty \end{pmatrix} \quad (3)$$

These equations are an example of the account which keeps the same status. For the transition probabilities to other statuses the equations should be transformed appropriately.

There are two concepts how many models we need. Multinomial regression is more convenient for a computation and it is not obligatory to build logistic regression model for each transition, but use ‘From’ status as a variable. However, we use an assumption that for each status the transition probabilities regression equation will have different slopes and trends for predictors.

Schema of the modelling



MODELLING RESULTS

Model 1 – Decision tree of the conditional logistic regressions with binary target

The problem can be presented as a binary decision tree where number of leaves is equal to number of states S and number of transition models is S-1. The result of regression is a set of the conditional logistic regressions with binary target. The general model can be presented as binary tree (see **Ошибка! Источник ссылки не найден.**).

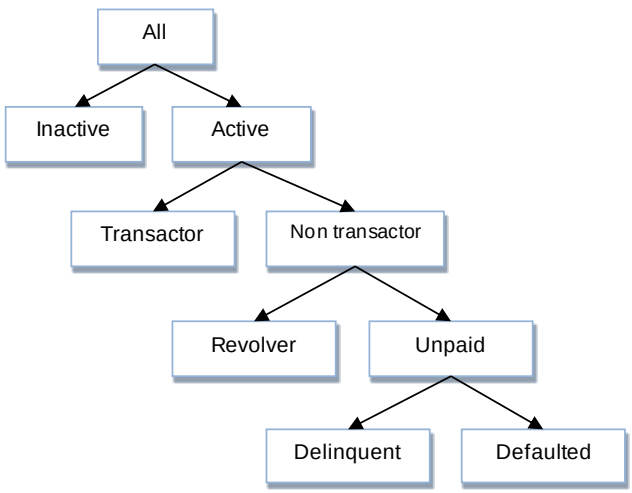


Figure 2. Multistage schema of the conditional logistic regression models

At each stage we predict the probability of transition to one of the states at the next level conditional on the state in the higher level.

For a full description of all stages we need four equations:

$$\Pr(s_{t+1} = NA) = \boldsymbol{\beta}_{NA}^{\mathbf{T}}\mathbf{x}$$

$$\frac{\Pr(s_{t+1} = TR)}{1-\Pr(s_{t+1} = NA)} = \boldsymbol{\beta}_{\mathbf{T}R}^{\mathbf{T}}\mathbf{x}$$

$$\frac{\Pr(s_{t+1} = RE)}{(1-\Pr(s_{t+1} = TR))(1-\Pr(s_{t+1} = NA))} = \boldsymbol{\beta}_{RE}^{\mathbf{T}}\mathbf{x}$$

$$\frac{\Pr(s_{t+1} = D1 \mid s_t \notin (NA, TR))}{(1-\Pr(s_{t+1} = NA))(1-\Pr(s_{t+1} = TR))(1-\Pr(s_{t+1} = RE))} = \boldsymbol{\beta}_{D1}^{\mathbf{T}}\mathbf{x}$$

$$\frac{\Pr(s_{t+1} = D2 \mid s_t \notin (NA, TR))}{(1-\Pr(s_{t+1} = NA))(1-\Pr(s_{t+1} = TR))(1-\Pr(s_{t+1} = RE))(1-\Pr(s_{t+1} = D1))} = \boldsymbol{\beta}_{D2}^{\mathbf{T}}\mathbf{x}$$

$$\frac{\Pr(s_{t+1} = DF \mid s_t \notin (NA, TR, RE, D1))}{(1-\Pr(s_{t+1} = NA))(1-\Pr(s_{t+1} = TR))(1-\Pr(s_{t+1} = RE))(1-\Pr(s_{t+1} = D1))(1-\Pr(s_{t+1} = D2))} = \boldsymbol{\beta}_{DF}^{\mathbf{T}}\mathbf{x}$$

where account status s is NA – non-active, T – transactor, R – revolver, D1 – delinquent 1 bucket, D2 – delinquent 2 bucket, Def – defaulted.

Generally logistic regression matches the log of the probability odds by a linear combination of the characteristic variables as

$$\mathrm{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \boldsymbol{\beta} \cdot \mathbf{x}_i^T,$$

where

- pi is the probability of particular outcome,
- β0 and β are regression coefficients,
- x are predictors.

The probability of event for i_{th} observation is calculated as

$$P_i = E(Y_i \mid \mathbf{x}_i, \beta) = \Pr(Y_i = 1 \mid \mathbf{x}_i, \beta)$$

In the program below the first stage is probability the customer is active or non-active. PROC LOGISTIC is used to run binary logistic regression step by step. The stepwise method has been applied (selection = stepwise) with significance levels slentry and slstray equal to 0.1. The SAS code example provided for current state Transactor modeling:

```
/* ------------------------- LOGISTIC MULTISTAGE -------------------------
*/
/* Transactors */

/* stage 1 - to be NA */

data &r.Tr_beh_dev_tr_st_m1;
set &r.Tr_ beh_dev_tr;
if Target_S1='NA' then Target_st1=1; else Target_st1=0;
run;

proc logistic data=&r.Tr _beh_dev_tr_st_m1;
model Target_st1(event='1') = &Predictors/
selection = stepwise
slentry = 0.1
slstay = 0.1;
output out=&r.Tr_ beh_dev_tr_st_m1 predicted=pr_st1 ;
run;
/* stage 2 - to be TR */

data &r.Tr _beh_dev_tr_st_m1;
set &r.Tr _beh_dev_tr_st_m1;
if Target_S1 = 'NA' then Target_st2 =_NULL_;
else if Target_S1='Re' then Target_st2=1; else Target_st2=0;
run;
proc logistic data=&r.Tr _beh_dev_tr_st_m1;
model Target_st2(event='1') = & Predictors/
selection = stepwise
slentry = 0.1
slstay = 0.1;
output out=&r.Tr_ beh_dev_tr_st_m1 predicted=pr_st2 ;
run;
```

```
/* --- Calculate the probabilities ------------ */

data &r.Tr_beh_dev_tr_st_m1;
set &r.Tr_beh_dev_tr_st_m1;
pr_na=pr_st1;
pr_re= (1-pr_st1)*pr_st2;
pr_tr = (1-pr_st1)*(1-pr_st2);
check=pr_na+pr_tr+pr_re;
run;
```

MODEL 2 – MULTINOMIAL LOGISTIC REGRESSION WITH NON-BINARY TARGET

However, this complicated procedure can be avoided in case of application of *ordered logistic regression*, or multinomial logistic regression.

The equation defined as $R_i^* = X_i\beta + \varepsilon_i$ with

$$R_i = \begin{cases} 1 & if\ \ R_i^* \le \mu_0 \\ 2 & if\ \ \mu_0 < R_i^* \le \mu_1 \\ 3 & if\ \ \mu_1 < R_i^* \le \mu_2 \\ \dots & \\ N & if\ \ \mu_N < R_i^* \end{cases}$$

where R_i are the observed scores that are given numerical values as follows: status 1, status 2,..., status N; R_i^* is unobserved dependent variable (the exact level of agreement with the statement proposed),

Xi is a vector of variables that explains the variation of status; β is a vector of coefficients; μ_i are the threshold parameters to be estimated along with β; and ε_i is a disturbance term that is assumed normally distributed.

The final parameter estimation is a system of equations:

$$\ln\frac{\Pr(Y_i=1)}{\Pr(Y_i=K)} = \boldsymbol{\beta}_1 \cdot \mathbf{X}_i$$
$$\ln\frac{\Pr(Y_i=2)}{\Pr(Y_i=K)} = \boldsymbol{\beta}_2 \cdot \mathbf{X}_i$$
$$\dots\dots$$
$$\ln\frac{\Pr(Y_i=K-1)}{\Pr(Y_i=K)} = \boldsymbol{\beta}_{K-1} \cdot \mathbf{X}_i$$

One of the applications of multinomial regression for credit cards usage states modelling has been proposed by Volker (1982). He defined four type of card usage (hold bankcard, use credit, use regularly, and use moderately) and compared how the same set of predictors (age, professional skills, marital status, region of residence etc.) impact on the customer probability to obtain one of the mentioned statuses.

For the multinomial regression we use the SAS PROC LOGISTIC with use of generalized logit parameter Link=glogit:

```
PROC LOGISTIC data=&r.Tr_beh_dev_tr outest = &r.est_Mult_tr;
model Target_S = &Predictors/
selection = stepwise
slentry = 0.1
slstay = 0.1
include=50
link=glogit;
output out=&r.Pr_beh_dev_tr (keep=tr_id month target_s6 _LEVEL_ pr_s)
predicted=pr_s;
run;
```

Estimation results from the Table 3. Multinomial regression parameters estimations show that the same predictors have different correlations and even opposite trends for the probability of transition. For example, behavioural characteristic b_TRsum_crd1_to_OB1 – ratio of a total amount of credit transactions to the average outstanding balance for the last month has positive coefficients for transitions from transactor state to inactive, from revolver to inactive and revolver, from delinquent to all other state and negative coefficients for transitions form transactor to revolver and from revolver to delinquent.

For categorical variables like applicant's characteristics such as education, marital status, position etc. the dummy variables approach is applied. So each value of categorical parameter is defined for a separate characteristic. For example, manager position has positive estimations values for the transition from delinquent state to another states and negative for all another transitions, but technical staff has negative one estimations for delinquent state transitions.

	NA		TR		Re			Dl		
Parameter	**NA**	**Re**	**NA**	**Re**	**NA**	**Re**	**Dl**	**Re**	**Dl**	**Df**
b_atm_flag_0	-1.301	0.0205	-0.7706	-0.1245	-0.0537	0.0352	0.0308	-0.6183	-0.4235	-1.0327
b_atm_flag_13	-1.3343	-0.7059	-0.2998	0.1353	-0.0187	0.0221	0.00914	3.6879	-4.6138	3.052
b_atm_flag_use13vs46	0.4835	0.2297	0.0756	-0.0313	0.0998	-0.085	-0.1558	-0.3405	-0.2942	-0.3465
b_atm_flag_used46vs1	-0.0875	0.0366	0.4119	0.3787	0.0675	-0.0084	-0.1594	-0.4549	-0.7909	-1.0524
b_avgNumDeb16	-0.00559	0.0333	-0.0872	-0.0005	-0.00396	0.000705	-0.00071	0.1384	0.1201	0.0951
b_AvgOB16_to_MaxOB16	0.1527	0.2086	-0.0495	0.0995	-0.579	0.4217	0.3277	1.178	1.8773	0.7239
b_DelBucket16	3.4238	3.1253	5.0093	4.7271	-0.2446	-0.3813	2.2051	-0.2102	-0.1823	0.9418
b_inactive13	0.0275	0.3162	0	0	0	0	0	0	0	0
b_NumDeb13to46ln	-0.0118	0.00741	0.0476	0.0252	0.0797	-0.0629	-0.0468	0.4529	0.4703	0.3598
b_OB_avg_to_eop1ln	0.0344	-0.00826	0.0115	0.0431	0.8299	-0.3465	-0.0919	-4.4728	-2.1966	-1.656
b_payment_lt_5p_1	-0.1898	-0.00331	-0.1892	-0.3846	-0.1389	0.0983	0.1286	1.0131	1.1619	1.8105
b_payment_lt_5p_13	-0.4752	-0.7944	-6.453	-2.6808	-0.0572	0.0486	0.2768	-0.3617	-0.00357	0.6378
b_pos_flag_0	-1.5643	-0.5033	-0.831	-0.322	-0.0995	-0.0685	0.1866	2.4833	21.3137	14.1346
b_pos_flag_13	-0.3446	-0.1653	-0.1981	0.1551	-0.0864	-0.0298	0.1367	-9.0955	-18.9315	-20.4139
b_pos_flag_use13vs46	0.3902	0.2865	0.2909	-0.0444	-0.0118	-0.0212	-0.1224	-0.9643	-1.1442	-0.8545
b_pos_flag_used46vs1	-0.1923	-0.0896	0.3059	0.3401	0.0404	0.0489	-0.1321	-3.2584	-3.46	-6.4911
b_pos_use_only_flag_	-0.9998	-0.6248	-0.7299	-0.2038	0.0632	-0.2897	0.000783	3.8619	-3.8421	4.0725
b_TRavg_deb16_to_avg	-0.1284	-0.074	0.0557	-0.0762	0.0382	-0.0349	-0.00817	0.6563	0.0651	0.3612
b_TRmax_deb16_To_Lim	0.0646	0.065	0.0228	0.0253	-0.0161	-0.0128	0.0239	-0.408	-0.0864	-0.1964
b_TRsum_crd1_to_OB1_	0	0	0.023	-0.0758	0.0786	0.0236	-0.1836	0.1181	0.1052	0.0155
b_TRsum_crd13_to_OB1	-0.0164	-0.0572	-0.3738	-0.2222	0.0758	-0.0565	-0.0215	-0.00607	0.0495	-0.0722
b_TRsum_deb16_to_TRs	-0.00654	-0.00039	-0.4588	-0.00104	-0.1935	0.0699	0.0344	-0.0735	0.0335	0.5135
b_UT1_to_AvgUT16ln	0	0	-0.1104	-0.0218	-0.5148	0.1992	0.1075	0.3374	-0.2789	-4.5139
b_UT1to2ln	0	0	0.0303	0.00174	0.0828	-0.0924	-0.0505	1.4581	1.3208	1.2898
max_dpd_60	5.7105	6.1331	5.6345	5.585	-1.087	16.3607	-10.5797	0.3432	0.9232	0.9855
mob	0.0224	-0.0107	-0.0151	-0.00369	0.00048	-0.00024	-0.00251	-0.057	-0.0183	-0.0618
no_dpd	3.6349	3.2844	4.9951	4.6152	0.0248	0.0291	-0.1942	0	0	0
l_ch1_ln	-0.2887	0.1436	-0.4325	-0.0795	-0.246	0.0266	-0.2603	14.819	14.1673	15.1089
l_ch6_ln	0.1795	0.2065	0.0169	0.1838	-0.1765	0.1068	-0.2039	3.2038	3.4672	4.1909
AgeGRP1	-0.0814	0.0277	0.2094	0.4295	-0.00821	-0.00847	0.1227	1.0288	1.0672	1.0648
AgeGRP3	0.0328	0.00357	0.1295	0.1611	0.00932	0.0145	-0.1592	0.5811	0.1383	0.6383

	NA		TR		Re			Dl		
Parameter	**NA**	**Re**	**NA**	**Re**	**NA**	**Re**	**Dl**	**Re**	**Dl**	**Df**
avg_balance_6	-0.2898	0.2767	-0.00005	-0.00005	-0.00002	0.000019	6.77E-06	0.000063	-0.00002	0.000292
customer_income_ln	-0.0131	-0.0352	-0.025	-0.0674	-0.0544	-0.0729	-0.0812	-0.741	-0.9324	-0.5747
Edu_High	-0.2924	-0.1898	0.481	0.1214	0.056	-0.0253	-0.1296	0.1671	0.1061	-0.1498
Edu_Special	-0.148	-0.0869	0.2095	-0.0154	-0.00646	0.0283	-0.0145	0.4795	0.3588	0.6204
Edu_TwoDegree	-0.3221	-0.204	0.6438	0.3011	0.00172	-0.0952	-0.1599	-1.9808	-1.8417	-2.2571
Intercept	0.9968	-0.5522	-3.021	-4.0378	-1.8266	3.3452	-0.6312	5.6449	0.9115	-6.3674
Marital_Civ	0.5746	0.6968	0.0216	0.1398	0.0308	0.00463	0.0315	1.377	1.2032	1.7662
Marital_Div	0.0567	0.0645	0.1267	0.223	0.0562	0.0316	-0.0624	-0.8603	-0.995	-0.8913
Marital_Sin	0.0577	-0.00502	0.1044	0.1965	-0.0139	0.000454	0.0513	0.0177	-0.2862	0.1961
Marital_Wid	0.3452	0.179	0.2966	0.3454	0.0252	0.0678	-0.0118	-0.8242	-1.1807	-0.9777
position_Man	-0.0765	-0.0544	-0.1188	-0.0198	-0.00512	-0.019	-0.055	1.2765	1.5022	1.4493
position_Oth	-0.2071	-0.0755	-0.1215	-0.1764	-0.0132	0.0283	-0.0348	-0.1081	-0.5273	0.4837
position_Tech	-0.1837	-0.0904	0.2715	0.00561	-0.0683	0.0321	0.0151	-0.3033	-0.366	-0.3285
position_Top	-0.1049	-0.0153	-0.0967	0.0422	-0.00694	-0.1784	-0.1479	-0.2757	0.0921	-0.3933
SalaryYear_lnyoy_6	5.4957	2.9357	-0.92	1.6574	0.0664	-0.0178	-0.3487	4.6512	5.7042	6.0895
UAH_EURRate_lnmom_6	4.8216	-0.3475	0.2254	-2.485	-1.1878	-0.00486	-0.3994	6.9056	8.8226	9.9647
UAH_EURRate_lnyoy_6	-2.6934	-1.8483	0.5456	-0.7113	-0.0179	-0.1594	0.5979	-0.305	-1.0516	1.588
Unempl_lnyoy_6	-0.6169	-1.1316	0.0276	-0.126	-0.0811	0.1245	0.153	-1.7341	-2.4043	-0.2159

Table 3. Multinomial regression parameters estimations

A COMPARATIVE ANALYSIS OF MODELS FOR TRANSITION PREDICTION

In multistage logistic regression approach the final performance result depends on the order of inclusion of status in the decision tree. In Table 4. Comparative analysis of multinomial and multistage binary logistic regression approaches the arrows show the order of states for conditional logistic regression binary tree building. The original model use the order from inactive state to delinguent one. The last column ‘Another order of the stages’ shows prediction order from the delinquent to inactive stages for revolver and delinquent current states.

Status		**Gini coefficient value**		
From	**To**	Multi nomial	Multistage binary logistic	Another order of the stages in logit model
NA	na	36%	36%	-
	tr	38%	-	-
	re	31%	30%	-
TR	na	47%	47%	56%
	tr	44%	36%	-
	re	38%	-	38%
Re	na	55%	49%	-
	tr	56%	67%	47%
	re	61%	68%	60%
	dl	64%	-	70%
DL	tr	44%	80%	-
	re	48%	60%	40%
	dl	38%	48%	48%
	df	79%	-	80%

Table 4. Comparative analysis of multinomial and multistage binary logistic regression approaches

The first or the last model in the set can have the best predictive power, but in is not a rule. However single binary model results are better than results of the multinomial logistic regression for the selected segment.

THE UTILIZATION RATE PREDICTION WITH TWO STAGE MODEL

The usage of credit limit may be changed during a lifetime period. The utilization rate (Ut) is defined as the outstanding balance (OB) divided by credit limit (L) Ut = OB/L.
For the full utilization rate model and more information about prediction methods see *Osipenko & Crook (2015).*

Two-stage model means that at the first stage the probability to get a boarder value as 0 and 1 is calculated, and then the proportion estimation in the interval (0;1) are applied. At the first stage the probability that an account has zero utilization (Pr (Ut=0) and then that an account has full utilization (Pr (Ut=1)) in the performance period is calculated with binary logistic regression. At the second stage the proportion between 0 and 1 excluding 0 and 1 values is calculated according to the set of the approaches used for one-stage direct estimation.

The two-stage model utilization rate is calculated with the following formula:

$$Ut = (1 - \Pr(Ut = 0))(\Pr(Ut = 1) + (1 - \Pr(Ut = 1)) \cdot E(Ut \mid Ut \neq 0, Ut \neq 1))$$

Where Pr(Ut=0) and Pr(Ut=1) are the probability the utilization rate is equal to 0 or 1 respectively.

$E(Ut \mid Ut \neq 0, Ut \neq 1)$ is the utilization rate proportion estimation for the utilization rates not equal zero and not equal to 1.

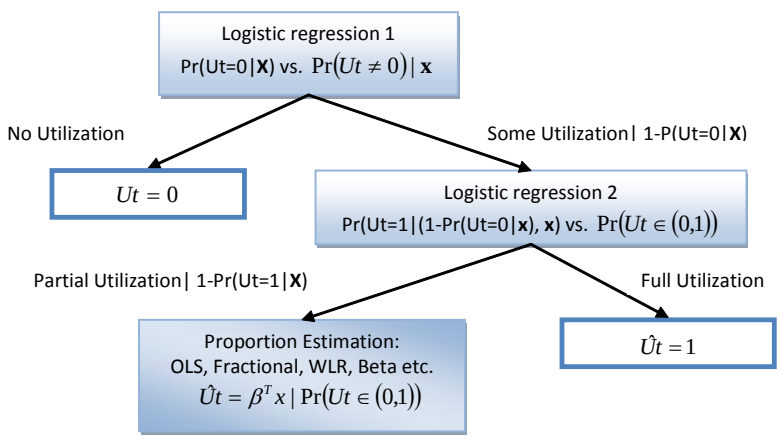


Figure 3. Two-stage regression model schema

Two-stage model consist of two parts: the probability of zero utilization and full utilization with use of logistic regression and the proportion estimation with use of the set of the same methods as for one-stage model.

Ttwo-stage models have shown better model accuracy and prediction results for development and validation samples, but the difference in forecasts errors are insignificant. For example, for Limit No Change model for OLS method for one-stage and two-stage approaches R^2 = 0.5498 and 0.5534, MAE= 0.1930 and 0.1913 respectively. However, if we compare Stage 2 model with one-stage direct estimation it can be seen that one-stage model gives better results.

Month on Book	Limit Changes	Stage	Method	Development Sample				Validation Out-of-sample			
MOB 6 or more	Limit NO change	*Stage 1*	*Probability*	*KS*	*Gini*	*ROC*		*KS*	*Gini*	*ROC*	
		Pr(UT=0)	Logistic Regression	0.6262	0.7479	0.8739		0.6331	0.7547	0.8774	
		Pr(UT=1)	Logistic Regression	0.5931	0.7243	0.8622		0.6036	0.7355	0.8678	
		Stage 2	*Proportion Estimation*	*R2*	*MAE*	*RMSE*	*MAPE*	*R2*	*MAE*	*RMSE*	*MAPE*
		0<UT<1	OLS	0.4310	0.1948	0.2462	4.9151	0.4235	0.1950	0.2462	4.8260
			Fractional(Quasi-Likelihood)	0.4309	0.1946	0.2463	4.9683	0.4235	0.1950	0.2462	4.8260
			Beta regression (nlmixed)	0.4183	0.2102	0.2506	5.0499	0.4108	0.2104	0.2507	4.9075
			Beta transformation + OLS	0.3680	0.1802	0.2673	2.7377	0.3618	0.1809	0.2673	2.6513
			Weighted Logistic Regression	**0.4325**	0.1945	0.2457	4.8937	0.4253	0.1948	0.2456	4.7564
		Two-stage	*Aggregate*	*R2*	*MAE*	*RMSE*	*MAPE*	*R2*	*MAE*	*RMSE*	*MAPE*
		0<= UT <=1	OLS	0.5534	0.1913	0.2535	3.1366	0.5536	0.1910	0.2526	3.0784
			Fractional(Quasi-Likelihood)	0.5527	0.1915	0.2536	3.1590	0.5529	0.1912	0.2528	3.0979
			Beta regression (nlmixed)	0.5366	0.2068	0.2581	3.2109	0.5364	0.2063	0.2574	3.1502
			Beta transformation + OLS	0.4720	0.1773	0.2754	1.7407	0.4724	0.1774	0.2745	1.7019
			Weighted Logistic Regression	**0.5548**	0.1914	0.2531	3.1116	0.5553	0.1910	0.2521	3.0532

Table 5. Two-stage models comparative analysis

NON-INTEREST RATE PROFITABILITY MODELLING

PANEL DATA

Econometrics data can be divided into two types: i) Cross-sectional which has dimensions by economic items at the same point of time (without any relation to the time), ii) Time series or observation of the economic values ranked in time.

In practice often this two dimensions is joined. The simplest join is the Independent one (not ranked in time) or pooled data. For example, data slices dated monthly as Balance as of end of Jan, Balance as of end of Feb etc. are added to the data sample as independent observations. The Panel data is two-dimension array both cross-sectional and time series where cross-sectional characteristics are ranked as time series.

Cross-sectional and time series data are joined but in different ways. As industrial standard it is often used independent join (not ranked in time) or pooled data. However, we take into account how predictors impact on outcome and for the same account it like independent cases (12 periods of time – 12 rows) or, for instance, average.

In general, researchers mark out the next advantages of the panel data:

i) Higher number of observations results increase in the levels of freedom, gives more efficient estimations

ii) Heterogeneity of the sample objects is under control

iii) Testing of the effects which is impossible to identify separately in cross-sections and time series

iv) Decrease in multicollinearity

v) It's possible to build more complicated behavioural models and decrease the influence of the missing values and incorrectly measured observations

It has been considered to use cross-sectional data only with behavioural characteristics calculation at

point in time for the initial investigation. The main assumption was that the customer behavioural characteristics are homogeneous in time and number of observation is a big enough to level all possible time and structure fluctuations. However, because of some changes in customer behaviour and accounts dynamics in period 2011-2012 years it has sense to apply panel data model approach to take into account cross-sectional changes.

PROC PANEL is used for a panel data generation

```
proc panel data=tr_final;
      id tr_id t;
        lag
      UT0 (1 2 3 4 5 6 7 8 9)
      amt_pos (1 2 3 4 5 6 7 8 9)
      amt_ir (1 2 3 4 5 6 7 8 9)
       …     /
      out=tr_final_lag;
   run;
```

Generally panel data model can be presented by the next equation:

$$y_{it} = \alpha + X'_{it}\beta + Z'_{it}\gamma + u_{it}, \; i = 1,...,N, \; t = 1,...,T$$

X is observed factors vector;

Z is unobserved factors vector, $Z_{it} = Z_t$.

$$y_{it} = \alpha_i + X'_{it}\beta + \gamma_t + u_{it};$$

N is number of cases;

T is number of time periods;

β and γ - regression slope coefficients;

$$u_{it} = \mu_i + \lambda_t + \upsilon_{it}$$

μ_i, λ_i – non-observed individual and time effects, u_{it} – residual idiosyncratic components.

Pooled model in fact is the same as general linear regression model and doesn't take into account time component:

$$y_{it} = X'_{it}\beta + \alpha + \varepsilon_{it}$$

α и β – intercept and slope is independent from observation and time

X_{it} - vector of regressors (predictors)

Approach with time slices is widely applied as industry standard, for instance, to create development and validation samples from the data set with not enough observation at the point in time or to take into account different seasons.

Use of pooled panel data approach requires the next assumption:

- ✓ dependence between factors is stable in time;
- ✓ correlation between observations is not taking into account.

However, in real practice these conditions often are not satisfied. Thus to consider the time component the fixed and random effects are used.

Random effect model

$$y_{it} = \alpha + X'_{it}\beta + (u_i + v_{it})$$

$(u_i + v_{it})$ is a random effect. Intercept is constant. Error variance is varying across groups and/or times

PROFITABILITY MODELLING WITH TWO-STAGE MODEL

1st stage – estimation of the probability that the client will use credit cards for POS/ATM transaction during the forecast period

$$\ln\left(\frac{P}{1-P}\right) = \phi \frac{1}{T}\sum_{l=1}^{T} UT_{i(t-l)} + \sum_{k=1}^{K}\beta_k \cdot B_{ki,t-1} + \sum_{l=1}^{L}\alpha_a \cdot A_{ai} + \sum_{m=1}^{M}\gamma_m M_{m,t-1}$$

2nd stage – income amount for the period

$$POS_{it} = \phi \frac{1}{T}\sum_{n=1}^{T} UT_{i(t-n)} + \sum_{k=1}^{K}\beta_k \cdot B_{bi,t-1} + \sum_{l=1}^{L}\alpha_a \cdot A_{ai} + \sum_{m}^{M}\gamma_m M_{m,t-1}$$

$$ATM_{it} = \phi \frac{1}{T}\sum_{n=1}^{T} UT_{i(t-n)} + \sum_{k=1}^{K}\beta_k \cdot B_{bi,t-1} + \sum_{l=1}^{L}\alpha_a \cdot A_{ai} + \sum_{m}^{M}\gamma_m M_{m,t-1}$$

φ, α, β, γ – regression coefficients (slopes)

B – vector of behavioural factors (for example, average balance to maximum balance, maximum debit turnover to average outstanding balance or limit)

A – vector of application factors - client's demographic, financial and product characteristics

M – vector of macroeconomic factors (GDP, FX, Unemployment rate changes, etc)

Expected income is equal to the product of two functions: the probability that customer will use cards for certain transaction (for example, pos transaction, atm cash withdrawal) and the estimation of income from this transaction.

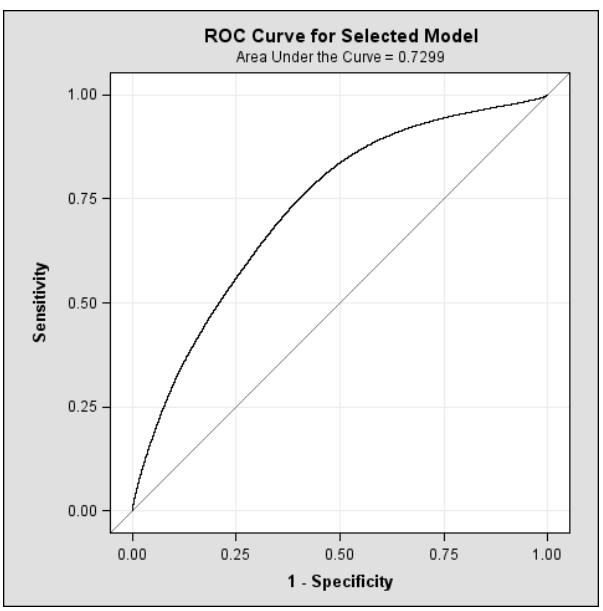
Stage 1 – Logistic regression – probability of ATM transaction

```
proc logistic data=&u.Tr_final_plus6m_atm_log outest = &u.r_atm_log_est;
model YN_ATM = &RegV /
selection = stepwise
slentry = 0.05
slstay = 0.05
include=25
outroc=&u.r_atm_log_roc1;
output out=&u.r_atm_log_out predicted=score;
run;
```

Parameter	Estimate	Standard	Wald Chi-Square	Pr > ChiSq
Intercept	9.2935	0.4657	398.215	<.0001
limit	-0.00005	1.46E-06	1073.447	<.0001
customer_income	0.000103	3.54E-06	842.0063	<.0001
other_income	0.000027	6.88E-06	15.4273	<.0001
spouse_income	0.000026	2.55E-06	103.5814	<.0001
UnemplRate_5	-115.4	5.9185	380.3775	<.0001
Unempl_lnyoy_3	-3.4027	0.1457	545.7368	<.0001
UAH_EURRate_lnyoy_3	-2.2556	0.1738	168.4545	<.0001
b_Avg_UT13	-0.9976	0.0416	575.8202	<.0001
b_Avg_UT16	-1.1048	0.0407	735.1118	<.0001
b_AvgOB13_TO_MaxOB13	1.4041	0.0202	4815.554	<.0001
b_TRmax_deb13_To_Lim	-0.1721	0.0188	83.9031	<.0001
b_TRmax_deb13_To_avg	-0.00579	0.0012	23.2232	<.0001
b_TRavg_deb13_to_avg	0.00636	0.00235	7.3382	0.0068

Parameter	Estimate	Standard	Wald Chi-Square	Pr > ChiSq
b_TRmax_deb16_To_Lim	0.2896	0.0154	353.3622	<.0001
b_TRmax_deb16_To_avg	0.0114	0.000849	181.5597	<.0001
b_TRavg_deb16_to_avg	-0.00436	0.00239	3.3257	0.0682
b_TRsum_deb16_to_TRs	0.000011	0.000019	0.3446	0.5572
b_DeltaUT13to46	5.47E-09	4.89E-07	0.0001	0.9911
b_UT1_to_AvgUT16	-0.1414	0.00426	1103.523	<.0001
b_avgNumDeb13	-0.0235	0.00439	28.7645	<.0001
b_avgNumDeb16	-0.1525	0.00446	1168.944	<.0001
b_DeltaNumDeb13to46	-0.1095	0.00464	556.5153	<.0001
b_max_dpd13	0.00275	0.00316	0.7543	0.3851
b_max_dpd16	0.0159	0.00159	100.4742	<.0001
b_DelBucket13	0.1576	0.0442	12.7119	0.0004
Edu_High	0.1027	0.0086	142.5155	<.0001
Edu_Secondary	-0.0447	0.0101	19.6682	<.0001
Edu_TwoDegree	0.3121	0.0243	165.4246	<.0001
Marital_Civ	-0.0839	0.0172	23.838	<.0001
Marital_Sin	-0.0807	0.00958	71.0351	<.0001
Marital_Wid	0.1878	0.0197	91.1336	<.0001
position_Man	0.0638	0.0117	29.6921	<.0001
position_Tech	-0.0449	0.00907	24.4913	<.0001
position_Top	0.1375	0.0234	34.5839	<.0001
sec_Agricult	-0.0582	0.0187	9.6822	0.0019
sec_Energy	0.0459	0.0163	7.9784	0.0047
sec_Fin	0.2461	0.0133	343.0326	<.0001
sec_Manufact	-0.0707	0.0252	7.8427	0.0051
sec_Service	0.0311	0.0091	11.6527	0.0006
sec_Trade	0.0836	0.0129	41.7864	<.0001

Table 6. Logistic regression for ATM transaction model



Stage 2 – ATM - Average income for 6 month

```
proc panel data=&u.Tr_final_plus6m_atm_reg_t
outtrans=&u.r_atm_avg_panel_randtwo;
id tr_id t;
model Target_atm_avg = &RegV
/rantwo plots=all; run;
```

Variable	Pooled				Random effect			
	Estimate	Standard Error	t Value	Pr > |t|	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-442.192	12.4774	-35.44	<.0001	-223.115	5.1615	-43.23	<.0001
limit	0.003616	0.00004	89.35	<.0001	0.004743	0.000036	131.54	<.0001
b_Avg_UT13	-10.9967	1.4095	-7.8	<.0001	-17.1383	0.5292	-32.38	<.0001
b_Avg_UT16	18.64621	1.2205	15.28	<.0001	6.16316	0.5281	11.67	<.0001
b_AvgOB13_TO_MaxOB13	-35.3299	1.0402	-33.97	<.0001	-6.46975	0.3288	-19.68	<.0001
b_TRmax_deb13_To_Limit	6.600914	0.5394	12.24	<.0001	-2.44306	0.2032	-12.03	<.0001
b_TRmax_deb13_To_avgOB13	-0.40976	0.0648	-6.32	<.0001	-0.00745	0.0139	-0.54	0.5917
b_TRavg_deb13_to_avgOB13	0.170551	0.1681	1.01	0.3104	-0.01533	0.0265	-0.58	0.5633
b_TRmax_deb16_To_Limit	1.421084	0.4231	3.36	0.0008	-6.33309	0.1763	-35.91	<.0001
b_TRmax_deb16_To_avgOB16	0.347462	0.0599	5.8	<.0001	-0.05018	0.0101	-4.99	<.0001
b_TRavg_deb16_to_avgOB16	-0.03497	0.2365	-0.15	0.8825	0.030792	0.0271	1.14	0.2561
b_TRsum_deb16_to_TRsum_crd16	-0.97891	0.05	-19.56	<.0001	-0.00019	0.000204	-0.93	0.3537
b_UT1_to_AvgUT16	-3.12585	0.2138	-14.62	<.0001	-0.72892	0.0506	-14.42	<.0001
b_avgNumDeb13	0.997632	0.0792	12.6	<.0001	0.038507	0.00864	4.46	<.0001
b_avgNumDeb16	0.741181	0.0824	9	<.0001	-0.03519	0.0105	-3.35	0.0008
b_max_dpd13	0.146617	0.1274	1.15	0.25	0.039974	0.0395	1.01	0.3119
b_max_dpd16	-0.23121	0.0505	-4.58	<.0001	-0.06325	0.0208	-3.03	0.0024
b_DelBucket13	0.151839	1.5025	0.1	0.9195	-0.26153	0.5764	-0.45	0.65
Edu_High	0	.	.	.	0	.	.	.
Edu_Secondary	0.575082	0.2609	2.2	0.0275	3.807464	0.436	8.73	<.0001
Edu_Special	0.874075	0.2338	3.74	0.0002	3.024464	0.3701	8.17	<.0001
Edu_TwoDegree	0.794615	0.7758	1.02	0.3057	-1.62078	1.0363	-1.56	0.1178
Marital_Civ	0.172336	0.4287	0.4	0.6877	2.627457	0.7143	3.68	0.0002
Marital_Div	0.646814	0.3113	2.08	0.0377	1.596742	0.4993	3.2	0.0014
Marital_Mar	0	.	.	.	0	.	.	.
Marital_Sin	1.401869	0.251	5.59	<.0001	3.531679	0.4182	8.45	<.0001
Marital_Wid	-0.27384	0.5556	-0.49	0.6221	0.642042	0.8575	0.75	0.454
position_Empl	0	.	.	.	0	.	.	.
position_Man	0.265037	0.3502	0.76	0.4492	-0.79598	0.5155	-1.54	0.1226
position_Oth	0.922934	0.2831	3.26	0.0011	0.919888	0.4551	2.02	0.0433
position_Tech	1.395485	0.2502	5.58	<.0001	1.841124	0.4156	4.43	<.0001
position_Top	6.204581	0.7949	7.81	<.0001	1.001446	1.038	0.96	0.3347
sec_Agricult	-1.3189	0.7009	-1.88	0.0599	0.464309	1.1239	0.41	0.6795
sec_Constr	-1.22052	0.8014	-1.52	0.1278	-0.59565	1.3208	-0.45	0.652
sec_Energy	-2.55654	0.6644	-3.85	0.0001	-2.93138	1.0626	-2.76	0.0058
sec_Fin	-5.88534	0.6529	-9.01	<.0001	-4.73793	1.0087	-4.7	<.0001
sec_Gov	-0.87934	0.539	-1.63	0.1028	-0.04731	0.8738	-0.05	0.9568
sec_Industry	-1.99687	0.8879	-2.25	0.0245	-1.23167	1.5228	-0.81	0.4186
sec_Manufact	-3.48977	0.7701	-4.53	<.0001	-0.81731	1.3085	-0.62	0.5322
sec_Mining	-0.69692	0.6694	-1.04	0.2978	0.903281	1.1086	0.81	0.4152
customer_income	-0.00162	0.000114	-14.28	<.0001	-0.00556	0.000142	-39.15	<.0001
UnemplRate_5	6075.604	158.4	38.36	<.0001	3004.845	64.4027	46.66	<.0001
Unempl_lnyoy_3	202.8223	3.9114	51.85	<.0001	106.3556	1.594	66.72	<.0001
UAH_EURRate_lnyoy_3	356.431	4.6549	76.57	<.0001	143.9031	1.914	75.19	<.0001

Table 7. Comparison of the coefficients estimation for pooled linear regression and random-effect model for ATM withdrawn amount prediction

Pooled estimation (as OLS or GLM) and random-effect estimations give different trends and significance for the same predictors. For example, average outstanding balance to maximum balance in the current month and maximum debit turnover to the credit limit have slopes for polled regression two times less than for random effect. The impact of Unemployment ln yoy lag3m has been reduced for the random effect. The majority of characteristics became less significant (t Value for random effect less than for

pooled). Thus the panel regression can be used also for the understanding of impact of the time component.

SUMMARY OF THE NON-INTEREST INCOME FUNCTIONS PERFORMANCE

Model	Regression equation	Target	Results
Probability of POS transaction	Logistic regression $\ln\left(\frac{P_i}{1-P_i}\right) = \phi \frac{1}{T}\sum_{l=1}^{T} UT_{i(t-l)} + + \sum_{k=1}^{K}\beta_k \cdot B_{ki,t-1} + \sum_{l=1}^{L}\alpha_a \cdot A_{ai} + \sum_{m=1}^{M}\gamma_m M_{m,t-1}$	POS transaction next 6 month	ROC = 0.74
Probability of ATM withdrawal	Logistic regression $\ln\left(\frac{P_i}{1-P_i}\right) = \phi \frac{1}{T}\sum_{l=1}^{T} UT_{i(t-l)} + + \sum_{k=1}^{K}\beta_k \cdot B_{ki,t-1} + \sum_{l=1}^{L}\alpha_a \cdot A_{ai} + \sum_{m=1}^{M}\gamma_m M_{m,t-1}$	ATM withdrawal next 6 month	ROC=0.73
POS income (interchange)	Panel regression: polled $POS_i = \phi \frac{1}{T}\sum_{n=1}^{T} UT_{in} + + \sum_{k=1}^{K}\beta_k \cdot B_{ik} + \sum_{l=1}^{L}\alpha_a \cdot A_{ai} + \sum_{m=1}^{M}\gamma_m M_m$	POS Income next 6 month	R^2~0.33 – accounts with all months transactions only
POS income (interchange)	Panel regression: random-effect $POS_{it} = \phi \frac{1}{T}\sum_{n=1}^{T} UT_{i(t-n)} + + \sum_{k=1}^{K}\beta_k \cdot B_{bi,t-1} + \sum_{l=1}^{L}\alpha_a \cdot A_{ai} + \sum_{m}^{M}\gamma_1 M_{m,t-1}$	POS Income next 6 month	R^2~ 0.30 - – accounts with all months transactions only
ATM withdrawal income	Panel regression: polled $ATM_i = \phi \frac{1}{T}\sum_{n=1}^{T} UT_{in} + + \sum_{k=1}^{K}\beta_k \cdot B_{ik} + \sum_{l=1}^{L}\alpha_a \cdot A_{ai} + \sum_{m=1}^{M}\gamma_m M_m$	ATM withdrawal income next 6 month	R^2 ~ 0.32 – accounts with all months transactions only
ATM withdrawal income	Panel regression: random-effect $ATM_{it} = \phi \frac{1}{T}\sum_{n=1}^{T} UT_{i(t-n)} + + \sum_{k=1}^{K}\beta_k \cdot B_{bi,t-1} + \sum_{l=1}^{L}\alpha_a \cdot A_{ai} + \sum_{m}^{M}\gamma_1 M_{m,t-1}$	ATM withdrawal income next 6 month	R^`~ 0.32 - – accounts with all months transactions only

Table 8. Performance quality of the income prediction functions

CONCLUSION

Two innovative model building approaches were used in this research:

i) Credit cards holders' multistatus transition probabilities model which allow to estimate future income depending not only on current status, but also on possible future statuses and use the transition probability as a weight for the expected income estimation.

ii) We apply assumption that the non-income profit is generated by each customer from the number of sources and use the probability of credit card usage type models as an income amount weights.

The comparative empirical analysis of multinomial logistic regression and conditional multistage binary logistic regression has shown that both methods do not have strict preferences or advantages and both of them give satisfactory validation results of transition prediction for different types of account statuses. Conditional binary logistic regression models efficiency depending on the order of stages and lengthy. Multinomial regression gives more convenient model in use and helps to avoid the problem of stage ordering choice. However, the order it can be useful if we know what is more critical segment in sense of quality prediction. Random-effect model shows lower prediction accuracy, but the estimations are more efficient.

The further steps: To achieve the higher predictive power of the transition probabilities in multistage conditional models it is recommended to try all possible variation, then to start from the best validation results segment and then descend to the less predictive one. However, we rely on the discrete choice models such as nested logit to use for multistates transition probabilities modelling.

REFERENCES

[1] P. Volker, "A note on factors influencing the utilization." Australian University, Canberra. Econ Record September 1982, pp. 281–289.

[2] J. N. Crook, R. Hamilton and L. C.Thomas, "Credit Card Holders: Characteristics of Users and Non-Users". The Service Industries Journal, Vol. 12, No. 2 (April 1992), pp. 251-262.

[3] Bellotti T. and Crook J, "Loss Given Default models for UK retail credit cards", CRC working paper 09/1, 2009.

[4] Banasik J., Crook J., Thomas L, "Scoring by usage". Journal of the Operational Research Society, 2001, 52, 997-1006

[5] P Ma, J Crook and J Ansell. "Modelling take-up and profitability". Journal of the Operational Research Society, 2010, 61, 430-442.

[6] So, M. C., Thomas, Lyn C. and Seow, Hsin-Vonn. "Using a transactor/revolver scorecard to make credit and pricing decisions". In, Credit Scoring and Credit Control XIII, Edinburgh, GB, 28 - 30 Aug 2013.

[7] Osipenko D., Crook J. (2015). The Comparative Analysis of Predictive Models for Credit Limit Utilization Rate with SAS/STAT®. SAS Forum, 2015. Paper 3328-2015

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Denys Osipenko
The University of Edinburgh Business School
29 Buccleuch Place, Edinburgh, Lothian EH8 9JS
denis.osipenko@gmail.com