Paper 3217-2015

Credit Card Holders' Behavior Modeling: Transition Probability Prediction with Multinomial and Conditional Logistic Regression in SAS/STAT®

Denys Osipenko, the University of Edinburgh;

Professor Jonathan Crook, the University of Edinburgh

ABSTRACT

Because of the variety of card holders' behavior patterns and income sources, each consumer account can change to different states. Each consumer account can change to states such as non-active, transactor, revolver, delinquent, and defaulted, and each account requires an individual model for generated income prediction. The estimation of the transition probability between statuses at the account level helps to avoid the lack of memory in the MDP approach. The key question is which approach gives more accurate results: multinomial logistic regression or multistage decision tree with binary logistic regressions. This paper investigates the approaches to credit cards' profitability estimation at the account level based on multistates conditional probability by using the SAS/STAT procedure PROC LOGISTIC. Both models show moderate, but not strong, predictive power. Prediction accuracy for decision tree is dependent on the order of stages for conditional binary logistic regression. Current development is concentrated on discrete choice models as nested logit with PROC MDC.

INTRODUCTION

Credit card profitability prediction is a complex problem because of variety of the card holders' behaviour patterns and different sources of the interest and transactional income. Each consumer account can move to a number of states like 'transactor', 'revolver', and 'delinquent' and requires an individual model for generated income prediction. Credit cards modelling to be more reliable and accurate need to take into account revolving products dual nature both as standard loan and payment tool. Thus scoring models should be split up according to customer behaviour segment and source of generated income for the bank. The state of the credit card depends on the type of card usage and payments delinquency. Thus 5 states can be defined: inactive, transactor, revolver, delinquent, default.

The estimation of status transition probability on account level helps to avoid the memorylessness property of Markov Chains approach. Proposed credit cards profit prediction model consists of five stages: account or consumer status prediction with conditional transition probabilities, outstanding balance and interest income estimation, non-interest income estimation, expected losses estimation, and profit estimation.

The main question of this paper is which approach to prediction of multistates transition probability gives more accurate results: multinomial logistic regression or decision tree with conditional binary logistic regressions. The first stage of the profit estimation model is the determination of the account status via transition probabilities on account level. Two approaches to predict the status have been investigated: i) conditional logistic regression as Bayesian network, ii) multinomial logistic regression.

This paper describes an approach to credit cards profitability estimation on account level based on multistates conditional probabilities model. The empirical investigation presents the comparative analysis of multinomial logistic regression and conditional probabilities model in application to credit card holders behaviour modelling..

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. Application data contains consumer's socio-demographic characteristics like age, gender, education, marital status, residence, region of residence, number of family members etc., and economic factors like monthly income, sources of income, spouse income etc. Behavioural data from DWH contains dynamic information about consumer transactions, balances and delinquencies like outstanding balance at the end of month and the average monthly outstanding balance, number of debit and credit transactions, average, maximum and minimum transaction amount per month, days past due counter, arrears amount at the end of month etc. National statistical bureau data contains macroeconomic indicators in dynamics like GDP, CPI, unemployment rate, foreign currency exchange rates etc. All required behavioural and macroeconomic data is collected on monthly basis. Also the model uses constants which are originated from application and behavioural data and primary describe the product characteristics like interest rate, loan amount, loan term etc. However, 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.

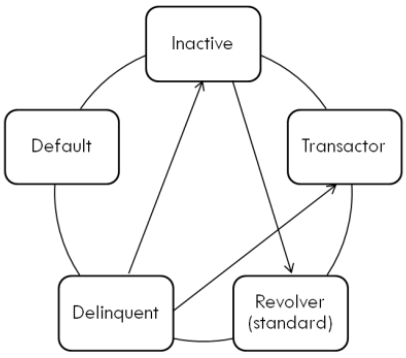

Figure 1. Transition between states

The key points of the research:

i. Credit cards processes modelling to be more reliable and accurate need to take into account revolving products dual nature both as standard loan and payment tool. Thus scoring models should be split up according to customer behaviour segment and source of generated income for the bank.

ii. The state of the credit card depends on the type of card usage and payments delinquency. Thus 5 states can be defined: inactive, transactor, revolver, delinquent, default.

iii. Ordered or multinomial logistic regression can be used for non-binary outcome estimation (instead of a decision tree of binary logistic regressions) and for transition matrix on account level

development (instead of pool level transition matrices). It used for state transition probability estimation (item ii).

The questions: which approach to prediction of multistatuses transition on account level gives more accurate results:

1) Multinomial regression, or

2) Decision tree with conditional binary logistic regressions?

Problem: In case we have more than 2 transition outcomes from the state to state it is necessary to build conditional probabilities system (Bayesian probabilities equations). However, for system of 4 statuses it will be 3-level conditional probability with 3 logistic regressions. Use of multinomial logistic regression can be reliable solution to avoid excessive complexity of the system of equations.

DATA SAMPLE

The data set for the current research contains information about credit card portfolio dynamics on account level and cardholders applications. The data sample is uploaded from the data warehouse one of the European commercial banks. 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 credit limit and interest rate time-dependent, 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 source and contain the main macroindicators such as GDP, CPI, unemployment rate, and foreign to local currency exchange rate. The data sample is available for 3 year period. The total number of accounts available for the analysis for whole lending period is 153 400, but not all accounts have been included into the data sample.

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 Table 1. List of original characteristics. The dictionary is not full.

Characteristic	Description
Behavioural characteristics (transactional) – Time Random	
b_AvgBeop13_to_AvgBeop46	Average Balance EOP in the last 3 month to Average Balance in months 4-6
b_maxdpd16	Maximum days past due in the last 6 months
b_Tr_Sum_deb_to_Crd_16	Sum of Debit transactions amounts to Credit transactions amounts for months 1-6
b_Tr_Sum_deb_to_Crd_13	Sum of Debit transactions amounts to Credit transactions amounts for months 1-3
b_Tr_Avg_deb_to_Crd_16	Average Debit transactions amounts to Average Credit transactions amounts for months 1-6
b_Tr_Avg_deb_to_Crd_13	Average Debit transactions amounts to Average Credit transactions amounts for months 1-3
b_TR_AvgNum_deb_16	Average monthly number of debit transactions for months 1-6
b_TR_AvgNum_Crd_16	Average monthly number of credit transactions for months 1-6
b_TR_MaxNum_deb_16	Maximum monthly number of debit transactions for months 1-6
b_TR_MaxNum_Crd_16	Maximum monthly number of credit transactions for months 1-6
b_TR_max_deb_to_Limit16	Amount of maximum debit transaction to limit for months 1-6
b_TR_sum_crd_to_OB13	Sum of credit transaction to average outstanding balance for month 1-3

Characteristic	Description
b_TRsum_deb16_to_TRcrd16	Sum of debit ti sum of credit transactions for month 1-6
b_NoAction_NumM_16	Number of month with no actions for months 1-6
b_NoAction_NumM_13	Number of month with no actions for months 1-3
b_pos_flag_0	POS transaction indicator for current month
b_pos_flag_13	POS transaction indicator for the previous 3 month
b_atm_flag_0	ATM transaction indicator for current month
b_atm_flag_13	ATM transaction indicator for the previous 3 month
b_pos_flag_used46vs13	POS transaction in month 4-6 but no transaction in month 1-3
b_pos_flag_use13vs46	POS transaction in month 1-3 but no transaction in month 4-6
b_atm_flag_used46vs13	ATM transaction in month 4-6 but no transaction in month 1-3
b_atm_flag_use13vs46	ATM transaction in month 1-3 but no transaction in month 4-6
No_dpd	Flag if account was in delinquency
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
Position	The position occupied by an applicant
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 original characteristics

MODEL BUILDING

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.

Credit cards dual nature and profitability researches: Crook, Hamilton, Thomas (1992), Banasik, Crook, Thomas (2001), Ma, Crook, Ansell (2010), So, Thomas (2008), Cheu, Loke (2010) , Tan, Steven, Yen (2011).

Generally the risk management approach define account on delinquency buckets like current, day past due 1-30 (Bucket 1), DPD 31-60 (Bucket 2) etc. We propose to define the credit card statuses subject to the revenue source and the revenue availability.

Account status	*Symbol*	*Definition*	*Risk assessment*	*Revenue assessment*	*Note*
closed	C	Account is closed or inactive more than 6 months	No	No	Exclude from analysis
inactive	IA	OB (1-6M) = 0 and Turnover (1-6M) = 0	No	No	No predicted revenue, but EL can be
transactor	TR	OB (1-6M) = 0 and Turnover (1-6M) <> 0	No	ProfitRate_TR*Limit	(avg interchange rate + fees rate)*TR
current	B0	OB > 0 and DPD = 0	Behavioural Score B0	Limit*ProfitRate	Beh. and Revenue Rate Scorecards for Current
delinquent	B1-3	OB > 0 and DPD > 0 and DPD <=90	Behavioural Score Delinq.	No	Beh. and Revenue Rate Scorecards for B1-3
defaulted	D	OB > 0 and DPD > 90	LGD	-	Recovery is not revenue. It's EL reduction

Table 2. Account state definition and related assessments

The main task of the first stage is to estimate the probability of transition from status to status on the client level. Transition matrix as classic approach is calculated the transition probabilities on the portfolio/pool level. However, the problem is that the number of statuses which account can move in is more than two (for example, revolver can be transactor, delinquent, stay revolver, or inactive).

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 doesn't generate interest income because of non-paid debt. However, delinquent account is not losses like defaulted one.

There are two concepts how many models we need for both approaches. Multinomial regression is more economic for computation (see Table 3. Multinomial logistic regression models covering).

Status	**To**				
From	**Non Active**	**Transactor**	**Revolver**	**Delinquent**	**Defaulted**
Non Active	Model S NA			X	X
Transactor	Model S Tr			X	X
Revolver	Model S Re				X
Delinquent	Model S Dl				
Defaulted	X	X	X	X	X

Table 3. Multinomial logistic regression models covering

On the other hand, it is not obligatory to build logistic regression model for each transition, but 'From' status can be used 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. So it is necessary to build N-1 model for each state, where N is number of possible transitions from the current state. For example, transactor can be non-active, transactor, and revolver, but cannot be delinquent of defaulted

next period and does not need a prediction model as it is shown by white cells in Table 4. Multi-stage logistic regression models covering.

Status	To				
From	**Non Active**	**Transactor**	**Revolver**	**Delinquent**	**Defaulted**
Non Active	Model_NA_NA	Model_NA_Tr	Model_NA_Re	X	X
Transactor	Model_Tr_NA	Model_Tr_Tr	Model_Tr_Re	X	X
Revolver	Model_Re_NA	Model_Re_Tr	Model_Re_Re	Model_Re_Dl	X
Delinquent	Model_Dl_NA	Model_Dl_Tr	Model_Dl_Re	Model_Dl_Dl	Model_Dl_Df
Defaulted	X	X	X	X	X

Table 4. Multi-stage logistic regression models covering

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 Figure 2. Multistage schema of the conditional logistic regression models).

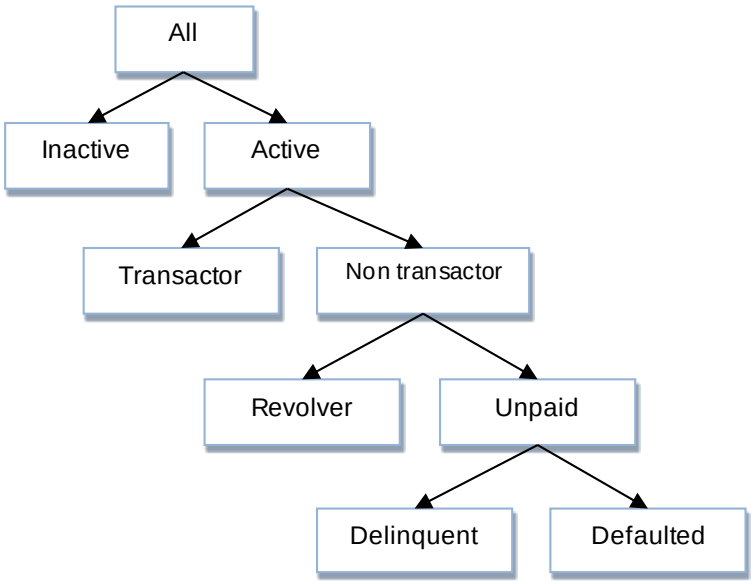

Figure 2. Multistage schema of the conditional logistic regression models

Each stage of the tree is the conditional probability of the next level transition. For full description of stage we need four equations:

$$\Pr(s = NA) = \boldsymbol{\beta}_{NA}\mathbf{x}$$

$$\Pr(s = T \mid s \neq NA) = \boldsymbol{\beta}_{T}\mathbf{x} \cdot (1 - \Pr(s = NA))$$

$$\Pr(s = R \mid s \notin (NA, T)) = \boldsymbol{\beta}_{R}\mathbf{x} \cdot (1 - \Pr(s = NA)(1 - \Pr(s = T)))$$

$$\Pr(s = Dlq \mid s \notin (NA, T, R)) = \boldsymbol{\beta}_{Dlq}\mathbf{x} \cdot (1 - \Pr(s = NA)(1 - \Pr(s = T))(1 - \Pr(s = R)))$$

where

account status s is NA – non-active, T – transactor, R – revolver, Dlq – delinquent, Def – defaulted.

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

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

,

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;
```

For other states the script is the similar, but with another start point.

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

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

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:

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

A hypothesis for an investigation is the next. This kind of the models can show weaker results than ordinal logistic regression, but better than tree of conditional logistic regressions. The estimation results generally weaker than with use of logistic regression, but that we talk about many outcomes and need to build multilevel decision tree, multinomial regression can give more confident results. The residuals (errors) on the tails of distribution for some untypical cases can be higher, but in general the model is less robust than multistage binary logistic regression.

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 5. Multinomial regression parameters estimations show that the same predictors have different correlations and even an 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
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 5. Multinomial regression parameters estimations

AN EXAMPLE OF THE PROBABILITY MODEL VALIDATION FOR MULTINOMIAL REGRESSION

We provide with the figures of the best and the worst predictive power models. The full aggregated results are in Table 6. Comparative analysis of multinomial and multistage binary logistic regression approaches.

We shoe the validation results for two stages opposite by both business cycle and predictive power

The inactive account has poor predictive power (KS=0.26, Gini=0.35) as shown in Figure 3. Kholmogorov-Smirnov characteristic for the probability to be Non-active from Non-Active, Figure 4. Gini characteristic for the probability to be Non-active from Non-Active, because generally inactive accounts have not enough behavioural history. However, for application scoring models those KS and Gini are really not poor results. But in our model all inactive customers: with history and new loans are presented.

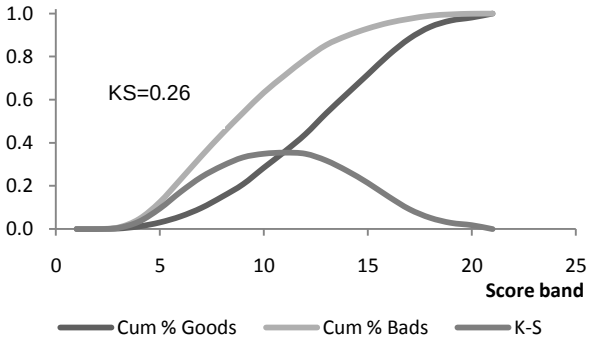

Figure 3. Kholmogorov-Smirnov characteristic for the probability to be Non-active from Non-Active

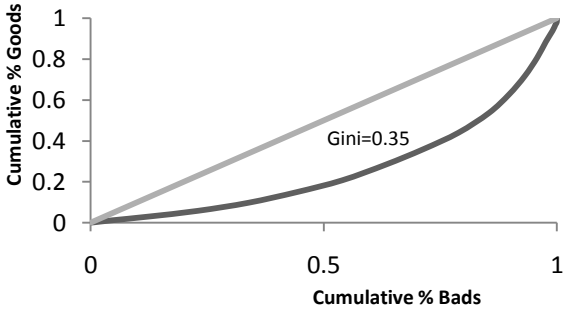

Figure 4. Gini characteristic for the probability to be Non-active from Non-Active

The delinquent account has extremely strong predictive model for the probability to be defaulted (KS=0.79, Gini=0.65) – see Figure 5. Kholmogorov-Smirnov characteristic for the probability to be Defaulted from Delinquent state, Figure 6. Gini characteristic for the probability to be Defaulted from Delinquent.

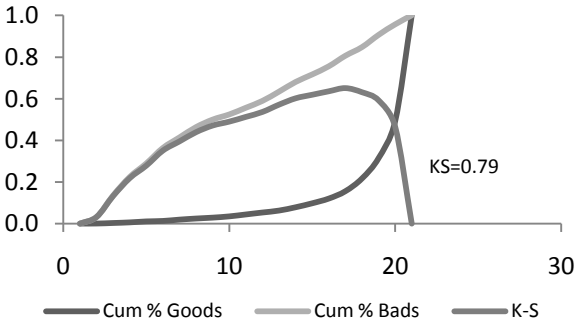

Figure 5. Kholmogorov-Smirnov characteristic for the probability to be Defaulted from Delinquent state

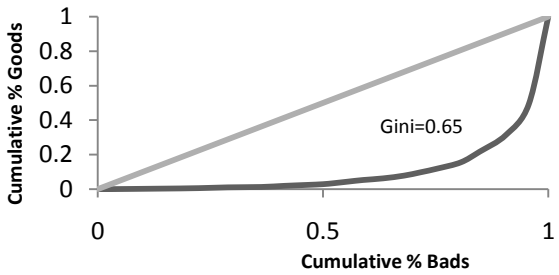

Figure 6. Gini characteristic for the probability to be Defaulted from Delinquent

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 6. 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 6. 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.

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.

Multinomial logistic regression is relatively innovative approach in risk modelling and there is lack of developed techniques for use in credit scoring. On the other hand the decision tree of conditional binary logistic regressions has given similar results. Both models have moderate, but not strong predictive power. Prediction accuracy for decision tree is depending on the order of stages for conditional binary logistic regression. An examination of possible options is complicated and long process.

The next 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, authors 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.

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