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 the 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 multinominal 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).

<table>
<thead>
<tr>
<th>Month name</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Month Num</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

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."
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoAction_NumM_16</td>
<td>Number of month with no actions for period 1-6</td>
</tr>
</tbody>
</table>

**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 \( N_{s,t+1} \) transition probabilities \( p_j \), where \( j \) is the transition index, and \( \sum_{j=1}^{N_{s,t+1}-1} p_j = 1 \),

where \( N_{s,t+1} \) is the number of states available for moving from the state \( s \) at time \( t+1 \).

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

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:
\[
I(i, t+1 | s_i = T) = \Pr(s_{i,t+1} = T | s_i \neq D) \times \\
(\Pr(a_{i,t+1} = POS | s_{i,t+1} = T) \cdot R(x_i | a = POS) + \\
+ \Pr(a_{i,t+1} = ATM | s_{i,t+1} = T) \cdot R(x_i | a = ATM) )
\]  \hspace{2cm} (1)

where \( R(.) \) is the revenue function.

Revolver income:
\[
I(i, t+1 | s_i = R) = \Pr(s_{i,t+1} = R | s_i \neq D) \times \\
\left( Ut(x_i | s_{i,t+1} = R) \times IR \times Limit \times P(s_{i,t+1} = R | s_i = R) + \\
+ \Pr(a_{i,t+1} = POS | s_{i,t+1} = R) \cdot R(x_i | a = POS) + \\
+ \Pr(a_{i,t+1} = ATM | s_{i,t+1} = R) \cdot R(x_i | a = ATM) \right)
\]  \hspace{2cm} (2)

where \( Ut(.) \) is the utilization rate function.

Delinquent income:
\[
I(i, t+1 | s_i = Dlq) = \Pr(s_{i,t+1} = Dlq | s_i = Dlq) \times \\
( \Pr(a_{i,t+1} = POS | s_{i,t+1} = Dlq) \cdot R(x_i | a = POS) + \\
+ \Pr(a_{i,t+1} = ATM | s_{i,t+1} = Dlq) \cdot R(x_i | a = ATM) )
\]  \hspace{2cm} (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

1 - Status – Transition probability on account level. Conditional logistic regression OR Multinomial regression

Initial STATUS: Inactive, Transactor, Revolver, Delinquent, Defaulted

2 - Outstanding balance – Interest Income and Expected Losses

3 - Debt Transactions – Non-interest Income

Probability to be Revolver: Logistic regression OR Multinomial

Utilization rate: Binary transformed logistic regression OR Beta-transf. x N Limit segments

Expected Losses (LGD is constant)

To Add: Nested logistic regression, Discrete Choice

Total transactional income

Total income from interest rate

Total profit

STATUS means 5 stages of the account: inactive, transactor, revolver (current), revolver (delinquent), defaulted. Probability to be in STATUS estimated by conditional probabilities with logistic regression OR multinomial regression and depends on 1) current status of the account; 2) behavioural history (dynamic); 3) application characteristics (static); 4) macroeconomic changes (dynamic).
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.).

![Decision Tree Diagram](image)

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:

\[
\begin{align*}
\Pr(s_{t+1} = NA) &= \beta_{NA}^T \mathbf{x} \\
\frac{\Pr(s_{t+1} = TR)}{1 - \Pr(s_{t+1} = NA)} &= \beta_{TR}^T \mathbf{x} \\
\frac{\Pr(s_{t+1} = RE)}{(1 - \Pr(s_{t+1} = TR))(1 - \Pr(s_{t+1} = NA))} &= \beta_{RE}^T \mathbf{x} \\
\frac{\Pr(s_{t+1} = D1 | s_i \notin (NA, TR))}{(1 - \Pr(s_{t+1} = NA))(1 - \Pr(s_{t+1} = TR))(1 - \Pr(s_{t+1} = RE))} &= \beta_{D1}^T \mathbf{x} \\
\frac{\Pr(s_{t+1} = D2 | s_i \notin (NA, TR))}{(1 - \Pr(s_{t+1} = NA))(1 - \Pr(s_{t+1} = TR))(1 - \Pr(s_{t+1} = RE))(1 - \Pr(s_{t+1} = D1))} &= \beta_{D2}^T \mathbf{x} \\
\frac{\Pr(s_{t+1} = DF | s_i \notin (NA, TR, RE, D1))}{(1 - \Pr(s_{t+1} = NA))(1 - \Pr(s_{t+1} = TR))(1 - \Pr(s_{t+1} = RE))(1 - \Pr(s_{t+1} = D1))(1 - \Pr(s_{t+1} = D2))} &= \beta_{DF}^T \mathbf{x}
\end{align*}
\]
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

$$\text{logit}(p_i) = \ln \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta \cdot x^T_i,$$

where

- $p_i$ is the probability of particular outcome,
- $\beta_0$ and $\beta$ are regression coefficients,
- $x$ are predictors.

The probability of event for $i$th observation is calculated as

$$P_i = E(Y_i | x_i, \beta) = \Pr(Y_i = 1 | x_i, \beta)$$

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 slstay equal to 0.1. The SAS code example provided for current state Transactor modeling:

```sas
/* ----------------------------- 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='1T') = &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='1T') = &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

\[
R_i = \begin{cases} 
1 & \text{if } R_i^* \leq \mu_0 \\
2 & \text{if } \mu_0 < R_i^* \leq \mu_1 \\
3 & \text{if } \mu_1 < R_i^* \leq \mu_2 \\
\vdots & \text{if } \mu_{N-1} < R_i^* \\
N & \text{if } \mu_N < R_i^*
\end{cases}
\]

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),

$X_i$ is a vector of variables that explains the variation of status; $\beta$ is a vector of coefficients; $\mu_i$ are the threshold parameters to be estimated along with $\beta$; and $\varepsilon_i$ is a disturbance term that is assumed normally distributed.

The final parameter estimation is a system of equations:

\[
\ln \frac{\Pr(Y_i = 1)}{\Pr(Y_i = K)} = \beta_1 \cdot X_i \\
\ln \frac{\Pr(Y_i = 2)}{\Pr(Y_i = K)} = \beta_2 \cdot X_i \\
\vdots \\
\ln \frac{\Pr(Y_i = K-1)}{\Pr(Y_i = K)} = \beta_{K-1} \cdot X_i
\]

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:
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 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.
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 border 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.

![Two-stage regression model schema](image)

**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.

Two-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.
Table 5. Two-stage models comparative analysis

<table>
<thead>
<tr>
<th>Month on Book</th>
<th>Limit Changes</th>
<th>Stage</th>
<th>Method</th>
<th>Development Sample</th>
<th>Validation Out-of-sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Stage 1</td>
<td>Probability</td>
<td>KS</td>
<td>Gini</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Probability</td>
<td>Logistic Regression</td>
<td>0.6262</td>
<td>0.7479</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Probability</td>
<td>Logistic Regression</td>
<td>0.5931</td>
<td>0.7243</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stage 2</td>
<td>Proportion Estimation</td>
<td>R²</td>
<td>MAE</td>
</tr>
<tr>
<td>MOB 6 or more</td>
<td>Limit NO change</td>
<td>Logistic Regression</td>
<td>OLS</td>
<td>0.4310</td>
<td>0.1948</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Logistic Regression</td>
<td>Fractional (Quasi-Likelihood)</td>
<td>0.4309</td>
<td>0.1946</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Logistic Regression</td>
<td>Beta regression (nlmixed)</td>
<td>0.4183</td>
<td>0.2102</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Logistic Regression</td>
<td>Beta transformation + OLS</td>
<td>0.3680</td>
<td>0.1802</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Logistic Regression</td>
<td>Weighted Logistic Regression</td>
<td>0.4325</td>
<td>0.1945</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Two-stage</td>
<td>Aggregate</td>
<td>R²</td>
<td>MAE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Logistic Regression</td>
<td>OLS</td>
<td>0.5534</td>
<td>0.1913</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Logistic Regression</td>
<td>Fractional (Quasi-Likelihood)</td>
<td>0.5527</td>
<td>0.1915</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Logistic Regression</td>
<td>Beta regression (nlmixed)</td>
<td>0.5366</td>
<td>0.2068</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Logistic Regression</td>
<td>Beta transformation + OLS</td>
<td>0.4720</td>
<td>0.1773</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Logistic Regression</td>
<td>Weighted Logistic Regression</td>
<td>0.5548</td>
<td>0.1914</td>
</tr>
</tbody>
</table>

NON-INTEREST RATE PROFITABILITY MODELLING

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

```sql
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:

\[ y_{it} = \alpha + X_{it}'\beta + Z_{it}'\gamma + u_{it}, \quad i = 1,...,N, \quad t = 1,...,T \]

- \( X \) is observed factors vector;
- \( Z \) is unobserved factors vector, \( Z_{it} = Z_{i,t} \);
- \( y_{it} \) is number of cases;
- \( T \) is number of time periods;
- \( \beta \) and \( \gamma \) - regression slope coefficients;
- \( u_{it} = \mu_{i,t} + \lambda_{i} + \upsilon_{it} \);
- \( \mu_{i,t}, \lambda_{i} \) – non-observed individual and time effects, \( \upsilon_{it} \) – residual idiosyncratic components.

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

\[ y_{it} = X_{it}'\beta + \alpha + \epsilon_{it} \]

- \( \alpha \) и \( \beta \) – 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

\[
\ln \left( \frac{P}{1-P} \right) = \phi \frac{1}{T} \sum_{t=1}^{T} UT_{i(t-1)} + \sum_{k=1}^{K} \beta_k \cdot B_{i_k(t-1)} + \sum_{l=1}^{L} \alpha_l \cdot A_{i_l} + \sum_{m=1}^{M} \gamma_m M_{m,t-1}
\]

2nd stage – income amount for the period

\[
POS_{it} = \phi \frac{1}{T} \sum_{n=1}^{T} UT_{i(t-n)} + \sum_{k=1}^{K} \beta_k \cdot B_{i_k(t-n)} + \sum_{l=1}^{L} \alpha_l \cdot A_{i_l} + \sum_{m=1}^{M} \gamma_m M_{m,t-1}
\]

\[
ATM_{it} = \phi \frac{1}{T} \sum_{n=1}^{T} UT_{i(t-n)} + \sum_{k=1}^{K} \beta_k \cdot B_{i_k(t-n)} + \sum_{l=1}^{L} \alpha_l \cdot A_{i_l} + \sum_{m=1}^{M} \gamma_m M_{m,t-1}
\]

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

```plaintext
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;
```

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard</th>
<th>Wald Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>9.2935</td>
<td>0.4657</td>
<td>398.215</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>limit</td>
<td>-0.00005</td>
<td>1.46E-06</td>
<td>1073.447</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>customer_income</td>
<td>0.000103</td>
<td>3.54E-06</td>
<td>842.0063</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>other_income</td>
<td>0.000027</td>
<td>6.88E-06</td>
<td>15.4273</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>spouse_income</td>
<td>0.000026</td>
<td>2.55E-06</td>
<td>103.5814</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>UnemplRate_5</td>
<td>-115.4</td>
<td>5.9185</td>
<td>380.3775</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Unempl_lnyoy_3</td>
<td>-3.4027</td>
<td>0.1457</td>
<td>545.7368</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>UAH_EURRate_lnyoy_3</td>
<td>-2.2556</td>
<td>0.1738</td>
<td>168.4545</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>b_Avg_UT13</td>
<td>-0.9976</td>
<td>0.0416</td>
<td>575.8202</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>b_Avg_UT16</td>
<td>-1.1048</td>
<td>0.0407</td>
<td>735.1118</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>b_AvgOB13_TO_MaxOB13</td>
<td>1.4041</td>
<td>0.0202</td>
<td>4815.554</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>b_TRmax_deb13_To_Lim</td>
<td>-0.1871</td>
<td>0.0188</td>
<td>83.9031</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>b_TRmax_deb13_To_avg</td>
<td>-0.00579</td>
<td>0.0012</td>
<td>23.2232</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>b_TRavg_deb13_to_avg</td>
<td>0.00636</td>
<td>0.00235</td>
<td>7.3382</td>
<td>0.0068</td>
</tr>
<tr>
<td>Parameter</td>
<td>Estimate</td>
<td>Standard Error</td>
<td>Wald Chi-Square</td>
<td>Pr &gt; ChiSq</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>----------</td>
<td>----------------</td>
<td>-----------------</td>
<td>------------</td>
</tr>
<tr>
<td>b_TRmax_deb16_To_Lim</td>
<td>0.2896</td>
<td>0.0154</td>
<td>353.3622</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>b_TRmax_deb16_To_avg</td>
<td>0.0114</td>
<td>0.000849</td>
<td>181.5597</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>b_TRavg_deb16_to_avg</td>
<td>-0.00436</td>
<td>0.00239</td>
<td>3.3257</td>
<td>0.0682</td>
</tr>
<tr>
<td>b_TRsum_deb16_to_TRs</td>
<td>0.000011</td>
<td>0.000019</td>
<td>0.3446</td>
<td>0.5572</td>
</tr>
<tr>
<td>b_DeltaUT13to46</td>
<td>5.47E-09</td>
<td>4.89E-07</td>
<td>0.0001</td>
<td>0.9911</td>
</tr>
<tr>
<td>b_UT1_to_AvgUT16</td>
<td>-0.1414</td>
<td>0.00426</td>
<td>1103.523</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>b_avgNumDeb13</td>
<td>-0.0235</td>
<td>0.00439</td>
<td>28.7645</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>b_avgNumDeb16</td>
<td>-0.1525</td>
<td>0.00446</td>
<td>1168.944</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>b_DeltaNumDeb13to46</td>
<td>-0.1095</td>
<td>0.00464</td>
<td>556.5153</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>b_max_dpd13</td>
<td>0.00275</td>
<td>0.00316</td>
<td>0.7543</td>
<td>0.3851</td>
</tr>
<tr>
<td>b_max_dpd16</td>
<td>0.0159</td>
<td>0.00159</td>
<td>100.4742</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>b_DelBucket13</td>
<td>0.1576</td>
<td>0.0442</td>
<td>12.7119</td>
<td>0.0004</td>
</tr>
<tr>
<td>Edu_High</td>
<td>0.1072</td>
<td>0.0086</td>
<td>142.5155</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Edu_Secondary</td>
<td>-0.0447</td>
<td>0.0101</td>
<td>19.6682</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Edu_TwoDegree</td>
<td>0.3121</td>
<td>0.0243</td>
<td>165.4246</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Marital_Civ</td>
<td>-0.0839</td>
<td>0.0172</td>
<td>23.838</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Marital_Sin</td>
<td>-0.0807</td>
<td>0.00958</td>
<td>71.0351</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Marital_Wid</td>
<td>0.1878</td>
<td>0.0197</td>
<td>91.1336</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>position_Man</td>
<td>0.0638</td>
<td>0.0117</td>
<td>29.6921</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>position_Tech</td>
<td>-0.0449</td>
<td>0.00907</td>
<td>24.4913</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>position_Top</td>
<td>0.1375</td>
<td>0.0234</td>
<td>34.5839</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>sec_Agricult</td>
<td>-0.0582</td>
<td>0.0187</td>
<td>9.6822</td>
<td>0.0019</td>
</tr>
<tr>
<td>sec_Energy</td>
<td>0.0459</td>
<td>0.0163</td>
<td>7.9784</td>
<td>0.0047</td>
</tr>
<tr>
<td>sec_Fin</td>
<td>0.2461</td>
<td>0.0133</td>
<td>343.0326</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>sec_Manufact</td>
<td>-0.0707</td>
<td>0.0252</td>
<td>7.8427</td>
<td>0.0051</td>
</tr>
<tr>
<td>sec_Service</td>
<td>0.0311</td>
<td>0.0091</td>
<td>11.6527</td>
<td>0.0006</td>
</tr>
<tr>
<td>sec_Trade</td>
<td>0.0836</td>
<td>0.0129</td>
<td>41.7864</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Table 6. Logistic regression for ATM transaction model

![ROC Curve for Selected Model](image)

Stage 2 – ATM - Average income for 6 month

```r
proc panel data=\u.Tr_finalplus6m_atm_reg_t;
outtrans=\u.r_atm_avg_panel_randtwo;
id tr_id t;
model Target_atm_avg = &RegV/rantwo plots=all; run;
```
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 pooled 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**

<table>
<thead>
<tr>
<th>Model</th>
<th>Regression equation</th>
<th>Target</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of POS transaction</td>
<td>$\ln \left( \frac{P_i}{1-P_i} \right) = \phi \frac{1}{T} \sum_{t=1}^{T} UT_{i(t-i)} + \sum_{k=1}^{K} \beta_k \cdot B_{ik,t-1} + \sum_{l=1}^{L} \alpha_{il} \cdot A_{il} + \sum_{m=1}^{M} \gamma_m M_{m,t-1}$</td>
<td>POS transaction next 6 month</td>
<td>ROC = 0.74</td>
</tr>
<tr>
<td>Probability of ATM withdrawal</td>
<td>$\ln \left( \frac{P_i}{1-P_i} \right) = \phi \frac{1}{T} \sum_{t=1}^{T} UT_{i(t-i)} + \sum_{k=1}^{K} \beta_k \cdot B_{ik,t-1} + \sum_{l=1}^{L} \alpha_{il} \cdot A_{il} + \sum_{m=1}^{M} \gamma_m M_{m,t-1}$</td>
<td>ATM withdrawal next 6 month</td>
<td>ROC = 0.73</td>
</tr>
<tr>
<td>POS income (interchange)</td>
<td>Panel regression: polled ($POS_i = \phi \frac{1}{T} \sum_{t=1}^{T} UT_{i} + \sum_{k=1}^{K} \beta_k \cdot B_{ik} + \sum_{l=1}^{L} \alpha_{il} \cdot A_{il} + \sum_{m=1}^{M} \gamma_m M_{m}$)</td>
<td>POS Income next 6 month</td>
<td>$R^2$ ~ 0.33 – accounts with all months transaction s only</td>
</tr>
<tr>
<td>POS income (interchange)</td>
<td>Panel regression: random-effect ($POS_i = \phi \frac{1}{T} \sum_{t=1}^{T} UT_{i(t-i)} + \sum_{k=1}^{K} \beta_k \cdot B_{ik,t-1} + \sum_{l=1}^{L} \alpha_{il} \cdot A_{il} + \sum_{m=1}^{M} \gamma_m M_{m,t-1}$)</td>
<td>POS Income next 6 month</td>
<td>$R^2$ ~ 0.30 – accounts with all months transaction s only</td>
</tr>
<tr>
<td>ATM withdrawal income</td>
<td>Panel regression: polled ($ATM_i = \phi \frac{1}{T} \sum_{t=1}^{T} UT_{i} + \sum_{k=1}^{K} \beta_k \cdot B_{ik} + \sum_{l=1}^{L} \alpha_{il} \cdot A_{il} + \sum_{m=1}^{M} \gamma_m M_{m}$)</td>
<td>ATM withdrawal income next 6 month</td>
<td>$R^2$ ~ 0.32 – accounts with all months transaction s only</td>
</tr>
<tr>
<td>ATM withdrawal income</td>
<td>Panel regression: random-effect ($ATM_i = \phi \frac{1}{T} \sum_{t=1}^{T} UT_{i(t-i)} + \sum_{k=1}^{K} \beta_k \cdot B_{ik,t-1} + \sum_{l=1}^{L} \alpha_{il} \cdot A_{il} + \sum_{m=1}^{M} \gamma_m M_{m,t-1}$)</td>
<td>ATM withdrawal income next 6 month</td>
<td>$R^2$ ~ 0.32 – accounts with all months transaction s only</td>
</tr>
</tbody>
</table>

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


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

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Other brand and product names are trademarks of their respective companies.