Banking retail consumer finance data generator - credit scoring data repository

Karol Przanowski
Warsaw School of Economics - SGH
Institute of Statistics and Demography
Event History Analysis and Multilevel Analysis Unit
ul.Madalinskiego 6/8
02-513 Warszawa
email: kprzan@sgh.waw.pl
url:
www.sgh.waw.pl/zaklady/zahziaw/english/
kprzan.w.interia.pl

Abstract

This paper presents two cases of random banking data generators based on migration matrices and scoring rules. The banking data generator is a new hope in researches aimed at finding the method to compare various credit scoring techniques. These data are very useful for various analyses to understand the complexity of the banking processes in a better way and also for students and their researches. The influence of one cyclic macro–economic variable on client characteristics and their stability over time is analyzed. Some stimulating conclusions for crisis behavior are presented, namely that if a crisis is impacted by both factors: application and behavioral then it is very difficult to clearly indicate these factors in a typical scoring analysis and the crisis becomes noticeable everywhere, in every kind of risk report.

Key words: credit scoring, crisis analysis, banking data generator, retail portfolio, scorecard building, predictive modeling.

Contents

1 Introduction

Currently predictive models and especially credit scoring models are very popular in the management of banking processes [Hua07]. It is generally the case that risk scorecards are used in credit acceptance processes to optimize and control any risk. Various forms of behavioral scorecards are also used for the management of repeat business (cross-up sell) and also for PD (probability of default) models in Basel RWA (Risk Weighted Assets) calculation [oBS05b]. It is often sufficient to obtain a list of about 10 account or client characteristics which, when combined, can predict their future behavior, their style of payments and their delinquency.

One may add that trivial fact scorecards are of use and their methodology is well known, but on the other hand credit scoring can be still developed and new techniques should always be tested. The main problem today is that there is no defined general testing ideology for new methods and techniques; there is no proven method to gauge their correctness. Many good articles are prepared based on one particular case study, on one example of real data coming from either one or more than one bank [Tho09], [HS07] and [Maj10]. From a theoretical point of view, even if good results are reached and sound arguments put forward to suggest choosing one method instead of another, these results are usually reached using that particular data which indicate the difference, but other data will often suggest a different conclusion; nobody can guarantee correctness for all cases.

Other important reasons remain for why real banking data are not available globally and it cannot be used by every analyst, namely reasons such as: legal constraints or new products with too short a data history. These two factors suggest finding a quite new approach for predictive modeling testing in banking usage.

It seems a sensible idea to start developing two parallel ways: real data and random–simulated data approaches. Even if the second one cannot replace real data, it can be very useful to understand better the relations among various factors in data; to imagine the complexity of the process and it can be an attempt to create a more general class of semi–real data.

Let us consider some of the advantages of random data:

1. Today many analysts try to understand and to analyze the most recent crisis [May09], among other things they develop methods of indicating

risk stable in sub–portfolios that remain stable over time. Is not an easy topic and it cannot be solved by typical predictive models based on target variables as in the case of a default risk. The notion of stability cannot be defined for every particular account or client. It is difficult to state that an account is stable, only the set of accounts can be tested, so this technique should be developed by a quite different method than a typical predictive modeling with target variable. It can be formulated by the simple conclusion: the more accounts, the more robust the stability testing. Various scenarios based on the random data generator can be tested to see and to understand the problem better.

2. Scoring Challenges or Scoring Games. From time to time different competitions are organized by various institutions: banks, universities, consultancy companies or associations; to find good modelers, or to test new techniques. Sometimes, data are used that have too many "real cases". Here "too real" means that some real processes are unpredictable, because they are influenced by many immeasurable factors. Even if scoring models are used in practice in these cases, it is not a good idea to use that data for competition. The best solution and the best choice is to use a random data generator directly predictable process.

3. Reject inference area [HS07]. This topic requires further development. Random data can also be generated in reality for testing rejected cases, so it can be used for better estimation of risk on missed area and in order to gain experience.

4. Today there are two or more techniques of scorecard building [Sid05]. We need to make some comparisons and carry out some analysis to define recommendations: it must be clear when to use one method rather than another. The same situation can be applied for different variable selection methods.

5. Product profitability, bad debts and cut–offs. For random data all the notions mentioned can be tested and analysts experience can be broadened.

6. Random data can also be a very important factor in the topic of data standardization or the idea of auditing. Let us imagine that the soft-

ware tools for MIS (Management Information Systems) and KPI (Key Performance Indicators), reporting on the generic data structure, which has firstly been uploaded by random data are already running. Then the auditing of any other data will be minimized only by the upload data process.

Simulation data are used in many areas, for example it is of use in the research of a telecommunication network by a system like OPNET [Inca]. Some simulated data in the banking area by [Sup09] and [Wat81] are also developed.

The simplest retail consumer finance portfolio is the fixed installment loan portfolio. Here the process can be simplified by the following assumptions:

- for all accounts one due date in the middle of the month is defined (every 15th),
- every client has only one loan,
- a client can pay the whole installment, a few installments or pay nothing. These can be categorized as either payment or no payment,
- there are measured delinquencies on states: at the end of the month by indicating the number of due installments,
- all customer and account properties are randomly generated by defined proper random distributions,
- if the number of due installments reaches 7 (180 past due days) the process is stopped and an account is given a bad account status; any further collection steps are omitted,
- if a number of paid installments reaches the total number of all Installments then the process is stopped and an account is given a closed account status,
- payments or missing payments are determined by three factors: the score calculated on account characteristics, migration matrix and adjustment of that matrix by one macro–economic variable time cycle,
- a score is calculated separately for every due installments group. In more general cases a different score for every status can be defined: due installments 0, 1, ..., and 6.

It is worth noticing that risk management today has very good tools for risk control. Even if the current crisis has occurred and was not predicted in the correct way, it could have been indicated promptly. It seems that the best tool of risk control is the migration matrix reporting.

The goal of this paper can also be formulated in the following way: to create random data with the aim of obtaining the same results as those observed in reality using typical reporting like migration matrix, flow-rates or roll-rates and vintage or default rates.

2 Detailed description of data generator

2.1 The main options

All data are generated from starting date T_s to ending T_e.

The migration matrix M_{ij} (transition matrix) is defined as a percent of one month transition from due installments i to due installments j.

There is only one macro–economic variable dependent on the time described by the formula: $E(m)$, where m is the number of months from T_s. It should fulfill the following simple condition: $0.01 < E(m) < 0.9$, because it is used as an adjustment of migration matrix, so it has an influence on the risk; in some months it produces a slightly greater risk and in some months a lower one.

2.2 Production dataset

The first dataset contains all applications with all available customer characteristics and credit properties.

Customer characteristics (application data):

- Birthday - T_{Birth} – with the distribution D_{Birth}
- Income – x^a_{Income} - D_{Income}
- Spending - $x^a_{Spending}$ – $D_{Spending}$

- Four nominal characteristics – $x^a_{Nom_1}, ..., x^a_{Nom_4}$ - $D_{Nom_1}, D_{Nom_2}, ..., D_{Nom_4}$, in practice they can represent variables such as: job category, marital status, home status, education level, or others.
- Four interval characteristics - $x^a_{Int_1}, ..., x^a_{Int_4}$ – $D_{Int_1}, D_{Int_2}, ..., D_{Int_4}$, represent variables such as: job seniority, personal account seniority, number of households, household spending or others.

Credit properties (loan data):

- Installment amount – x^l_{Inst} - with the distribution D_{Inst}
- Number of installments – $x^l_{N_{inst}}$ - $D_{N_{inst}}$
- Loan amount – $x^l_{Amount} = x^l_{Inst} \cdot x^l_{N_{inst}}$
- Date of application (year, month) – T_{app}
- Id of application

The number of rows per month is generated based on the distribution $D_{Applications}$.

2.3 Transaction dataset

Every row contains the following information (transaction data):

- Id of application
- Date of application (year, month) – T_{app}
- Current month – T_{cur}
- Number of due installments (number of missing payments) – $x^t_{n_{due}}$
- Number of paid installments – $x^t_{n_{paid}}$
- Status – x^t_{status} - Active (A) - remains unpaid, Closed (C) is paid, or Bad (B) – when $x^t_{n_{due}} = 7$
- Pay days – x^t_{days} – number of days from the interval $[-15, 15]$ before or after due date in a current month when payment was made, if there is missing payment, then pay days equal to missing value.

2.4 Inserting the Production dataset into the Transaction dataset

Every month of the Production dataset updates the Transaction dataset with the following formulas:

$$T_{cur} = T_{app}, \quad x^t_{n_{due}} = 0, \quad x^t_{n_{paid}} = 0, \quad x^t_{status} = A, \quad x^t_{days} = 0.$$

This is the process of inserting the starting points of new accounts.

2.5 Analytical Base Table – ABT dataset

The history of payments for every account is dependent on behavioral data, or, in other words, on the behavior of previous payments. This is, of course, the assumption of the data generator.

There are many theories on how to create behavioral characteristics. Here are presented some simple methods to consider their last available time stamps (actual states) and to indicate their evaluations over time. All data are prepared in ABT datasets, the notion of Analytical Base Table (ABT) is used by SAS Credit Scoring Solution [Incb].

Let us set the current date T_{cur} as a fixed value. The actual states are calculated for that date by the formulas (actual data):

$$\begin{aligned}
x^{act}_{days} &= x^t_{days} + 15, \\
x^{act}_{n_{paid}} &= x^t_{n_{paid}}, \\
x^{act}_{n_{due}} &= x^t_{n_{due}}, \\
x^{act}_{utl} &= x^t_{n_{paid}} / x^l_{N_{inst}}, \\
x^{act}_{dueutl} &= x^t_{n_{due}} / x^l_{N_{inst}}, \\
x^{act}_{age} &= years(T_{Birth}, T_{cur}), \\
x^{act}_{capacity} &= (x^l_{Inst} + x^a_{Spending}) / x^a_{Income}, \\
x^{act}_{dueinc} &= (x^t_{n_{due}} \cdot x^l_{Inst}) / x^a_{Income}, \\
x^{act}_{loaninc} &= x^l_{Amount} / x^a_{Income}, \\
x^{act}_{seniority} &= T_{cur} - T_{app} + 1,
\end{aligned}$$

where $years()$ calculates the difference between two dates in years.

Let us consider two time series of pay days and due installments for the last 11 months from a fixed current date by the formulas:

$$
\begin{aligned}
x^{act}_{days}(m) = &\quad x^{act}_{days}(T_{cur} - m), \\
x^{act}_{n_{due}}(m) = &\quad x^{act}_{n_{due}}(T_{cur} - m),
\end{aligned}
$$

where $m = 0, 1, ..., 11$.

The characteristics indicated by the evaluation over the time can be calculated by the formulas:

If all the elements of the time series for the last t-months are available then (behavioral data):

$$
\begin{aligned}
x^{beh}_{days}(t) = &\quad (\textstyle\sum_{m=0}^{t-1} x^{act}_{days}(m))/t, \\
x^{beh}_{n_{due}}(t) = &\quad (\textstyle\sum_{m=0}^{t-1} x^{act}_{n_{due}}(m))/t,
\end{aligned}
$$

where $t = 3, 6, 9, 12$.

If all the elements of the time series are not available then (missing imputation formulas):

$$
\begin{aligned}
x^{beh}_{days}(t) = &\quad 15, \\
x^{beh}_{n_{due}}(t) = &\quad 2.
\end{aligned}
\tag{2.1}
$$

In other words, behavioral variables represent average states for the previous 3, 6, 9 or 12 months. Without any difficulties a user can add many other variables by replacing the average statistic by another like MAX, MIN or other.

2.6 Migration matrix adjustment

Macro–economic variable $E(m)$ influences the migration matrix by the formula:

$$
M^{adj}_{ij} = \begin{cases}
M_{ij}(1 - E(m)) & \text{for} \quad j \leq i, \\
M_{ij} & \text{for} \quad j > i + 1, \\
M_{ij} + \sum_{k=0}^{i} E(m) M_{ik} & \text{for} \quad j = i + 1.
\end{cases}
$$

2.7 Iteration step

This step is running to generate the next month of transactions, from T_{cur} to $T_{cur} + 1$. For new accounts the Transaction dataset is only updated by

the ideas described in subsection 2.4. Other accounts, which are not new, change the status by the formula:

$$x^t_{status} = \begin{cases} C & \text{when} \quad x^{act}_{n_{paid}} = x^l_{N_{inst}}, \\ B & \text{when} \quad x^{act}_{n_{due}} = 7, \end{cases}$$

and these accounts are not continued in the next months.

For other active accounts in the next month there are two events which may be generated: payment or missing payment. This is based on two scorings:

$$\begin{aligned} Score_{Main} = & \sum_\alpha \beta^a_\alpha x^a_\alpha + \sum_\gamma \beta^l_\gamma x^l_\gamma + \sum_\delta \beta^{act}_\delta x^{act}_\delta \\ + & \sum_\eta \sum_t \beta^{beh}_\eta(t) x^{beh}_\eta(t) + \beta_r \varepsilon + \beta_0, \qquad (2.2) \\ Score_{Cycle} = & \sum_\alpha \phi^a_\alpha x^a_\alpha + \sum_\gamma \phi^l_\gamma x^l_\gamma + \sum_\delta \phi^{act}_\delta x^{act}_\delta \\ + & \sum_\eta \sum_t \phi^{beh}_\eta(t) x^{beh}_\eta(t) + \phi_r \epsilon + \phi_0, \qquad (2.3) \end{aligned}$$

where $t = 3, 6, 9, 12$, $\alpha = Income, Spending, Nom_1, ..., Nom_4, Int_1, ..., Int_4$, $\gamma = Inst, N_{Inst}, Amount$, $\eta = days, n_{due}$, $\delta = days, n_{paid}, n_{due}, utl, dueutl$, $age, capacity, dueinc, loaninc, seniority$, ε and ϵ are taken from the standardized normal distribution N.

Let us consider the following migration matrix:

$$M^{act}_{ij} = \begin{cases} M^{adj}_{ij} & \text{when} \quad Score_{Cycle} \leq \text{Cutoff}, \\ M_{ij} & \text{when} \quad Score_{Cycle} > \text{Cutoff}, \end{cases}$$

where Cutoff is another parameter like all βs and ϕs.

For fixed T_{cur} and fixed $x^{act}_{n_{due}} = i$ all active accounts can be segmented by $Score_{Main}$ to satisfy the same proportions such as the appropriate elements of migration matrix M^{act}_{ij}: the first group $g = 0$ by the highest scores having share equals to M^{act}_{i0}, the second $g = 1$ having share M^{act}_{i1}, ..., and the last group $g = 7$ share – M^{act}_{i7}.

For a particular account assigned to the group g the payment is done in month $T_{cur} + 1$ when $g \leq i$, in other cases payment is considered missing.

For any missing payment Transaction dataset is updated by the following information:

$$x^t_{n_{paid}} = x^{act}_{n_{paid}},$$
$$x^t_{n_{due}} = g,$$
$$x^t_{days} = \text{Missing}.$$

For existing payment by formulas:

$$x^t_{n_{paid}} = \min(x^{act}_{n_{paid}} + x^{act}_{n_{due}} - g + 1, x^l_{N_{inst}}),$$

$$x^t_{n_{due}} = g,$$

and x^t_{days} are generated from the distribution D_{days}.

The steps described are repeated for all months between T_s and T_e.

2.8 Default definition

A Default is a typical credit scoring and Basel II notion. Every account from the observation point T_{cur} which is tested during the outcome period equals 3, 6, 9 and 12 months. During this time the maximal number of due installments is analyzed, namely:

$$\text{MAX} = \text{MAX}^{t-1}_{m=0}(x^{act}_{n_{due}}(T_{cur} + m)),$$

where $t = 3, 6, 9, 12$. Dependent on the value of MAX there are defined three values of default statuses Default_t:

Good: When $\text{MAX} \leq 1$ or during the outcome period was $x^t_{status} = C$.

Bad: When $\text{MAX} > 3$ or during the outcome period $x^t_{status} = B$. In the case $t = 3$ when $\text{MAX} > 2$.

Indeterminate: for all other cases.

The existence of Indeterminate status can at times be questionable. In some analysis only two statuses are preferable, for example in Basel II. This may be a good topic for further research and can be solved due to the data generator described in this paper.

2.9 Portfolio segmentation and risk measures

Typically credit scoring is used for the control of the following sub–portfolios or processes:

Acceptance process – APP portfolio: This is the set of all starting points of credits, where it is decided which ones are accepted or rejected. Acceptance sub–portfolio is defined as the set of rows of Transaction dataset with the condition: $T_{cur} = T_{app}$. Every account belongs to that set only once.

Cross–up sell process – BEH portfolio: This is the set of all accounts with a history longer than 2 months and in a good condition (without delinquency). Cross–up sell or Behavioral sub–portfolio is defined as the set of

rows of Transaction dataset with the condition: $x^{act}_{seniority} > 2$ and $x^{act}_{n_{due}} = 0$. Every account can belong to that set more than once for different observation points T_{cur}.

Collection process – COL portfolio: This is the set of all accounts with the delinquency, but at the beginning of the collection process. Collection sub–portfolio is defined as the set of rows of Transaction dataset with the condition: $x^{act}_{n_{due}} = 1$. Every account can belong to that set more than once.

For every sub–portfolio mentioned one can calculate and test risk measurements called bad rates, defined as the share of **Bad** statuses for every observation point and outcome period.

Definitions of mentioned sub–portfolios in reality can be more complex. Reference to simpler versions of the cases studies presented in section 4 are suggested for further analysis

3 General theory

3.1 The main assumption and definition

Definition. The layout

$$(T_s, T_e, M_{ij}, E(m), \beta^a_\alpha, \beta^l_\gamma, \beta^{act}_\delta, \beta^{beh}_\eta(t), \beta_r, \beta_0,$$

$$\phi^a_\alpha, \phi^l_\gamma, \phi^{act}_\delta, \phi^{beh}_\eta(t), \phi_r, \phi_0, \varepsilon, \epsilon, D_{Birth}, D_\alpha, D_\gamma, D_{Applications}, D_{days}, \text{Cutoff})$$

with all the rules and symbols, relations and processes described in the section 2 is called **The Retail Consumer Finance Data Generator in the case of fixed installment loans** with the abbreviation **RCFDG**.

Theorem – assumption. Every consumer finance portfolio with the fixed installment loans can be estimated using **RCFDG**.

The theorem can be always done correctly due to parts: $\beta_r\varepsilon$ and $\phi_r\epsilon$ in the formulas 2.2 and 2.3. From an empirical point of view credit scoring is always used in portfolio control, so the above–mentioned theorem can be considered correct, but the problem is with the goodness of fit. For the time being it is too early to define a good measures of fit. However, it is a proper

starting point in the next development of the general theory of consumer finance portfolios.

The similar ideas and researches are presented in [Tho09].

3.2 Open questions

The next steps probably will concentrate on:

- Finding the correct goodness of fit statistics measuring the distance between the real consumer finance portfolio and **RCFDG**. The properties of these statistics should also be tested.
- Analyzing the additional constraints to satisfy for example properties such as: the predictive power, measured for example by Gini [oBS05a], of characteristic $x^{beh}_{days}(3)$ on Default_6 should be equaled to 40%.
- Creating more general case for all collection processes, more than one loan per customer, more than one macro–economic factors and other detailed issues.
- Analyzing various existing real consumer finance portfolios and finding the set of parameters describing each of them. Only then can the theory of principal component analysis (PCA) for all consumer finance portfolios in a particular country or in the world be developed.
- Defining the notion of a consumer finance portfolio which contains almost all the properties of real portfolios (generalization of the notion).
- Using that notion in researches on the development of scoring methods in order to use that notion as a general idea of method proving. For example, the theorem: *Scoring models build on* $Default_3$ *and on* $Default_{12}$ *produce the same results* can be solved by the additional condition: betas for $t = 3$ and for $t = 12$ should be similar. It is very probable that any future researches will discover many properties and relations among betas, as well as the coefficients of the migration matrix and their consequences.

4 Two case studies

4.1 Common parameters

All random numbers are based on two typical random generators: uniform U and standardized normal N distributions, in detail: the distribution U returns a number from the interval $(0, 1)$ with equal probability.

All common coefficients present as follows: $T_s = 1970.01$ (January 1970), $T_e = 1976.12$ (December 1976),

$$M_{ij} = \begin{bmatrix} & j=0 & j=1 & j=2 & j=3 & j=4 & j=5 & j=6 & j=7 \\ i=0 & 0.850 & 0.150 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ i=1 & 0.250 & 0.450 & 0.300 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ i=2 & 0.040 & 0.240 & 0.190 & 0.530 & 0.000 & 0.000 & 0.000 & 0.000 \\ i=3 & 0.005 & 0.025 & 0.080 & 0.100 & 0.790 & 0.000 & 0.000 & 0.000 \\ i=4 & 0.000 & 0.000 & 0.010 & 0.080 & 0.090 & 0.820 & 0.000 & 0.000 \\ i=5 & 0.000 & 0.000 & 0.000 & 0.000 & 0.020 & 0.030 & 0.950 & 0.000 \\ i=6 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.010 & 0.010 & 0.980 \end{bmatrix},$$

$E(m) = 0.01 + (1.5 + \sin((5 \cdot \pi \cdot m)/(T_e - T_s)) + N/5)/8$, $D_{Applications} = 300 \cdot 30 \cdot (1 + N/20)$, if T_{app} is December then $D_{Applications} = D_{Applications} \cdot 1.2$. To define D_{Birth} distribution of age is defined first: $D_{Age} = ((75 - 18) \cdot (N + 4)/7 + 10 + 20 \cdot U)$ if $Age > 75$ then $Age = 75$, if $Age < 18$ then $Age = 18$. $D_{Birth} = T_{app} - D_{Age} \cdot 365.5$, $D_{Income} = int((10000 - 500)/40 \cdot 10 \cdot abs(N) + 500)$, $D_{Inst} = int(Income \cdot abs(N)/4)$, $D_{Spending} = int(Income \cdot abs(N)/4)$, $D_{N_{Inst}} = int(30 \cdot abs(N)/4 + 6)$ if $N_{Inst} < 6$ then $N_{Inst} = 6$, $D_{Nom_i} = int(5 \cdot abs(N))$ and $D_{Int_i} = 10 \cdot U$, for $i = 1, 2, 3, 4$, if $x^{act}_{n_{due}} < 2$ then $D_{days} = -int(15 \cdot (abs(N)/4))$ else $D_{days} = int(15 \cdot (N/4))$, where $int()$ and $abs()$ – integer value and absolute value are suitable.

To avoid a scale or unit problem for every individual variable it is suggested to make a simple standardization step for any ABT table for every T_{cur} before score calculation. This idea is quite realistic, because even some customers who are reliable may experience more problems during a time of crisis. So the general condition of the current month can influence all customers. On the other hand, in order to present two interesting cases it has been decided to standardize the variables by the global parameters.

Scoring formula for $Score_{Main}$ is calculated basing on the table 1, namely:

$$Score_{Main} = \sum_{index=1}^{28} \beta(x - \mu)/\sigma.$$

Table 1: Scoring formula for $Score_{Main}$.

Index	x – variable	μ	σ	β
1	$x^a_{Nom_1}$	3.5	3	1
2	$x^a_{Nom_2}$	3.5	3	2
3	$x^a_{Nom_3}$	3.5	3	1
4	$x^a_{Nom_4}$	3.5	3	3
5	$x^a_{Int_1}$	5	2.89	1
6	$x^a_{Int_2}$	5	2.89	-4
7	$x^a_{Int_3}$	5	2.89	1
8	$x^a_{Int_4}$	5	2.89	-2
9	x^{act}_{days}	13	2.42	-5
10	x^{act}_{utl}	0.36	0.28	-4
11	$x^{act}_{due utl}$	0.12	0.2	-6
12	$x^{act}_{n_due}$	1.3	2	-2
13	x^{act}_{age}	53	9.9	4
14	$x^{act}_{capacity}$	0.4	0.21	-2
15	$x^{act}_{due inc}$	0.3	0.6	-1
16	$x^{act}_{loaninc}$	2.4	2.1	-2
17	x^a_{Income}	2395	1431	2
18	x^l_{Amount}	5741	6804	-1
19	$x^l_{N_{inst}}$	12.3	4.63	-4
20	$x^{beh}_{n_{due}}(3)$	1.4	1.6	-4
21	$x^{beh}_{days}(3)$	14.15	1.4	-6
22	$x^{beh}_{n_{due}}(6)$	1.6	1.13	-5
23	$x^{beh}_{days}(6)$	14.57	1.02	-6
24	$x^{beh}_{n_{due}}(9)$	1.78	0.75	-5
25	$x^{beh}_{days}(9)$	14.78	0.72	-6
26	$x^{beh}_{n_{due}}(12)$	1.89	0.48	-5
27	$x^{beh}_{days}(12)$	14.91	0.49	-6
28	ε	0	0.02916	1

All beta coefficients can be recalculated without the standardization step, but in that case it would be more difficult to interpret them. By a simple study of table 1 it can be indicated that the most significant variables have an absolute value equal to 6.

4.2 The first case study – unstable application characteristic – APP

In this case it is assumed that only customers with low incomes will be influenced by a crisis. Application characteristic income in the data generator is a stable variable during this time, and the migration matrix is adjusted by

the macro–economic $E(m)$ only for cases:

$$x^{a}_{Income} < 1800.$$

The relation presented can easily be transformed into a general form 2.3.

4.3 The second case study – unstable behavioral characteristic – BEH

Here, the condition for migration matrix adjustment is as follows:

$$x^{beh}_{n_{due}}(6) > 0 \quad \text{and} \quad x^{act}_{seniority} > 6,$$

the rule for the seniority variable is added to the unadjusted accounts with the missing imputation based on 2.1. This case presents a situation when a crisis has an impact on customers who have experienced a delinquency during the previous 6 months.

4.4 Stability problem

Let us consider the typical scoring models building process, for example on the behavioral sub–portfolio. Because both cases are based on two variables: one application and one behavioral, let only the set of these two variables be considered. To indicate the extreme instability of the models they are being analyzed with the target variable Default_9.

Every variable is segmented or binned for the attributes described in tables 2 and 3.

In the case of an unstable application variable (APP) studying figure 6 confirms, what may be expected, that attribute 2 is very stable during this time and accounts from that group are not oversensitive to crisis changes. In contrary, attribute 1 is very unstable. The same groups, in the case of unstable behavioral variable (BEH), are both unstable, see figure 7. The same group accounts from attribute 2, which are presented in figure 5, allow both cases to indicate in a better way that APP case can really choose accounts that are not sensitive to a crisis. Even data generator is a simplification of the real data, a conclusion that is extremely useful. Some application data can be profitable in risk management to indicate sub–segments with a stable risk over time.

Quite different conclusions can be formulated for the behavioral variable $x^{beh}_{n_{due}}(6)$. In figure 3 risk evolutions for three attributes of that variable are presented . All of them are unstable. The most stable attribute is number 3. In the case of BEH, that attribute is also unstable, see figure 4. In an attempt to prove this only attributes 3 for both cases are also presented in figure 2. The reader may say that both cases have unstable risk. Even in the case of BEH attribute 3 is expected to have a stable risk, however, due to the rule of migration matrix adjustment, our expectations have not been met. The reason can be found in the correct understanding of the process. A typical scoring approach is based on the principal idea that information available up to the observation point is able to predict the behavior during the outcome period. Up to the observation point if an account has not had any delinquency, so the variable $x^{beh}_{n_{due}}(6) = 0$. After that point that account may have due installments in the next months. It may be adjusted by the macro–economic variable with the result that the group can become unstable.

This idea is very important for further research of the crisis. It should be emphasized that typical scoring methods used on three types of sub–portfolios: APP, BEH and COL cannot reveal in the correct way the rule of crisis adjustment and cannot indicate some sub–segments that are stable over time. Of course, scoring can also be used just as in this paper for the prediction of migration states; namely to be precise, not for default statuses prediction but for transition prediction. The best method is probably the survival analysis [BC09] or [Cro08] with time covariates (time dependent variables), where in a natural way the factor of being a better or a worse payer is indicated in a set period of time, namely, in the typical scoring model the factor is considered only up to the observation point. In the survival model, however it can be also taken into the account after that observation point, so in what may be considered a more realistic way.

Many other cases of data generators with more complex rule for $Score_{Cycle}$ are made. If both types of variables: application and behavioral are taken together then the case becomes too complicated and there is widespread unstable property. In that case it is not possible to find a stable factor. That conclusion is also very important for crisis analysis, because it describes the nature of crisis: if it is a major event and it has an impact on both types of characteristics: behavioral and application, then the risk management can only try to find some sub–segments more stable than others or with a maximum risk not exceeding the expected boundary.

Table 2: Simple binning for two variables in the case APP.

Characteristic	Attribute number	Condition	Bad rate on Default$_9$	Population percent	Gini on Default$_9$
$x^{beh}_{n_{due}}(6)$	1	$x^{act}_{seniority} < 6$	16.77%	37.09%	51.34%
	2	$x^{beh}_{n_{due}}(6) > 0$ and $x^{act}_{seniority} \geq 6$	6.48%	22.49%	
	3	otherwise	1.07%	40.42%	
x^{a}_{Income}	1	$x^{a}_{Income} < 1800$	20.11%	18.32%	36.29%
	2	$x^{a}_{Income} \geq 1800$	4.72%	81.68%	

Table 3: Simple binning for two variables in the case BEH.

Characteristic	Attribute number	Condition	Bad rate on Default$_9$	Population percent	Gini on Default$_9$
$x^{beh}_{n_{due}}(6)$	1	$x^{act}_{seniority} < 6$	19.49%	40.05%	46.54%
	2	$x^{beh}_{n_{due}}(6) > 0$ and $x^{act}_{seniority} \geq 6$	14.04%	16.52%	
	3	otherwise	1.74%	43.43%	
x^{a}_{Income}	1	$x^{a}_{Income} < 1800$	12.09%	39.49%	5.04%
	2	$x^{a}_{Income} \geq 1800$	10.09%	60.51%	

4.5 Various types of risk measures

Let us define crisis as a time where risk is the highest. The most popular reporting for risk management is based on bad rates, vintage and flow rates. Figure 1 presents bad rates for three different sub–portfolios application, behavioral and collection. One flow rate is also presented. There is a simple conclusion to be drawn, that crisis does not occur at the same time. Some curves indicate local maximum risk earlier than others. The difference in time is significant and can be as much as 6 months, so it is very important to remember the nature of reports that can indicate a crisis as quickly as possible. It should be emphasized that bad rates reports present, in a standard way, the evaluation of risk by observation points and a crisis time can occur between the observation point and the end of outcome period. It seems that flow rates report indicates the crisis time in better way.

4.6 Implementation

All data were prepared by the SAS System [Incb] by manual codes written in SAS 4GL used units: Base SAS and SAS/STAT. For the case of unstable behavioral variable – BEH: Production dataset has 779 993 rows (about

Figure 1: Risk measures on Default_9 comparison on sub–portfolios: APP, BEH and COL and also with one flow rate M_{23}.

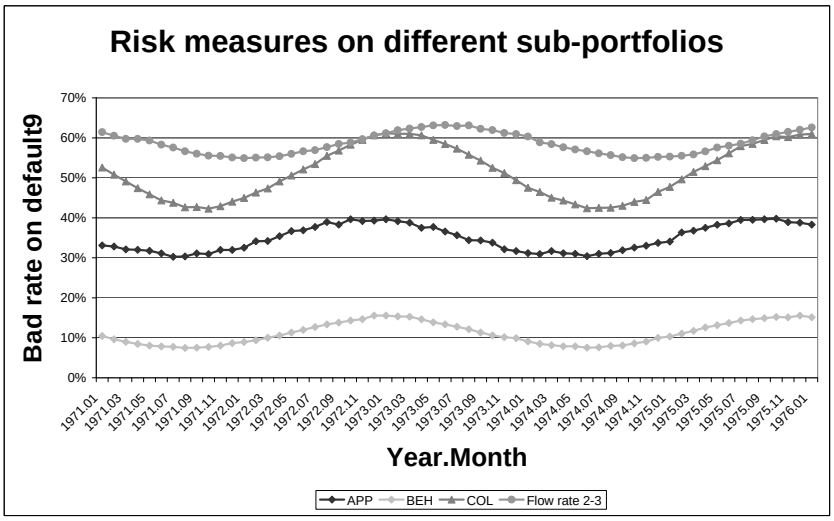


Figure 2: Risk measures on Default_9 on attribute 3 of a variable $x^{beh}_{n_{due}}(6)$ for two cases APP and BEH.

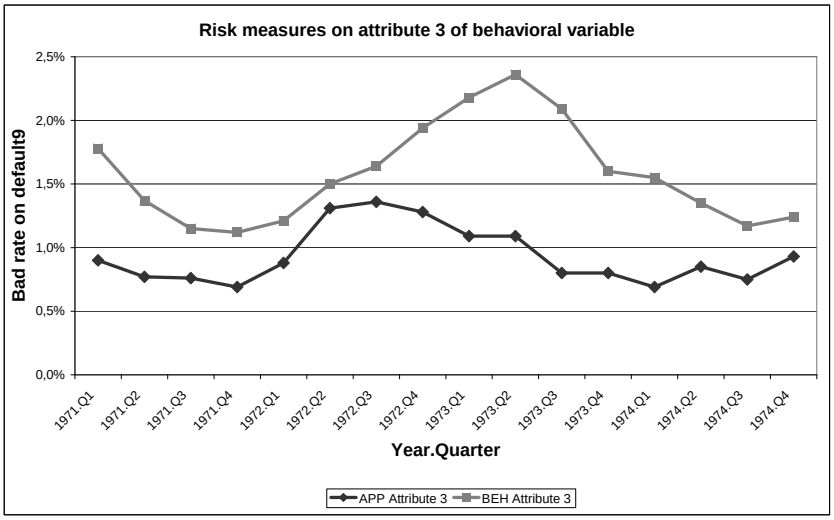


Figure 3: Risk measures on Default_9 on attributes of a variable $x^{beh}_{n_{due}}(6)$ for the case APP.

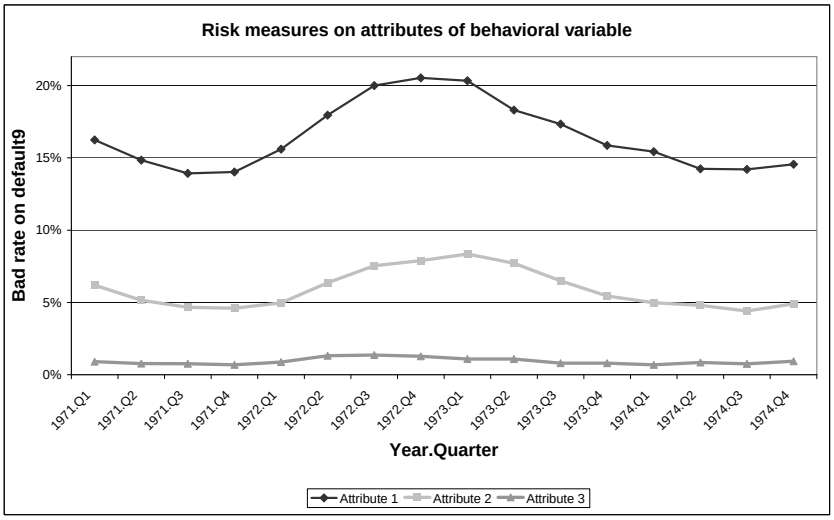


Figure 4: Risk measures on Default_9 on attributes of a variable $x^{beh}_{n_{due}}(6)$ for the case BEH.

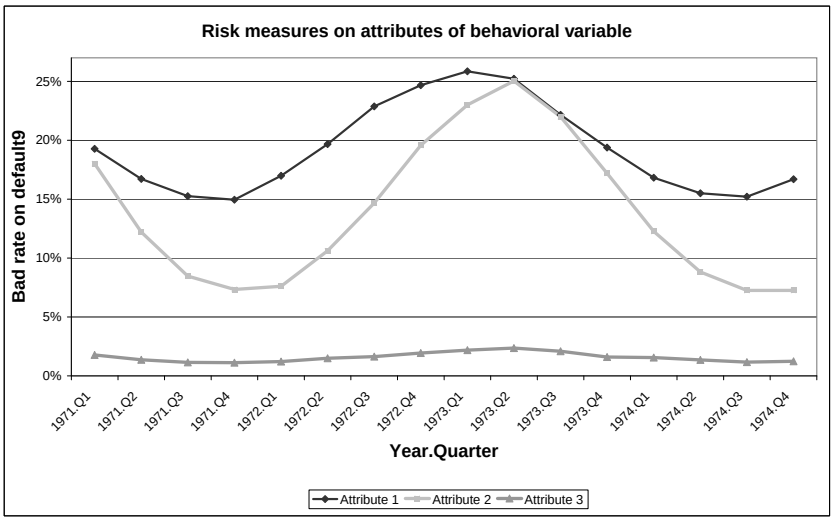


Figure 5: Risk measures on Default_9 on attribute 2 of a variable x^a_{Income} for two cases APP and BEH.

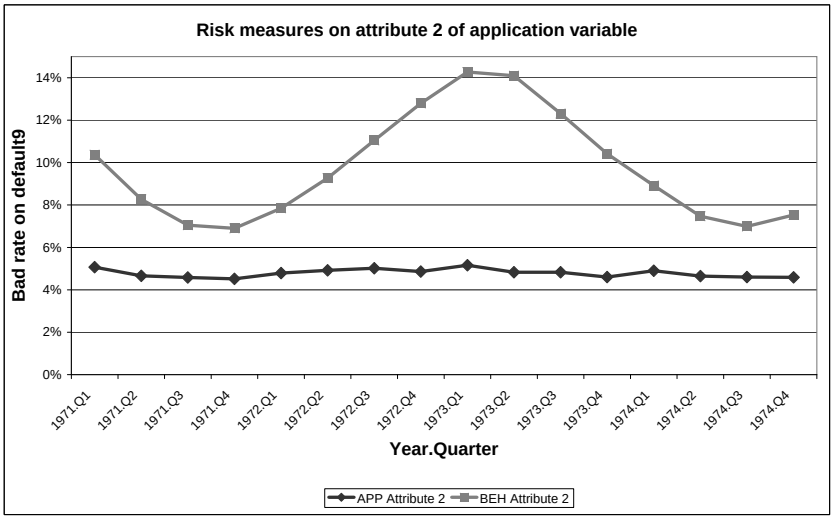


Figure 6: Risk measures on Default_9 on attributes of variable x^a_{Income} for the case APP.

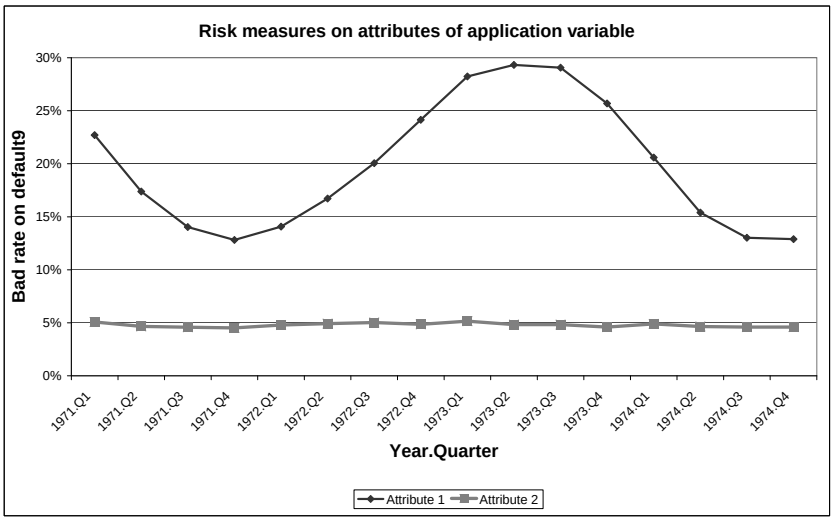
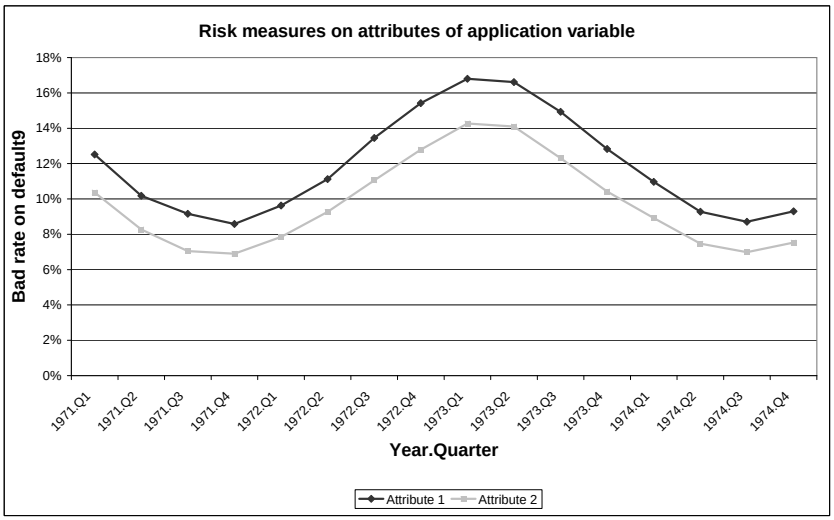


Figure 7: Risk measures on Default_9 on attributes of variable x^a_{Income} for the case BEH.



90MB) and Transaction dataset – 8 969 413 rows (about 400MB). Total time of calculation per case takes about 4 hours.

5 Conclusions

Even if data are generated by a random–simulated process, which is not realistic, the conclusions give the possibility to better understand the nature of the crisis.

The banking data generator is a new hope in researches aimed at finding the method of comparisons of various credit scoring techniques. It is probable that in the future many randomly generated data will become the new repository for testing and comparisons.

In the first case, an unstable application variable like income is possible to split portfolio into two parts: a stable and an unstable one over a period of time. In the second case, an unstable behavioral characteristic, the task is more complicated and it is not possible to split it in the same way. Some sub–

segments may have better stability but they always fluctuate. Moreover, if a crisis is impacted by many factors, both from an application from customer characteristics and from a customer behavioral, it is very difficult to indicate these factors and the crisis is widespread in all reports.

The generated data are very useful for various analyses and researches. There are many rows, many bad default statuses, so an analyst can make many good exercises to improve his experience.

References

[BC09] Tony Bellotti and Jonathan Crook. Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, 60:1699–1707, 2009.

[Cro08] Jonathan Crook. Dynamic consumer risk models: an overview. *Credit Scoring Conference CRC, Edinburgh*, 2008. `http://www.crc.man.ed.ac.uk/publications/papers/`.

[HS07] Edward Huang and Christopher Scott. Credit risk scorecard design, validation and user acceptance: A lesson for modellers and risk managers. *Credit Scoring Conference CRC, Edinburgh*, 2007. `http://www.crc.man.ed.ac.uk/publications/papers/`.

[Hua07] Edward Huang. Scorecard specification, validation and user acceptance: A lesson for modellers and risk managers. *Credit Scoring Conference CRC, Edinburgh*, 2007. `http://www.crc.man.ed.ac.uk/publications/papers/`.

[Inca] OPNET Technologies Inc. `http://www.opnet.com`.

[Incb] SAS Institute Inc. `http://www.sas.com`.

[Maj10] Izabela Majer. Application scoring: logit model approach and the divergence method compared. *Warsaw School of Economics – SGH, Working Paper No. 10-06*, 2010.

[May09] Elizabeth Mays. Systematic risk effects on consumer lending products. *Credit Scoring Conference CRC, Edinburgh*, 2009. `http://www.crc.man.ed.ac.uk/publications/papers/`.

[oBS05a] Basel Committee on Banking Supervision. Validation of internal rating systems. *Working Paper No. 14*, February 2005. `http://www.bis.org`.

[oBS05b] Basel Committee on Banking Supervision. International convergence of capital measurement and capital standards. *A Revised Framework*, Updated November 2005. `http://www.bis.org`.

[Sid05] Naeem Siddiqi. Credit risk scorecards: Developing and implementing intelligent credit scoring. *Wiley and SAS Business Series*, 2005.

[Sup09] Bala Supramaniam, Mahadevan; Shanmugam. Simulating retail banking for banking students. *Reports – Evaluative, Practitioners and Researchers ERIC Identifier: ED503907*, 2009.

[Tho09] Madhur Malik & Lyn C Thomas. Modelling credit risk in portfolios of consumer loans: Transition matrix model for consumer credit ratings. *Credit Scoring Conference CRC, Edinburgh*, 2009. `http://www.crc.man.ed.ac.uk/publications/papers/`.

[Wat81] H. J. Watson. Simulating retail banking for banking students. *Computer simulation in business. New York: John Wiley & Sons.*, 1981.