Paper 2524-2015

Methodology of Model Creation

Mgr. Peter Kertys, VÚB a. s.

ABSTRACT

The goal of this session is to describe the whole process of model creation from the business request through model specification, data preparation, iterative model creation, model tuning, implementation, and model servicing. Each mentioned phase consists of several steps in which we describe the main goal of the step, the expected outcome, the tools used, our own SAS codes, useful nodes, and settings in SAS® Enterprise Miner™, procedures in SAS® Enterprise Guide®, measurement criteria, and expected duration in "man-days." For three steps, we also present deep insights with examples of practical usage, explanations of used codes, settings, and ways of exploring and interpreting the output. During the actual model creation process, we suggest using Microsoft Excel to keep all input metadata along with information about transformations performed in SAS Enterprise Miner. To get faster information about model results, we combine an automatic SAS® code generator implemented in Excel, and then we input this code to SAS Enterprise Guide and create a specific profile of results directly from the nodes output tables of SAS Enterprise Miner. This paper also focuses on an example of a binary model stability check-in time performed in SAS Enterprise Guide through measuring optimal cut-off percentage and lift. These measurements are visualized and automatized using our own codes. By using this methodology, users would have direct contact with transformed data along with the possibility to analyze and explore any semi-results. Furthermore, the proposed approach could be used for several types of modeling (for example, binary and nominal predictive models or segmentation models). Generally, we have summarized our best practices of combining specific procedures performed in SAS Enterprise Guide, SAS Enterprise Miner, and Microsoft Excel to create and interpret models faster and more effectively.

INTRODUCTION

There are plenty of documents describing the process of model creation or data preparation but when it comes to actual model creation, you find need of going through a series of steps and try not to forget anything. We try to summarize all these steps in an "algorithm"-like structure from the very beginning of business request through model servicing and implementation. This paper is written from a retail point of view with customer centric orientation. Methods we use are fully usable for predictive models with binary target, however almost every part could be used for other types of models, including cluster analysis. In the paper we assume the reader has appropriate knowledge of statistical terms and methods used for modeling in SAS®.

OVERVIEW ON METHODOLOGY

The presented methodology of modeling consists from 7 phases described below. Specific steps are described on the following pages; here is the summary of phases with main goals. Along with a description of each phase we provide short demonstration of usage when creating binary model for churn prediction. To avoid abuse, some details are blurred.

BUSINESS REQUEST

In this phase the main goal is a YES/NO decision whether a specific business idea is worth modeling. This decision is based on expected results, costs of model development, business case on usage of model, added value to customer knowledge and potential risks during modeling and small feasibility study.

Steps: Business request, Study of materials by analyst, Request clarification, Feasibility study, Decision.

Example: Business wants to decrease churn of clients. We studied several materials and concluded we should create a binary model. After several meetings we agreed that the model should predict probability to churn (probability to leave) for clients with income above 500€. Feasibility study has shown that if we

would prevent churn of 10% clients, the activity would be profitable. Business decided that they wanted to build and use the model.

MODEL SPECIFICATION

In this phase we need to understand customer behavior associated with model request. Afterwards, we define all relevant data for modeling, i.e. definition of target, list of variables we want to use, definition of time scopes and identification whether we need to create new variables/measurements criteria, size of target group.

Steps: Process understanding, Variables and target, Snapshot and dataset, Variable check.

Example: We researched motivations of clients that want to churn. We identified several "moments of truth" which we could derive from the data. We decided to use 9 months of history of behavior of these clients to predict the target.

DATA PREPARATION

This phase consists of actual data creation including programming variables, merging datasets into a consolidated Analytical Base Table (further as "ABT") on client/product, definition of roles and levels for use in SAS Enterprise Miner (EM), statistical analysis, replacement of missing/extreme values and variable transformations (e.g. binning, standardization). Optional part of data preparation is simple segmentation of dataset based on several variables according to target.

Steps: ABT creation, Dataset stability check, ABT analysis, Roles & levels, Distribution exploration, Missing values, Extreme values, Segmentation.

Example: We created a specific SQL script to generate variables which are not in our regular ABT. Then we merged these datasets and added the target. Afterwards we ran summary statistics along with some specific measures. Then we identified variables which are irrelevant and for the others we defined roles and levels.

ITERATIVE MODEL CREATION

Here comes the core of modeling – model creation by using nodes in SAS EM. In this phase we begin with prepared ABT and perform model creation by sampling, variable selection, correlation analysis, variable transformations, regression, decision trees and model evaluation. To achieve maximum model performance, we usually iterate the process several times with various node settings and therefore we obtain dozens of different results. Some of these results are used just for variable selection or to get a basic idea what outcome we could expect. Others are considered as pre-final models which we test in the following phase.

Steps: Sampling, Transformations, Correlations, Worth on target, Merge variables, 2nd stage of clustering/transformations, Model creation.

Example: We split the dataset into training and validation parts. Afterwards we run all suggested imputations, transformations and correlation analyses. Next we run variable clustering with different settings followed by various regression models. To identify the most relevant variables we also used decision trees. Finally, we got 4 candidate models.

MODEL TUNNING

To get an actionable predictive model we need to select one which would be stable in time and still has high precision. From the previous phase we have several models which now we compare through model comparison. To avoid an "over-trained" model and sampling disproportions we also create various combinations of train & test datasets with different random seeds. Finally, we test the model on "out of sample" data. According to all these measures we could now select a final model.

Steps: Model stability check, Out of sample test, Un-sampling, Creation of scoring rule.

Example: In this phase we checked created model on datasets from different snapshots. We found that model no.3 has the most stable results and then we created a scoring rule based on it.

IMPLEMENTATION

Most of a data-miner's job ends before implementation with model tuning. However, we think data-miners and data analysts along with CRM departments should work together on implementation and therefore they need to define effective cut-off for models with probabilities. The next step is to perform a pilot project to validate hypothesis of model strength. For further usage, newly created variables should be programmed and automated on a snapshot basis (e.g. monthly, weekly). Update of variable catalogue (document with information about all variables in ABT except newly create variables) and documentation of model and learned best practices could be useful as in every other project. After a successful pilot we get the model fully automated and migrate it into production.

Steps: Cut-off, Pilot, Update ABT & variable catalogue, Documentation/Presentation, Migration and automation

Example: To achieve maximum precision, we decided to set the cut-off on probability to churn above 24%. During the following 3 months we performed a pilot project on usage of this model. As the activity based on the model has a positive outcome, we then implemented the model into production. In the preparation phase we created 3 new variables. After implementation we have updated our variable catalogue along with the addition of these variables to automation process of ABT creation.

MODEL SERVICING

As we now have a fully implemented model we need to regularly check its stability as well as precision. Another important thing is to keep an eye on stability of input variables and target. After some time of using the model we also need to think about recalibration or redefining target due to market changes and development.

Steps: Stability check, Recalibration.

Example: After 5 months of usage of this model the precision rapidly decreased. We found it was due to two variables which became obsolete as there was a change in product coding. Due to this reason we needed to recalibrate the model omitting these kinds of variables.

USAGE OF METHODOLOGY

Once we have a general overview of all of the phases of modeling and tools we could start describing how to use it. Usually we go step by step and we use it as checklist to not forget anything. An important part is also tracking the duration of each step. In Table 1 in the Appendix we state detailed description of each step of the methodology. There are also all mentioned steps with expected outcomes, node description/setting, name of used node/procedure and corresponding table from the SAS EM project. Stated settings of nodes are informative and based on general suggestions or our experience. We use this as guideline, the decision of which one setting to apply is on the data-miner.

By using this methodology in our environment we obtained following improvements:

- We have improved and clarified the decision process on model creation and it helped the business to understand data analyst's point of view.
- We reduced misunderstood assignments.
- We could plan model production in a more precise way and estimate a real expectation of model delivery.
- We are able to produce models faster along with deeper statistic insight.
- We are ready to migrate any produced models into production and in a short time we could compute expected results of activity that should be based on model prediction.
- We reduced the risk of information loss as the process is mapped and could easily be adopted by new employees.

SUPPORTIVE TOOLS

To maximize performance of each phase we use 3 different tools for the whole process described above – MS Excel, SAS Enterprise Guide (EG) and SAS EM. On these tools we have built another 6 important "things" which helps the whole process run faster: Metadata model documentation, ABT, Variable catalogue, Variable stability test, our own macro function for computing summary statistics (MST macro), Models scoring and stability functions.

METADATA MODEL DOCUMENTATION

When creating the model we begin with a huge dataset which we analyze. Then we transform it and afterwards we want to create a profile of obtained results. To not get lost in all kinds of transformations we suggest having a summarized Metadata model documentation in MS Excel. We could always look there and find what transformations were made to the data and where we could also generate automatic codes for usage in SAS EG or SAS EM. By using these codes we could define roles faster, not by clicking on each variable in SAS EM´s "Metadata" node. Another use is for creating profiles and statistics in SAS EM directly from SAS EM´s output tables.

ABT (DATASET)

On a regular basis we create the analytical base table of about 1000 variables/characteristics for each client including aggregated values for 3,6,12 months. We compute this table with data valided at the end of the month.

VARIABLE CATALOGUE

The variable catalogue is a list of variables in the ABT with detailed information containing their names, labels, description, link on script by which they are computed, type of variable (nominal, interval, binary), validity flag (some of variables can become unstable due to change in source systems or implementation of new channels), flag of usage in other models and other important notes for keeping track of all available variables.

VARIABLE STABILITY TEST

To be ready for modeling at any time we regularly test all variables in the variable catalogue to record changes in distribution (e.g. if new channel is implemented, channel score of each client increase and if this variable is used in model, we need to recalibrate the model).

MST MACRO

The MST macro (based on PROC MEANS) is our own function which performs an initial complex analysis of all input variables and computes basic statistics including suggestions of unusable variables for modeling.

MODEL SCORING AND STABILITY FUNCTIONS

After we put the model into production we use a project in SAS EG for model scoring and stability check. Here we compute several basic measures oriented on model stability (e.g. Lift, distributions of predicted probabilities or precision of prediction).

HIGHLIGHTS OF SOME STEPS

In this part we describe in detail some parts of using the methodology above. First we look closer at the Metadata documentation mentioned in supportive tools. Secondly, we provide an example of Stability check of a binary model. Finally, we suggest setting up optimal cut-off based on model precision.

METADATA MODEL EVIDENCE

As the model creation takes usually more than one day, we make an evidence of all executed steps in our MS Excel Metadata document. Into the first 2 rows we write date of performed analysis and order in which

columns were added. In this document we have in one row the following information for each variable used in the original dataset:

- Variable name
- Variable description
- Type of variable (binary/nominal/interval)
- Characteristic of variable (numeric/character)
- Count & % of distinct values
- Count & % of missing values
- Mean, N, Min, Max, P5, P10, P25, P50, P75, P90, P95, Skewness
- Variable Role & Label for use in SAS EM
- Code for setting variable role & level
- After these basic statistics we fill the next columns with the following information which may differentiate model from model (for column naming we use the source table from Enterprise Miner)
    - Name after replacement node
    - Name after grouping node
    - Flag notification of rejection by correlation node
    - Name after transformation node
- For profiling the results we could use codes for boxplot creation, which is described at the end

To get a faster overview of this document we use different colors for specific columns: grey for binary flags, orange for SAS codes, blue for computed values, green for manual inserted values. This table may look as Figure 1 below.

	A	B	C	D	E	F	G
1	Date	13.3.	13.3.	13.3.	13.3.	13.3.	13.3.
2	Order	1	2	5	10	11	12
3	No.:	Name	Label	NOMINAL	Level	Role	Code
4	198	cif_cust_marital_status	Marital status of client	1	nominal	input	%em_metachange(name=cif_cust_marital_status, role=input, level=nominal);
5	254	fl_employer	Flag indicating client is employed.	1	nominal	input	%em_metachange(name=fl_employer, role=input, level=nominal);

Figure 1 Example of Metadata evidence

We use this metadata evidence during the whole process of modeling. After the model is finished it is easier to create model documentation as we know which steps we did and in what order.

STABILITY CHECK

Once after we create the model we need to regularly check its stability and precision as the market or behavior of customers change. First we should check source data for unexpected errors. Afterwards we check the precision of prediction and compare related measures to see whether the model is still performing well. We keep the computed measures in the “stability table” and for better overview we plot it into a graph. In this section we present an example of stability check of binary churn model. If the model and market is stable we expect similar results in consequent periods (months) of scoring.

In Figure 2 we display stability and performance of these measures:

- Overall churn rate on whole base: we use this to see whether the clients churn in the same way as in previous months. In some ways it is just standard reporting but we display it along with other measures to keep all related information in one place.
- Churn rate on top (5%,10%,15%,20%) clients with highest probability to churn: by using the model we want to prevent clients with highest probability from churning by performing agreed

activities. Usually before we start some activity we check whether the model is precise in these groups. Sometimes we want to target only small groups, sometimes larger and therefore we display different sizes of groups from top 5% to top 20% of clients sorted from highest probability. In these categories we measure churn rate and if it is stable, we could assume that in the following period (month) there would be a similar churn rate. We expect the churn to scale down after performing the activity.

- Lift of model on top 10% clients with highest probability to churn: this is probably the most popular measure of model precision and easy to explain for business managers as we could say something like “our model is 3 times better than random choice”. In this case we are interested in lift on the top 10% of clients.

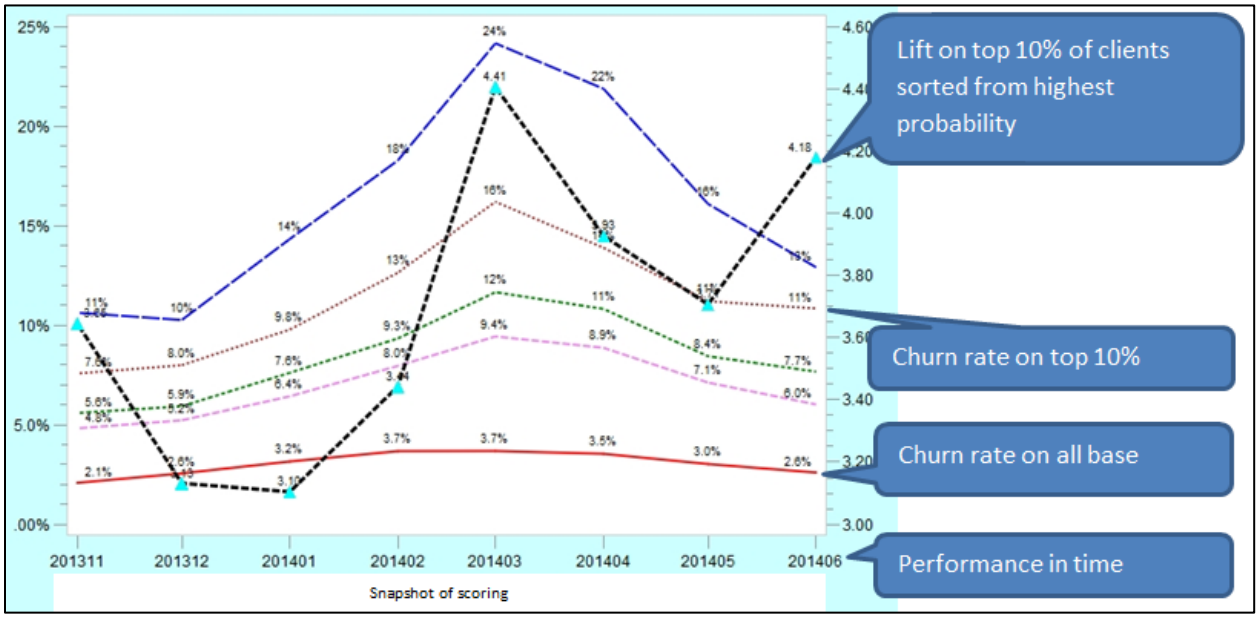

Figure 2 Stability check through Lift and precision

As we can see above, our measures are displayed monthly from 11/2013 to 6/2014. During these months overall churn rate was from 2.1% to 3.7% and the shape of course indicates higher churn in 2/2014 and 3/2014. This increment could be due to many things, e.g. seasonality – spring, new product released by rival. On the other side, the decrease in 6/2014 could be due to successful activity or again by aversion of clients to change something before holidays.

Now we discuss model precision. The dotted brown line represents the history of churn rate on top 10% clients. Churn on this group of clients (each month there could be different clients) is not very stable. We see values from 7.6% in 11/2013 to 16% in 3/2014 and a decrease in 6/2014 to 11%. Under the assumption that we have started performing activities on this group since 4/2014 a decrease means that we targeted right group.

We could notice the lift bounce a lot: from 3.6 to 3.1 to 4.41 to 3.7 to 4.18. We observe that until 3/2014 the lift was relatively stable as the difference was about 0.5. In 3/2014 something special has happened and lift escalated. It is important to reconsider this change according to the whole group, which has a churn rate of 3.7 in the previous month. By watching the lift curve we easily see some kind of “lag” in the model, as it reacts on the market one month later. In following periods beginning with 4/2014 we see lower lift, probably due to activities we performed.

As mentioned above, a disadvantage of these measures is that we do not know what would have happened if we do not start preventive activities on these clients. It just displays reality as it is and we need to think about all possible internal/external impacts and explain them.

SETTING OPTIMAL CUT-OFF

Prior model usage for some activities we need to quantify expected results and find how many clients we should address with the offer. We have previously mentioned a binary model for churn prediction. We look on this task through as setting up an optimal cut-off probability to maximize yield of activity for churn prevention. In our example for simplification we omit the yield and focus only on maximum model precision, i.e. we are looking for probability threshold below which the model is not effective. The restrictions are as follows:

- Maximum number of clients successfully prevented from churn
- Maximum precise prediction in targeted group
- Minimum number of clients to address (this is bounded with previous variables and is not a parameter)

In this example we do not state computation; rather we explain the logic behind it in Figure 3.

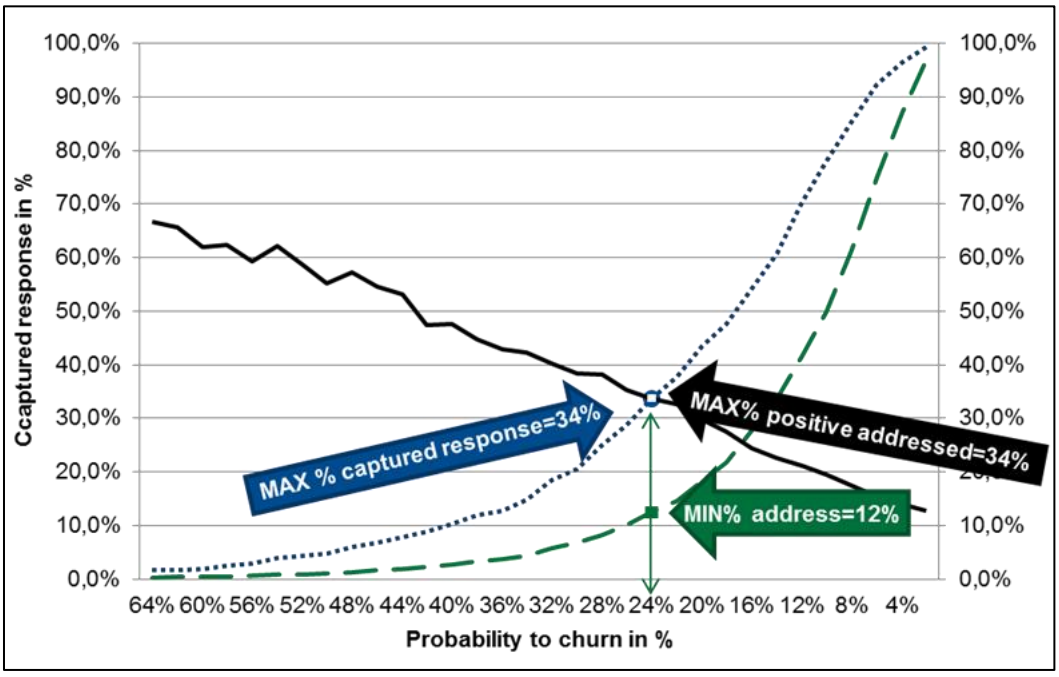

Figure 3 Graphic explanation of setting up cut off

On the horizontal axis we have different levels of probability to churn. For each probability level we compute the measures mentioned above: cumulative captured response in blue dot line (left vertical axis), precision in target group in black solid line (right vertical axis). For an overall view we also display proportion of target group according to all clients in green dash line (right vertical axis).

We see that to obtain maximum precision in target group we need to address all clients with probability of churn higher than 24% and for our information it is 12% of all clients for whom we have computed the probability. Afterwards we expect that churn prevention activity is adequate for 34% of targeted clients. If the activity would succeed 100% we would prevent 34% of clients who would otherwise churn.

Sometimes we perform further augmentation of this approach by adding 2 new parameters: expected success and costs of activity.

USEFUL CODES

During some steps we suggest using following codes which speed the analyzing process.

CODE FOR DEFINING VARIABLE ROLES AND LEVELS IN SAS EM,

By this code we create a simple statement which we directly insert into the SAS Code node editor. Example of the code we insert into MS Excel:

- =CONCATENATE("%em_metachange(name=";B2;", role=";C2;",level=";D2;");")

In Figure 4 we see the insert in Enterprise Miner.

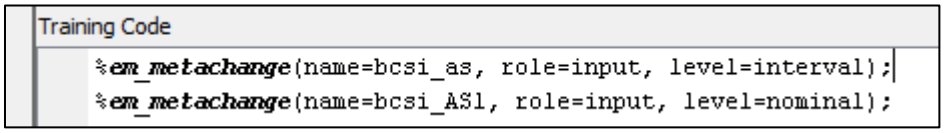

Figure 4 Example of code insert into Enterprise Miner

CODE FOR GRAPH CREATION – BOXPLOTS

During graphic visualization we sometimes need to display many boxplots and this is especially useful when profiling segments through many variables. We suggest usage of the following code to create SAS code which we insert into the program in SAS EG. This would generate many graphs in a much shorter time than the run of the Segment profile node would last in SAS EM.

- =CONCATENATE("TITLE1 '";B2;" vs. Segment";"""";"PROC BOXPLOT DATA=&syslast; plot (";B2;")*_SEGMENT_ / BOXCONNECT=MEAN GRID;RUN;")

CODE FOR GRAPH CREATION – BARCHARTS

Similar code as above designed for bar charts. We use this code to profile binary variables vs. binary target variable.

- =CONCATENATE("TITLE1 ""Segment vs. ";B2;"""; PROC GCHART DATA=&syslast; VBAR _SEGMENT_/ SUBGROUP=";B15;" GROUP=_SEGMENT_ G100 NOZERO TYPE=PCT INSIDE=PCT RAXIS=AXIS2 LEGEND=LEGEND1;RUN;")

Example of this code for two variables is in Figure 5.

```
TITLE1 "Segment vs. fl_regular_payment_1m"; PROC GCHART DATA=&syslast; VBAR
new_subsegment4/ SUBGROUP=fl_regular_payment_1m GROUP=new_subsegment4 G100
NOZERO TYPE=PCT INSIDE=PCT RAXIS=AXIS2 LEGEND=LEGEND1;RUN;
TITLE1 "Segment vs. fl_abroad_PK_txn_1m"; PROC GCHART DATA=&syslast; VBAR
new_subsegment4/ SUBGROUP=fl_abroad_PK_txn_1m GROUP=new_subsegment4 G100 NOZERO
TYPE=PCT INSIDE=PCT RAXIS=AXIS2 LEGEND=LEGEND1;RUN;
```

Figure 5 Example of generated SAS code

Example of generated graph is shown in Figure 6.

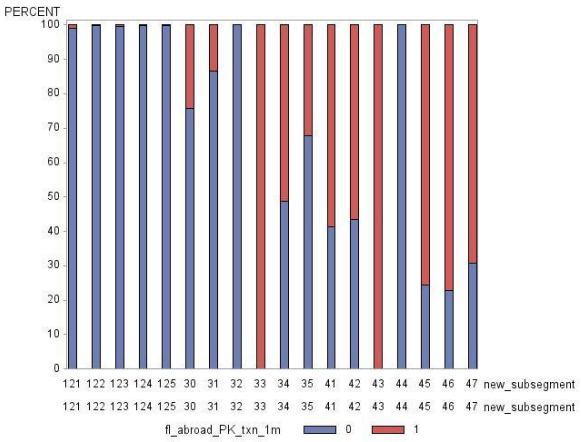

Figure 6 Example of generated bar chart

CODE FOR GRAPH CREATION – NOMINAL VARIABLES

Similar code as above designed for profiling nominal variables vs. nominal target.

- =CONCATENATE("TITLE1 ""Segment vs. ";B2;""""; PROC GCHART DATA=&syslast; VBAR ";B2;"/ DISCRETE GROUP=_SEGMENT_ NOZERO TYPE=FREQ INSIDE=PCT RAXIS=AXIS3 LEGEND=LEGEND1 FRAME PATTERNID=MIDPOINT;RUN;")

PART OF MST MACRO CODE

In this section we state part of our own MST macro code for identifying number of specific values among variables. In our example we start from file “MST_BASE” and we want to identify mainly how many “0” values are in each variable in this dataset. We use for it following code:

```
/*  CREATE CONTENTS          */
proc contents data MST_BASE out=content;
/*  CREATE MACRO CODES       */
data content_codes;
set content;
   format code_is_null code_is_dist code_is_0 $111.;
   format code_tabulate_is_null $143.;
   code_is_null = catx(" ","sum(case when",name,"is null then 1 else 0 end)
as",name);
   code_is_dist = catx(" ","(COUNT(DISTINCT(",name,"))) AS",name);
   code_is_0 = catx(" ","sum(case when",name,"=0 then 1 else 0 end)
as",name);
   code_means_is_0 = catx(" ","if",name,"=0 then ",name,"=.;");
run;
/*  CREATE VARIABLE FROM CODES    */
proc sql noprint; select code_is_0 into :code_is_0_1 separated by ',' from
content_codes;
/*  CREATE TABLE WITH VALUES BASED ON CODES         */
proc sql; create table is_0 as select &code_is_0_1 from MST_BASE;
/*  TRANSPOSE OUTPUT INTO 1 COLUMN FOR BETTER ANALYSIS    */
%let ownoutput=%cmpres(&syslast.)_trans;
%let characteristic=%substr(&ownoutput,6,%length(&ownoutput)-10);
    PROC TRANSPOSE DATA=&syslast
          OUT=&ownoutput
          PREFIX=&characteristic
          NAME=Variable ;
          VAR _numeric_;
    RUN;
```

In the first part we use PROC CONTENTS procedure to get a variable list. Then we create 4 codes for each variable. In consequent code we create the sum of specific values and later we transpose this information into a table from which we copy these values into our Metadata evidence. Example of the information we put into Metadata is in Figure 7.

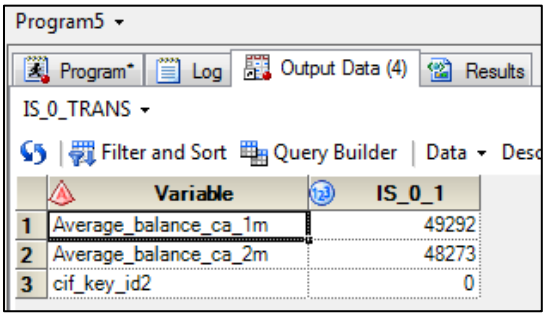

Figure 7 Example of information for input into Metadata evidence

As we can see, by analyzing the variable “Average_balance_ca_1m” (average balance on all of clients current accounts during past 1 month) we have 49,292 clients with an average balance of 0€ in this particular dataset. This would highly skew the mean and we need to keep this information in mind.

CONCLUSION

The goal of this paper is to discuss aspects of modeling with a focus on presenting best practice methodology and useful tools which could enhance model creation. We give the reader an overview of what steps a data miner/data analyst has to go through and we provide a simple tool that could be used as a checklist. In each step we define expected outcomes and duration which helps keeping the project delivered on time and with proper analysis. For actual model creation we mentioned several settings of SAS EM nodes or suggested useful procedures in SAS EG. Afterwards we discussed a few highlights of the specific procedures/steps from the methodology with practical examples of usage. Finally, we provided our SAS code which can to define roles and levels for modeling, create a statistic report and some graphical representation of results.

Generally, this paper summarizes our best practices of creating predictive models with useful tips how to use the tools more effectively and faster.

ACKNOWLEDGMENTS

In this part we would like to thank VÚB for permission to share this knowledge and for providing a creative environment to develop such methodology on CRM department.

We also express gratitude to SAS Slovakia which supported creation of this paper by providing consultations.

CONTACT INFORMATION

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

Mgr. Peter Kertys
VÚB a.s.
+421 2 5055 8608
+421 910 292 876
pkertys@vub.sk

APPENDIX

Phase	Who	Step	Description of step	Tool	Output	Step/node description	Node/Procedure	Table in SAS EM	Duration
Business request	Business	Business case	Defining request, expected results, costs, usage and risk.	Human	Business case				0,5-1
	Analyst	(Study)	Study of materials, research if someone has not already developed the same thing	Human	Insight look				0,5-1
	Bus. & An.	Request understanding	Clarification of business request and possibilities of realization.	Client	Definition of Feasibility study				0,5-1
	Analyst	(Feasibility study)	Simple case study for decision making, quantified time costs, expected yield, etc.	Human	Feasibility study				0,5-2,5
	Business	Decision	Decision of realization based on Business Case and Feasibility study.	Human	Go/no GO + deadlines				0,1-0,5
Model specification	Business & Analyst	Process understanding	Understanding of process in real world, defining which situations could happen from customer point of view.	Other departm.	Understood process				0,2-0,5
		Variables & target	Defining appropriate variables and target, and specification of new variables.	Human	ABT definition				0,2-0,5
		Snapshot and datasets	Defining ABT snapshots used for dataset creation, test and validation set, business segmentation if needed, creating plan of model creation.	Human	List of snapshots, business segments, plan of model creation		PROC FREQ		0,2-0,5
		Variable check	Check variables in variable catalogue if they are not in black list due to changes in infrastructure. Check basic response to target.	Human	Appropriate variables				0,1-0,2
Data preparation	Analyst	Dataset creation	Dataset creation, programming of new variables. Merging snapshots.	SQL/EG	Dataset in SAS EG/EM				0,5-5
		Dataset stability	Dataset stability in time check in case dataset is merged from several snapshots.	EG/Excel	Stability check for different snapshots				0,1-1
	Data miner	ABT analysis	Primary analysis of variables and selection according to missing or unary values. Using MST macro.	EG/Excel	Summary statistics, rejection of irrelevant variables	reject: <5%!=0, irrelevant, >95%=., target>5%, <5%distinct	PROC MEANS, PROC FREQ		0,5-1,5
		Roles and levels	Manual definition of roles and levels of variables, methods of replacement and types of imputations in case of missing values.	Excel	Code for roles and level definition, rules for missing and extreme values replacement /imputation, list of data quality issues	ID, unary, binary, ordinal, nominal, interval / input, target, segment, rejected; min, max, distribution, 0, "U"	Metadata node	META_EMTRAINVARIABLE	0,2-0,5
		(Distributions)	Graphical distributions of selected variables vs. target variable.	EG/Excel	Check of output, decision for binning		PROC GBARLINE, StatExplore node		0,5-1
		Missing values	Replacement of missing values according to suggestion from Metadata document.	EM	Dataset without missing values	Interval / nominal VS. numeric / character, default constant value, minimum, maximum, distribution, 0, "U"	Impute node	IMPT_RESULT	0,1-0,2
		Extreme values	Replacement of extreme values according to suggestion from Metadata document.	EM	Dataset without extreme values	Replacement editor: Extreme percentiles cut off: 95%/99%, or: 2/3 sigma (st.dev.)	Replacement node	REPL_LIMITS, REPL_CLASSINFO	0,1-0,2
		(Segmentation)	Segmentation according to target based on several manual selected business oriented variables in order to homogenize target groups.	EM	Dataset split into segments		Cluster node		1-2
		Transformation	Creating several variable transformations, normalization of data, etc.	EM	Set of variables for model input	interval<->nominal, maximum normal	Transform variables, Interactive grouping, Replacement, Principal components node	TRANS_RESULT, Trans/EMPUBLISHSCORE.sas	1-2
Iterative model creation	Data miner	Correlations	Multivariate correlation analysis of variables.	EG/EM	Uncorrelated variables with highest influence on target	Pearson / Spearman <0,6	PROC CORR, Variable clustering, StatExplore node	VARCLUS_OUTSQUARE	0,5-1
		Sampling	Revision of target penetration and then oversample or undersample the dataset. Creation of data partitions for train and validate.	EM	TRAIN-VALIDATE-SCORE		Sample, Data partition node	PART_TRAIN, PART_VALIDATE	0,1-0,2
		Worth on target	Variable selection according to worth on target.	EM	Important variables		Variable selection, StatExplore node	VARSEL_OUTEFFECT, VARSEL_OUTGVAR, STAT_CHIMEASURE	0,1-0,3
		(New variables)	Creation of transformed/new variables through different methods.	EM	New and transformed variables	Use interactive grouping. For binary target use WOE		BINNING_VARMAPPINGS, IGN_VARMAPPINGS, TRANS_RESULT, TRANS_STATISTICS	0,1-0,3
		(Correlations 2)	Optional step in case of too many variables.	EG/EM		Varsel-chisquare for nominal, r-square for interval, check through manual selector	Variable clustering, Variable selection node, PROC CORR	VARCLUS_OUTSQUARE	0,1-0,3
		(Transformations 2)	Same as above. Optional step.	EM			Startgroup, EndGroup node	ENDGRP_TRAIN	0,1-0,3
		Model	Model creation and selection from various alternatives, iterative approach during variable selection, different methods of model selection.	EM	Best models		PROC LOGISTIC, Regression, Clustering, Neural network, Decision tree node	REG_EFFECTS	0,5-3
Model tuning	Data miner	Model stability	Model selection based on target precision and reduction of number of input factors. Stability check on different partitions.	EM/Excel	Final model with interpretable coefficients	60-40/70-30/80-20 / score; comparison with CAPC, LIFT>1,5, GINI>0,3, Train & Validate should be similar, chis-square for most important variables, odds ratio	Model comparison node	MDLCOMP_COMPAREFIT	1-2
		Model test	Model test "out of sample" - stability check on different ABT data snapshots.	EM/EG	Validated model				0,1-1
		Unsampling	Usage of "Decision" node in case of under/oversampling.	EM	Model ready for usage in reality	Unsample through prior probabilities	Decisions node		0,1-0,2
		Scoring	Creation of scoring rules for use in SAS Enterprise Guide.	EM/EG	Score code, upload into production		Score node	Score/PATHPUBLISHSCORECODE.sas	0,1-0,2
Implementation	Data miner	(Cut off)	Defining way of model usage, setting cut-off.	Human	Efficient frontier setting		Decisions node		0,1-0,2
	Business	(Pilot)	Pilot of model - validation if activities based on model create expected results.	Human	Scored clients, pilot evaluation				0,1-0,2
	Data miner	Update ABT	Dataset with appropriate variables for model and target computation automation.	SQL/EG	Sql script				0,2-0,5
		Update variable catalogue	Input of newly created variables into variable catalogue with notes and codes.	Excel	Variable catalogue with actual data				0,1-0,2
		Documentation/Presentation	Create documentation, explain used variables and defining client conditions, tips for model usage, update model evidence, upload outputs on intranet.	MS Office	Documentation, PowerPoint presentation				1-2
		Best practice	Summarize experiences of model creation, tips and tricks, process optimization.	MS Office	Best practice document				0-0,5
	IT	Migration in production	Reduction of diagrams in SAS EM and EG from unused nodes, copy model on production server.	Human	Model in production				0,1-0,1
	Data miner	Automation	Setting on automatic process of time stability checks, publication of results.	SQL/EG	EG project for automation of scoring				0,5-1
Model service	Data miner	Stability check	Manual control of test and scoring log for warnings, comparison of different results in time through stats and graphs per segments.	EG/Excel	Validation of model	LIFT, CAPC	PROC BOXPLOT, PROC GCHART		monthly - 0,2-0,5
		Recalibration	Model recalibration and validation.	EM	Recalibrated model				½/yearly

Table 1 Methodology of Model Creation