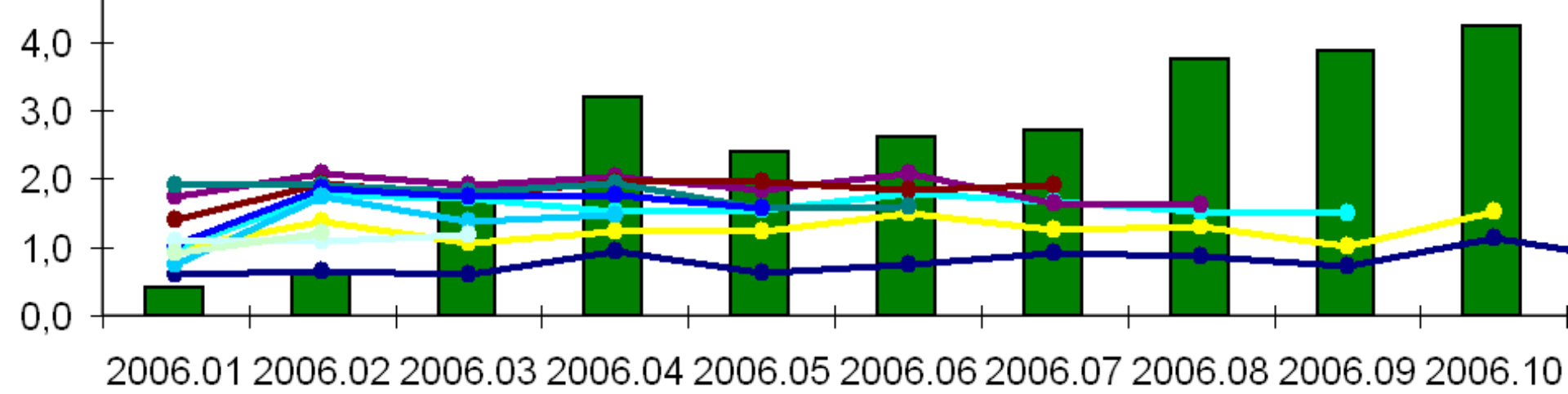
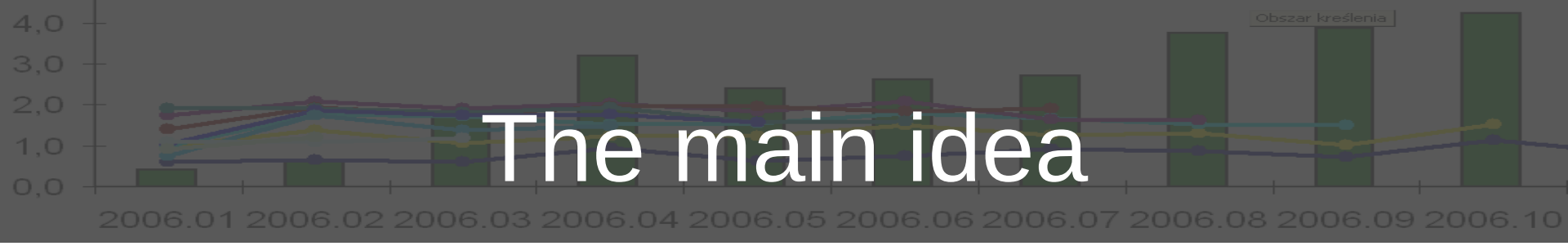


Advance Scorecard Builder



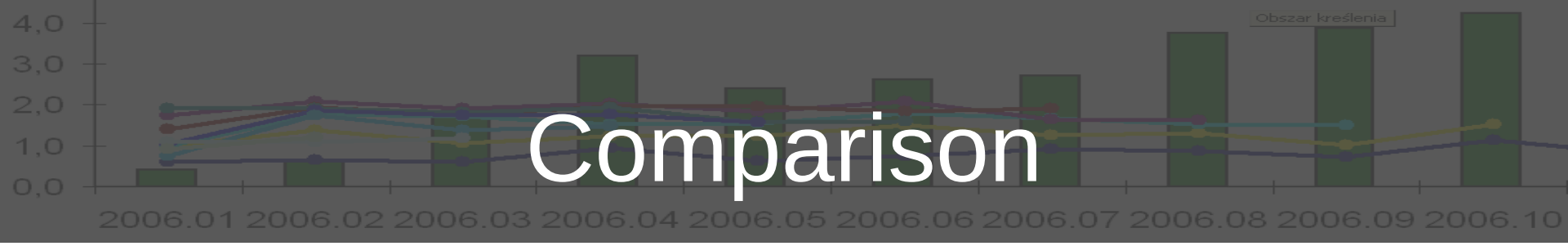
The main idea

- Scorecard building is not an automatic process!
- Always is needed to tune, customize some elements
- Tool for building should be very flexible

- The best solution:
  - all open source codes
  - open architecture
  - easy for modification and customization for certain project

Why automation ±

- Cannot be automated every step
- But many steps take a time and are rather similar in every project
- Advantages:
  - save time for difficult steps
  - quick model in shorter time
  - positive criterion for Regulator
- Disadvantages:
  - first usage needs courses
  - some standards are needed



Comparison

- SAS Credit Scoring Solution
- SAS Enterprise Miner

  ⇕

- SAS Analytics = Base SAS, SAS/STAT, SAS/GRAPH, SAS Interface to PC formats
- plus set of SAS 4GL codes for Advance Scorecard Builder (ASB)

Different properties

Criterion	SAS Enterprise Miner	SAS Analytics + ASB
Special nodes	Interactive Grouping, Scorecard, Reject Inference	All nodes implemented in open source codes
Flexibility	only interface, "by clicking"	Every code can be changed
Key performance	Fixed list of statistics: KS, GIni, IV, WOE	KS, Gini, Gains, KLD, VIF, CI and always can defined new one
Stability validation	No	KLD, many special reports
Variable reports	No	many attribute reports
Evaluation in the time	No	Yes
Automatic monitoring	No	Yes
Score calibration	Only one method	More than one
Bining methods	only typical rules	every SAS 4GL expression
Beta calculation	Dummy and WOE approaches	Dummy, Logit, WOE and full control

Modeling steps

1. Data structure
2. ABT preparation
3. System options
4. Sampling and data partition
5. Variables and its types
6. Binning, comparison and manual corrections
7. Coding
8. Variable pre-selection
9. Variable reports (HTML)
10. Multifactor analysis and simple model validation
11. Manual exercises
12. Full model validation and reporting (HTML)
13. Scoring code
14. Model monitoring

Data structure

id client or account	default	period	var1	var2	var3
	0	200901			
	1	200902			
	.i	200901			
	.d				

Good

Bad

Indeterminate

Dormant

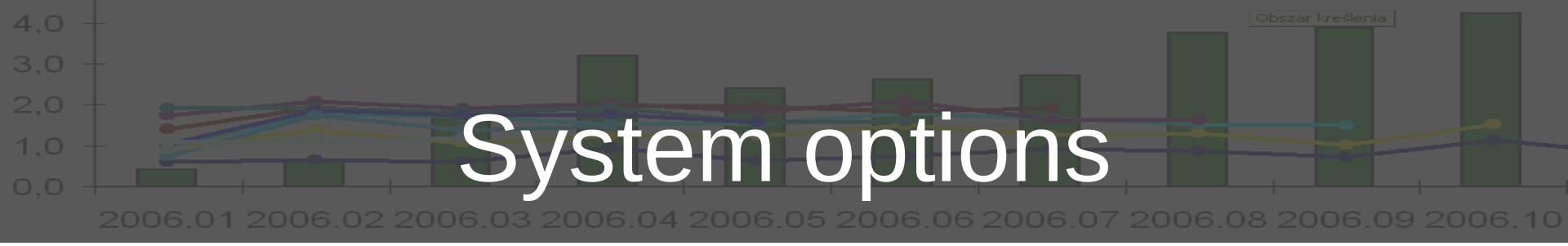
ABT tables

- Code abt_behavioral_columns.sas

Example variable:

agr3_Trend_Max_Bal –

1. for every account of the same client is calculated maximal balance on period
2. for three last periods is calculated trend on previous numbers (maximal balances per period)



System options

```
%let prefix_dir=E:\moje\modele\ex_behavioral\;
%let em_nodedir=&prefix_dir.modele\;
%let em_import_data=abt.train_woe;
%let em_import_validate=abt.valid_woe;
%let em_lib=wyj;
/*definicje zbiorów traningowego i walidacyjnego*/
%let zb=abt.train;
%let zb_v=abt.valid;
%let zb_vg=;
/*definicja zmiennych*/
%let em_data_variableset=wyj.Zmienne_definicja;
/*zmienna ze scorami*/
%let zm=SCORECARD_POINTS;
/*zmienna z defaultem*/
/*%let porz_tar=descending; dla response*/
%let porz_tar= ;
/*dla risk*/
%let tar=default6_pcmc;
%let nr_mod=1;
%let the_best_model=1;
libname abt "&prefix_dir.abt" compress=yes;
libname wyj "&prefix_dir.wyj";
libname modele "&prefix_dir.modele";
libname freq "&prefix_dir.freq";
libname adj "&prefix_dir.adj";
libname inlib "&prefix_dir.data";
```

Sampling

- Input: ABT tables, percent – share of all data
- Output: abt.train and abt.valid datasets

period	train	valid
200801		
200802		
200803		
200804		
200805		
200806		
200807		
200808		
200809		
200810		
200811		
200812		
200901		
200902		
200903		
200904		
200905		
200906		

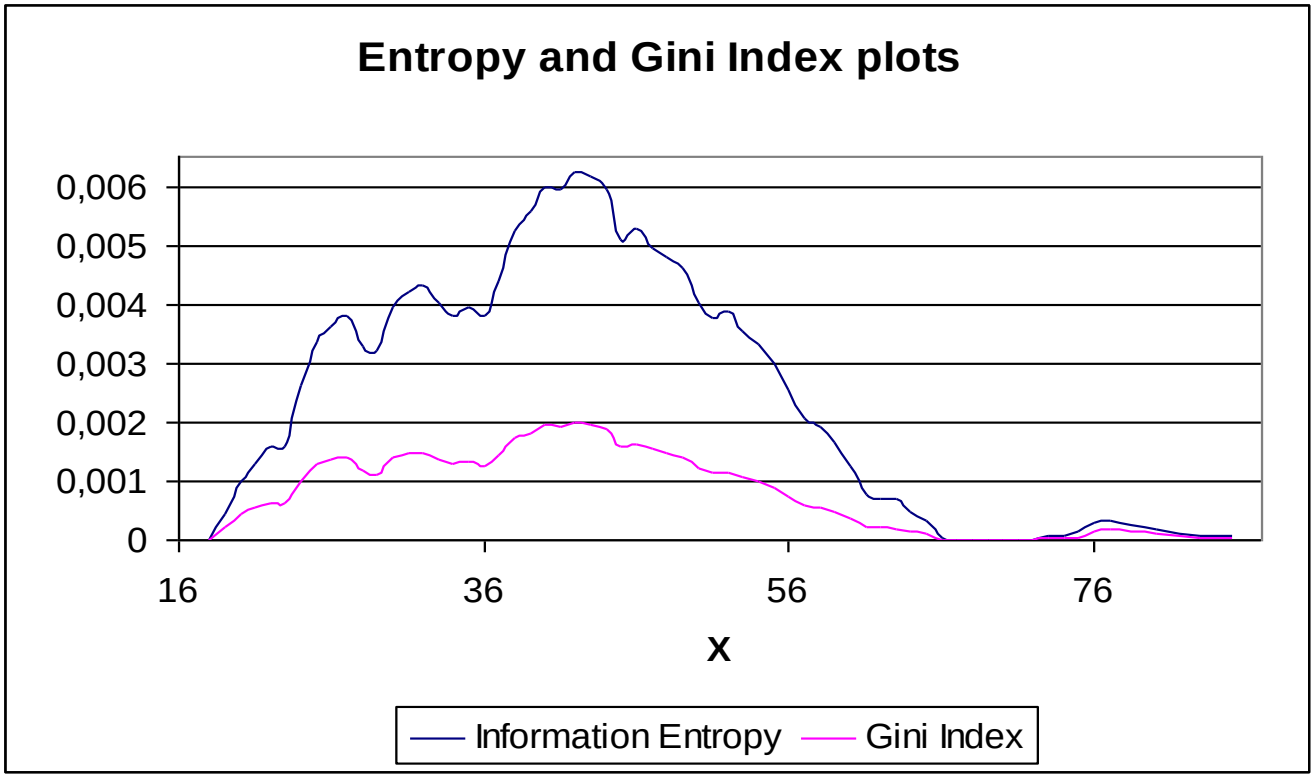
period	
200801	
200802	
200803	
200804	
200805	
200806	
200807	
200808	
200809	
200810	
200811	
200812	
200901	
200902	
200903	
200904	
200905	
200906	

Variables and types

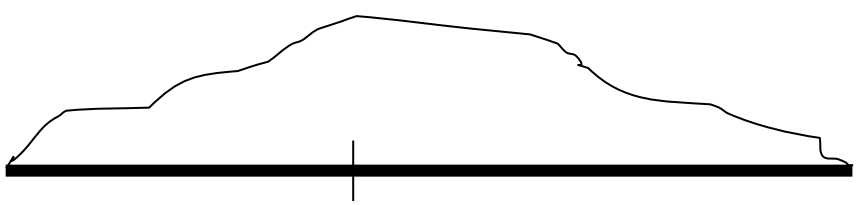
Variable	Type
act_age	interval
aggr3_max_max_bal	interval
act_status	nominal

- Type is automatically indicated, but can be manually updated

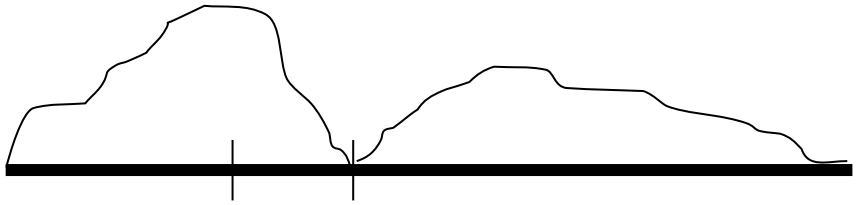
Binning - theory



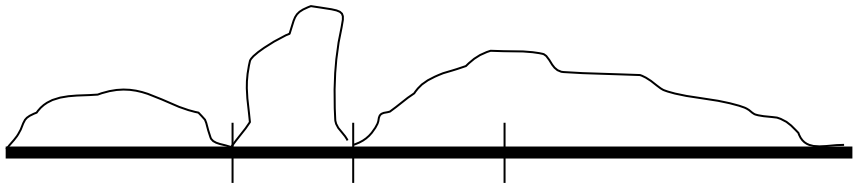
Binning - algorithm



1. First splitting point is indicated



2. Second splitting point



3. Third is chosen from wider interval

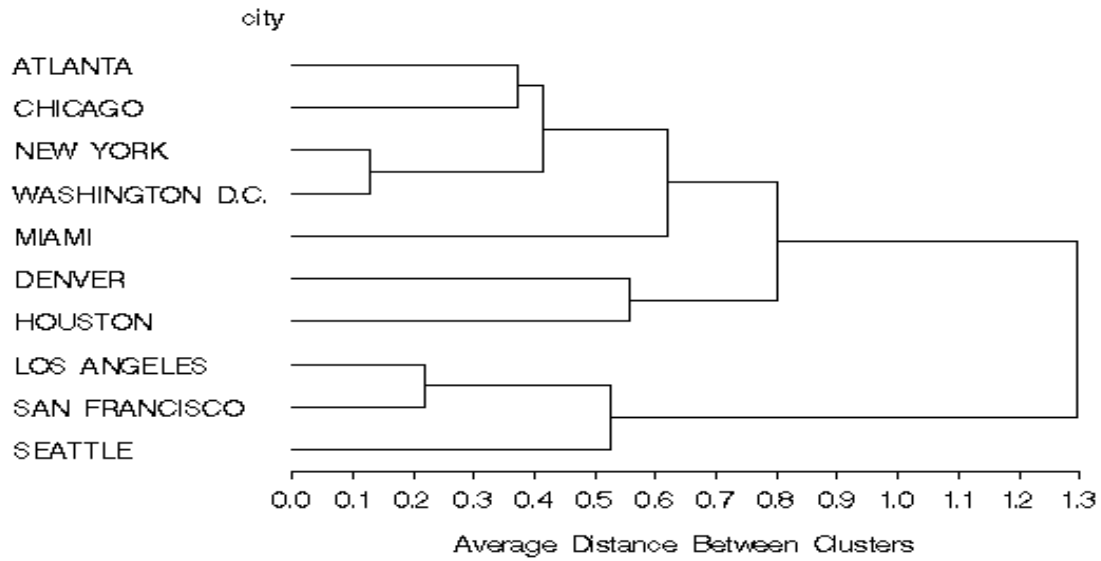
Binning - versions

- Non-monotonic
- Monotonic (additional constraint)
- Maximization by Gini
- Equal intervals, shares

Number	Variable name	Gini_before	Gini_NonMon	Gini_MonNew	Gini_MonOld
1	AGGR6_MEAN_S_CASHUTL_EM	64,72%	63,31%	7,57%	63,04%
2	AGSP6_MAX_BAL_EMCL	60,09%	60,02%	.	59,65%
3	ACT_S_CASHUTL_EM	58,38%	60,03%	20,71%	59,81%
4	AGGR3_MEAN_S_RBAL_EMCL	38,44%	36,11%	39,35%	36,11%
5	AGSP3_MIN_PMT	29,33%	40,36%	31,39%	36,87%
6	AGSP6_MAX_NOTPAID	28,32%	17,85%	17,85%	.
7	ACT_PMT	27,59%	43,33%	32,33%	41,03%
8	ACT_S_RBAL_EM	26,41%	.	26,00%	.
9	AGGR6_MAX_CYCLE_DD	14,36%	7,58%	7,58%	7,58%
10	AGSP3_MAX_PMT	8,88%	.	19,71%	24,65%
11	ACT_NBR_LCF	1,75%	27,51%	27,54%	.

Binning for nominal variables

- Are indicated every unique values with the assumption: minimal percent >= 1%
- Every category is joined used the Cluster procedure by the similar bad rates



Binning - code

- Input: abt.train, variables types, manual attributes relations
- Output: data set with splitting points

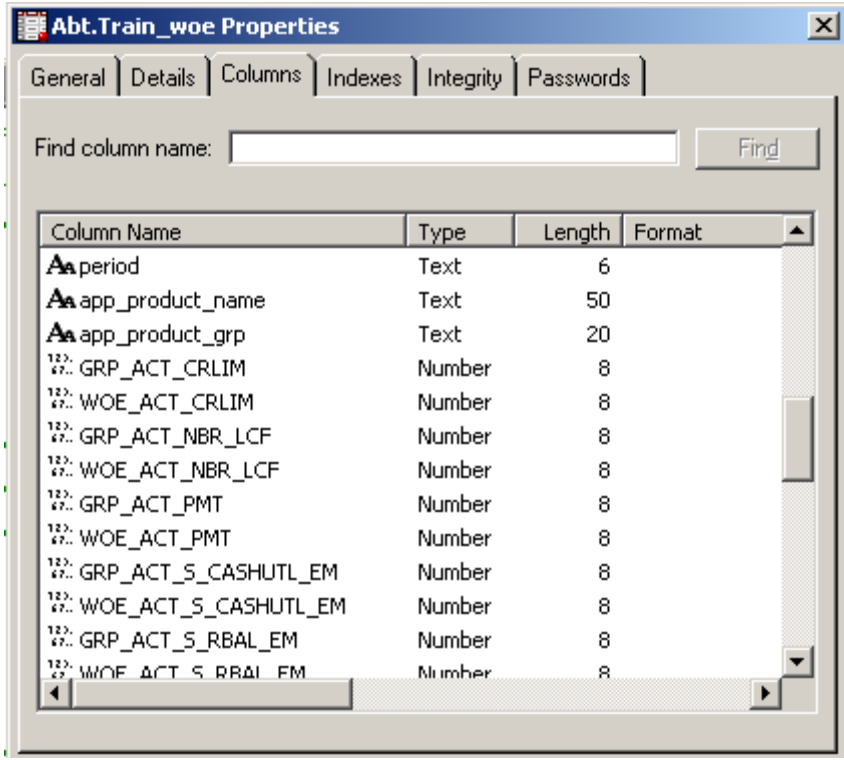
VIEWTABLE: Wyj.Podzialy_interwalowe

	grp	war	zmienna
1	1	11.41052146 < ACT_CRLIM	ACT_CRLIM
2	2	1.9148542155 < ACT_CRLIM <= 11.41052146	ACT_CRLIM
3	3	not missing(ACT_CRLIM) and ACT_CRLIM <= 1.9148542155	ACT_CRLIM
4	4	missing(ACT_CRLIM)	ACT_CRLIM
5	1	11.41052146 < ACT_NBR_LCF	ACT_NBR_LCF
6	2	1.9148542155 < ACT_NBR_LCF <= 11.41052146	ACT_NBR_LCF
7	3	not missing(ACT_NBR_LCF) and ACT_NBR_LCF <= 1.9148542155	ACT_NBR_LCF
8	4	missing(ACT_NBR_LCF)	ACT_NBR_LCF

```
data adj.ACT_NBR_LCF;
length grp 8 war $300 zmienna $32;
zmienna="ACT_NBR_LCF";
input;
war=_infile_;
grp=_n_;
cards;
ACT_NBR_LCF>=1
ACT_NBR_LCF<=0
;
run;
```

Coding - code

- Input: abt.train, abt.valid, datasets with splitting points
- Output: abt.train_woe, abt.valid_woe, dataset with splitting points and many attribute statistics, sas code for coding

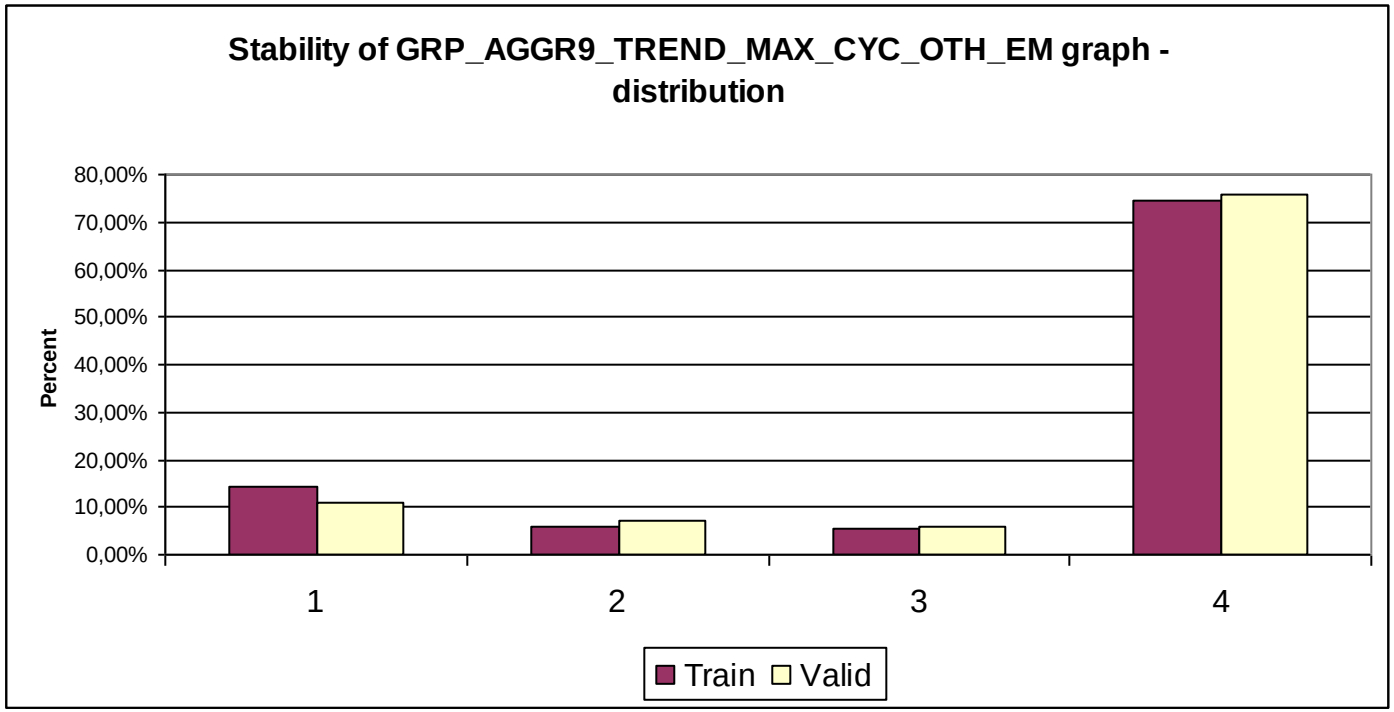


Variable pre-selection

- Every variable is validated separately by:
  - predictive power
  - distribution stability in the time (KLD)
  - Gini stability in the time
  - missing value
  - typical descriptive statistics like min, max and the most frequent value
- Input: abt.train_woe, abt_valid_woe
- Output: dataset with many variable statistics

Stability in the time

- KLD for attribute distributions and for only bad cases distributions: H_GRP_TV and H_Br_GRP_TV



Variable reports

- In html: simple usage, interactive links
- In excel: printable version (not automated yet)
- The main reports:
  - descriptive statistics
  - attribute indicators including: iv, woe, br
  - evaluation in the time of attributes: bad rates, shares, balances and credit limits
  - clustered variables

Variable reports

- All variables are grouped in clusters
- Variables are correlated inside cluster
- Correlation between clusters is rather small
- Cumulative proportion inform us about number of orthogonal dimensions in the space of variables

Obs	Number	Eigenvalue	Difference	Proportion	Cumulative
1	1	2,83858046	0,8983386	31,54%	31,54%
2	2	1,94024187	0,90106143	21,56%	53,10%
3	3	1,03918044	0,11832011	11,55%	64,64%
4	4	0,92086032	0,10011509	10,23%	74,88%
5	5	0,82074524	0,26466194	9,12%	84,00%
6	6	0,5560833	0,22605046	6,18%	90,17%
7	7	0,33003284	0,01355258	3,67%	93,84%
8	8	0,31648026	0,07868499	3,52%	97,36%
9	9	0,23779527	_	2,64%	100,00%

Multifactor analysis

- There are used various variable selection methods in logistic procedure:
  - stepwise, forward, backward, score.
- Every automatically generated model is validated by:
  - colinearity
  - stability
  - different predictive powers statistics

Manual exercises

- Despite automatic model generation can be also validated any combination of variables inputted manually
- After choosing the small list of candidate models, there are calculated statistics per model like:
  - gains, lifts, stability of score distributions
  - and in the final step are prepared scorecard tables

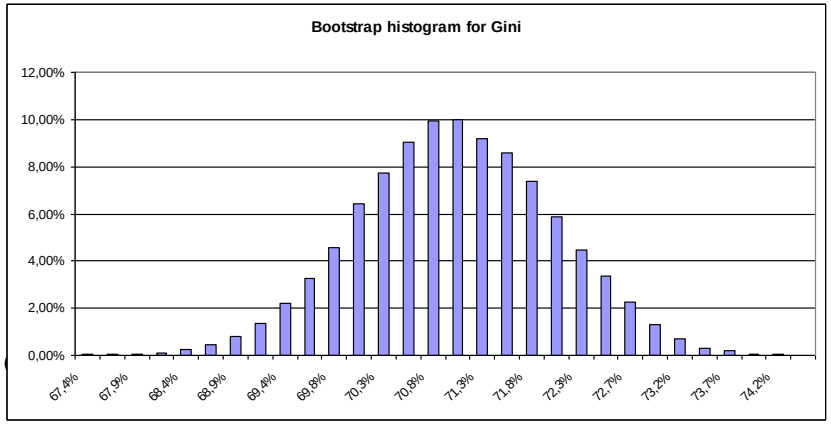
Full model validation

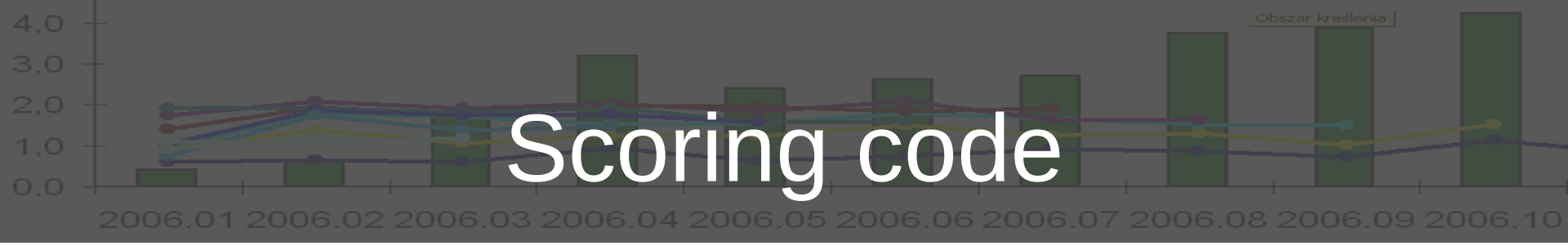
- In html: all partial steps of modeling:
  - variable statistics and graphs
  - splitting points
  - scorecard and partial score indicators
  - bootstrap tests
  - colinearity and stability tests
- In excel: can be fully automated

Full model validation

- Bootstrap and asymptotic confidence intervals

Gini statistic for the model		
Modeling process	Traininng data set	Validating data set
	71,03%	69,56%
Asymptotic	Lower confidence limit	Upper confidence limit
95%	69,20%	72,85%
99%	68,63%	73,43%
Bootstrap	Lower confidence limit	Upper confidence limit
95%	71,01%	71,04%





Scoring code

- Automatically generated based on the chosen final model (based on its scorecard)
- Code is useful in implementation process and in additional model analysis

Model monitoring

- In html: useful interactive reports like:
  - Gini in the time
  - bad rates in the time
  - shares of score bands and variable attributes in the time
  - are presented reports for various default indicators, with longer and shorter outcome periods
- There is used whole available dataset, not only random sample like in modeling
- Can be validated properties of the model on special subsets - segments