Subject

In some circumstances, the goal of the supervised learning is not to classify examples but rather to organize them in order to point up the most interesting individuals. For instance, in the direct marketing campaign, we want to detect the customers which are the most likely to respond to the solicitation. In this context, the confusion matrix is not really suitable for the evaluation of the predictive model. It is more valuable to use another tool, more appropriate for the evaluation of the respondents corresponding to the number of reached individuals: this is the “lift curve” (“gain chart”).

In this tutorial, we use the binary logistic regression for the construction of the gain chart. We show also that the variable selection is really useful in the context of dealing with large number of predictive variables.

Dataset

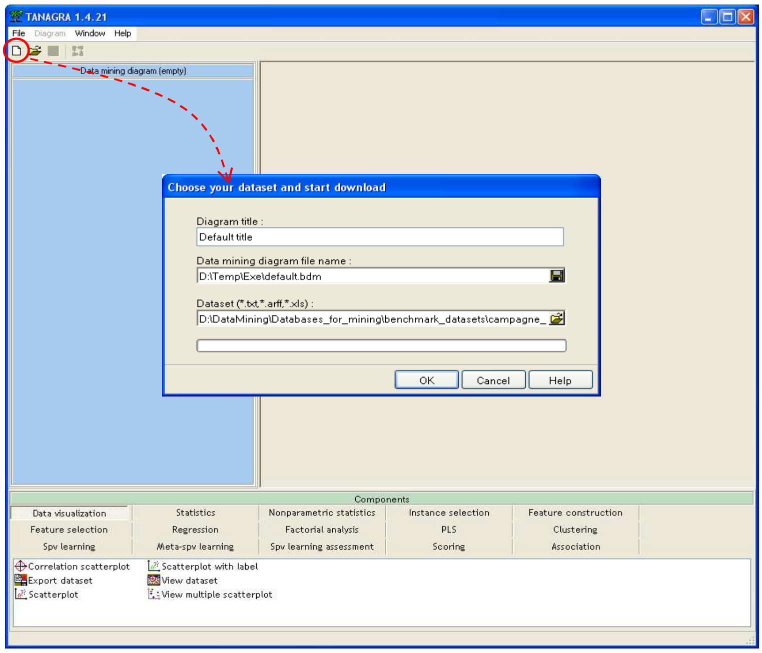
In this tutorial, we use a real/realistic dataset from the following website http://www.ssc.ca/documents/case_studies/2000/datamining_e.html. It contains 2158 examples and 200 predictive attributes. The objective variable is a response variable indicating whether or not a consumer responded to a direct mail campaign for a specific product.

We transform the dataset in a XLS spreadsheet file format[1]. We add a new attribute (EXSTATUS) which specify whether an instance belong or not to the training sample part (train sample: 1158 examples, test sample: 1000 examples).

Binary logistic regression and lift curve

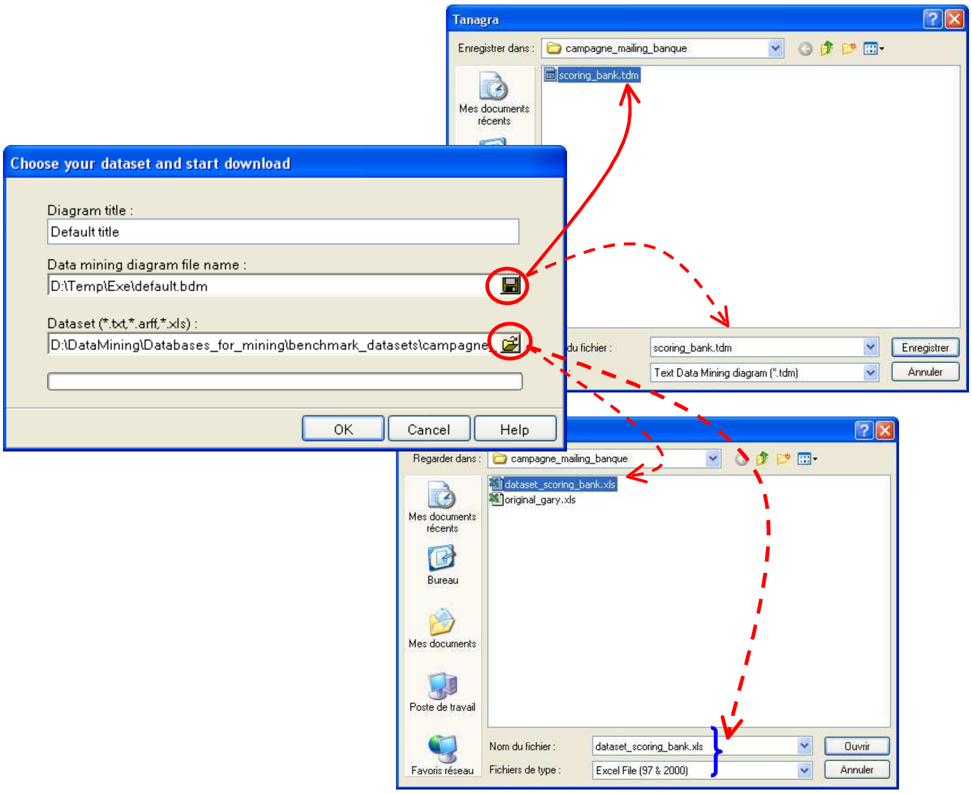
Accessing to the dataset and creating a new diagram

After starting TANAGRA, we create a new diagram by activating the FILE/NEW menu.

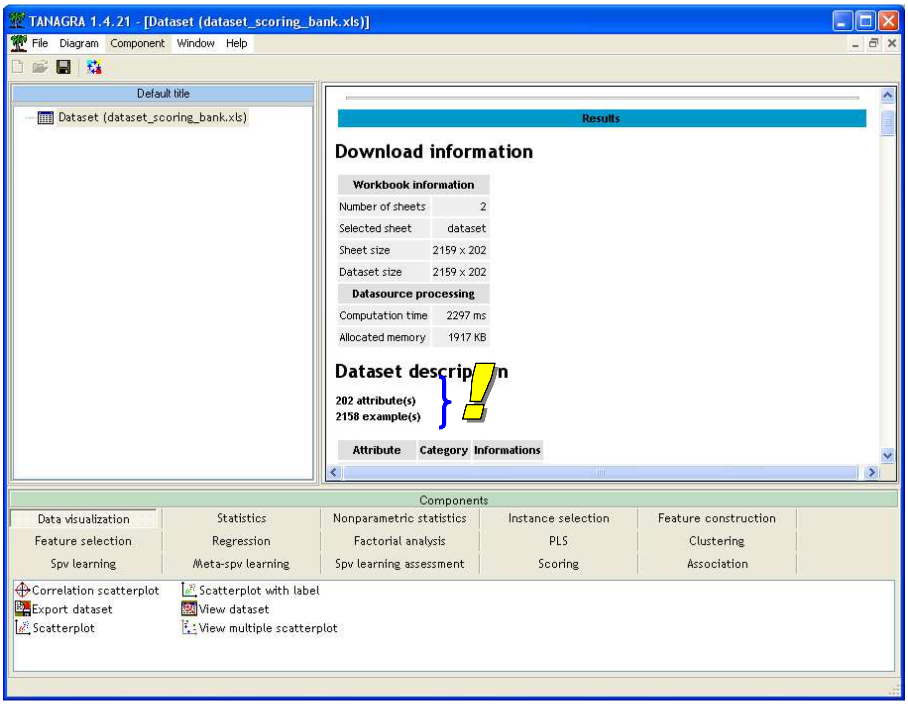


[1] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/dataset_scoring_bank.xls

In the dialog box, we choose the data file DATASET_SCORING_BANK.XLS and then we specify the name of the diagram. For XLS files, the importation functions properly if the folder is not being edited further, and that the data are located in the first sheet.



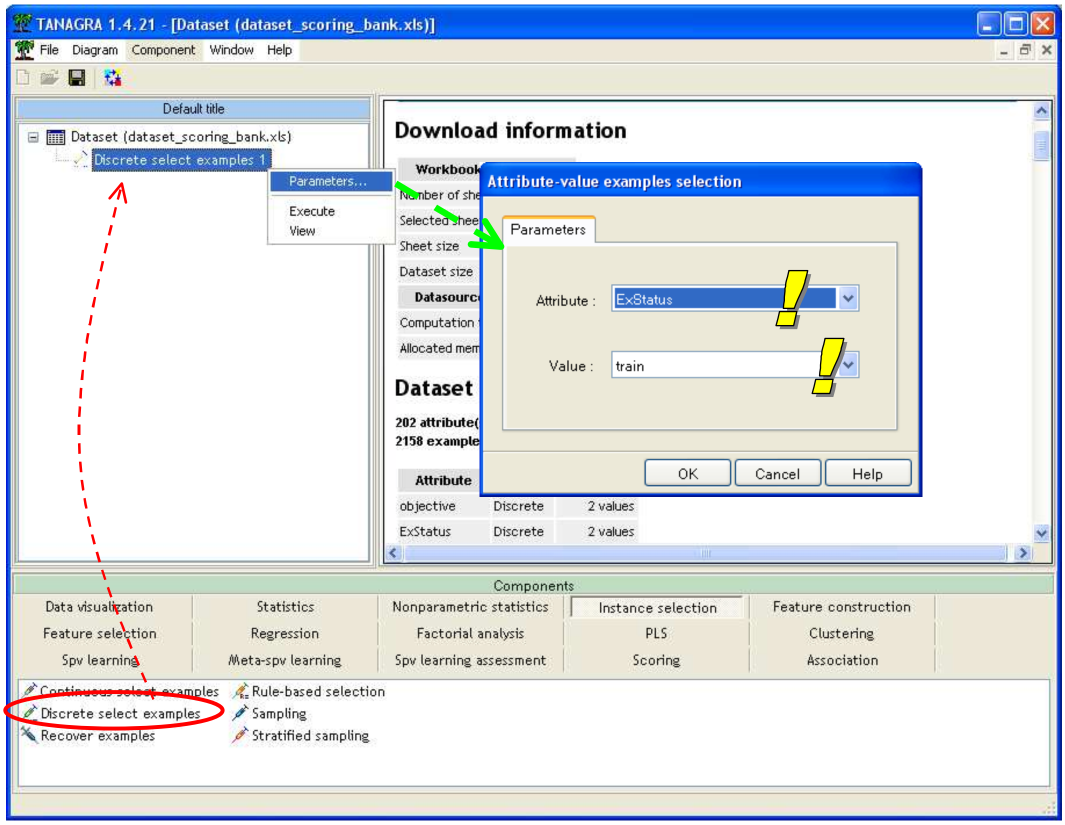
We check the number of examples (2158 examples) and variables (202 attributes) downloaded.



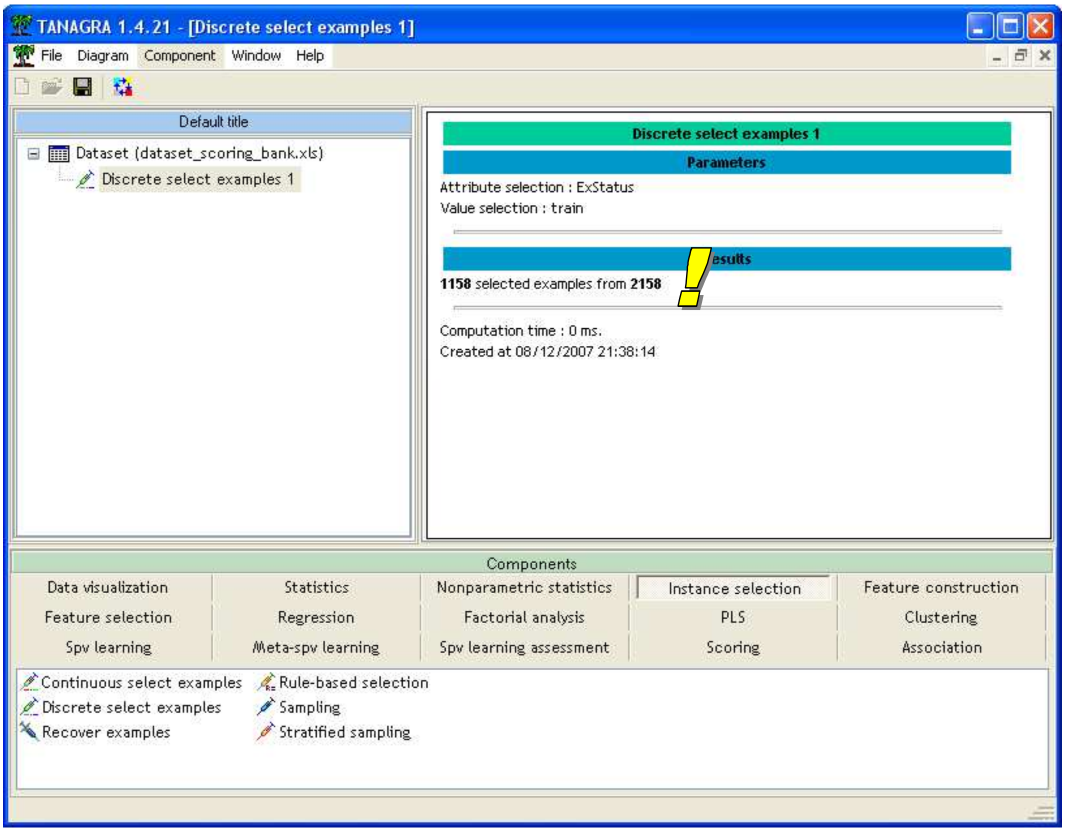
Partitioning the dataset into train and test set

In order to obtain an honest evaluation of the model, we must partition the data into a train set, for the construction of the model, and a test set, for the validation.

We add the DISCRETE SELECT EXAMPLES component (INSTANCE SELECTION tab) into the diagram. We activate the contextual PARAMETERS menu, we set EXSTATUS as the reference variable, and the train set corresponds to the TRAIN value of EXSTATUS.

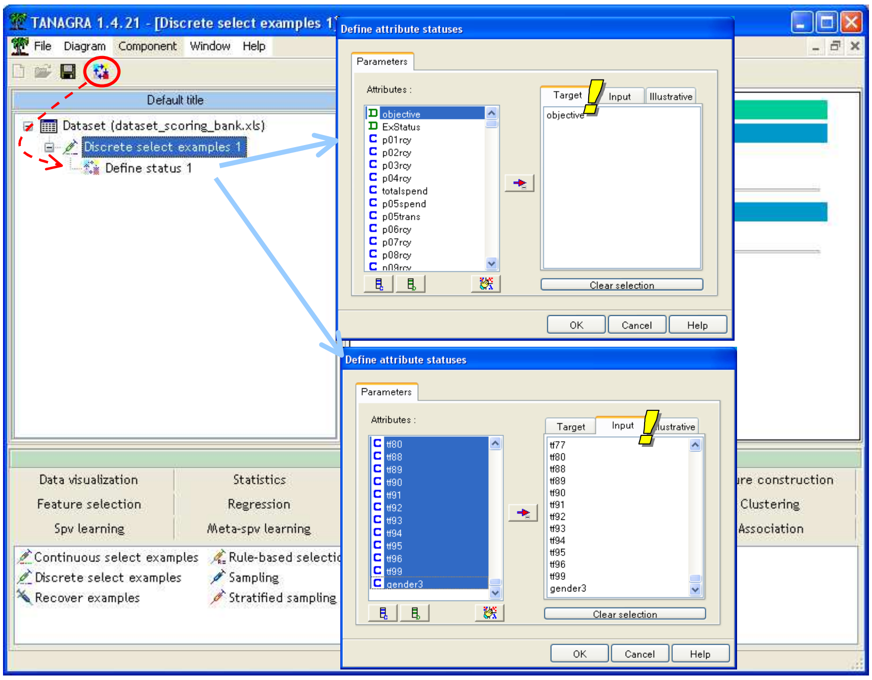


1158 observations are selected for the learning phase.



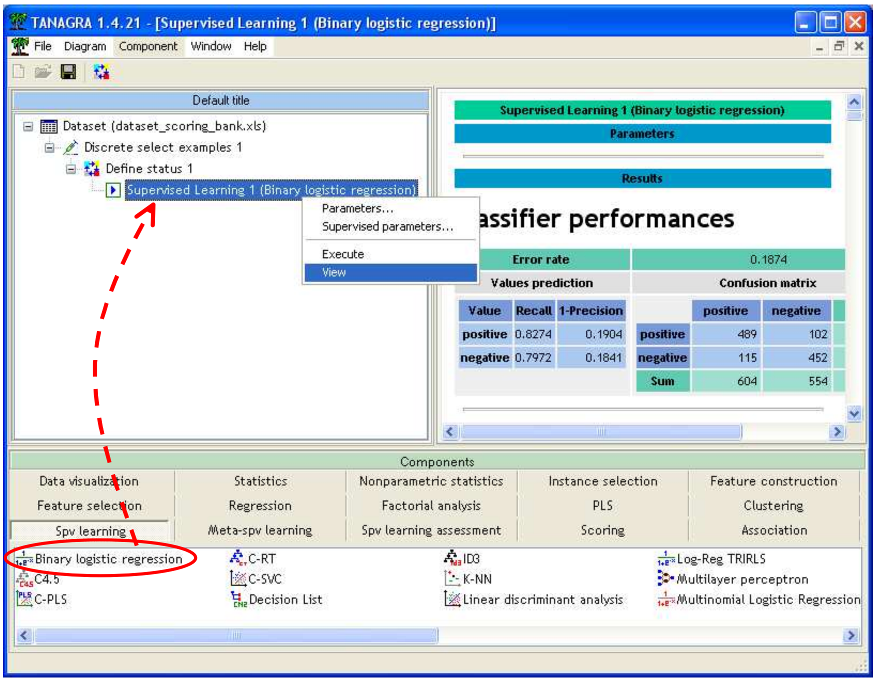
Defining the role of the variables

We must specify the role of the variables. OBJECTIVE is the target attribute, all the continuous attributes, from P01RCY to GENDER3 are the input ones. We do not use the EXSTATUS attribute here. We add the DEFINE STATUS component, using the toolbar shortcut, into the diagram.



Learning algorithm: logistic regression

Because of various theoretical and practical reasons, the binary logistic regression is a very popular method. We add this component (BINARY LOGISTIC REGRESSION, SPV LEARNING tab) into our diagram. We activate the VIEW menu in order to obtain the results. According the dataset size (number of examples and variables), the computation can be more or less high (5 seconds on my computer for this dataset). The window result comprises several sections.



The confusion matrix

Results

Classifier performances

Error rate			0.1874			
Values prediction			**Confusion matrix**			
Value	**Recall**	**1-Precision**		**positive**	**negative**	**Sum**
positive	0.8274	0.1904	**positive**	489	102	591
negative	0.7972	0.1841	**negative**	115	452	567
			Sum	604	554	1158

The confusion matrix is not really useful for our study.

Global evaluation

The most of the indicators presented in this part rely on the deviance or, more precisely, the likelihood ratio statistic. For further information about these indicators, see this very valuable reference http://www2.chass.ncsu.edu/garson/pa765/logistic.htm. We note that, according the Schwartz criterion (SC), that there are too many variables in our regression.

Adjustement quality

Predicted attribute		objective
Positive value		positive
Number of examples		1158
Model Fit Statistics		
Criterion	Intercept	Model
AIC	1606.831	1371.488
SC	1611.886	2387.433
-2LL	1604.831	969.488
Model Chi² test (LR)		
Chi-2		635.3431
d.f.		200
P(>Chi-2)		0.0000
R²-like		
McFadden's R²		0.3959
Cox and Snell's R²		0.4223
Nagelkerke's R²		0.5631

LOGIT coefficients (LOGITS)

Like the widely diffused statistical software, TANAGRA gives the estimated coefficients, their standard error, the Wald statistic and its p-value. We can check the significance of the variables.

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-5.454800	-	-	-
p01rcy	0.603947	2.7649	0.0477	0.8271
p02rcy	0.877664	2.3705	0.1371	0.7112
p03rcy	0.459307	1.6344	0.0790	0.7787
p04rcy	0.740081	3.7462	0.0390	0.8434
totalspend	-0.000010	0.0004	0.0006	0.9809
p05spend	-8.702667	18.1687	0.2294	0.6319
p05trans	0.490166	12.3199	0.0016	0.9683

It seems that none variable is significant at 1%. It does not mean that all the variables are irrelevant, but rather some irrelevant variables are probably correlated with the relevant ones, included in the regression. Variable selection is an important process in this context.

Odds-ratios

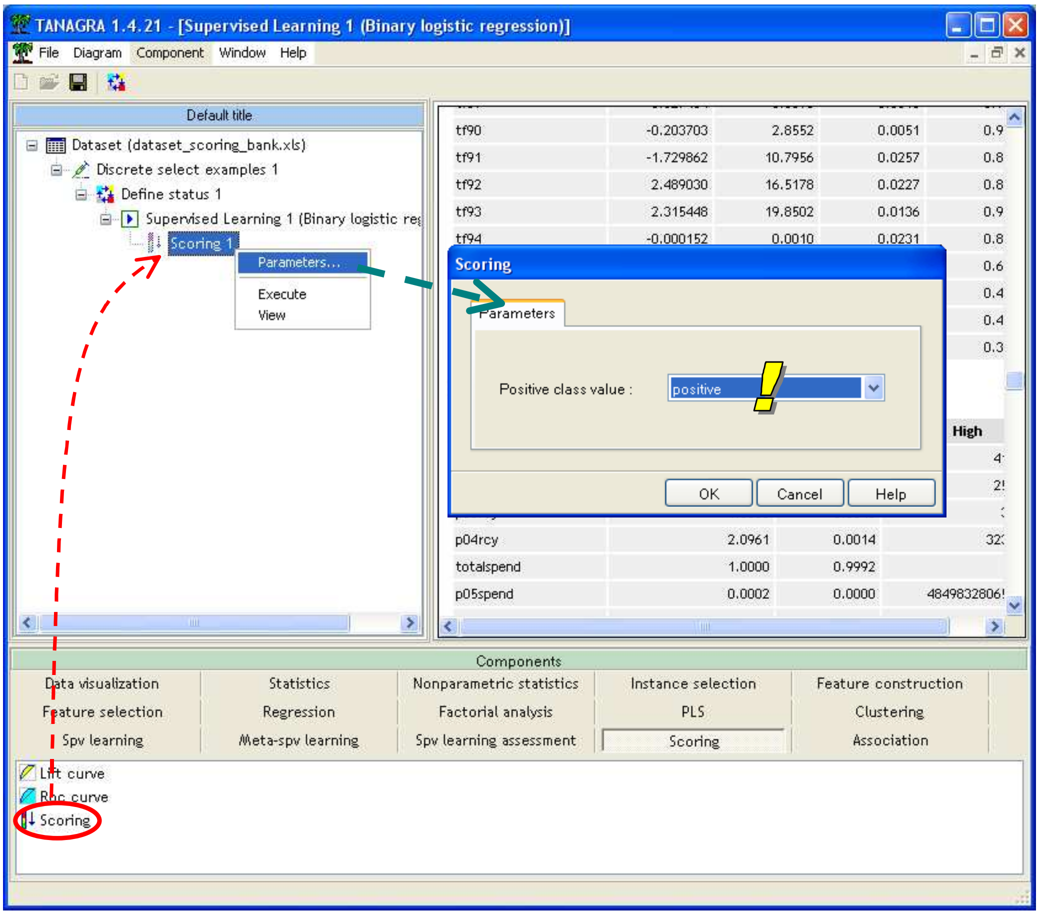
TANAGRA shows also odds-ratio for each variable and their confidence interval.

Odds ratios and 95% confidence intervals

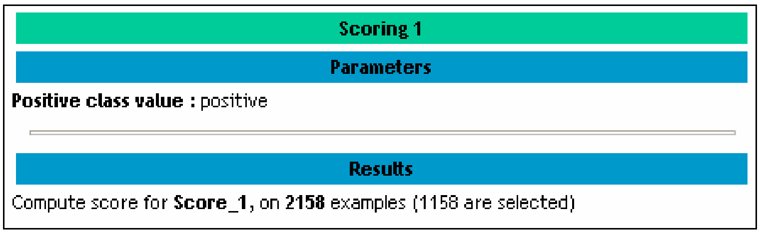
Attribute	Coef.	Low	High
p01rcy	1.8293	0.0081	412.8672
p02rcy	2.4053	0.0231	250.5605
p03rcy	1.5830	0.0643	38.9645
p04rcy	2.0961	0.0014	3237.1737
totalspend	1.0000	0.9992	1.0008

Computing the lift curve (Gain chart)

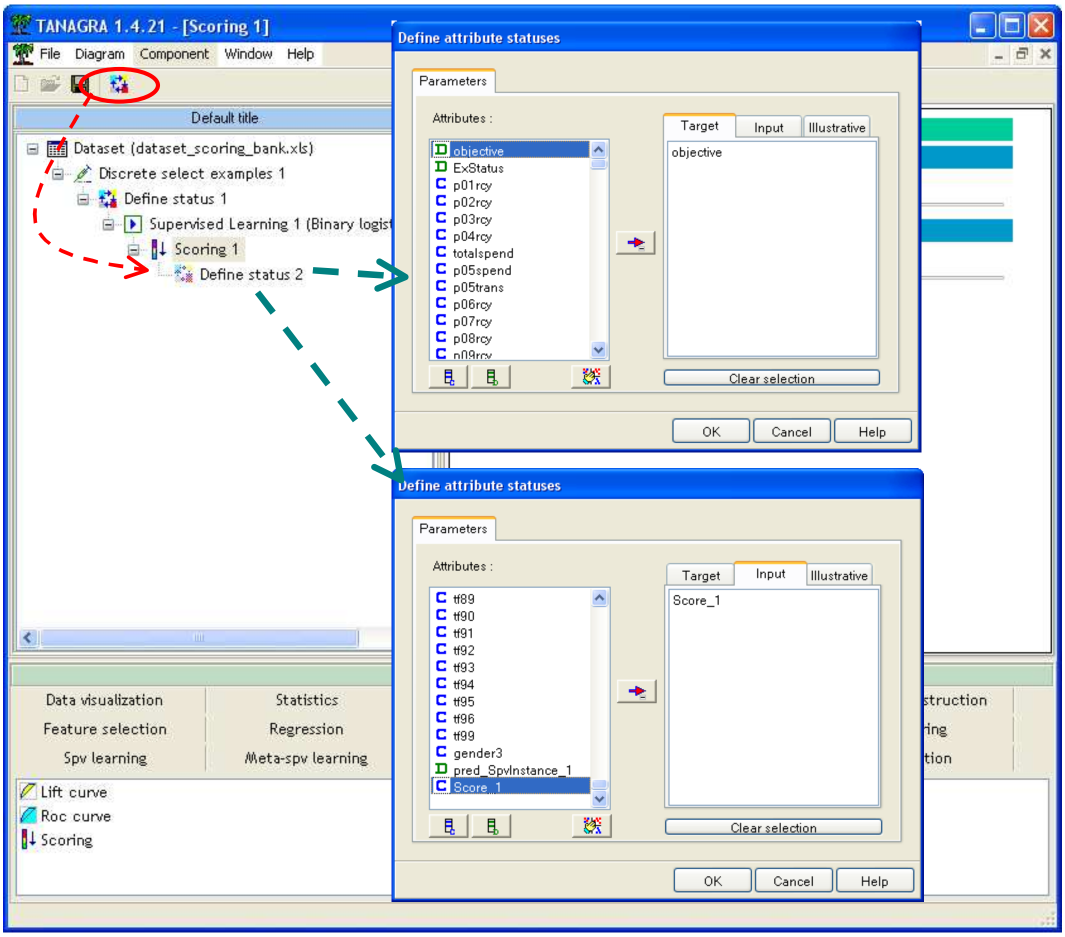
Firs, we must compute the "positive" probability of each individual. We add the component and we define the adequate parameter.



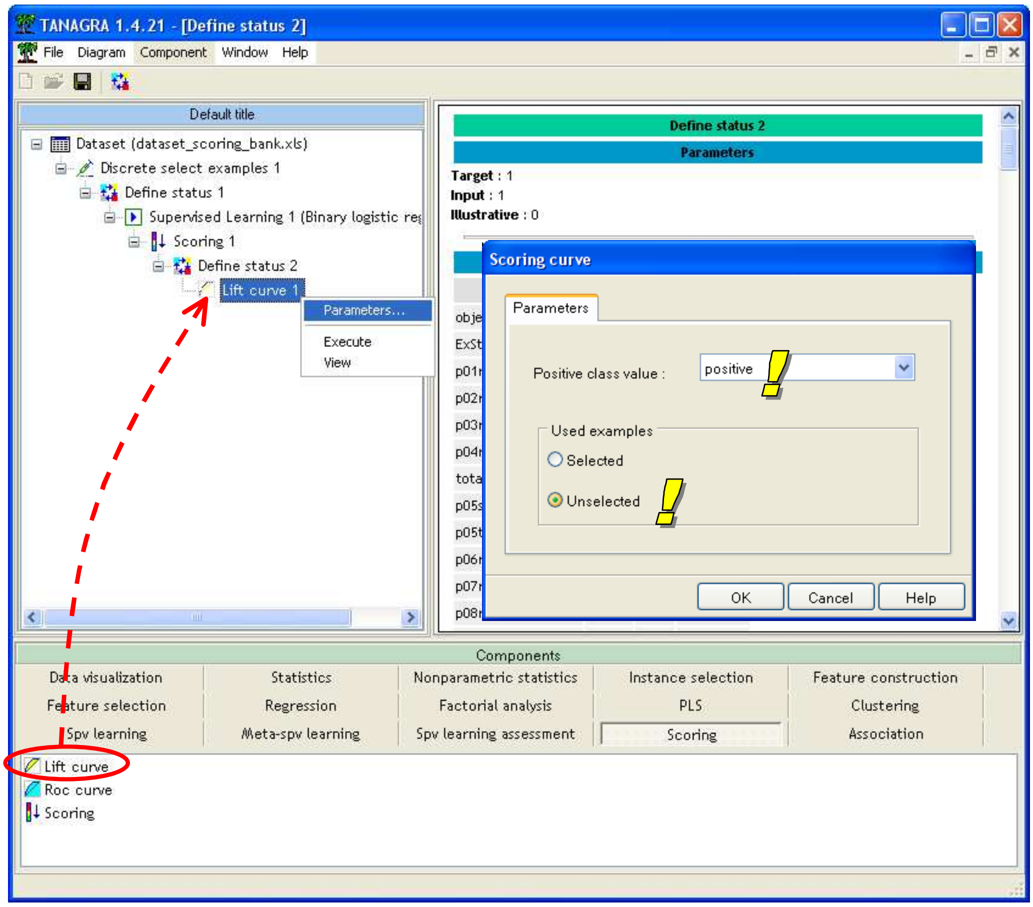
A new variable SCORE_1 is inserted in the dataset. The probability to be "positive" is computed for each example, even if it is not selected i.e. computed also for the test sample that will be used below.



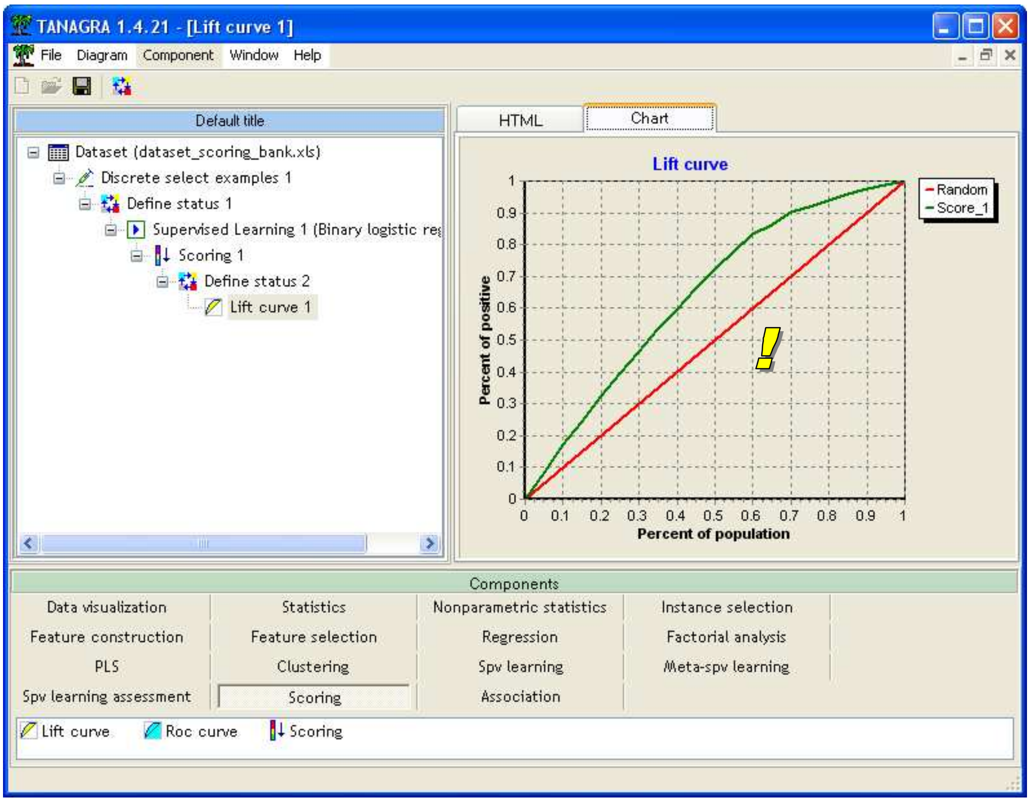
In order to compute the gain chart, we must indicate to TANAGRA the TARGET attribute (OBJECTIVE) and the attribute used for organize the individuals (INPUT attribute = SCORE_1). We insert again the DEFINE STATUS component using the toolbar shortcut and we set the adequate parameters.



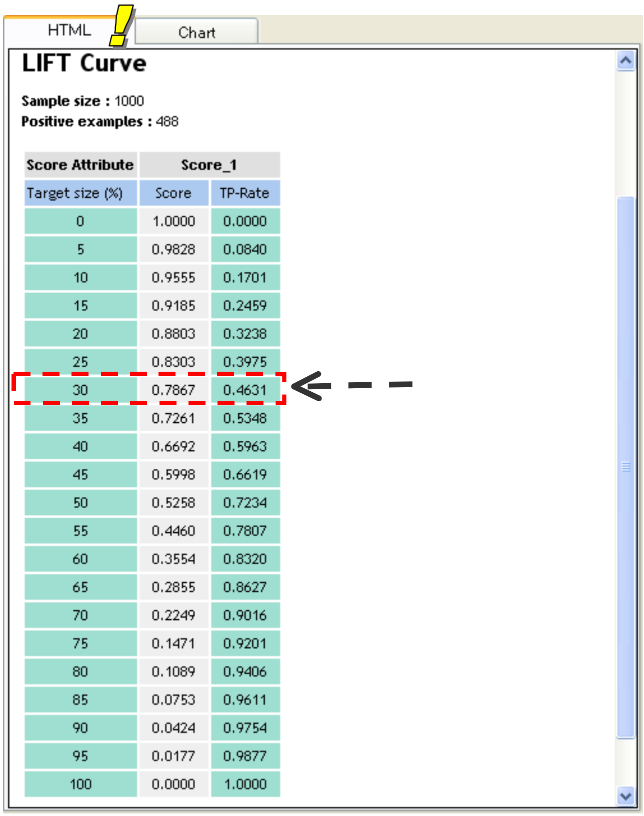
We must insert now the LIFT CURVE (SCORING tab) into the diagram. We must specify: the "positive" value of the target attribute and the examples used, i.e. the test set, for the gain chart.



We activate the VIEW menu and we obtain the following gain chart.



Into the HTML tab, we have the details of results.



HTML Chart

LIFT Curve

Sample size : 1000
Positive examples : 488

Score Attribute	Score_1	
Target size (%)	Score	TP-Rate
0	1.0000	0.0000
5	0.9828	0.0840
10	0.9555	0.1701
15	0.9185	0.2459
20	0.8803	0.3238
25	0.8303	0.3975
30	0.7867	0.4631
35	0.7261	0.5348
40	0.6692	0.5963
45	0.5998	0.6619
50	0.5258	0.7234
55	0.4460	0.7807
60	0.3554	0.8320
65	0.2855	0.8627
70	0.2249	0.9016
75	0.1471	0.9201
80	0.1089	0.9406
85	0.0753	0.9611
90	0.0424	0.9754
95	0.0177	0.9877
100	0.0000	1.0000

Among 1000 examples of the test set, there are 488 positives ones. We can reach 46% of the positives i.e. 46% x 488 # 225 individuals if we sent the mail to the 300 first examples (according their probability to be positive). If we had sent randomly the mails, the positive responses would be 30% x 488 # 146. The data mining process enables us to reach (225- 146) = 79 additional positive responses.

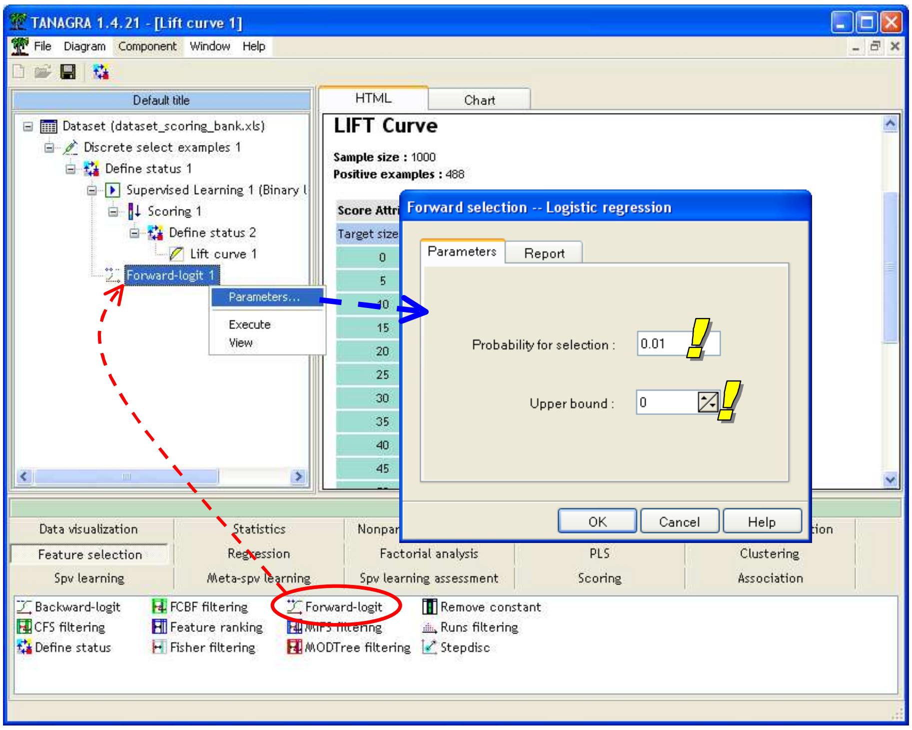
Logistic regression and variable selection

Variable selection – The FORWARD LOGIT component

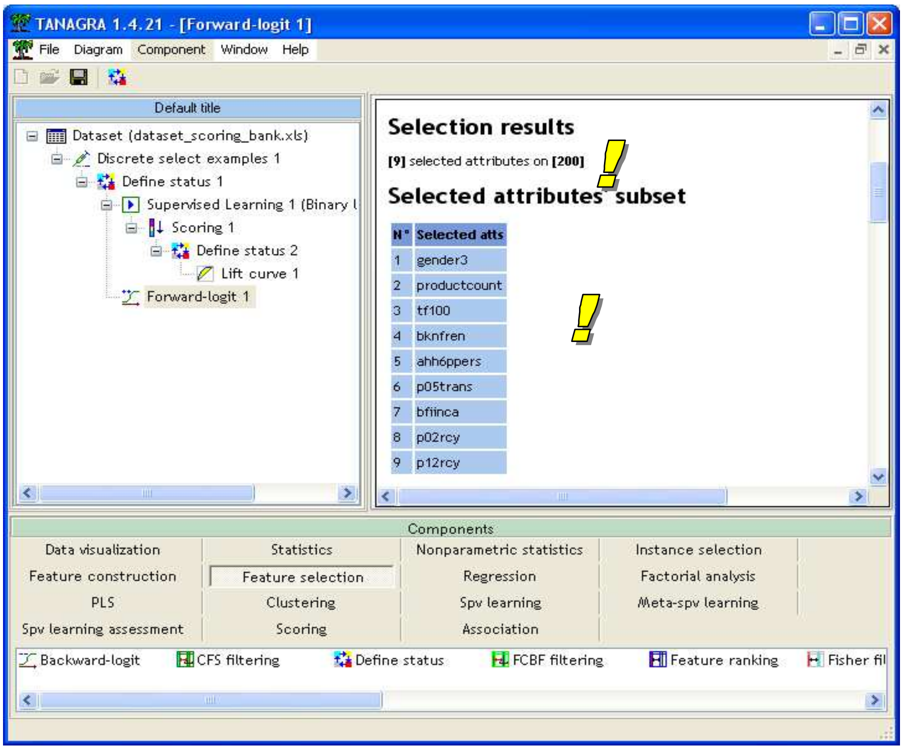
There are too many variables in our first regression. They all seem not relevant. An important task of the data miner is to reduce drastically the number of variables in order to retain only the relevant predictive attributes. The result will be then more interpretable, the model is often more robust, and it is more easy to use the model in an industrial context.

There are various variable selection strategies. In this tutorial, we study essentially FORWARD selection and BACKWARD elimination. In the forward (backward) approach, we add (remove) the most relevant (irrelevant) variable if their significance is lower (upper) than a user defined significance level (e.g. 1%).

We add the FORWARD-LOGIT (FEATURE SELECTION tab) component after DEFINE STATUS 1 into our diagram. We click on the menu parameter, we see that we may specify the probability for selection; we may also define the upper bound of the number of selected variables. If we set 0, only the first parameter (probability for selection) is activated.



We click on the VIEW menu in order to start the computation. According the number of predictive variables, the computation time may be high. On this dataset, it is about 5 seconds.



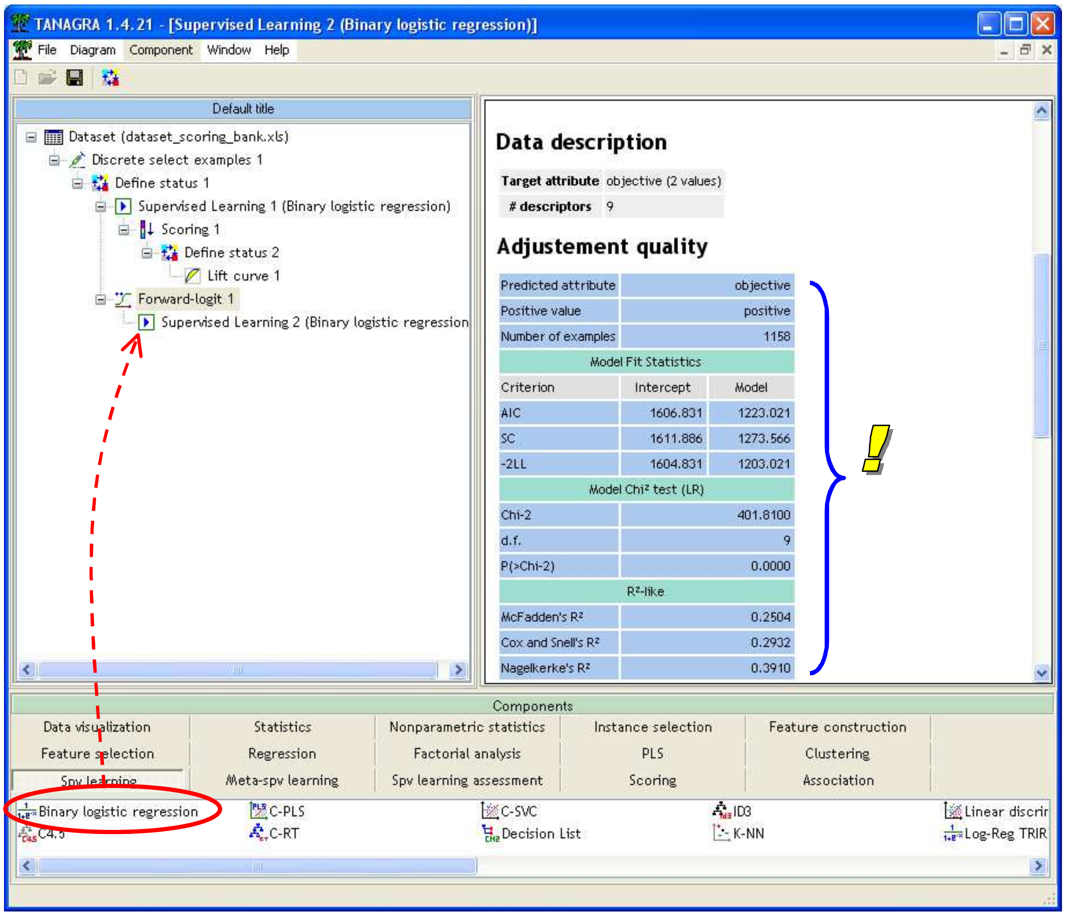
Nine variables are selected. We have the list of selected variables. Below, we have the details of computation. We can see at each step the best ones among the predictive variables. We voluntarily limit the table to the 5 first variables, because the read-out may become too blurred. We may modify the number of columns to display, if we set 0, all the variables are displayed.

Detailed results

N°	Current Reg.	Moved	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5
1	AIC : 1606.83 CHI-2 : 0.00 d.f. : 0 p-value : 0.0000	**gender3** Chi-2 : 217.468 p : 0.0000	**gender3** Chi-2 : 217.468 p : 0.0000	**productcount** Chi-2 : 100.896 p : 0.0000	**productcount6** Chi-2 : 98.436 p : 0.0000	**gender2** Chi-2 : 73.042 p : 0.0000	**tf33** Chi-2 : 56.896 p : 0.0000
2	AIC : 1380.08 CHI-2 : 228.75 d.f. : 1 p-value : 0.0000	**productcount** Chi-2 : 62.605 p : 0.0000	**productcount** Chi-2 : 62.605 p : 0.0000	**productcount6** Chi-2 : 56.182 p : 0.0000	**tf100** Chi-2 : 39.914 p : 0.0000	**tf37** Chi-2 : 39.800 p : 0.0000	**tf33** Chi-2 : 39.757 p : 0.0000
3	AIC : 1315.20 CHI-2 : 295.63 d.f. : 2 p-value : 0.0000	**tf100** Chi-2 : 30.484 p : 0.0000	**tf100** Chi-2 : 30.484 p : 0.0000	**bknfren** Chi-2 : 29.337 p : 0.0000	**tf37** Chi-2 : 28.987 p : 0.0000	**tf38** Chi-2 : 28.766 p : 0.0000	**tf33** Chi-2 : 28.116 p : 0.0000
4	AIC : 1286.09 CHI-2 : 326.74 d.f. : 3 p-value : 0.0000	**bknfren** Chi-2 : 21.123 p : 0.0000	**bknfren** Chi-2 : 21.123 p : 0.0000	**amtenglish** Chi-2 : 20.556 p : 0.0000	**bhlenglish** Chi-2 : 19.557 p : 0.0000	**brlprotest** Chi-2 : 18.602 p : 0.0000	**brlanglic** Chi-2 : 18.076 p : 0.0000
5	AIC : 1263.28 CHI-2 : 351.55 d.f. : 4 p-value : 0.0000	**ahh6ppers** Chi-2 : 11.637 p : 0.0006	**ahh6ppers** Chi-2 : 11.637 p : 0.0006	**amttagalog** Chi-2 : 11.284 p : 0.0008	**p05trans** Chi-2 : 10.671 p : 0.0011	**p05spend** Chi-2 : 10.241 p : 0.0014	**bimprovres** Chi-2 : 9.602 p : 0.0019
6	AIC : 1253.31 CHI-2 : 363.52 d.f. : 5 p-value : 0.0000	**p05trans** Chi-2 : 10.933 p : 0.0009	**p05trans** Chi-2 : 10.933 p : 0.0009	**p05spend** Chi-2 : 10.353 p : 0.0013	**p02rcy** Chi-2 : 9.134 p : 0.0025	**bimprovres** Chi-2 : 8.727 p : 0.0031	**bfi50plus** Chi-2 : 8.226 p : 0.0041
7	AIC : 1243.20 CHI-2 : 375.63 d.f. : 6 p-value : 0.0000	**bfiinca** Chi-2 : 9.631 p : 0.0019	**bfiinca** Chi-2 : 9.631 p : 0.0019	**bfiincm** Chi-2 : 9.042 p : 0.0026	**p02rcy** Chi-2 : 8.455 p : 0.0036	**bfi50plus** Chi-2 : 8.418 p : 0.0037	**binminca** Chi-2 : 8.045 p : 0.0046
8	AIC : 1235.68 CHI-2 : 385.15 d.f. : 7 p-value : 0.0000	**p02rcy** Chi-2 : 8.781 p : 0.0030	**p02rcy** Chi-2 : 8.781 p : 0.0030	**p12rcy** Chi-2 : 7.892 p : 0.0050	**amttagalog** Chi-2 : 6.754 p : 0.0094	**brlanglic** Chi-2 : 6.162 p : 0.0130	**tf68** Chi-2 : 5.591 p : 0.0181
9	AIC : 1228.53 CHI-2 : 394.31 d.f. : 8 p-value : 0.0000	**p12rcy** Chi-2 : 7.248 p : 0.0071	**p12rcy** Chi-2 : 7.248 p : 0.0071	**amttagalog** Chi-2 : 6.542 p : 0.0105	**brlanglic** Chi-2 : 6.269 p : 0.0123	**gender1** Chi-2 : 5.923 p : 0.0149	**gender2** Chi-2 : 5.923 p : 0.0149
10	AIC : 1223.02 CHI-2 : 401.81 d.f. : 9 p-value : 0.0000	-	**amttagalog** Chi-2 : 6.045 p : 0.0139	**brlanglic** Chi-2 : 5.720 p : 0.0168	**gender1** Chi-2 : 5.703 p : 0.0169	**gender2** Chi-2 : 5.703 p : 0.0169	**tf68** Chi-2 : 5.126 p : 0.0236

Regression with the selected variables

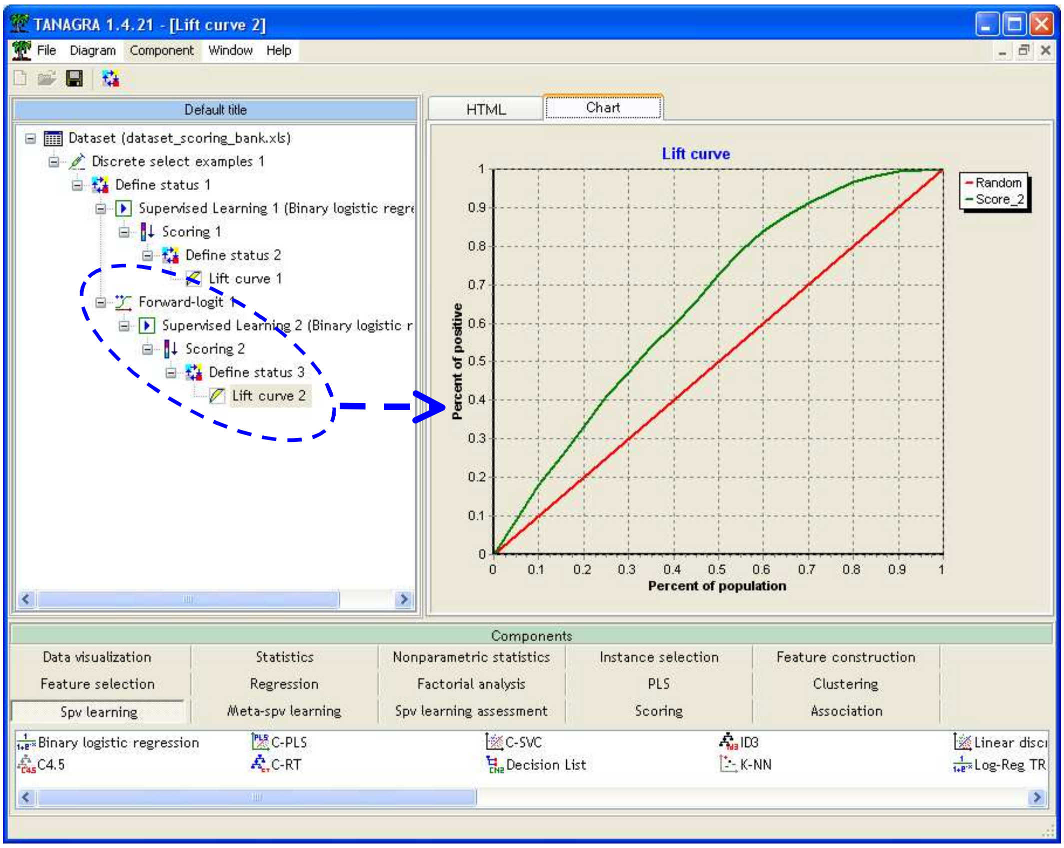
We add again the BINARY LOGISTIC REGRESSION (SPV LEARNING tab) after the FORWARD LOGIT 1 component. The computation is now realized on the 9 selected variables.



According to the confusion matrix and pseudo-R2, this new regression seems less powerful. But when we consider the criteria which take into account the complexity of the model such as AIC or SC, the last logistic regression is in reality preferable.

Criterion	Intercept only	Intercept + 200 variables	Intercept + 9 variables
AIC	1611.886	1371.488	**1223.021**
CS (or BIC)	1604.831	2387.433	**1273.566**

We insert again the same components as above with the adequate parameters: SCORING + DEFINE STATUS (SCORE_2 is the INPUT attribute, OBJECTIVE is always the TARGET) + LIFT. We obtain the following curve.



The curve is very similar to the previous. If we read the gain table (HTML tab), we see that for 30% of the first examples, we obtain 47% of positive individuals. The accuracy is the same one, but the new model now comprises only 9 variables. The interpretation of coefficients is easier.

BACKWARD elimination

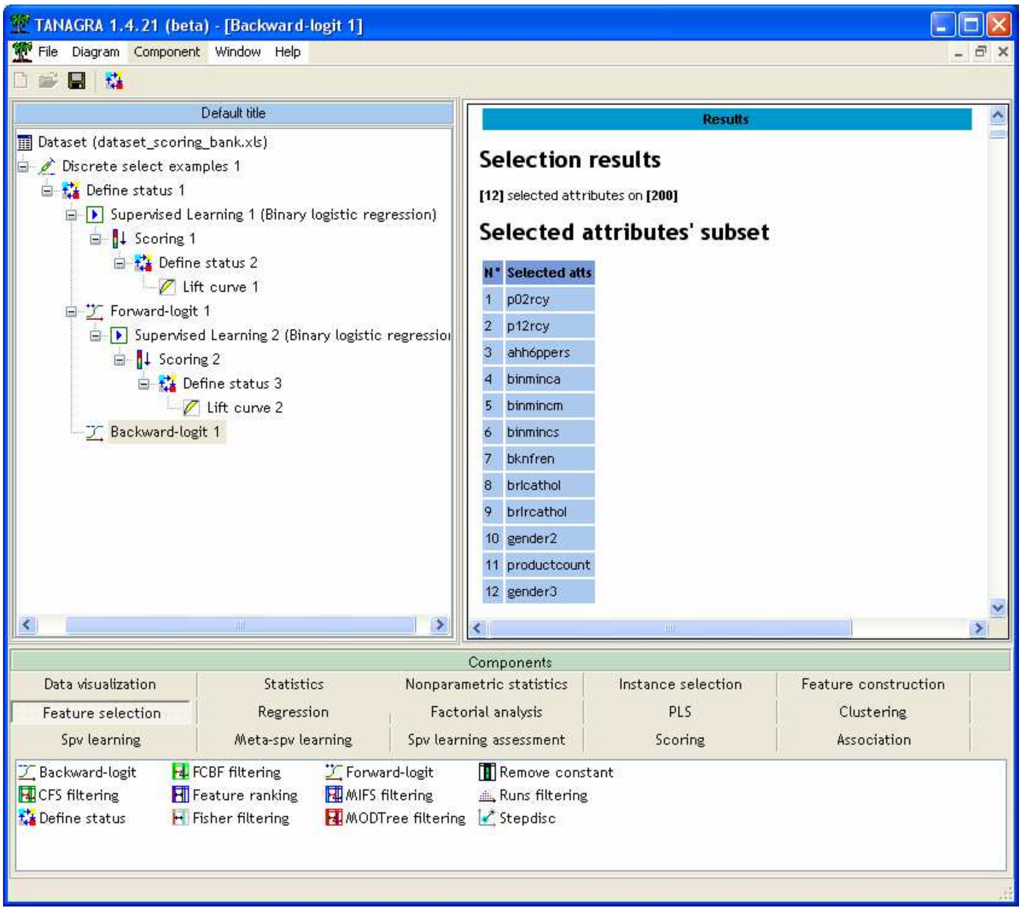
TANAGRA implements also the BACKWARD elimination strategy (BACKWARD LOGIT – FEATURE SELECTION tab). In some circumstances, some authors[2] claim that this approach is more efficient. This is surely true, but when we deal with very large dataset, the computation time becomes prohibitive because we perform many optimizations of the likelihood[3].

In my opinion, I think that these automatic variable selection processes (forward, backward and the other ones) enable us mostly to study deeply the relations between variables and to choose manually, according to the domain knowledge, the most adequate set of variables.

Just to give an idea (save the diagram before to start the process), in our dataset, the backward elimination selection takes a long time i.e. 872 second # 14 minutes on my computer. At the end, 12 variables are selected.

[2] S. Menard, « Applied Logistic Regression Analysis - Second Edition », Quantitative Applications in the Social Sciences Series, Sage Publications, 2002; page 64.

[3] We use the LEVENBERG-MARQUARDT algorithm, a variant of NEWTON-RAPHSON.



Six (6) variables are common between the 12 selected variables with the backward elimination process and the 9 selected with forward selection.

Backward	Forward
ahh6ppers	ahh6ppers
-	bfiinca
binminca	-
binmincm	-
binmincs	-
bknfren	bknfren
brlcathol	-
brlrcathol	-
gender2	-
gender3	gender3
p02rcy	p02rcy
-	p05trans
p12rcy	p12rcy
productcount	productcount
-	tf100

Conclusion

In this tutorial, we presented the construction of the lift curve with the logistic regression method. It was an opportunity to introduce two new components (version 1.4.21 of TANAGRA) dedicated to the supervised variable selection for logistic regression.