Objectif

Sélection de variables pour la régression logistique. Application au ciblage clientèle. Construction de la courbe lift (Gain Chart).

Le ciblage marketing (ou scoring) est certainement une des applications les plus populaires du Data Mining. Prenons un exemple pour fixer les idées : un établissement bancaire souhaite promouvoir un nouveau produit auprès de ses clients. Son budget est limité. Il ne peut pas, et de toute manière ne souhaite pas, solliciter tous ses clients. Il doit en priorité cibler les personnes les plus susceptibles de répondre positivement à son offre.

Il s'agit bien d'un apprentissage supervisé. La variable à prédire est la réponse positive ou non à la sollicitation. Les variables prédictives sont les différents descripteurs qui caractérisent les prospects (ex. revenu, âge, profession, comportement par rapport autres produits, etc.). Mais l'idée n'est pas tant de classer les individus, il s'agit plutôt de les hiérarchiser selon leur appétence c.-à-d. leur aptitude à répondre de manière positive à l'offre. Par la suite, en fonction de ses contraintes (budget) et de ses objectifs (parts de marché), le décideur pourra définir le nombre de client qu'il convient de contacter, il nous revient de lui indiquer le nombre de réponses positives qu'il peut espérer obtenir. Nous disposons pour cela d'un outil dédié : la courbe lift ou courbe de gain.

Dans ce didacticiel, nous présentons la mise en œuvre de la régression logistique dans le cadre du scoring marketing. L'objectif est double : (1) comment construire et lire la fameuse courbe lift à l'aide de TANAGRA ; (2) montrer l'intérêt et l'efficacité des techniques de sélection de variables associées à la régression logistique dans ce contexte.

Données

Pour illustrer notre propos, nous utilisons des données réelles/réalistes en provenance du site http://www.ssc.ca/documents/case_studies/2000/datamining_f.html. Il contient 2158 observations et 200 variables prédictives. Si le nombre d'observations est relativement faible, le nombre de variables correspond à peu près à ce que l'on rencontre souvent dans les études réelles : l'entrepôt de données de l'entreprise est à même de nous fournir un nombre élevé de variables, peu ou prou pertinentes, charge au data miner d'y discerner les variables appropriées pour le ciblage.

Pour une manipulation aisée, le fichier a été transformé au format EXCEL. Nous avons rajouté une variable indicatrice *ExStatus* (example status). Elle permet de le subdiviser aléatoirement en 1158 observations pour l'apprentissage et 1000 observations pour la construction de la courbe lift. Nous disposons ainsi d'un dispositif pour comparer des modèles avec des degrés de liberté différents. Le fichier est disponible en ligne[1].

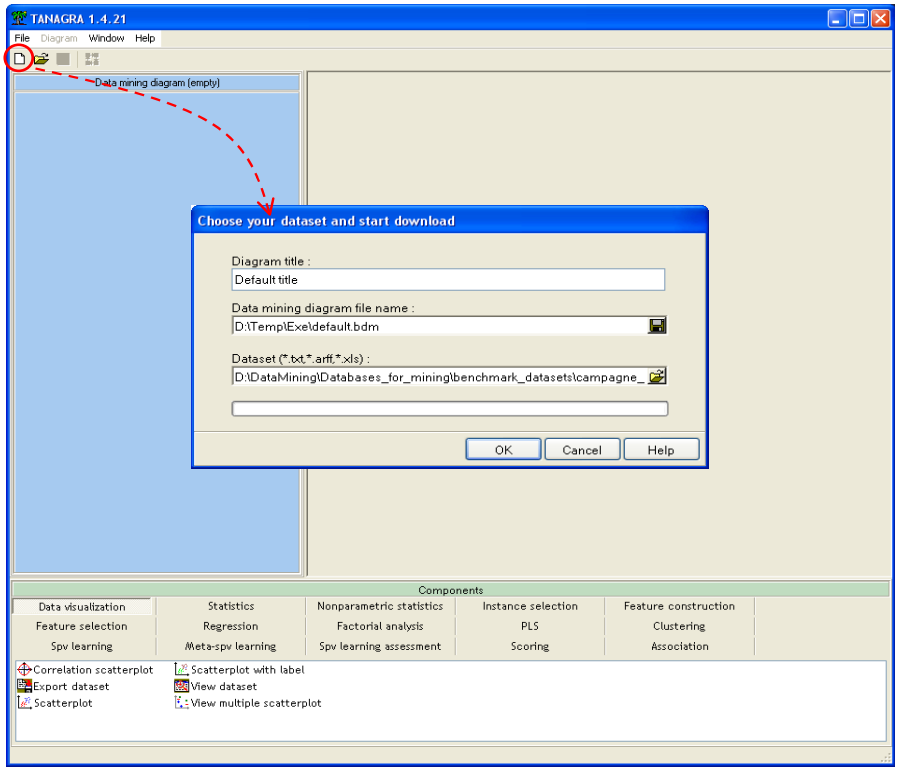
Régression logistique et courbe lift

Création du diagramme et importation des données

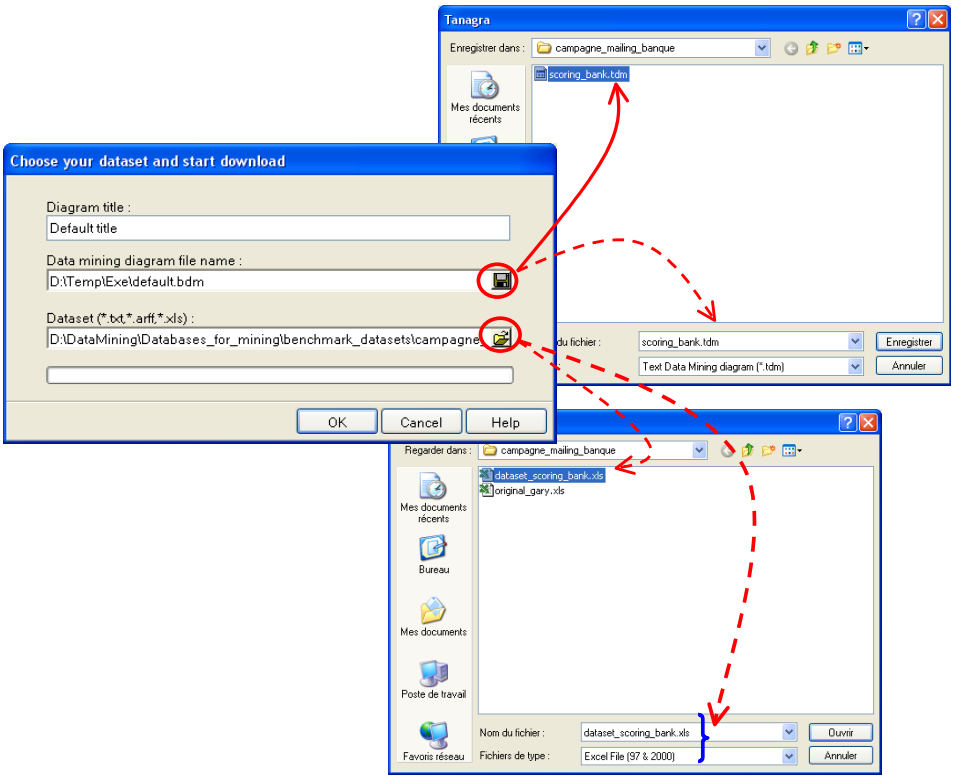
Il est possible d'ouvrir le fichier dans le tableur EXCEL et de lancer TANAGRA en lui transmettant directement les données en utilisant la macro complémentaire TANAGRA.XLA. Dans ce didacticiel, nous préférons importer directement le fichier dans TANAGRA. Ce faisant, nous bénéficions principalement de 2 avantages : le temps d'importation est réduit ; nous pouvons accéder aux données même si le tableur EXCEL n'est pas installé sur notre ordinateur.

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

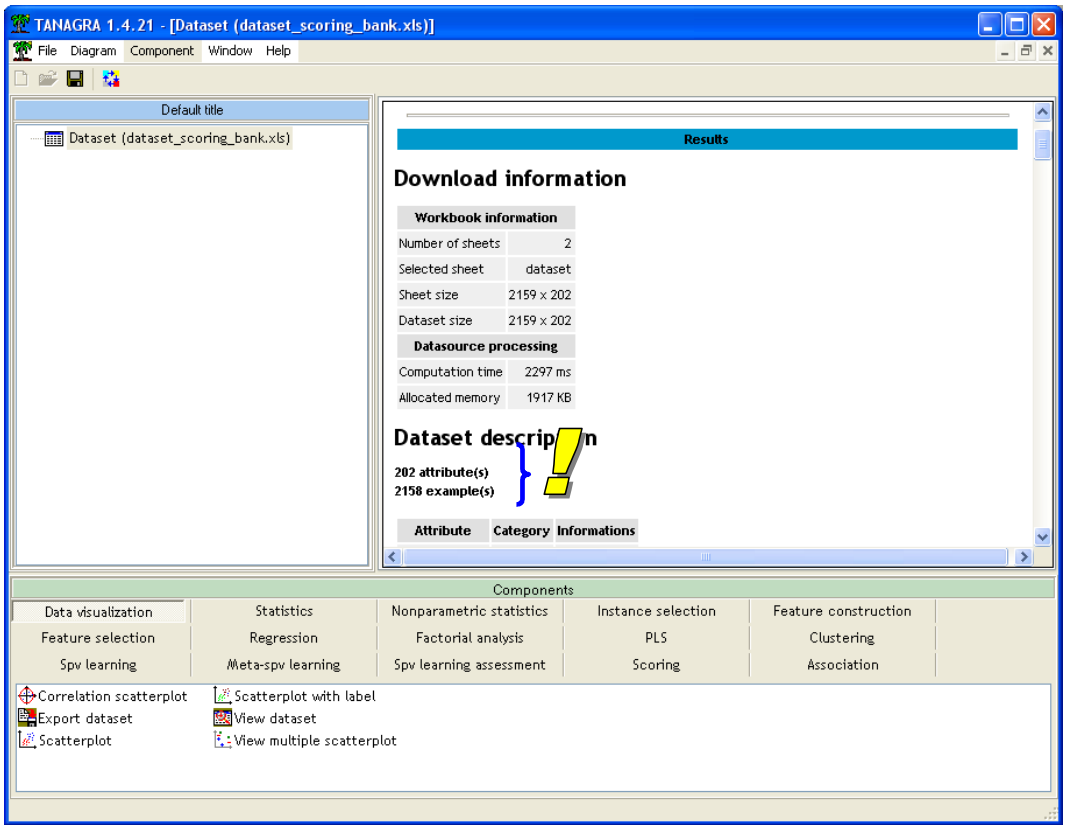
Après avoir lancé TANAGRA, pour créer un nouveau diagramme, nous activons le menu FILE / NEW. Une boîte de dialogue apparaît nous invitant à désigner le fichier de données et le nom du diagramme que nous sommes en train de créer.



Nous sélectionnons le fichier de données DATASET_SCORING_BANK.XLS. Attention, pour que l'importation de ce type de fichier se déroule correctement, le fichier ne doit pas être en cours d'édition dans le tableur EXCEL, les données doivent être situées dans la première feuille du classeur. Il nous faut également spécifier le nom du diagramme et son répertoire de destination.



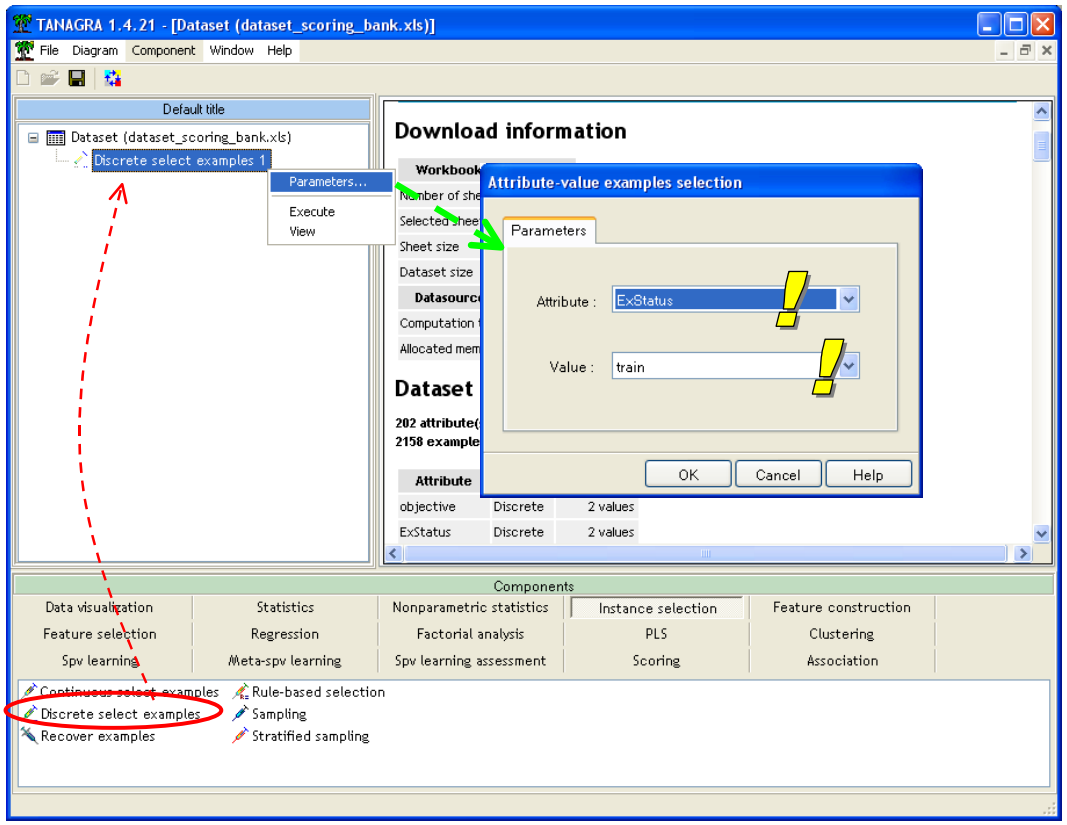
Nous validons en cliquant sur le bouton OK, les données sont chargées et un nouveau diagramme est créé. Vérifions que 202 variables et 2158 observations ont bien été importées.



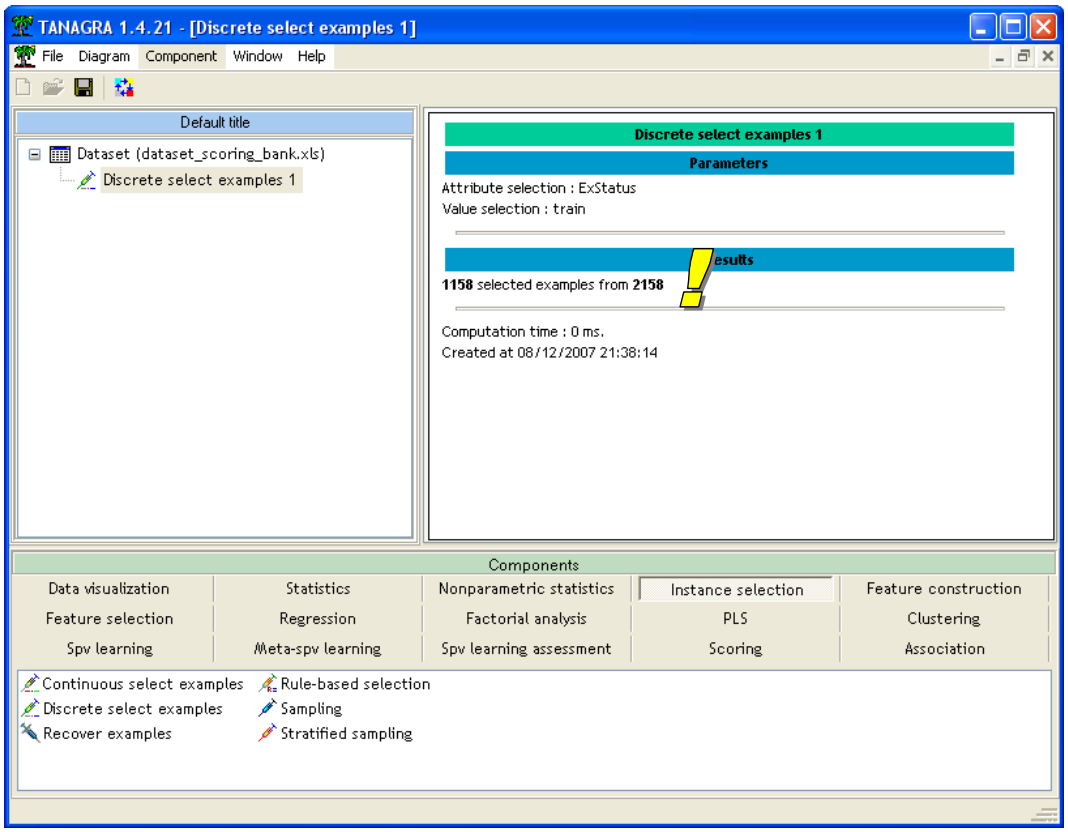
Subdivision apprentissage et test

Dans un premier temps, nous devons partitionner le fichier de données : une première partie, dite « apprentissage », est utilisée pour la construction des modèles ; une seconde partie dite « test » est réservée pour leur évaluation. Cette subdivision est toujours souhaitable dès lors que nous voulons obtenir une évaluation crédible des performances. Elle n'est malheureusement réalisable que lorsque nous disposons d'une base comportant un nombre relativement important d'observations. En effet, nous courons le risque de compromettre l'apprentissage en lui soustrayant une partie des données porteuses d'informations.

Nous insérons le composant DISCRETE SELECT EXAMPLES (onglet INSTANCE SELECTION) dans le diagramme. Nous le paramétrons en activant le menu contextuel PARAMETERS : le rôle de chaque observation est défini par la variable EXSTATUS, les individus à sélectionner correspondent à la modalité TRAIN.

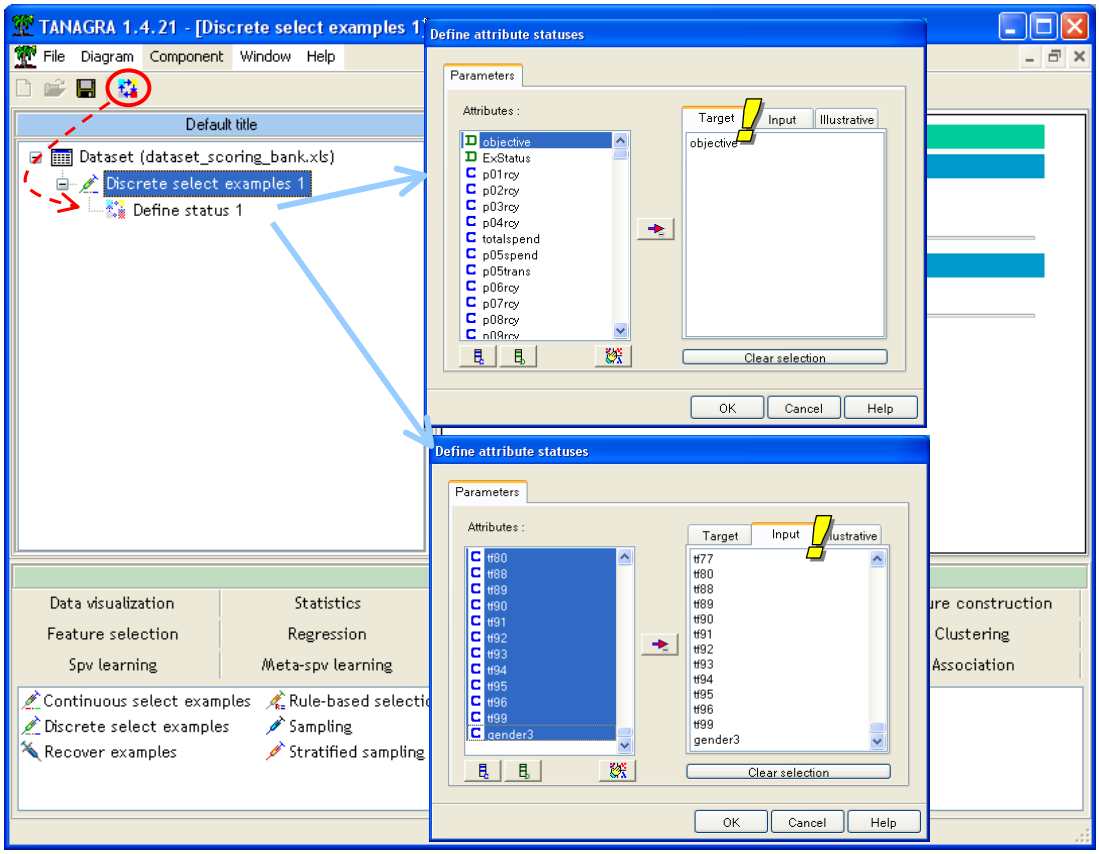


1158 observations sont sélectionnées pour l'apprentissage, les 1000 restantes seront mises de côté pour le moment.



Définition du problème

L'étape suivante consiste à choisir la variable à prédire, OBJECTIVE (TARGET), et les variables prédictives (INPUT), toutes les variables continues allant de P01RCY à GENDER3. Nous introduisons pour cela le composant DEFINE STATUS dans le diagramme. Le plus simple est de passer par le raccourci de la barre d'outils.

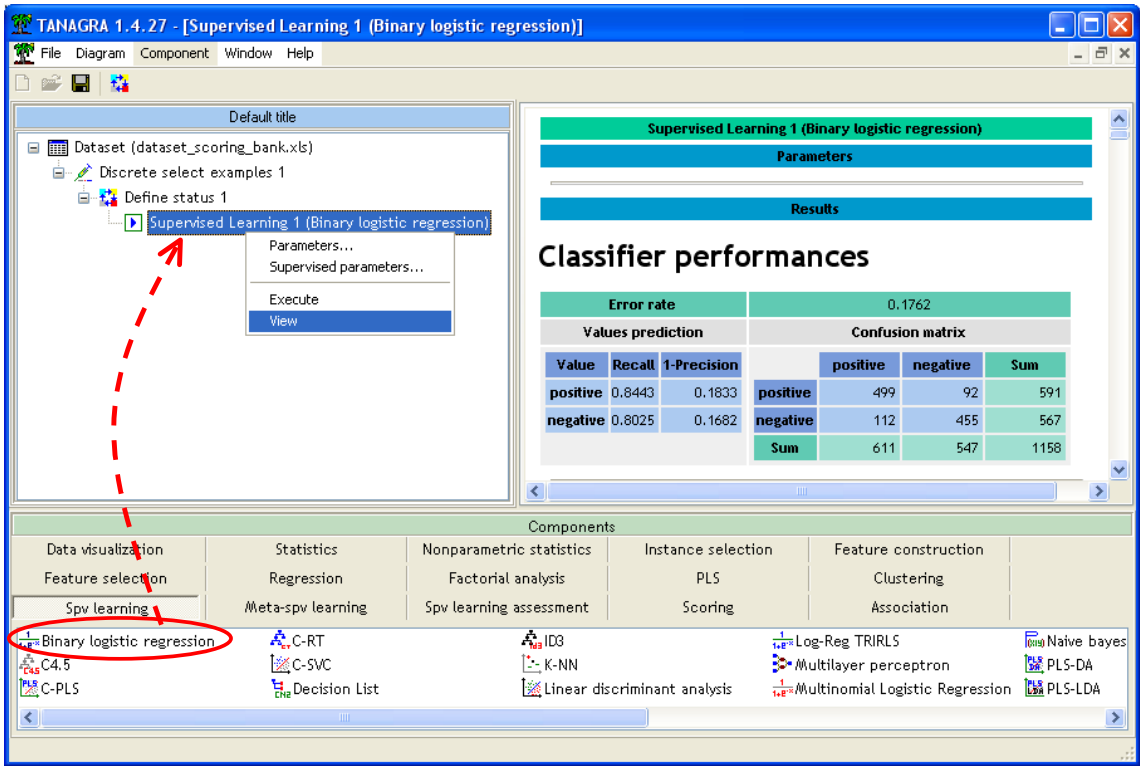


Choix de la méthode d'apprentissage : la régression logistique

Nous décidons de mettre en œuvre la régression logistique. C'est une méthode très populaire auprès des praticiens pour différentes, plus ou moins bonnes, raisons. Nous en retiendrons principalement deux : ses fondements théoriques sont directement adaptés au traitement des variables explicatives constituées d'un mélange de variables continues et de variables indicatrices 0/1 ; les coefficients issus de la régression s'interprètent comme un surcroît de risque d'appartenir à la modalité positive, ce sont les fameux odds-ratio.

En revanche, en termes de performances en prédiction, par rapport aux autres techniques induisant une séparation linéaire dans l'espace de représentation, l'analyse discriminante par exemple, la régression logistique ne se démarque pas vraiment.

Nous plaçons le composant BINARY LOGISTIC REGRESSION dans le diagramme. Nous activons le menu contextuel VIEW pour accéder aux résultats. La dimensionnalité étant assez élevée, le calcul prend un peu de temps, mais cela reste raisonnable (5 secondes sur notre machine).



La fenêtre de résultats est fractionnée en plusieurs parties.

La première partie comporte la matrice de confusion

Results

Classifier performances

Error rate			0.1762		
Values prediction			Confusion matrix		
Value	Recall	1-Precision		positive	negative	Sum
positive	0.8443	0.1833	positive	499	92	591
negative	0.8025	0.1682	negative	112	455	567
			Sum	611	547	1158

Elle est peu utile dans notre contexte. Nous ne cherchons pas affecter absolument tel ou tel individu à telle ou telle catégorie, nous cherchons plutôt à les hiérarchiser de manière à ce que les individus « intéressants », avec une propension élevée à être positif, soient classés premiers.

De plus, étant calculée sur les données en apprentissage, le taux d'erreur d'affectation qui en est issu est souvent optimiste, surtout au regard du faible ratio nombre d'individus positifs / nombre de variables (591 / 200 → 2.955) de notre fichier. Pour que les résultats en apprentissage soient réellement instructifs, et les coefficients interprétables, certains auteurs recommandent un ratio de 10 observations positives par co-variable[2]. Dans notre cas, l'objectif étant avant tout la prédiction, le véritable juge de paix sera la partie test des données que nous avons mise de côté.

[2] Voir P. Taffé, « Cours de Régression Logistique Appliquée », page 40, accessible en ligne à l'URL http://www.tesser-pro.org/stat/Cours_regression_logistique.pdf

La seconde partie indique la qualité globale de la régression

Plusieurs indicateurs sont proposés. Tous reposent sur la comparaison entre le modèle constitué de la seule constante et le modèle intégrant les variables explicatives. Certains indicateurs sont des ratios, telles les R^2 qui peuvent se lire, très approximativement, comme le coefficient de détermination de la régression linéaire. D'autres introduisent des tests statistiques basés sur le ratio de vraisemblance (LR). Dans notre cas, il semble que la régression soit globalement significative. D'autres enfin mettent en balance la qualité de l'ajustement (-2LL = -2 x log-vraisemblance que l'on cherche à minimiser) et la complexité du modèle. Si on se réfère au critère de Schwartz (SC que l'on retrouve sous l'appellation BIC dans d'autres logiciels), qui est très restrictif, il semble bien que le modèle soit trop complexe (SC du modèle avec la constante seule = 1604.831 vs. SC du modèle comportant les 200 variables = 2387.433).

Adjustement quality

Predicted attribute		objective
Positive value		positive
Number of examples		1158
Model Fit Statistics		
Criterion	Intercept	Model
AIC	1606.831	1361.431
SC	1611.886	2377.375
-2LL	1604.831	959.431
Model Chi² test (LR)		
Chi-2		645.4007
d.f.		200
P(>Chi-2)		0.0000
R²-like		
McFadden's R²		0.4022
Cox and Snell's R²		0.4273
Nagelkerke's R²		0.5698

La troisième partie comporte les coefficients du modèle

En plus des coefficients, nous disposons de l'estimation de leur écart-type, de la statistique de Wald destinée à évaluer leur significativité, c.-à-d. tester si le coefficient s'écarte significativement de 0, et de la probabilité critique du test. A ce stade commence réellement le travail d'analyse. A la lumière du signe, de la valeur et de la significativité des coefficients, l'expert du domaine sera à même d'interpréter les résultats, comprendre le sens des causalités, de proposer des études alternatives en supprimant manuellement certaines variables ou en rajoutant d'autres variables synthétiques, notamment pour mettre en évidence les interactions.

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-1.200958	-	-	-
p01rcy	0.607144	0.3711	2.6765	0.1018
p02rcy	1.052330	0.3809	7.6308	0.0057
p03rcy	0.540862	0.2763	3.8328	0.0503
p04rcy	0.763058	0.4569	2.7894	0.0949
totalspend	-0.000009	0.0001	0.0186	0.8917
p05spend	-10.168483	9.2515	1.2081	0.2717
p05trans	1.23	1535	0.0890	0.7654

Pour notre part, nous nous bornerons à remarquer que de nombreuses variables sont redondantes. Elles se gênent mutuellement dans la régression. La sélection de variables que nous introduirons plus loin jouera un rôle primordial. C'est souvent le cas dans le contexte des études réelles où les variables sont mises en vrac dans le but d'obtenir une prédiction aussi performante que possible, à charge pour la technique de sélectionner celles qui sont les plus pertinentes.

La quatrième partie comporte les odds-ratios

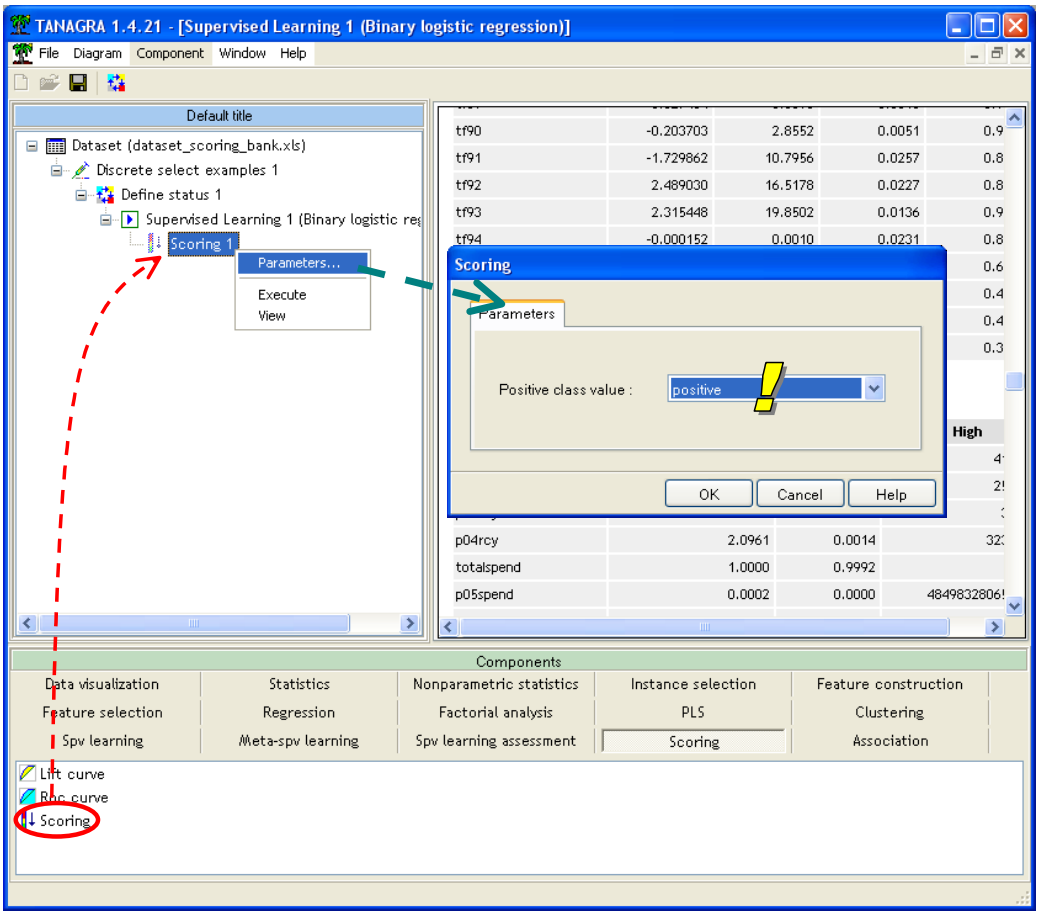
Il s'agit de l'exponentielle des coefficients. TANAGRA fournit aussi les intervalles de confiance à 5%.

Odds ratios and 95% confidence intervals

Attribute	Coef.	Low	High
p01rcy	1.8352	0.8867	3.7981
p02rcy	2.8643	1.3576	6.0435
p03rcy	1.7175	0.9994	2.9516
p04rcy	2.1448	0.8760	5.2516
totalspend	1.0000	0.9999	1.0001

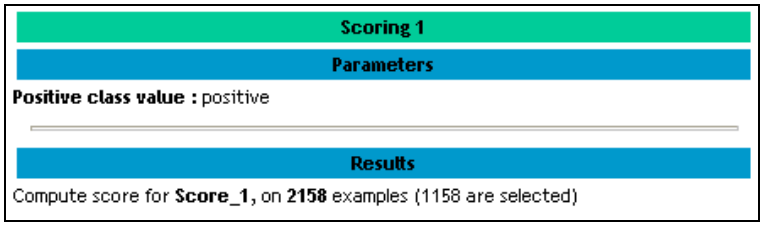
Scoring et construction de la courbe lift

Il nous faut maintenant attribuer à chaque individu sa probabilité d'être positif. Pour ce faire, nous insérons le composant SCORING (onglet SCORING) dans le diagramme. Nous le paramétrons de manière à calculer la probabilité de la modalité « positive ».

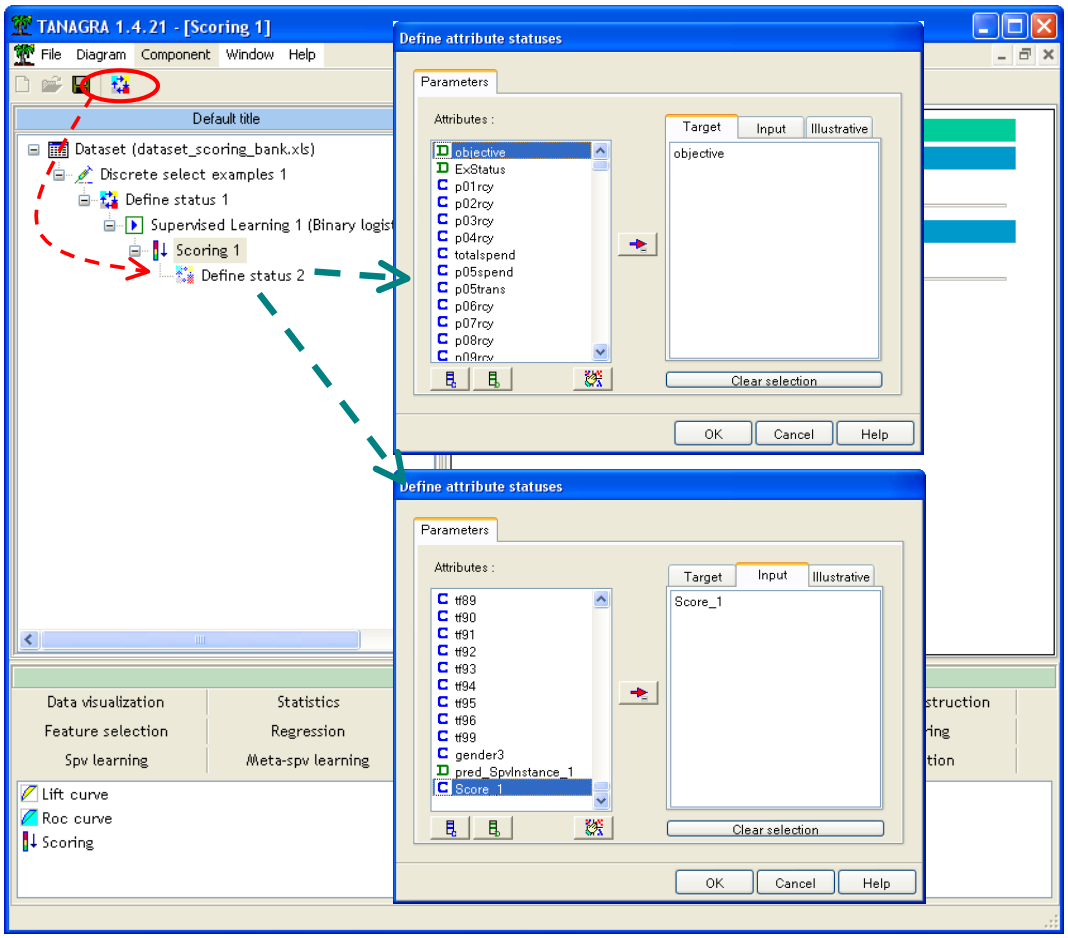


Nous activons le menu VIEW. TANAGRA indique qu'une nouvelle variable SCORE_1 a été créée. L'opération a été réalisée sur l'ensemble de la base, y compris les individus en test. Cette

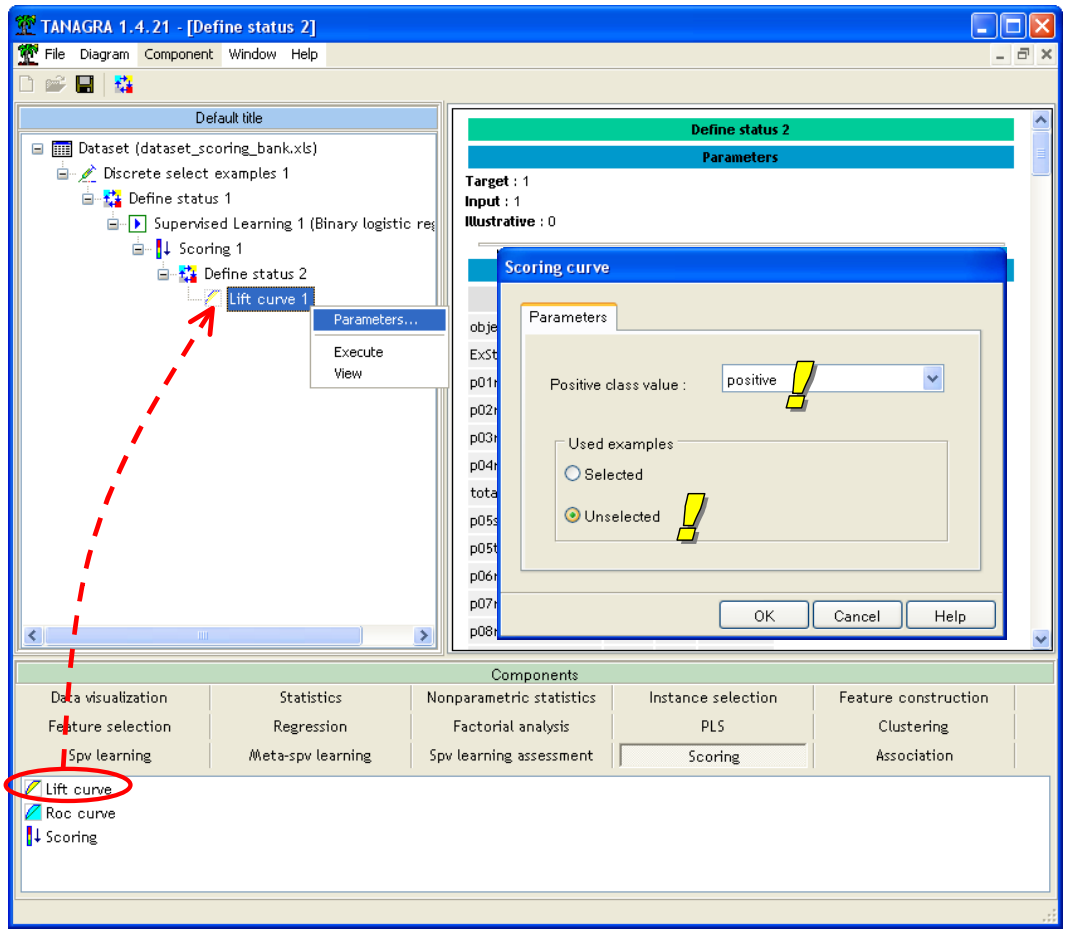
information est importante car c'est sur ces derniers que nous évaluerons la qualité du modèle par la suite.



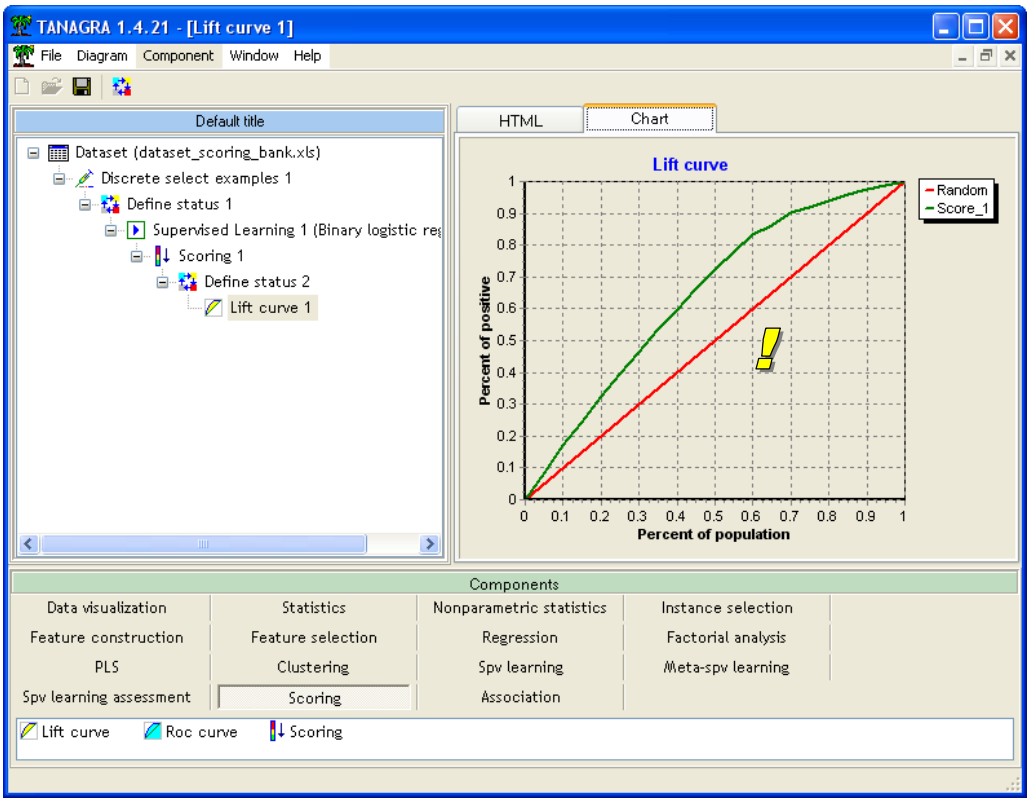
Pour construire le courbe lift, nous devons indiquer à TANAGRA la variable cible de référence et la variable qui sert à ordonner les observations. Nous introduisons de nouveau le composant DEFINE STATUS, toujours en utilisant le raccourci dans la barre d'outils. Nous plaçons en TARGET la variable OBJECTIVE, en INPUT la variable SCORE_1 construite précédemment.



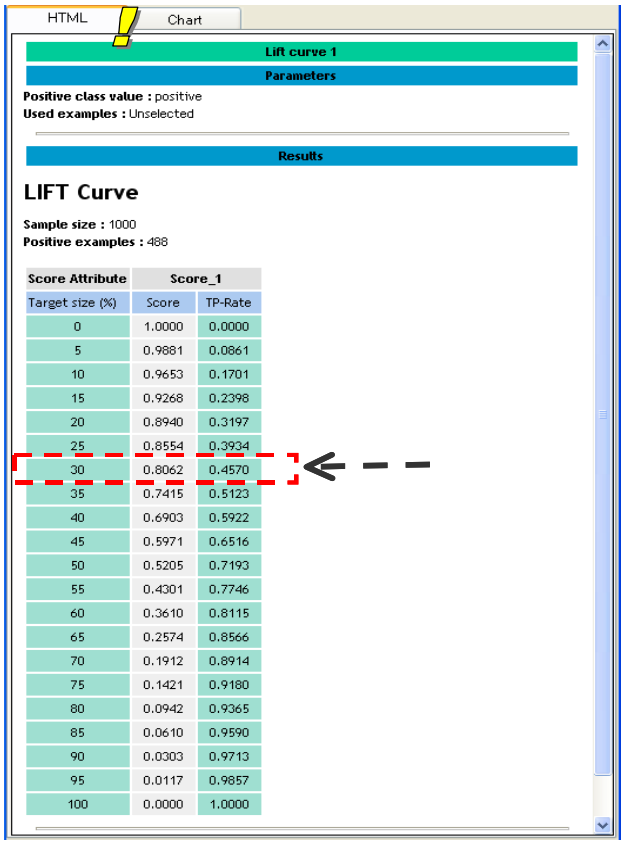
Il ne reste plus qu'à insérer le composant LIFT CURVE (onglet SCORING) dans le diagramme. Nous actionnons le menu PARAMETERS afin de spécifier : la modalité positive de la variable cible, les individus sur lesquels sera construite la courbe. Nous choisissons la partie test c.-à-d. les 1000 individus que nous avions mis de côté au départ.



Nous activons le menu VIEW pour accéder aux résultats, la courbe LIFT s'affiche directement.



Un graphique est toujours plaisant mais nous disposons de plus de détails dans l'onglet HTML.


HTML Chart

Lift curve 1

Parameters

Positive class value : positive
Used examples : Unselected

Results

LIFT Curve

Sample size : 1000
Positive examples : 488

Score Attribute	Score_1	
Target size (%)	Score	TP-Rate
0	1.0000	0.0000
5	0.9881	0.0861
10	0.9653	0.1701
15	0.9268	0.2398
20	0.8940	0.3197
25	0.8554	0.3934
30	0.8062	0.4570
35	0.7415	0.5123
40	0.6903	0.5922
45	0.5971	0.6516
50	0.5205	0.7193
55	0.4301	0.7746
60	0.3610	0.8115
65	0.2574	0.8566
70	0.1912	0.8914
75	0.1421	0.9180
80	0.0942	0.9365
85	0.0610	0.9590
90	0.0303	0.9713
95	0.0117	0.9857
100	0.0000	1.0000

Sur les 1000 individus en test, 488 sont positifs. Si nous ciblons 300 individus (30% de 1000), nous pouvons espérer atteindre 46% des positifs, soit 46% x 488 # 225 individus. Si nous avions envoyé les lettres au hasard, sans ciblage, nous aurions obtenu 30% x 488 # 146 réponses positives. C'est l'écart (225 – 146) = 79 individus supplémentaires conquis qui justifient notre présence dans les entreprises.

Régression logistique et sélection de variables

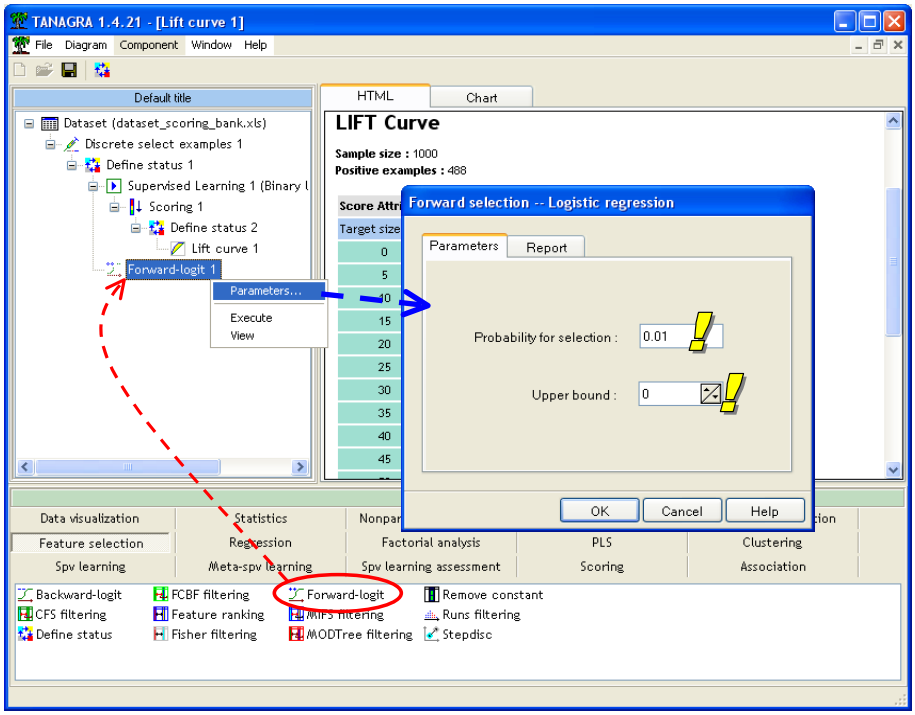
Sélection de variables – Le composant FORWARD LOGIT

Pour intéressante qu'elle soit, notre première analyse comporte une lacune importante. Le nombre de variables est trop important pour espérer extraire une interprétation intéressante des coefficients. D'autant plus qu'aucune variable ne semble significative au seuil de 1% que l'on s'est choisi. Nous devons réduire leur nombre.

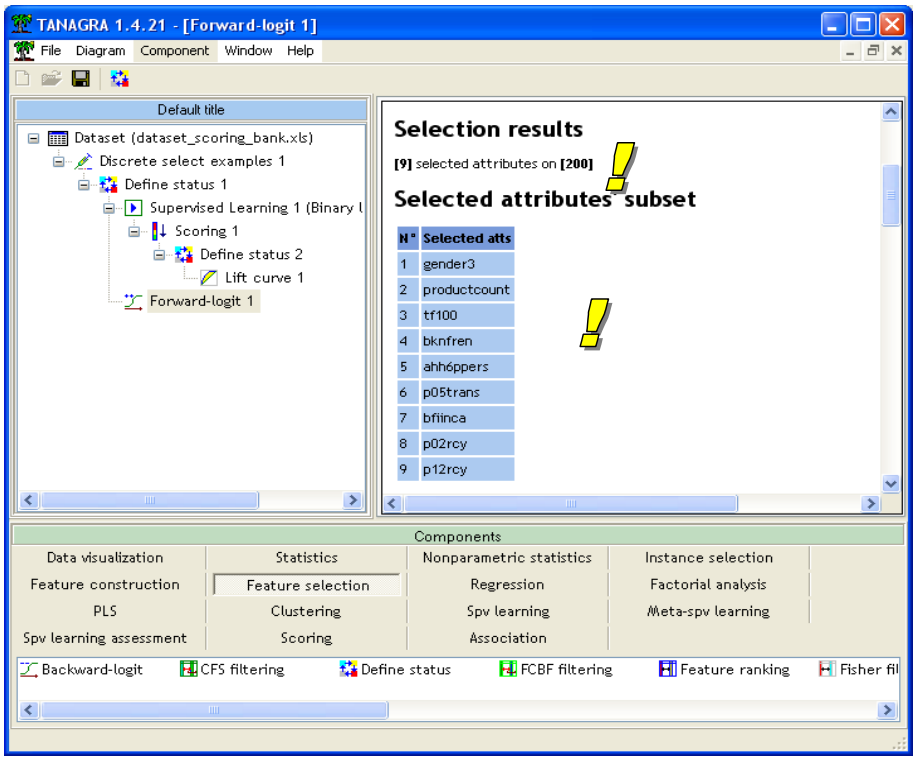
La sélection de variables est une étape primordiale. Elle facilite grandement la lecture des résultats et, de plus, le modèle est bien souvent plus performant. Le ratio nombre d'observations / nombre de variables étant amélioré, les estimations sont nettement plus fiables. De toute manière, même si les performances stagnaient, une réduction du nombre de variables est toujours un plus en termes de portabilité et d'industrialisation du modèle.

Il existe plusieurs stratégies de réduction de la dimensionnalité. Dans ce didacticiel, nous nous contenterons d'une approche purement mécanique en utilisant la sélection par avant (FORWARD SELECTION). Elle consiste à démarrer avec le modèle ne comportant que la constante, puis d'ajouter, au fur et à mesure, la variable la plus performante au sens du test du Score (Pour plus de détails, voir – http://eric.univ-lyon2.fr/~ricco/cours/slides/regression_logistique.pdf). La règle d'arrêt naturelle consiste à stopper l'adjonction lorsque, à une étape donnée, la meilleure variable n'est plus significative au sens du seuil de significativité (de 1%) que l'on s'est choisi.

Nous insérons le composant FORWARD-LOGIT (onglet FEATURE SELECTION) juste après le composant DEFINE STATUS 1 de notre diagramme. En activant le menu PARAMETER, nous constatons que nous avons la possibilité, en introduisant une valeur seuil strictement positive, de limiter la recherche en définissant un nombre maximum de variables à sélectionner. Cette option se révèle particulièrement pratique lorsque nous travaillons sur des bases comportant un très grand nombre de variables candidates (de l'ordre de plusieurs milliers). Dans notre cas, nous laissons ce paramètre à 0 c.-à-d. seul le paramètre « probabilité pour la sélection » est activé.



Nous activons le menu VIEW pour accéder aux résultats. Selon le nombre de variables et le nombre d'observations, le calcul peut être relativement long. Dans notre étude, il dure 5 secondes.



9 variables ont été sélectionnées. Elles sont directement proposées en INPUT à la sortie du composant. Dans la partie basse de la fenêtre, nous disposons du détail des calculs à chaque étape. Pour éviter la surabondance des informations, l'affichage est volontairement limité aux 5 premières variables, nous pouvons le paramétrer.

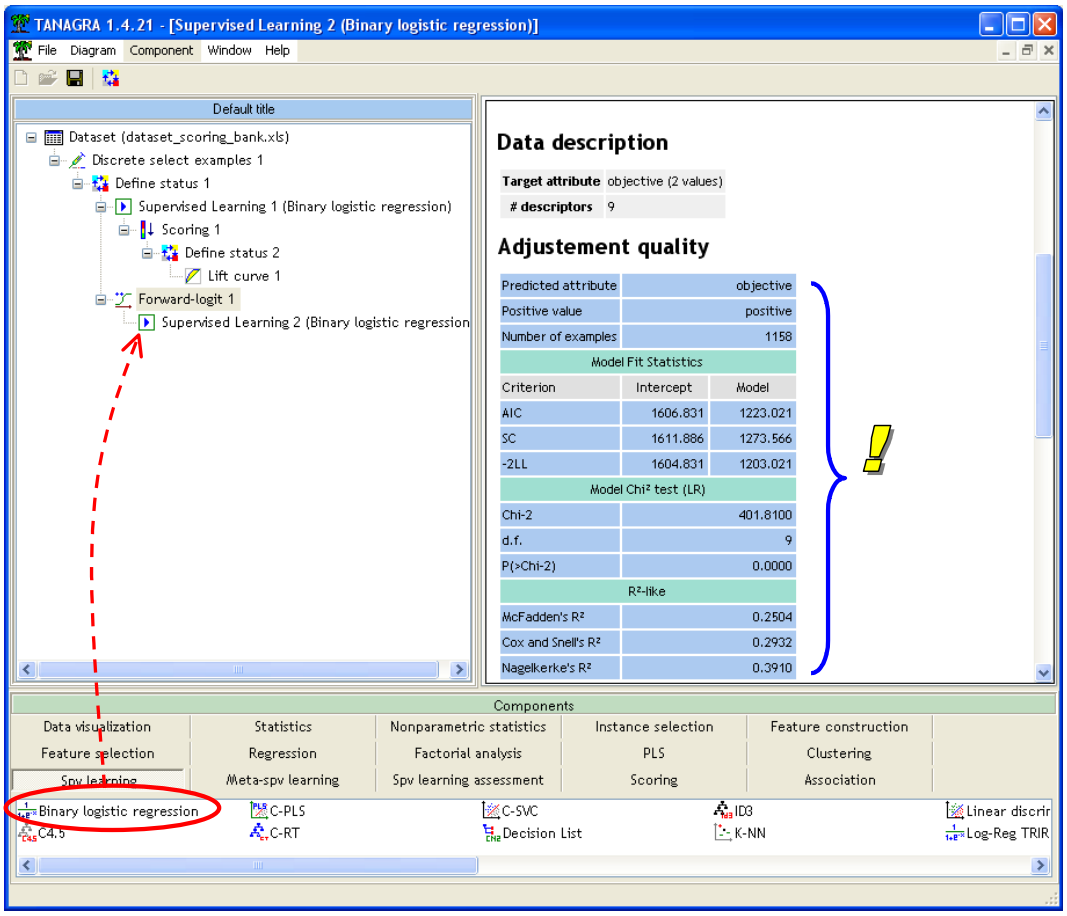
Detailed results

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

Le détail des résultats permet déjà de contrôler le processus. Il permet aussi de diagnostiquer finement le rôle des variables. A l'étape n°2, nous remarquerons par exemple que PRODUCTCOUNT et PRODUCTCOUNT6 sont en compétition. Une fois la première introduite, la seconde disparaît totalement des meilleures places. Il est vraisemblable que ces variables soient fortement redondantes.

Régression logistique sur les variables sélectionnées

De nouveau, nous introduisons le composant régression logistique binaire (BINARY LOGISTIC REGRESSION – onglet SPV LEARNING) à la suite du composant FORWARD-LOGIT 1. Il opère la régression sur les 9 variables sélectionnées.

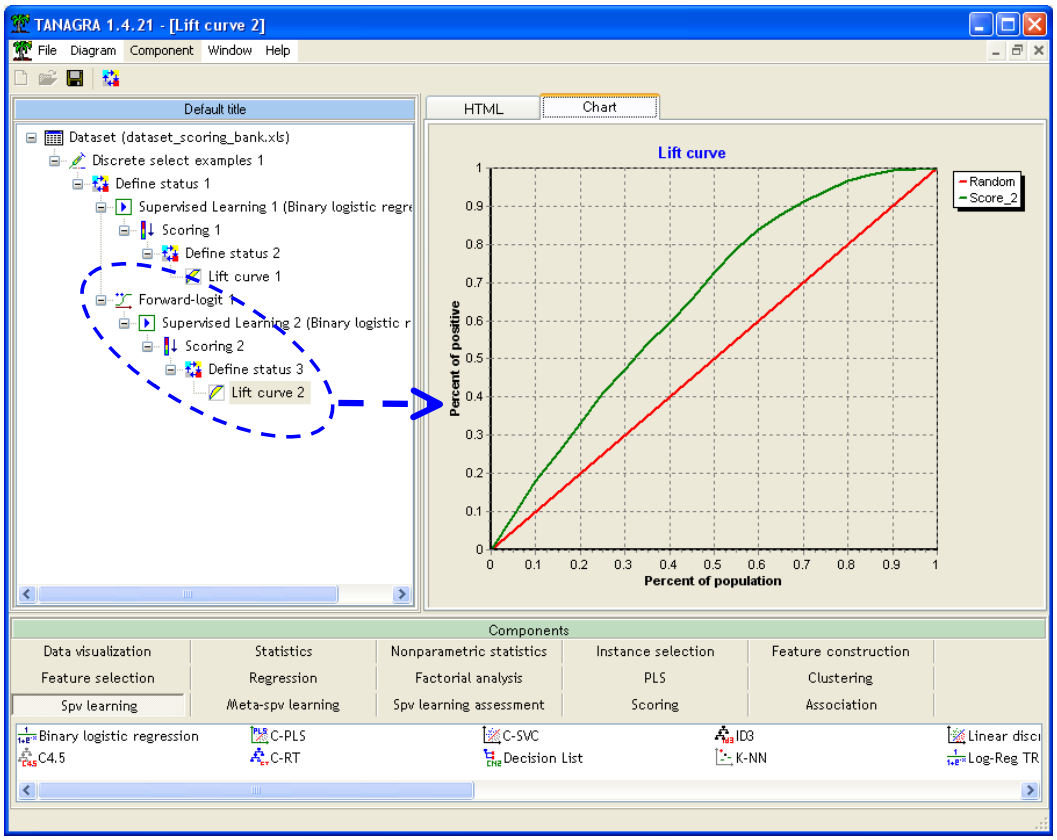


Si l'on s'en tient à la matrice de confusion et les pseudo-R², la régression semble de moins bonne qualité. Lorsqu'on se tourne vers les indicateurs tenant compte de la complexité (AIC et SC), on se rend compte que la réduction du nombre de variables améliore la qualité du modèle. Nous reprenons dans un tableau ci-dessous ces indicateurs.

Indicateur	Constante seule	Const. + 200 variables	Const. + 9 variables
AIC	1611.886	1361.431	**1223.021**
SC (ou BIC)	1604.831	2377.375	**1273.566**

Les deux critères s'accordent pour désigner le modèle à 9 variables comme le plus intéressant. Notons que le critère -2LL n'est absolument pas approprié dans le contexte de la sélection, il diminue mécaniquement lorsque nous augmentons le nombre de variables.

Il ne nous reste plus qu'à insérer les mêmes composants que précédemment : SCORING + DEFINE STATUS (SCORE_2 cette fois-ci en INPUT, OBJECTIVE toujours en TARGET) + LIFT avec les paramétrages adéquats. Nous obtenons la courbe suivante.



La courbe est quasiment identique. En nous penchant sur les détails (onglet HTML), nous constatons que lorsque nous ciblons les 30% premiers individus, nous pouvons espérer atteindre 47% des positifs, soit 47% x 488 # 230 individus. Le gain est faible par rapport à la régression précédente, il faut en convenir. A la différence que nous avons maintenant un modèle à 9 variables. L'interprétation des coefficients, et par là, des odds-ratios, est autrement plus aisée.

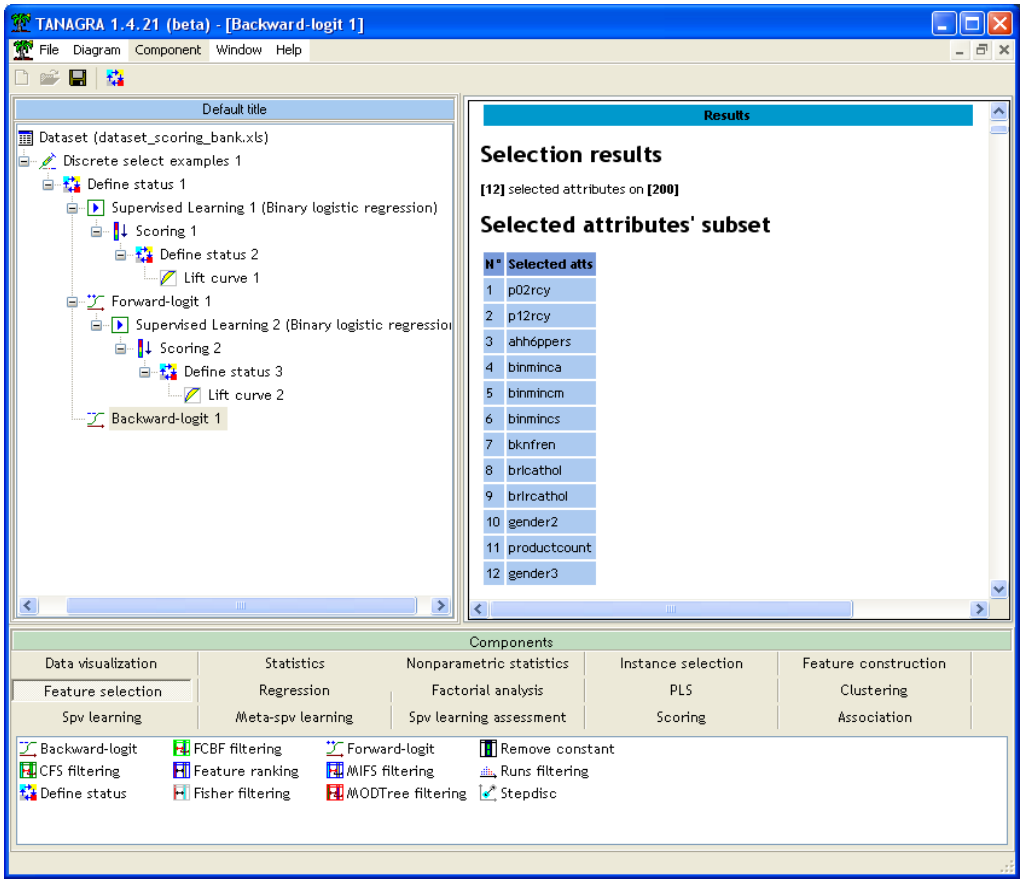
Sélection BACKWARD

TANAGRA intègre un second composant de sélection de variables basée sur la régression logistique (BACKWARD LOGIT – onglet FEATURE SELECTION). Il procède par éliminations successives à partir du test de Wald. Certains auteurs pensent que cette stratégie est plus performante car elle permet de tenir compte des relations entre les variables[3]. Certes, certes. On notera cependant que les calculs sont autrement plus longs. On notera surtout qu'il s'agit là avant tout de procédures purement mécaniques de sélection. Elles nous proposent des scénarios. Il nous revient d'inspecter consciencieusement les différentes solutions proposées pour les valider.

Voici, à titre indicatif (sauvegardez votre travail avant de lancer le traitement), les résultats fournis par la sélection BACKWARD. La durée des calculs est plus élevée (872 secondes # 14 minutes). En effet, puisque 12 variables sont sélectionnées à la sortie, le composant a procédé à 189 (200 - 12 + 1) régressions logistiques, et autant d'optimisation de la log-vraisemblance à l'aide de l'algorithme de NEWTON-RAPHSON[4]. Chaque régression durant approximativement 5 secondes, le temps de calcul est vite déduit.

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

[4] A l'ancien LEVENBERG-MARQUARDT (version 1.4.21) a été substitué un NEWTON-RAPHSON dans la version 1.4.27 de TANAGRA. Robuste depuis qu'un terme de régularisation a été introduit dans la matrice hessienne.



Parmi les 12 variables finalement sélectionnées, 6 sont communes au deux méthodes BACKWARD et FORWARD. En construisant la courbe LIFT, on constate que le modèle proposé n'est pas meilleur (46% de positifs parmi les 30% premiers ciblés). La méthode BACKWARD se justifie avant tout sur les petites bases où l'on essaie d'analyser finement les relations et interactions entre les variables.

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

Conclusion

Dans ce didacticiel, nous avons présenté la construction de la courbe lift dans le cadre du Scoring marketing. Nous en avons profité pour présenter deux nouveaux composants (version 1.4.21 de TANAGRA) de sélection de variables pour la régression logistique.