Objectif

Le ciblage marketing est une des utilisations privilégiées du data mining. Il s'agit d'isoler parmi un ensemble d'individus, ceux qui sont les plus à même de répondre positivement à une offre, l'objectif est de proposer à bon escient un produit que l'on veut promouvoir.

Dans la littérature, on parle généralement de SCORING. Il est à noter que cette démarche peut être étendue à d'autres domaines telles que le dépistage en épidémiologie, etc.

Nous utiliserons deux nouveaux composants : SCORING et LIFT.

Fichier

Les données proviennent d'une compétition qui a été organisée en 2000 (CoIL Challenge 2000 -- http://www.liacs.nl/~putten/library/cc2000/report2.html) : il s'agissait de repérer parmi les clients d'une compagnie d'assurance, ceux qui vont prendre une police d'assurance pour leur caravane.

Les fichiers étaient organisés de la manière suivante :

1. Un fichier d'apprentissage comprenant 5822 observations, outre la variable cible (prendre ou pas une police d'assurance pour sa caravane), il y avait 85 autres descripteurs. Les 43 premiers décrivent l'environnement socio-économique du prospect en utilisant comme repère son code postal ; les variables suivantes décrivent le comportement du client par rapport à d'autres produits.
2. Un fichier d'évaluation non étiqueté comprenant 4000 observations, l'objectif est de prédire la propension à consommer -- l'appétence diraient les spécialistes -- le produit « police d'assurance pour caravane ».

L'évaluation mise en place par le comité d'organisation était assez simple : isoler les 800 individus les plus appétents dans le fichier de validation (taille de la cible : 800 = 20% x 4000), le critère d 'évaluation est le nombre de positifs que l'on aura réussi à inclure parmi ces 800 individus. On sait par ailleurs que le nombre total de positifs dans le fichier de validation est de 238 individus.

Dans ce didacticiel, nous avons réuni l'ensemble des individus dans un seul fichier au format XLS, nous avons ajouté un descripteur supplémentaire (STATUS) qui permet de discerner la partie apprentissage de la partie évaluation.

Nous avons, de plus, récupéré les vraies étiquettes des individus du fichier de validation, ***ce qui n'était pas possible lors de la compétition***. Dans notre cas ça nous permettra de réaliser simplement tout le processus d'évaluation sans avoir à manipuler plusieurs fichiers.

Ciblage marketing avec TANAGRA

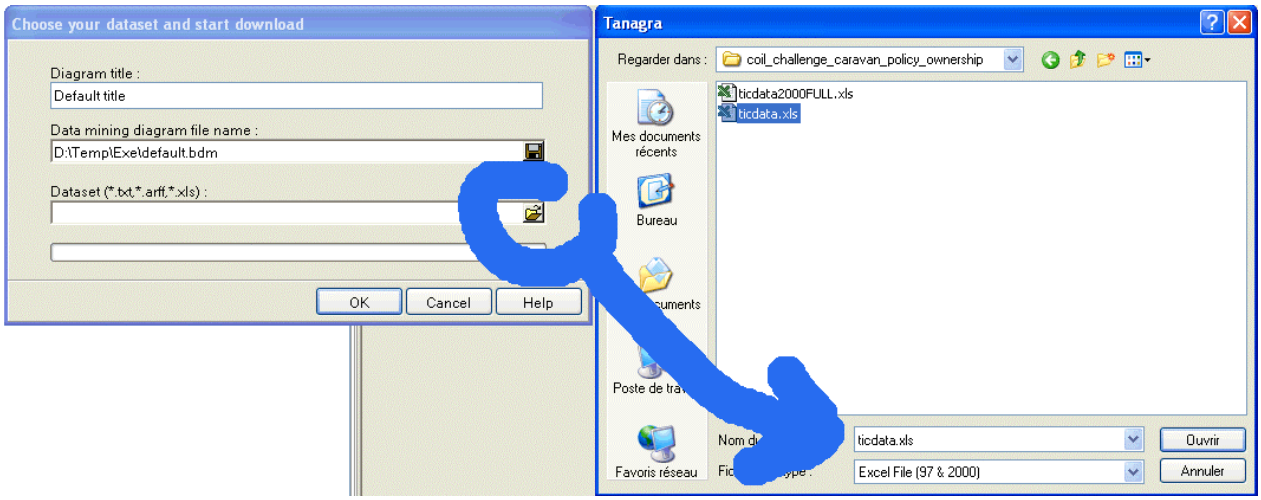
Préparer le fichier

Le fichier TICDATA.XLS contient 9822 observations : 5822 pour l'apprentissage et 4000 pour l'évaluation. La variable STATUS permet de les distinguer. Vous pouvez le visualiser dans n'importe quel tableur qui gère ce type de fichier.

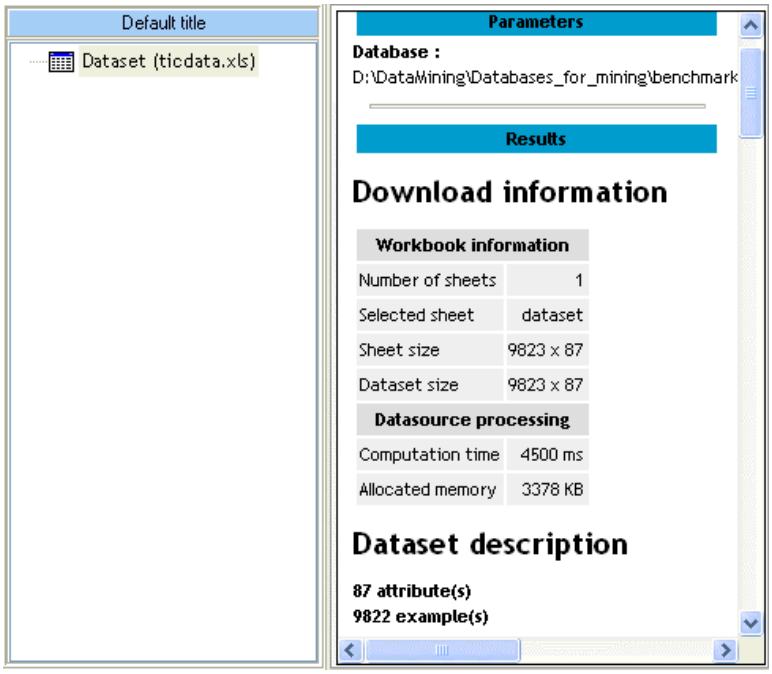
	A	CE	CF	CG	CH	CI	CJ
1	SD1	PO83	PO84	PO85	CLASS	STATUS	
5817	13	0	0	0	No	Learning	
5818	3	0	0	0	No	Learning	
5819	36	0	0	0	No	Learning	
5820	35	0	0	0	No	Learning	
5821	33	0	0	0	Yes	Learning	
5822	34	0	0	0	No	Learning	
5823	33	0	0	0	No	Learning	
5824	33	0	0	0	No	Test	
5825	6	0	0	0	Yes	Test	
5826	39	0	0	0	No	Test	
5827	9	0	0	0	No	Test	
5828	31	0	0	0	No	Test	

Charger le fichier dans TANAGRA

Pour importer le fichier dans TANAGRA, activez le menu « FILE / NEW » et sélectionnez le fichier ci-dessus ***après vous être assuré qu'il n'est plus ouvert dans votre tableur***.

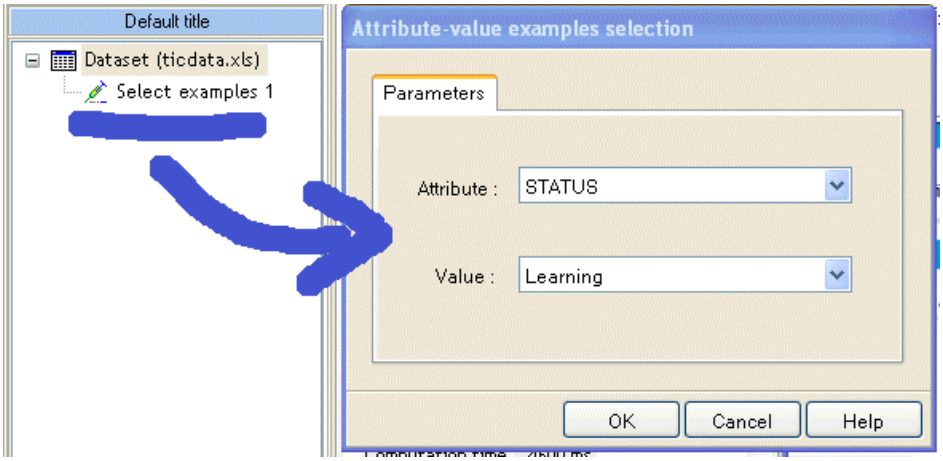


Vérifiez que les données chargées correspondent bien à l'affichage ci-dessous : 9822 observations et 87 attributs.



Subdiviser les données en « Apprentissage » - « Test »

Pour nous replacer dans les conditions de la compétition, il faut distinguer les données d'apprentissage, sur lesquelles nous construirons le modèle de prédiction, des données tests auxquelles nous attribuerons un score d'appétence. Utilisons pour ce faire le composant SELECT EXAMPLES (INSTANCE SELECTION) et paramétrons-le en mettant à contribution l'attribut STATUS.

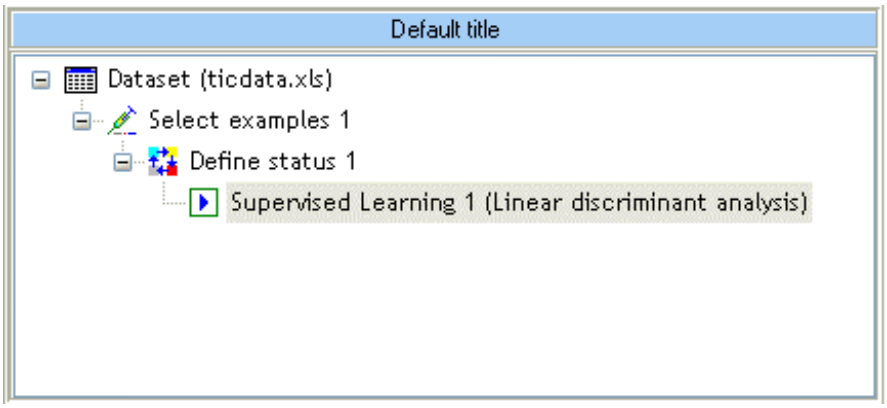


Analyse discriminante linéaire

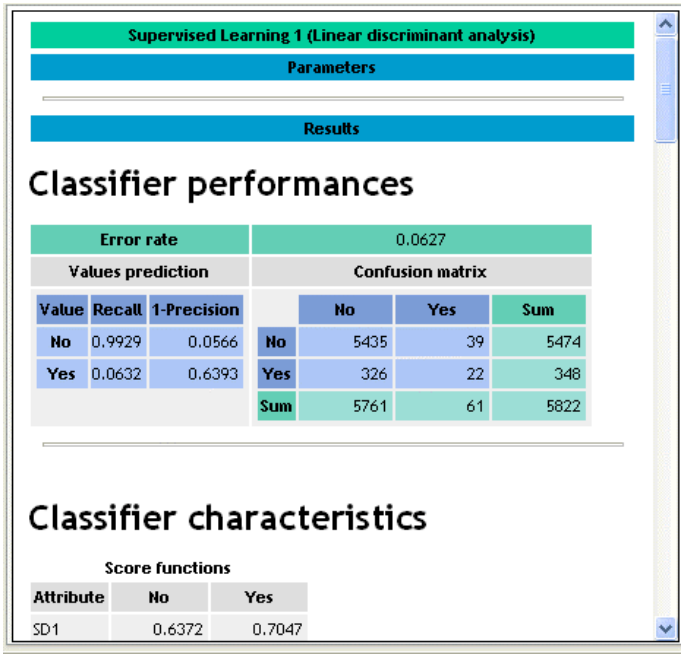
Reste alors à choisir les attributs puis lancer l'apprentissage.

Mettons en INPUT tous les attributs continus. Nous considérons qu'ils sont tous continus bien que dans certains cas, la première variable par exemple, cela n'est peut être pas tout à fait justifié. Nous plaçons en TARGET la variable CLASS. La variable STATUS ne doit pas être utilisée ici.

Placez alors l'apprentissage supervisé à l'aide de la méthode LDA (Analyse Discriminante Linéaire). Le diagramme de traitement est le suivant.



Les résultats semblent décevants, le taux d'erreur (6.27%) n'est pas très fameux comparé au taux d'erreur du classifieur par défaut (5.97% = 348/5822), ceci est notamment dû au fait que les classes sont très déséquilibrées.



Supervised Learning 1 (Linear discriminant analysis)

Parameters

Results

Classifier performances

Error rate	0.0627

Values prediction

Value	Recall	1-Precision
No	0.9929	0.0566
Yes	0.0632	0.6393

Confusion matrix

	No	Yes	Sum
No	5435	39	5474
Yes	326	22	348
Sum	5761	61	5822

Classifier characteristics

Score functions

Attribute	No	Yes
SD1	0.6372	0.7047

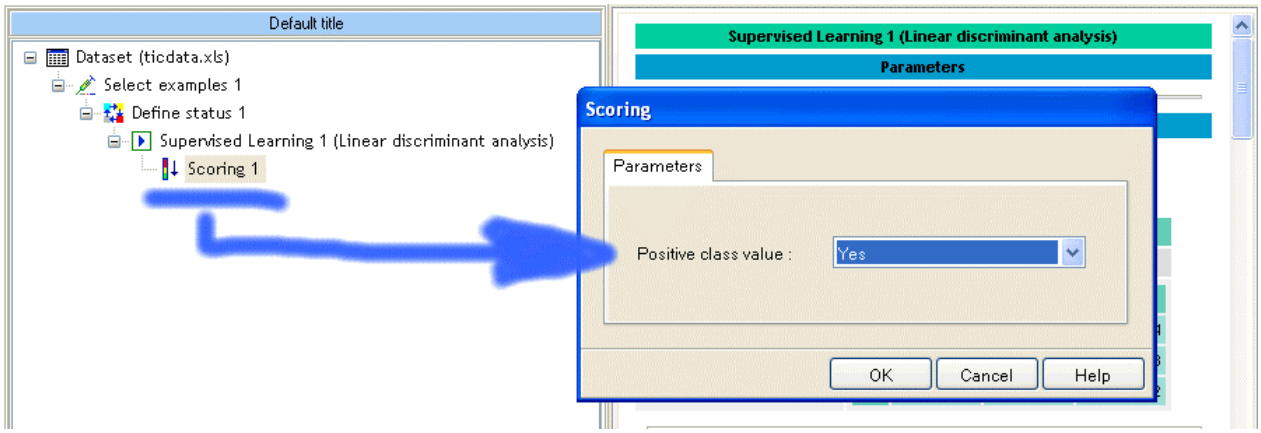
En réalité, le taux d'erreur n'est pas très pertinent pour juger de la qualité de notre apprentissage. Notre objectif n'est pas tant de classer globalement les individus mais plutôt

d'isoler, à coûts fixés – càd une taille de cible de 800 individus – les clients les plus à même de souscrire à l'offre.

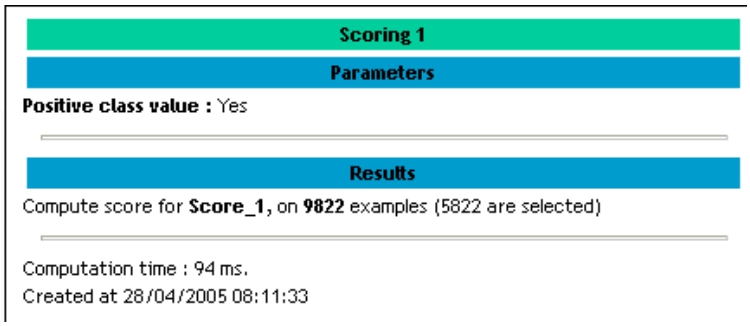
Attribuer un score aux individus

Pour ce faire, nous devons donc classer les individus selon leur appétence. C'est le rôle du composant SCORING qui calcule pour tous les individus de la base, qu'ils aient participé à l'apprentissage ou non, la probabilité de souscrire la police d'assurance. Il est à noter que si certaines méthodes produisent effectivement une probabilité, d'autres en revanche proposent un score qui n'est pas à proprement parler une probabilité mais qui induit le même classement des observations.

Placez le composant SCORING à la suite du diagramme et paramétrez-le en spécifiant que dans notre cas, les positifs correspondent aux individus qui présentent la valeur « YES » pour l'attribut TARGET.



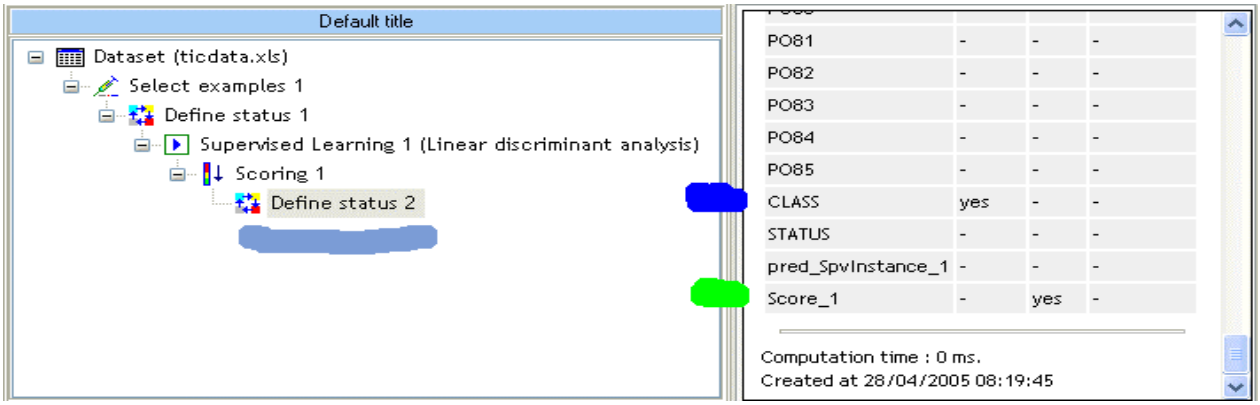
Les résultats montrent bien qu'un nouvel attribut « SCORE_1 » a été généré, le score a été attribué à tous les individus de la base même si le classifieur a été construit uniquement sur la partie apprentissage.



Construire la courbe LIFT

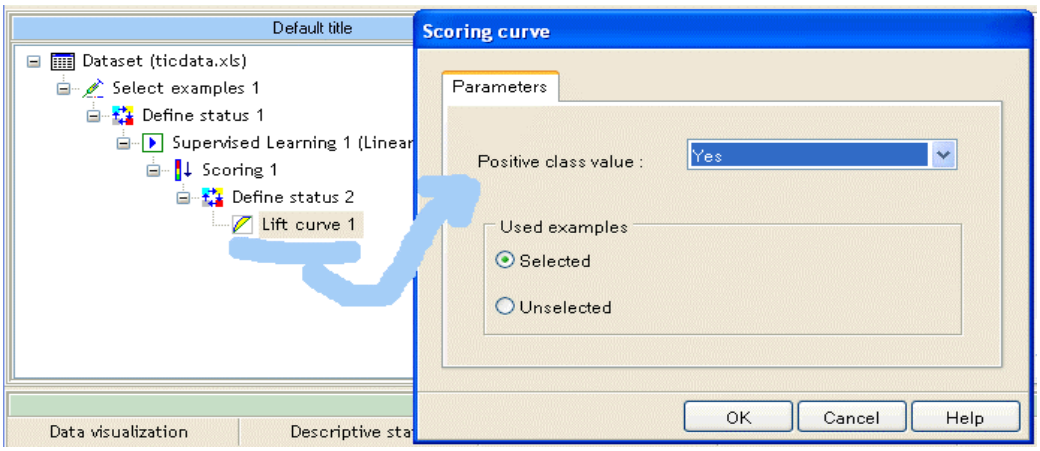
Pour évaluer la performance du ciblage, nous allons utiliser un outil différent de la matrice de confusion : la courbe lift. Elle décrit le pourcentage de positifs retrouvés (taux de vrais positifs) pour différentes tailles de cible.

Il faut dans un premier temps sélectionner les attributs que nous allons utiliser. En TARGET nous remettons la variable CLASS, en INPUT nous plaçons la variable SCORE_1. Cette étape supplémentaire peut paraître répétitive, en réalité elle améliore la souplesse du logiciel, il est possible en effet de placer plusieurs attributs en INPUT et comparer ainsi différents scores, notamment ceux qui auraient été fournis par un expert, en dehors de tout processus de modélisation.

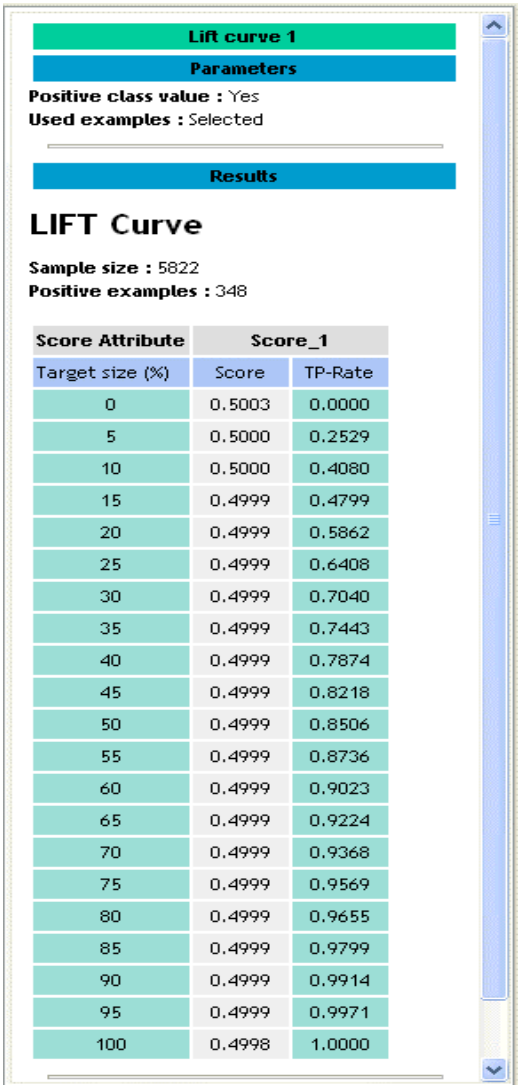


Dans un deuxième temps, il s'agit de placer le composant LIFT en spécifiant la modalité de l'attribut classe correspondant aux positifs.

Nous remarquerons au passage qu'il est donc possible dans TANAGRA de procéder à un ciblage pour les problèmes où la variable à prédire prend plus de deux valeurs. Il suffit de spécifier lors du SCORING et de la construction de la courbe LIFT laquelle de ces valeurs correspond à la modalité positive.



Plutôt que de fournir un graphique, TANAGRA fournit un tableau recensant le taux de vrais positifs pour chaque taille de cible. Il affiche également pour information la valeur du score qui a été utilisée pour chaque palier.



Lift curve 1

Parameters

Positive class value : Yes
Used examples : Selected

Results

LIFT Curve

Sample size : 5822
Positive examples : 348

Score Attribute	Score_1	
Target size (%)	Score	TP-Rate
0	0.5003	0.0000
5	0.5000	0.2529
10	0.5000	0.4080
15	0.4999	0.4799
20	0.4999	0.5862
25	0.4999	0.6408
30	0.4999	0.7040
35	0.4999	0.7443
40	0.4999	0.7874
45	0.4999	0.8218
50	0.4999	0.8506
55	0.4999	0.8736
60	0.4999	0.9023
65	0.4999	0.9224
70	0.4999	0.9368
75	0.4999	0.9569
80	0.4999	0.9655
85	0.4999	0.9799
90	0.4999	0.9914
95	0.4999	0.9971
100	0.4998	1.0000

Plusieurs résultats sont recensés :

- La modalité positive correspond à la valeur « YES » de la variable à prédire.
- Nous utilisons les données d'apprentissage pour construire la courbe LIFT.
- Il y a 5822 observations dans cet ensemble de données.
- 348 correspondent à la modalité positive.
- Pour chaque taille de cible, nous disposons de la proportion de positifs retrouvés.

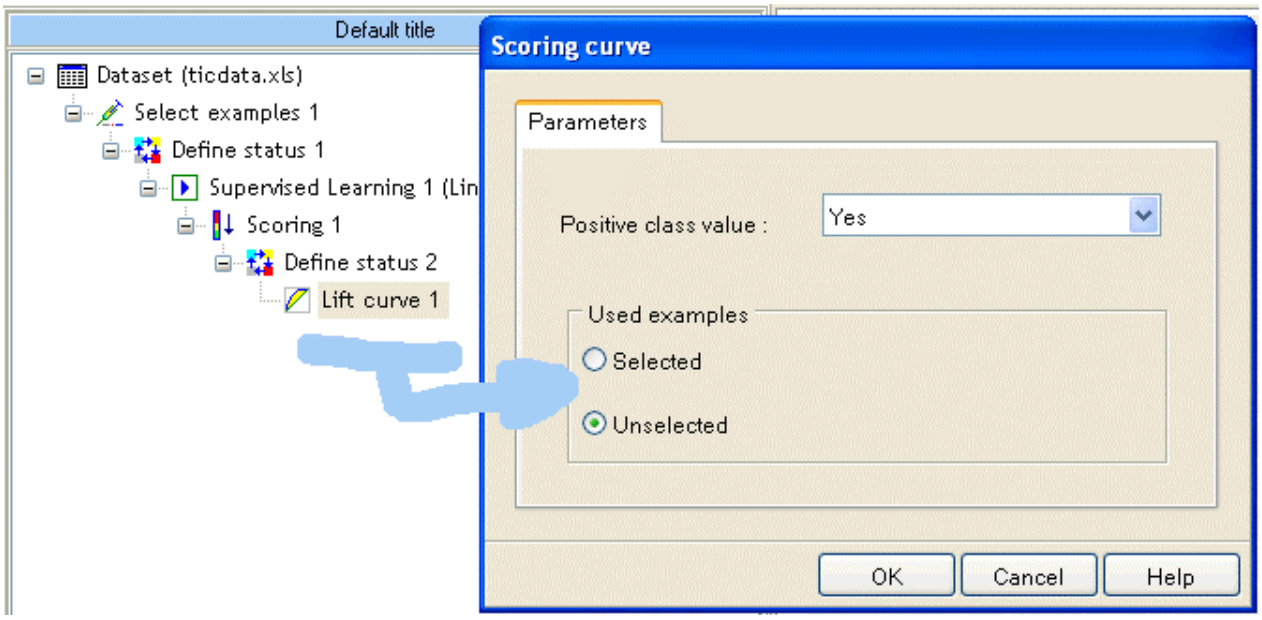
Prenons l'exemple d'une taille de cible égale à 20% de la taille de l'échantillon (20 % x 5822 # 1164 observations), nous pouvons espérer retrouver 58.62% des positifs càd 58.62 % x 348 # 204 positifs).

Transposons ce raisonnement sur la partie validation. La taille de la cible est égale à 800 observations (800 = 20% x 4000), il y a 238 positifs en tout, nous pouvons donc espérer retrouver 58.62 % x 238 # 139 positifs dans le fichier de validation.

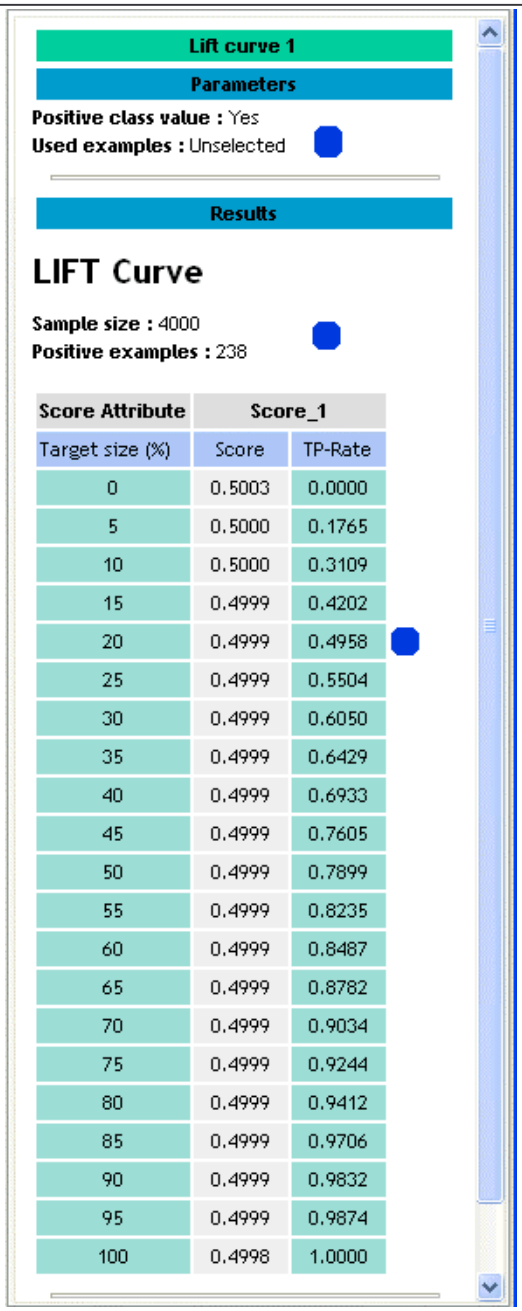
Pour encourageant qu'il soit, ce résultat est néanmoins biaisé car nous avons utilisé le même fichier pour la construction du modèle de classement (LDA) et son évaluation (construction de la courbe LIFT). Nous devons plutôt utiliser un fichier qui n'a pas participé à l'apprentissage pour obtenir une estimation « honnête » des performances du modèle de ciblage.

Courbe LIFT sur données de validation

Construisons la courbe LIFT sur des données qui n'ont pas participé à l'apprentissage. Ceci est possible en modifiant le paramètre « USED EXAMPLES » du composant LIFT.



La courbe est dès lors calculée sur les observations de la partie validation.



Lift curve 1

Parameters

Positive class value : Yes
Used examples : Unselected

Results

LIFT Curve

Sample size : 4000
Positive examples : 238

Score Attribute	Score_1	
Target size (%)	Score	TP-Rate
0	0.5003	0.0000
5	0.5000	0.1765
10	0.5000	0.3109
15	0.4999	0.4202
20	0.4999	0.4958
25	0.4999	0.5504
30	0.4999	0.6050
35	0.4999	0.6429
40	0.4999	0.6933
45	0.4999	0.7605
50	0.4999	0.7899
55	0.4999	0.8235
60	0.4999	0.8487
65	0.4999	0.8782
70	0.4999	0.9034
75	0.4999	0.9244
80	0.4999	0.9412
85	0.4999	0.9706
90	0.4999	0.9832
95	0.4999	0.9874
100	0.4998	1.0000

Il y a bien 4000 observations dans cette portion du fichier, dont 238 positifs. Pour une taille de cible à 20% (800 observations), nous retrouvons 49.58% des positifs, soit 49.58% x 238 # 118 personnes ayant souscrit effectivement au contrat d'assurance.

Ce résultat correspond bien à ce qui a été décrit dans les actes de la conférence, c'est le meilleur résultat que l'on puisse obtenir à l'aide d'un modèle linéaire sans pré-traitement préalable des données.

Pour la petite histoire, le gagnant du concours a utilisé un modèle bayesien naïf (modèle d'indépendance conditionnelle) après avoir éliminé la quasi-totalité des variables socio-économiques et essayé des combinaisons des autres variables. Il a réussi à intégrer 121

positifs dans sa cible de 800 individus. Ce résultat est d'autant plus remarquable que, précisons-le encore une fois, lors du concours, les compétiteurs ne disposaient pas de l'étiquette sur la partie validation.

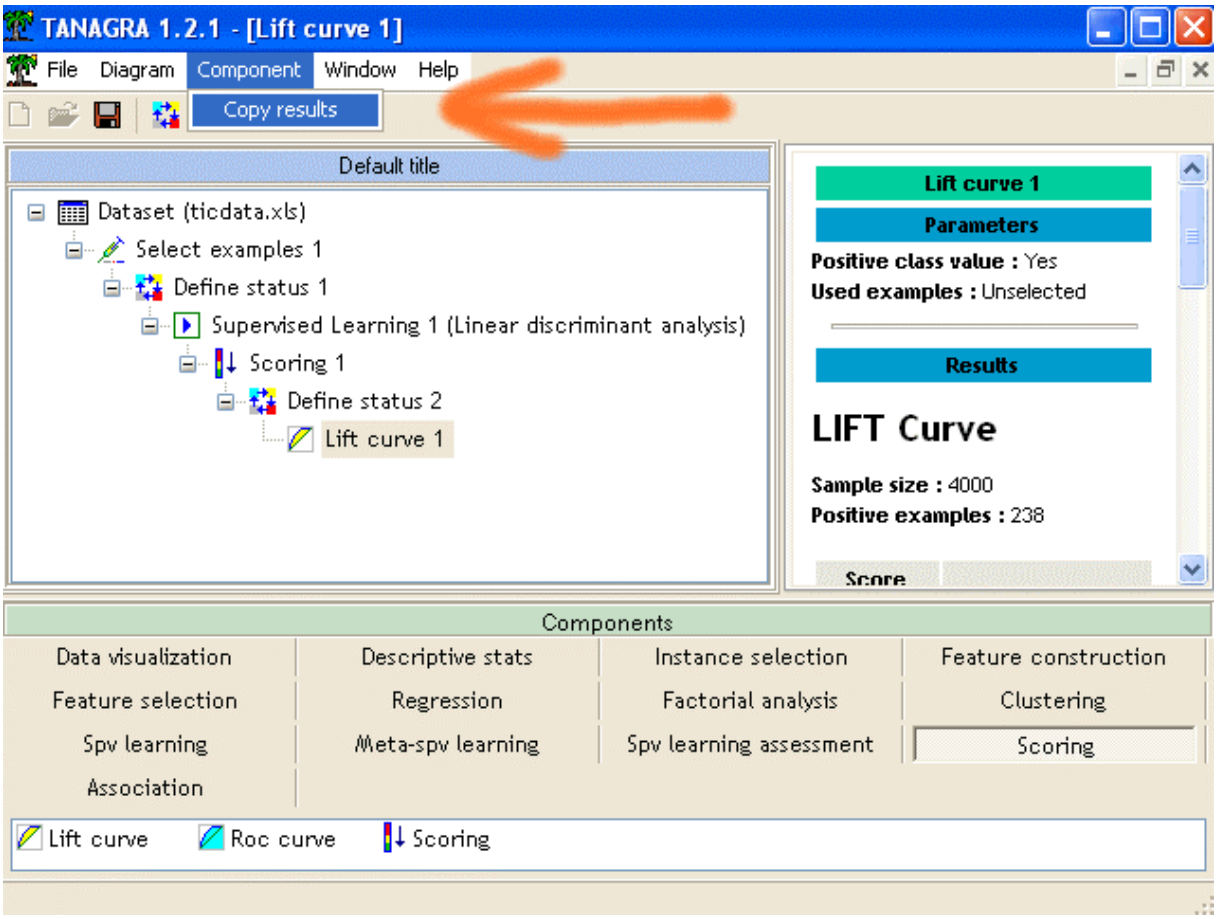
On aura noté d'ailleurs que tous les classifieurs compliqués, prétendument surpuissants, auront été surclassés par des modèles linéaires très simples (on peut montrer que le Bayesien Naïf induit une séparation linéaire sous certaines conditions).

Notons également que la courbe LIFT construite sur la partie apprentissage conduisait à une sur-estimation manifeste de la qualité du modèle, ce qui confirme encore une fois s'il en était encore besoin, que l'évaluation du modèle en resubstitution n'est pas du tout appropriée dans un processus supervisé.

Dessiner la courbe LIFT

TANAGRA présente les résultats sous forme de tableau. Il est possible de construire simplement la courbe sous forme graphique en exportant le résultat vers un tableur.

Il faut pour cela, copier les résultats via le menu « COMPONENT / COPY RESULTS ».



Le graphique peut être dès lors élaboré simplement dans le tableur de votre choix.

LIFT Curve

Sample size : 4000
Positive examples : 238

Score Attribute	Score_1	
Target size (%)	Score	TP-Rate
0	0.5003	0
5	0.5	0.1765
10	0.5	0.3109
15	0.4999	0.4202
20	0.4999	0.4958
25	0.4999	0.5504
30	0.4999	0.605
35	0.4999	0.6429
40	0.4999	0.6933
45	0.4999	0.7605
50	0.4999	0.7899
55	0.4999	0.8235
60	0.4999	0.8487
65	0.4999	0.8782
70	0.4999	0.9034
75	0.4999	0.9244
80	0.4999	0.9412
85	0.4999	0.9706
90	0.4999	0.9832
95	0.4999	0.9874
100	0.4998	1