J.S.C.
Jean Sousselier Conseil

Rescut

Modèle prédictif

1. Problème posé.

Pour un fichier d'individus donné, contenant :

- une variable à reconstituer (quantitative ou booléenne)
- des variables explicatives (indépendantes) nominales,

il s'agit de construire un modèle réalisant cette reconstitution avec une bonne approximation.

JSC a mis au point **une méthode originale** dans ce but, avec les caractéristiques suivantes :

- grande simplicité d'usage
- autorisation de pondération individuelle
- indifférence quant aux données manquantes : la méthode fonctionne avec les données effectivement présentes par individu

2. Méthode.

2.1 Forme canonique.

C'est le cas où la variable dépendante est **numérique**, et les variables indépendantes sont **booléennes**. On doit donc remplacer chaque variable indépendante nominale en autant de variables booléennes qu'elle a de modalités.

2.2 Principe.

Soit Y la variable dépendante, et (x_1 ,….,x_n) les variables indépendantes. (x_j =0 ou 1).
La méthode consiste à rechercher des « utilités » U_j , (coefficients analogues à des coefficients de régression) à associer à chaque variable :

Pour chaque individu i, la valeur prédite $\mathbf{Z_i}$ sera :

$$\mathbf{Z_i = \Sigma_j (U_j\, x_{ij}) / \Sigma\, x_{ij}} \qquad (1)$$

On choisit les utilités U_j de manière à minimiser les écarts entre les moyennes de Y et de Z pour toutes les populations telles que $x_j = 1$:
On cherche donc à minimiser les $\boldsymbol{\varepsilon_j}$ de chaque variable j :

$$\mathbf{\Sigma_i (Z_i\, x_{ij}) / \Sigma\, x_{ij} = \Sigma_i (Y_i\, x_{ij}) / \Sigma\, x_{ij} + \varepsilon_j}$$

La méthode RESCUT, créée dans ce but, conduit à avoir, dans la plupart des cas, des valeurs très faibles pour les $\boldsymbol{\varepsilon_j}$.

2.4 Justification.

- S'il n'a qu'une variable explicative, par exemple SEXE (0=Hommes, 1=Femmes), il est clair que la meilleure prédiction est de prendre pour les hommes, leur moyenne observée, et pour les femmes, leur moyenne observée (on pourrait aussi prendre dans certains cas la médiane, mais cela n'est pas généralisable).

- S'il y a plusieurs variables explicatives, on pourrait être tenté de suivre le même procédé, en appliquant à chaque individu la moyenne de la cellule élémentaire à laquelle il appartient, définie par la configuration de son vecteur (x_1 ,….,x_n) ; mais cela est en fait inapplicable, le nombre de telles cellules étant très grand (2^n), il y a énormément de cellules vides, empêchant toute prédiction, et d'autres avec très peu de représentants, avec donc des résultats peu significatifs.

- Il faut donc se contenter de considérer les moyennes dans les marges, c'est-à-dire indépendamment pour chaque variable.

2.5 Remarques :

1) Les utilités $\mathbf{U_j}$ ainsi calculées ont vocation à représenter l'influence propre de la variable j, alors que les moyennes observées $\mathbf{Y_j}$ sont entachées de l'influence des autres modalités, compte-tenu des inter-corrélations entre les variables.
2) L'équation (1) obtenue ne comprend ni terme constant, ni coefficients négatifs.
3) Contrairement à une équation de régression, $\mathbf{Z_i}$ est calculée comme la moyenne des U_j et non comme leur somme, cela est nécessaire car **le nombre de variables telles que $x_j = 1$ n'est pas constant d'un individu à l'autre** (il peut y avoir des non réponses dans les variables originelles par exemple).
4) Contrairement à une équation de régression, la méthode ne cherche pas à minimiser les **écarts individuels** entre valeurs observées et calculées, mais les écarts **entre les moyennes** observées et calculées dans chaque population telle que x_j =1. On ne cherche donc pas tant une reconstitution individuelle qu'une **reconstitution statistique de la variable dépendante dans ses effets** au niveau des variables indépendantes. Cela ne doit pas être choquant, dans la mesure où l'on considère la valeur de la variable dépendante non pas comme une valeur certaine, mais comme une probabilité.
5) **Si les coefficients d'utilité calculés sont nuls** pour certaines modalités des variables explicatives, (alors que ces modalités sont non vides), le modèle ne permet pas alors d'ajuster correctement les individus correspondant à ces modalités, ce qui entraîne que la moyenne calculée pour ces modalités sera plus forte que la moyenne observée, ce sera également le cas pour la moyenne générale.
6) Si un individu n'a aucune variable avec une valeur 1, il n'y a pas de valeur résultat pour la variable à expliquer, ce qui est normal.
7) S'il n'y a qu'une variable explicative, le résultat obtenu est manifestement le seul résultat possible.
8) Si on soupçonne une influence non négligeable des **interactions entre 2 variables** indépendantes, rien n'empêche de les remplacer par une troisième variable, construite comme le croisement des 2 premières.

2.6 Extension au cas où la variable dépendante est nominale ou multiple.

On fait tourner le modèle sur toutes les modalités de ces variables, on obtient ainsi des probabilités pour chacune d'elles. Des tirages aléatoires effectués en fonction de ces probabilités permettent de trouver les valeurs de variables nominales ou multiples.

3 Procédure à suivre.

Pour chaque étude à mener, le client fournit un fichier ASCII de données individuelles avec :

- éventuellement le poids de l'individu
- la variable à reconstituer (numérique, éventuellement 0/1)
- les variables nominales candidates pour être explicatives. Si possible, ces variables doivent avoir des valeurs de 1 à n, si ce n'est pas le cas, JSC procédera à une recodification.

JSC appliquera alors son modèle, après recodification éventuelle. Les résultats fournis seront :

- le fichier initial enrichi de la **valeur reconstituée**
- les **coefficients** (utilités) permettant de calculer le score pour tout nouvel individu, au moyen d'une formule simple, apparentée à une formule de régression.

Remarques :

- Si la variable à reconstituer est 0/1, le résultat fourni est sous forme d'une probabilité comprise entre 0 et 1; pour le transformer en une variable 0/1, on effectue un tirage aléatoire fonction de cette probabilité.

- En utilisant les coefficients fournis, il est possible de développer un **logiciel** demandant les valeurs des variables explicatives pour un individu et donnant comme résultat immédiat la valeur du score reconstitué.

05/04/02

Société XXX
Test méthode JSC de pronostic des Churners

Le modèle a d'abord été ajusté en utilisant le fichier « training ».

Puis, il a été appliqué au fichier « validation », conduisant à une note pour chaque individu, tendant à prévoir la variable « churners ».

Les résultats sont donnés dans les tableaux ci-joints, croisant la variable « churn » observée avec cette note mise en tranches (premier tableau) et en tranches cumulées (deuxième tableau).

Ils montrent que si on retient une note :

- supérieure à 0.8 (soit 2.7% de la population), on capte 2.4 % des non churners et **17.7%** des churners

- supérieure à 0.7 (soit 6.7% de la population), on capte 6.2 % des non churners et **32.2%** des churners

- supérieure à 0.6 (soit 18.4% de la population),, on capte 17.7% des non churners et **56.1%** des churners.

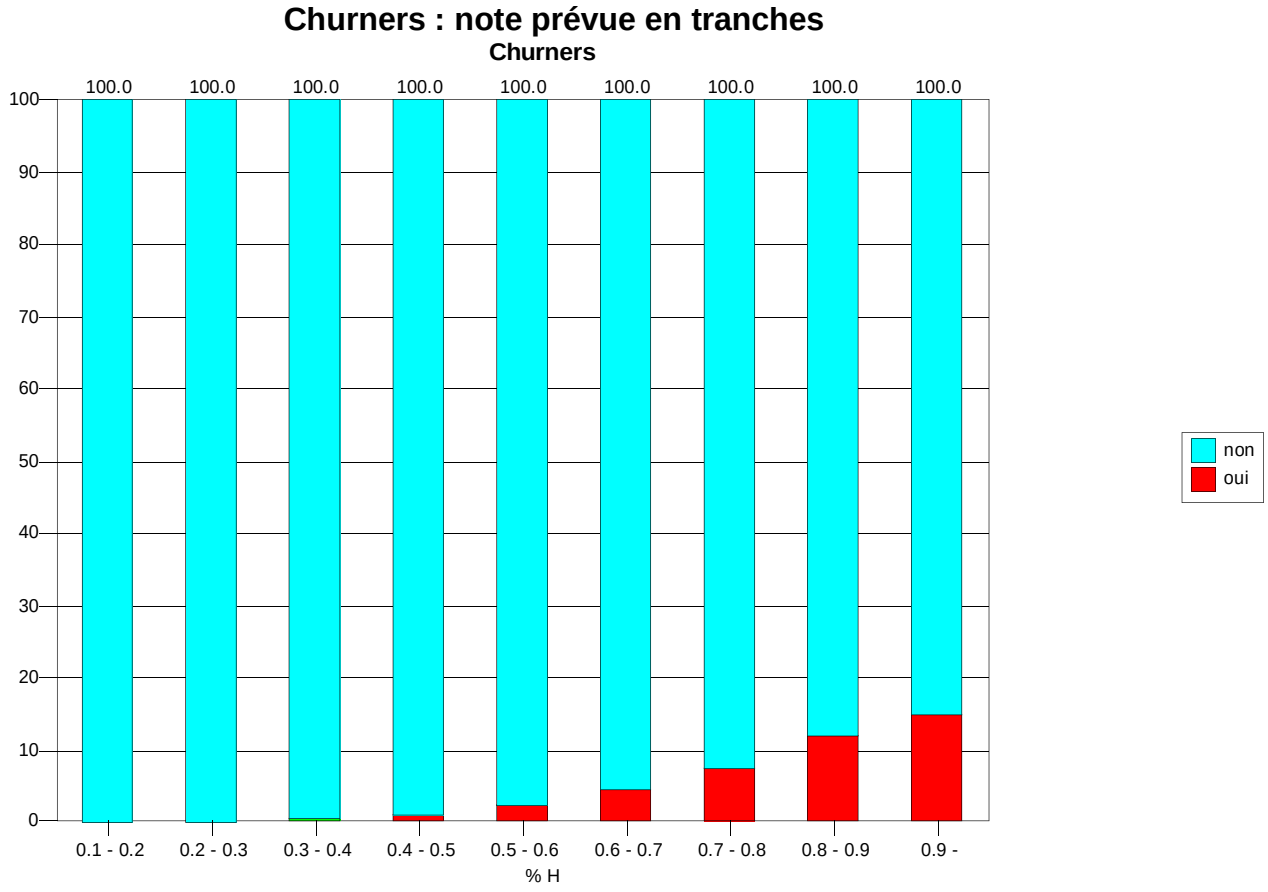
XXX - Test méthode JSC sur fichier Validation

Effectif % V
% H

	Churners					
Churners : note prévue en tranches	Total		Non		Oui	
TOTAL	24048 100.0	100.0	23567 98.0	100.0	481 2.0	100.0
0.1 - 0.2	1512 100.0	6.3	1511 99.9	6.4	1 0.1	0.2
0.2 - 0.3	5832 100.0	24.3	5820 99.8	24.7	12 0.2	2.5
0.3 - 0.4	3088 100.0	12.8	3066 99.3	13.0	22 0.7	4.6
0.4 - 0.5	4009 100.0	16.7	3963 98.9	16.8	46 1.1	9.6
0.5 - 0.6	5584 100.0	23.2	544 9	23.1	140 2.5	29.1
0.6 - 0.7	2568 100.0	10.7	51 95.4	10.4	117 4.6	24.3
0.7 - 0.8	870 100.0	3	804 92.4	3.4	66 7.6	13.7
0.8 - 0.9	393 100.0		345 87.8	1.5	48 12.2	10.0
0.9 -	197 100	0.8	163 84.9	0.7	29 15.1	6.0

0.1% des personnes ayant une note comprise entre 0.1 et 0.2 sont churners

15.1% des personnes ayant une note égale ou supérieure à 0.9 sont churners


Churners : note prévue en tranches
Churners
100.0
100.0
100.0
100.0
100.0
100.0
100.0
100.0
100.0
100
90
80
70
60
50
40
30
20
10
0
0.1 - 0.2
0.2 - 0.3
0.3 - 0.4
0.4 - 0.5
0.5 - 0.6
0.6 - 0.7
0.7 - 0.8
0.8 - 0.9
0.9 -
% H
non
oui

XXX - Test méthode JSC sur fichier Validation

Effectif % V % H

	Churners					
Churners : note prévue en tranches cumulées	**Total**		**Non**		**Oui**	
TOTAL	**24048** **100.0**	**100.0**	**23567** **98.0**	**100.0**	**481** **2.0**	**100.0**
0.9 -	**214** **100.0**	**0.9**	**182** **85.0**	**0.8**	**32** **15.0**	**6.7**
0.8 -	**645** **100.0**	**2.7**	**560** **86.8**	**2.4**	**85** **13.2**	**17.7**
0.7 -	**1605** **100.0**	**6.7**	**1450** **90.3**	**6.2**	**155** **9.7**	**32.2**
0.6 -	**4433** **100.0**	**18.4**	**4163** **93.9**	**17.7**	**270** **6.1**	**56.1**
0.5 -	**10238** **100.0**	**42.6**	**9828** **96.0**	**41.7**	**410** **4.0**	**85.2**
0.4 -	**13794** **100.0**	**57.4**	**13347** **96.8**	**56.6**	**447** **3.2**	**92.9**
0.3 -	**17202** **100.0**	**71.5**	**16732** **97.3**	**71.0**	**470** **2.7**	**97.7**
0.2 -	**23010** **100.0**	**95.7**	**22530** **97.9**	**95.6**	**480** **2.1**	**99.8**
0.1 -	**24048** **100.0**	**100.0**	**23567** **98.0**	**100.0**	**481** **2.0**	**100.0**
0 -	**24048** **100.0**	**100.0**	**23567** **98.0**	**100.0**	**481** **2.0**	**100.0**

17.7% de churners sont dans le groupe des personnes ayant une note égale ou supérieure à 0.8

Churners : note prévue en tranches cumulées