Grille de score

Construction à partir d'une régression logistique

Ricco RAKOTOMALALA

Université Lumière Lyon 2

PLAN

1. Position du problème – Grille de score ?
2. Construction à partir des coefficients de la régression logistique
3. Processus d'affectation via le score
4. Traitement des variables explicatives quantitatives
5. Grille de score à partir du couplage AFCM et ADL (DISQUAL)
6. Bibliographie

Position du problème

Construire une grille de notation des individus
indiquant leur « degré de positivité » (propension à être positif)

Contexte du « scoring »

Contexte : apprentissage supervisé,

- une variable cible Y binaire {+, - }
- des descripteurs tous qualitatifs (codés 0/1, codage disjonctif complet)

Exemple : apprécier les chances d'acceptation d'une demande de financement (un crédit) d'un achat effectuée par un client [oui = +, non = -]

Motif_AppMenager	Motif_Mobilier	Motif_HiFi	Assurance_oui	Assurance_non	Acceptation
0	0	1	1	0	oui
0	1	0	0	1	non
0	1	0	0	1	non
0	1	0	1	0	oui
0	0	1	1	0	non
0	1	0	1	0	non
0	0	1	0	1	non
0	0	1	1	0	oui
0	0	1	1	0	oui
0	0	1	1	0	oui

Motif : {App. Ménager, Mobilier, Hi Fi} Assurance : {oui, non}

Variable cible : Y = Acceptation {+, -}

On dispose de n = 944 observations

Grille de score = grille de notation

Permettant d'apprécier les chances du client de se voir octroyer un crédit

La grille de notation doit être calibrée, par ex. de 0 à 100.

	Note
Motif_AppMenager	20
Motif_Mobilier	0
Motif_HiFi	7
Assurance_oui	80
Assurance_non	0

Ex. 1 : client effectuant une demande pour « motif = mobilier » et ne prenant pas d'assurance « assurance = non » se voit attribuer la note 0 + 0 = 0 → il a un minimum de chances de voir acceptée sa demande de crédit (pire cas).

Ex. 2 : client effectuant une demande pour « motif = appareil ménager » et prenant une assurance « assurance = oui » se voit attribuer la note 20 + 80 = 100 → il maximise ses chances de voir acceptée sa demande crédit (meilleur cas).

Point de départ : résultats de la régression logistique [Equation LOGIT]

« Acceptation = oui » est la modalité « positive »

Attribute	Coef.
constant	1.12037
Motif_Mobilier	-0.50059
Motif_HiFi	-0.32038
Assurance_non	-1.98367

« Motif = App.Ménager » est la modalité de référence

« Assurance = oui » est la modalité de référence

(1) EXP(Coef) = Odds-ratio → surcroit de chances d'être positif

Par rapport à ceux qui prennent une assurance, ceux qui n'en prennent pas ont 7.26 [1/exp(-1.98)] fois plus de chances d'essuyer un refus que de voir leur demande acceptée.

(2) Calcul de la probabilité d'être positif (1ère version du « score »)

X : (Motif = mobilier, Assurance = non)

LOGIT = 1.12037 + (-0.50059) + (-1.98367) = -1.36389

P(Acceptation = oui / X) = 1/(1+ exp(-LOGIT)) = 0.204

NB. Le « pire cas » n'équivaut pas à une probabilité nulle !

On dispose déjà d'un système d'évaluation et de notation, mais il est peu intuitif, totalement abscons pour un non spécialiste…

Les deux représentations sont équivalentes…

Coefficients de la régression logistique

Attribute	Coef.
constant	1.12037
Motif_Mobilier	-0.50059
Motif_HiFi	-0.32038
Assurance_non	-1.98367

Grille de score

(Notation calibrée 0 à 100, ou 0 à 1000, etc.)

	Note
Motif_AppMenager	20
Motif_Mobilier	0
Motif_HiFi	7
Assurance_oui	80
Assurance_non	0

Mais…

Accessible aux non-spécialistes

Directement exploitable en déploiement

La grille est invariante par rapport au choix de la modalité de référence

Construction d'une grille de score

A partir des coefficients de la régression logistique

Réécriture du LOGIT : caler la note minimale à 0

Attribute	Coef.
constant	1.12037
Motif_AppMenager	0.00000
Motif_Mobilier	-0.50059
Motif_HiFi	-0.32038
Assurance_oui	0.00000
Assurance_non	-1.98367

Etape 1 : faire apparaître les modalités de référence

Etape 2 : Détection des valeurs min des coefficients par variable (la constante est mise de côté)

Attribute	Coef.
Motif_AppMenager	0.00000
Motif_Mobilier	-0.50059
Motif_HiFi	-0.32038
Assurance_oui	0.00000
Assurance_non	-1.98367

$Min_{motif} = -0.50059$

$Min_{assurance} = -1.98367$

Coef + $|Min_{variable}|$

Etape 3 : Correction par variable pour rendre positifs tous les coefficients

Attribute	Coef.
Motif_AppMenager	0.50059
Motif_Mobilier	0.00000
Motif_HiFi	0.18021
Assurance_oui	1.98367
Assurance_non	0.00000

Ainsi...

Les points attribués seront toujours positifs

Le minimum des points est égal à 0

Mise à l'échelle : caler la note maximale à 100 (ou 1000, ou 10000, etc.)

Attribute	Coef.
Motif_AppMenager	0.50059
Motif_Mobilier	0.00000
Motif_HiFi	0.18021
Assurance_oui	1.98367
Assurance_non	0.00000

Etape 4 : identifier le maximum des points

(attention si des coefs. sont > 0, $max_{variable}$ est différent de $|min_{variable}|$)

$Max_{motif} = 0.50059$

$Max_{assurance} = 1.98367$

$MAX_{points} = 0.50059 + 1.98367 = 2.48426$

Etape 5 : Calculer le facteur de correction η

$$\eta = \frac{100}{MAX_{points}} = \frac{100}{2.48426} = 40.25342$$

Etape 6 : Multiplier les points modalités par le facteur de correction η

$0.50059 \times 40.25342 \approx 20$

	Score
Motif_AppMenager	20
Motif_Mobilier	0
Motif_HiFi	7
Assurance_oui	80
Assurance_non	0

Les notes par modalité sont arrondies pour faciliter la lecture

Le score est calibré, il est compris entre 0 et 100

Affectation à partir du score

Reproduire le processus d'affectation basé sur le LOGIT

Calculer la valeur seuil du SCORE

Règle d'affectation basée sur le LOGIT

Pour un individu ω à classer, on s'appuie sur le LOGIT

$$a_1 x_1(\omega) + a_2 x_2(\omega) + \cdots \begin{cases} > -a_0 \Rightarrow \hat{Y}(\omega) = + \\ \leq -a_0 \Rightarrow \hat{Y}(\omega) = - \end{cases}$$

Comment déterminer la valeur **seuil** si on s'appuie sur le score ?

$$SCORE(\omega) \begin{cases} > seuil \Rightarrow \hat{Y}(\omega) = + \\ \leq seuil \Rightarrow \hat{Y}(\omega) = - \end{cases}$$

Il faut transformer la constante a_0 du LOGIT en respectant le schéma de constitution du score.

Calcul du seuil d'affectation

1. Chaque variable a été corrigée de $|Min_{variable}|$
2. Somme des corrections : $S = \Sigma |Min_{variable}|$
3. Seuil d'affectation avant calibrage : $C = S - a_0$
4. Seuil d'affectation après mise à l'échelle du score : $SEUIL = \eta \times C$

Attribute	Coef.
constant	1.12037
Motif_AppMenager	0.00000
Motif_Mobilier	-0.50059
Motif_HiFi	-0.32038
Assurance_oui	0.00000
Assurance_non	-1.98367

(Min_{motif} = -0.50059 ; $Min_{assurance}$ = -1.98367)

- S = 0.50059 + 1.98367 = 2.48426
- C = 2.48426 – 1.12037 = 1.36389
- SEUIL = 40.25342 x 1.36389 = **54.9**

On reproduit à l'identique le comportement de la régression logistique avec la règle de décision :

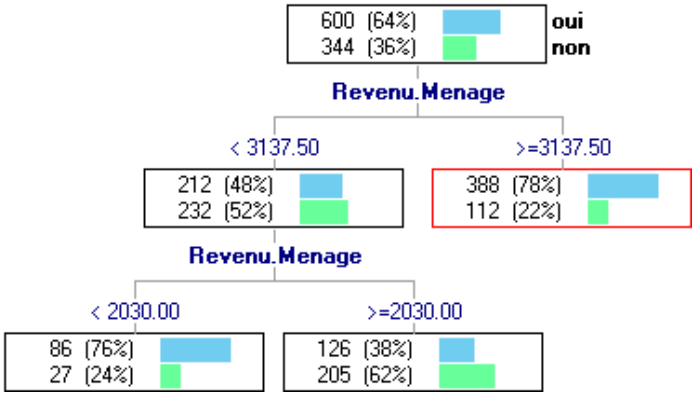
$$SCORE(\omega)\begin{cases} > 54.9 \Rightarrow \hat{Y}(\omega) = + \\ \leq 54.9 \Rightarrow \hat{Y}(\omega) = - \end{cases}$$

Traitement des variables explicatives quantitatives

Discrétisation (découpage en classes) des variables quantitatives

Transformation des variables quantitatives en indicatrices (1)

Un arbre de décision permet de répondre à ces spécifications



<u>Etape 1 :</u> découpage en classes

Comment choisir le nombre de classes ?

Comment choisir les bornes de découpage ?

Découpage en fonction de la variable cible Y !

3 intervalles avec les bornes (2030 et 3137.5)

NB. La méthode MDLPC (Fayyad & Irani, 1993) disponible dans de nombreux logiciels (Tanagra, Weka, R [package ''discretization''], etc.) est un arbre de décision avec une règle d'arrêt spécifique à la discrétisation.

Transformation des variables quantitatives en indicatrices (2)

[2030 ; 3137.5[[3137.5 ; +∞[

Revenu.Menage	REV.B	REV.C
2264	1	0
2181	1	0
4265	0	1
4431	0	1
3008	1	0
3042	1	0
4237	0	1
8454	0	1
3797	0	1
5193	0	1

Etape 2 : codage disjonctif complet à partir des intervalles. Attention (1), codage non imbriqué parce qu'on ne sait pas si l'effet est monotone ; (2) le premier intervalle sert de modalité de référence.

Régression logistique

Attribute	Coef.
constant	1.59696
REV.A : [0 ; 2030[	0.00000
REV.B : [2030 ; 3137.5[	-1.72488
REV.C : [3137.5 ; +00[	0.02628
Motif_AppMenager	0.00000
Motif_Mobilier	-0.27986
Motif_HiFi	-0.10055
Assurance_oui	0.00000
Assurance_non	-2.07249

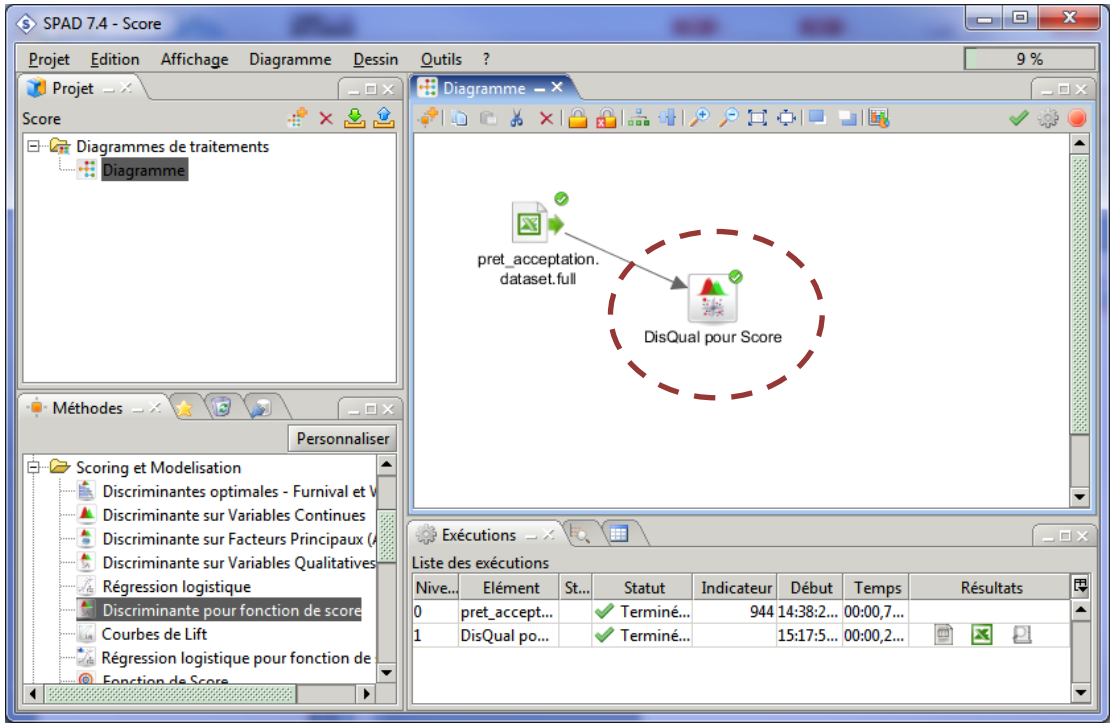
Grille de score

Attribute	Score
REV.A : [0 ; 2030[	42
REV.B : [2030 ; 3137.5[	0
REV.C : [3137.5 ; +00[	43
Motif_AppMenager	7
Motif_Mobilier	0
Motif_HiFi	4
Assurance_oui	51
Assurance_non	0

Grille de score via l'analyse discriminante

Couplage AFCM (analyse factorielle des correspondances multiples) et ADL (analyse discriminante linéaire)

La méthode DISQUAL (Saporta, 1975)



Calcul en 3 étapes :

(1) AFCM sur les descripteurs (catégoriels ou discrétisés)

(2) ADL sur une sélection (*) des facteurs de l'AFCM

(3) Reconstitution de la fonction de classement sur les indicatrices originelles

(*) Il est possible de les prendre tous

(*) En ne sélectionnant que « q » premiers les plus pertinents, on peut obtenir des résultats plus stables (on a une meilleure régularisation, c'est une forme de « nettoyage » des données)

DISQUAL sur acceptation de crédit

1 AFCM : Coefficients des fonctions permettant d'obtenir les coordonnées factorielles des individus

2 ADL : Fonction SCORE (obtenue par différenciation des fonctions de classement [oui – non]) définie sur les facteurs

3 Fonction SCORE définie sur les indicatrices des variables originelles

Coefficients

Appliqués aux indicatrices des variables actives

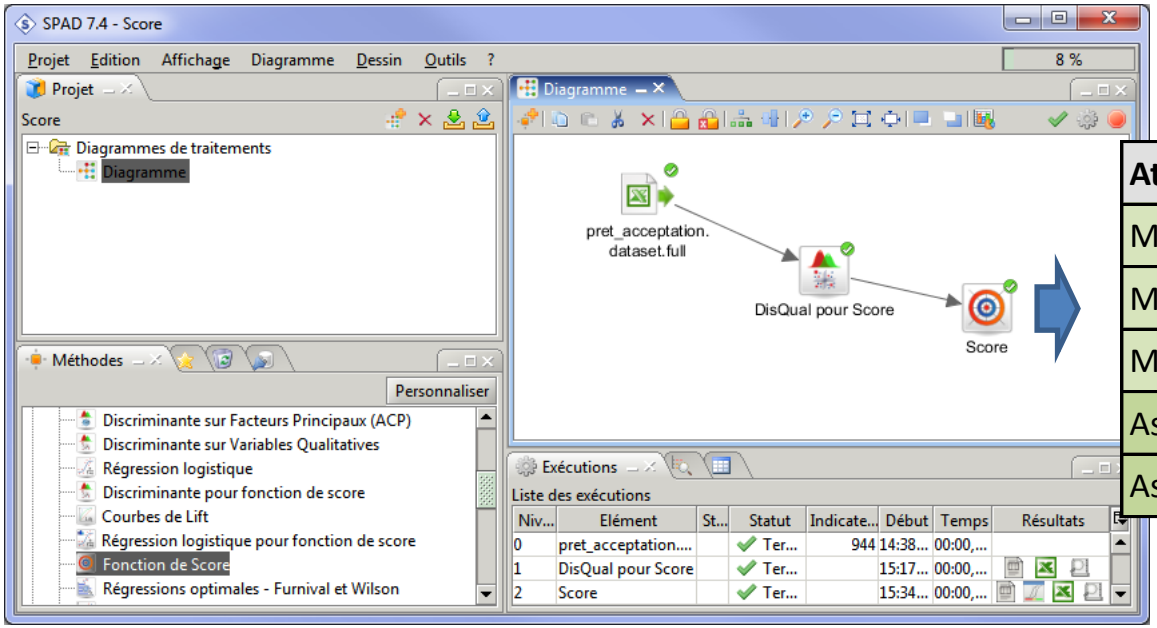
Attribute = Value	Axis_1	Axis_2	Axis_3
Motif = AppMenager	0.9750	-1.3746	-0.9750
Motif = Mobilier	-0.4900	-0.3314	0.4900
Motif = HiFi	0.2617	0.7349	-0.2617
Assurance = oui	-0.1633	0.0000	-0.1633
Assurance = non	1.5308	0.0000	1.5308

Attribute	Coef.
MCA_1_Axis_1	-0.4750
MCA_1_Axis_2	-0.0402
MCA_1_Axis_3	-0.7749
constant	0.6071

Attribute = Value	Coef.
Constant	0.6071
Motif = AppMenager	0.3478
Motif = Mobilier	-0.1336
Motif = HiFi	0.0489
Assurance = oui	0.2041
Assurance = non	-1.9133

Ex. $a_{Motif=AppMenager}$ = -0.4750 x (0.9750) + (-0.0402) x (-1.3746) + (-0.7749) x (-0.9750) = 0.3478

Grille de score via DISQUAL



Grille de score

Attribute = Value	Note (/100)
Motif = AppMenager	19
Motif = Mobilier	0
Motif = HiFi	7
Assurance = oui	81
Assurance = non	0

Au final, la grille de score est très proche de celle de la régression logistique… ce n'est pas étonnant… ce sont là deux classifieurs linéaires.

Bibliographie

G. Saporta, « Probabilités, Analyse de données et Statistique », Technip, 2006 ; pp. 462 à 467, section 18.4.3 « Un exemple de ‘’credit scoring’’ ».

J.P. Nakache, J. Confais, « Statistique explicative appliquée », Technip, 2003 ; pp. 58 à 60, section 2.2.2 « SCORE : construction d’un score ».