Cours de Scoring - Séance 1

Ibrahim TOURE,
Ingénieur Statisticien

Université d'Evry Val d'Essone

24 Novembre 2016

Sommaire

Introduction (1/3)

- Le scoring est un domaine de la statistique décisionnelle dont le but est de discriminer, de sélectionner, de classer, de segmenter, de prévoir le comportement d'un client conformément à un critère donné
- Les techniques de scoring peuvent s'appliquer à d'innombrables domaines de l'activité économique dont les principaux sont : le risque bancaire et le marketing
- Ces techniques s'inscrivent dans une problématique de recherche d'efficacité (rapidité), de rentabilité pour l'entreprise

Introduction (2/3)

- Dans le domaine bancaire : via le déploiement du dispositif Bâle 2, les notions de risque et de perte économique sont de rigueur
- L'objectif pour une banque est de maitriser son risque en adaptant les fonds propres qu'elle provisionne en fonction du risque observé sur ses propres clients
- De plus, les autorités de contrôle (ACPR) et de supervision (BCE) exige une documentation rigoureuse sur les modèles de score lors des inspections

Introduction (3/3)

- Dans le domaine du marketing : souhait pour l'entreprise d'optimiser ses actions commerciales en envoyant ses offres à des clients sélectionnés
- Intérêt de cette étude préalable : plus l'entreprise connaitra ses clients, plus elle sera susceptible de lui proposer des produits personnalisés
- Les analyses de « ciblage »font désormais partie intégrante des missions des équipes marketing
- D'autres études de scoring orientées « Marketing » : les **scores d'appétence** et les **scores d'attrition**

Objectifs du cours

- Présentation des différentes étapes à mettre en oeuvre pour construire un modèle de score
- Analyse et suivi des performances d'un modèle de score
- Sans viser l'exhaustivité, ce cours se limitera à une présentation des principales techniques, en introduisant leurs fondements théoriques, mais en établissant toujours le lien avec les applications concrètres qui en résultent

Sommaire

Les fondements

- Le but est de modéliser un évènement correspondant à une variable binaire (dichotomique) ou une variable ordinale (polytomique)
- Exemples : *un impayé sur un crédit, le remboursement anticipé d'un client, la réponse d'un client à une opération commerciale ou marketing, la montée au contentieux d'un client au recouvrement (phase pendant laquelle le client doit de l'argent à la banque*
- Mais le cadre standard de la modélisation en scoring se réfère essentiellement à l'analyse d'une variable dichotomique (*bon client vs mauvais client*
- Ce cours cours est focalisé sur l'analyse dichotomique essentiellement mais une introduction aux modèles polytomiques sera présentée

Exemple pratique : mise en place d'une opération de crédit

- Objectif : la banque veut prévoir en fonction des caractéristiques du client et de l'opération en question, le risque du client à terme et ainsi octroyer ou non le crédit
- A cette étape, la banque cherche à se prémunir d'un quelconque impayé de son nouveau client
- Ainsi, nous noterons par convention dans nos modèles Y=1 si aucun incident de paiement ne se produits lors de la vie du prêt et Y=0 sinon

$$Y = \begin{cases} 1 & \text{si le client est un bon payeur} \\ 0 & \text{sinon.} \end{cases}$$

- Objectif mathématique : prévoir la probabilité de survenance d'un impayé compte tenu de caractéristiques intrinsèques au client, notées **X**
- Autrement dit, on cherche à définir une relation d'ordre en terme de risque séparant de façon **optimale** kes bons des mauvais clients

Sommaire

Définition d'un score

- Le principe d'un **score linéaire** est de rechercher une décomposition des effets des facteurs, i.e. attribuer à chacune des variables explicatives du modèle une pondération en fonction des valeurs prises par cette variable
- Autrement dit, un score est une application de $\mathbb{R}^p$ vers $\mathbb{R}$
- La note globale d'un individu (noté S(X)) sera la somme des pondérations attribuées à chacune des covariables du modèle :

$$S(X) = S(X_1, ..., X_p) = \sum_{j=1}^{p} S_j(X_j) \tag{1}$$

- Une fois le core calculé, il convient de définir à partir de quel niveau ou seuil la banque acceptera ses clients.

Une règle parmi d'autres (1/2)

- D'un point de vue système d'aide à la décision, le score vient en appui d'autres règles métiers appelées *systèmes experts*
- On pourra retrouver des règles de **Fichage externe ou interne** (qui indique à la banque si le client a été fiché pour des raisons de surendettement (fichage banque de France), des tentatives de fraude), des **normes** (interdire des prêts aux mineurs, chômeurs, la règle du 33%) et enfin les **scores** qui indique si le client est en dessous ou au dessus de la barre de score *(s)*
- Ainsi le score apporte un niveau de contrôle théorique et statistique parmi un ensemble de règles plus opérationnelles

Une règle parmi d'autres (2/2)

On dira que le client est :

- **Accepté Score** si sa note de score est supérieure à la barre d'acceptation *(s)*
- **Refusé Score** si sa note de score est inférieure à la barre d'acceptation *(s)*

Un client sera accepté s'il passe toutes les règles du système d'aide à la décision, y compris les score. Mais, il se peut que la décision système soit négative mais que l'attaché commercial (chargés clientèles ou chargés d'affaire) accepte le dossier (exemple : une personne dont les parents se portent garants...) : on parle alors de **refus repris**.

Exemples de scores (1/2)

Exemple fictif de score sur lequel 3 covariables interviennent (p=3)

Variables du score	Modalités	Coefficients estimé
Age en années	Moins de 25	10
	25 à 40	20
	40 à 55	30
	Plus de 55	35
Situation familiale	Divorcé	15
	Célibataire	20
	Concubinage	25
	Marié	30
Profession	Sans emploi	5
	Autre	15
	Ouvrier	25
	Cadre	35

Table : Exemple de score

Exemple de scores (2/2)

- Lecture du tableau : Si le client a 30 ans, célibataire et qu'il est cadre, sa fonction de score s'écrira :

$$S(X) = S(X_{age}, X_{sitfam}, X_{profession}) = 20 + 20 + 35 = 75 \quad (2)$$

- Ainsi si la barre d'acceptation est fixée à 50, ce client sera **Accepté Score**
- En revanche, un client de 20 ans, célibataire et sans emploi serait **réfusé score**

Score d'octroi aux particuliers (1/2)

- Les crédits aux particuliers servent à financer l'achat d'un bien immobilier, d'une automobile, des travaux de logement, ...
- Deux types de variables interviennent dans ce type de modèles : les variables intrinsèques au client (pour évaluer sa solvabilité) et les variables du prêt proprement dit (i.e. spécifiques à l'opération)

Score d'octroi aux particuliers (2/2)

Exemple de variables de score

Variables du client	Variables Opérations
Age du demandeur	Montant du prêt
CSP du demandeur	Durée du prêt
Situation familiale et nombre d'enfants	Objet financé
Situation au logement locataire, accédant, propriétaire	Apport du client
Revenus bruts du demandeur	
Reste à Vivre du client	
Ancienneté dans le dernier emploi	

Table : Exemple de variables

La cotation Banque de France (1/2)

- On peut aussi construire des scores d'entreprises qui sont effectués autour d'informations de bilan (fonds propres ; endettement) et de ses ratios financiers
- C'est le cas du score BDFI de la Banque de France qui s'applique aux sociétés de l'industrie afin de détecter les défaillances d'entreprises
- Ce score se compose de 7 ratios : R1, R2, ..., R7
- La Banque de France a mis en place une cotation composée de 7 classes de risques. Les banques peuvent s'appuyer sur ces cotations pour les notations internes ou encore les scores d'octroi

La cotation Banque de France (2/2)

Ratios BDFI	Signification du ratio BDFI
R1 et R2	Rentabilité
R3	Importance des dettes fiscales et sociales
R4	Délai crédit fournisseur
R5	Importance de l'endettement
R6	Structure de l'endettement
R7	Coût de l'endettement financier

Table : Ratios score BDFI de la BDF

Sommaire

Notations (1/4)

- Les notations en majuscules (Y, X) représentent les grandeurs aléatoires alors que les minuscules (y, x) correspondent aux données observées
- Y représente le critère à modéliser, la survenue d'un défaut de paiement. Par convention, Y=0 pour qualifier un risque (impayé, remboursement anticipé, attrition) et Y=1 dans le cas contraire

Notations (2/4)

- On note p_1 et p_0 les probabilités marginales. Elles sont inconnues avec :

$$p_1 = \mathbb{P}[Y = 1] \tag{3}$$

$$p_2 = \mathbb{P}[Y = 0] = 1 - p_1 \tag{4}$$

- Y étant dichotomique, sa loi peut être modélisée par une bernoulli

Notations (3/4)

- Son espérance (la probabilité d'être sain) est donc donnée par $E(Y) = p_1$
- Sa variance (erreur de prédiction) est donnée par $V(Y) = E(Y^2) - E(Y)^2$
- Les covariables seront supposées continues et de densité $f(x)$. Les densités conditionnelles à $Y = 1$ et $Y = 0$ seront notées respectivement $f_1(x)$ et $f_0(x)$
- Important : les possibilités de discrimination seront d'autant plus fortes que ces densités seront différentes ou, autrement dit, que les populations des bons et des mauvais risques présentent des caractéristiques particulières

Notations (4/4)

- Enfin, nous notons $p_1(x)$ et $p_0(x)$ les probabilités conditionnelles de Y sachant X
- $P[Y = 1|X = x] = p_1(x)$ et $P[Y = 0|X = x] = p_0(x) = 1 - p_1(x)$
- On en déduit facilement les espérances conditionnelles

Choix du seuil (1/3)

- Soit A la région d'acceptation ou encore **seuil de score** ou encore le **calage de barre**
- Pour la détermination de A, la banque va chercher par exemple de minimiser les erreurs d'affectation afin de maximiser la rentabilité de son système d'octroi

Choix du seuil (2/3)

Notations

- On note g, le gain issu d'une bonne décision, à savoir accepter un crédit à un individu bon. celui se produit avec la probabilité $P[1_A(X); Y = 1]$
- On note c_0 le coût résultant de l'octroi d'un crédit à un individu « mauvais », cela se produit avec la probabilité $P[1_A(X); Y = 0]$
- On note c_1 le coût (manque à gagner) consistant à refuser un individu « bon », cela se produit avec la probabilité $P[1_{A^c}(X); Y = 1]$

Choix du seuil (3/3)

- La zone d'acceptation optimale, compte tenu des objectifs de rentabilité R fixés est : $A^* = argmax R(A)$
- La rentabilité s'écrivant :
$R = g * P[1_A(X)|Y = 1] * P[Y = 1] - c_0 * P[1_A(X)|Y = 0] * P[Y = 0] - c_1 * P[1_A(X)|Y = 1] * P[Y = 1]$

Sommaire

Choix du critère à modéliser

- Ce choix n'est pas anodin et doit répondre aux applications futures du score
- Le critère est très souvent sous forme binaire (achat/non achat, réponse / non réponse, impayé / sans impayé)
- Le critère peut être une variable aléatoire positive, assimilable à une durée : durée avant impayé, durée avant rupture de relation commerciale, durée avant remboursement anticipé, ...

Choix des données et de l'échantillon (1/2)

L'échantillon de données doit être suffisant et il convient de s'assurer de plusieurs choses au rang desquelles

- L'échantillon de construction doit être représentatif de la population sur laquelle le score est appliqué
- Cette structure de population doit être stable dans le temps
- Le score doit être stable dans le temps au niveau du risque observé par modalité de chaque variable
- Le score doit faire l'objet d'un suivi qualité fréquent et doit être refondu ou au moins réestimé dans le cas où il perd en discrimination

Choix des données et de l'échantillon (2/2)

Dans la pratique, on procède de la façon suivante :

- Une scission aléatoire de l'échantillon en deux parties. La plus grosse (75% environ) est utilisée pour estimer le score et l'autre partie (25%) pour tester ses performances (problème de surapprentissage en statistique)
- On peut aussi s'assurer de la stabilité du score (en structure et en risque) sur une population plus récente afin de contrôle sa qualité de discrimination

Traitement des variables

Les variables sont rarement des variables brutes issues des bases de données (défaut de renseignement, valeurs manquantes, valeurs non significatives) et nécessitent de traitements préalables :

- Le découpagne des variables pour prendre en compte des effets par morceaux
- Le recodage des variables
- La sélection des variables

Estimation du modèle

Une fois le critère à modéliser est choisi et les variables explicatives retenues, le modèle peut être écrit ou estimé. Voici les modèles couramment utilisés

- Les modèles LOGIT et PROBIT (les plus classiques). On les résout par maximisation de la vraisemblance
- L'analyse discriminante, appropriée dans le cas d'une approche explicative
- Les réseaux de neuronnes
- Les modèles de durée (dans le cas de la modélisation d'une variables positive)

NB : ces méthodes d'estimation ne sont pas forcément équivalentes et ne s'appliquent pas toujours dans les mêmes circonstances