LA NOTATION STATISTIQUE DES EMPRUNTEURS OU « SCORING »

Gilbert Saporta
Professeur de Statistique Appliquée
Conservatoire National des Arts et Métiers

Dans leur quasi totalité, les banques et organismes financiers utilisent l'analyse statistique pour prédire si un emprunteur sera un bon ou un mauvais payeur et prendre ensuite la décision appropriée : acceptation sans condition, prise de garantie, refus.

La modélisation et la décision se fondent sur l'observation du passé : on connaît pour un certain nombre de prêts attribués la qualité payeur qui est donc une variable qualitative Y à deux modalités (« bon » ou « mauvais ») ainsi que les données recueillies lors du dépôt du dossier de prêt : ce sont les variables X (X_1, …, X_p) . Typiquement pour des particuliers on trouvera l'age, la profession , le statut matrimonial, le fait d'être ou non propriétaire, donc majoritairement des variables qualitatives, alors que pour des entreprises on aura plutôt des variables numériques comme des ratios issus de la comptabilité.

Formellement il s'agit de trouver une fonction f(X_1, …, X_p) permettant de prédire Y.

Dans ce qui suit nous décrirons les diverses étapes et les problèmes qui se posent depuis la collecte des données jusqu'à la mise en œuvre en donnant à chaque fois des indications sur les méthodologies à utiliser.

I. La collecte de l'information

Le premier travail consiste à constituer un fichier qui contient des informations complètes sur des dossiers de prêts. Il se présentera sous la forme d'un tableau rectangulaire individus-variables où les n individus sont partagés en deux groupes d'effectifs n_1 et n_2 : les bons et les mauvais.

Ce travail essentiel est maintenant facilité par le stockage informatique, mais cela n'a pas toujours été le cas : les variables du dossier de demande n'étaient pas forcément saisies car elles n'étaient pas toutes jugées utiles pour la gestion du prêt. Il fallait alors retrouver les dossiers papiers.

Les n individus constituent en fait un échantillon de l'ensemble des N données disponibles : nous verrons plus loin qu'il est indispensable de garder de côté un certain nombre de dossiers afin de valider les résultats obtenus. Il faut donc prélever aléatoirement n individus parmi les N : comme il faut s'assurer d'avoir un nombre suffisant et non aléatoire (ce qui introduirait une source de variabilité supplémentaire, donc une moindre précision) d'observations dans chacun des deux groupes, on procède à un sondage stratifié avec tirage séparé des n_1 et n_2 individus. Deux questions se posent alors : quel effectif global et quelle répartition de n_1 et n_2 . Une idée naturelle consisterait à prélever n_1 et n_2 en respectant les proportions de bons et mauvais dossiers, d'autant plus que l'on sait que le sondage stratifié à répartition proportionnelle est toujours meilleur que l'échantillonnage simple sans stratification. Cette méthode est cependant à déconseiller ici car les deux groupes ont des proportions très différentes : le groupe à risque (les mauvais payeurs) qu'il faut détecter est très minoritaire (mettons 10%) et serait mal représenté. On a pu démontrer qu'une répartition équilibrée $n_{1=} n_2$ est bien meilleure, sinon optimale sous des hypothèses assez générales. Les vraies proportions p_1 et p_2 servent ultérieurement pour les calculs de probabilités *a posteriori*.

Quant au nombre total n, il est typiquement de quelques milliers.

Un problème plus complexe est celui du biais de sélection : en fait les dossiers dont on connaît l'issue (bons ou mauvais) résultent d'un choix effectué en général par des analystes de crédit ; tous les dossiers de prêt n'étaient évidemment pas acceptés et ceux qui l'ont été ne constituent pas un échantillon représentatif de toutes les demandes. Même si la méthode antérieure de sélection n'était pas scientifique, il est clair que les dossiers acceptés n'ont pas les mêmes caractéristiques que les dossiers refusés. Or pour construire une règle de décision valable pour tous les nouveaux dossiers, il aurait fallu savoir ce que seraient devenus les dossiers refusés si on les avait acceptés…Il faut alors recourir à des techniques assez élaborées (estimation en deux phases, modèle Tobit). Sans entrer dans les détails, disons seulement que l'on modélise également le processus de sélection.

Le problème du biais de sélection n'intervient pas dans d'autres domaines où des techniques similaires de scoring sont utilisées comme l'assurance automobile (pour la détection des conducteurs à risque) ou la sélection d'adresses pour optimiser l'envoi de propositions commerciales (dans ce dernier cas on effectue un scoring à partir des résultats d'un premier courrier ; les « bons « étant les répondants, les « mauvais » les non-répondants) .

II Les analyses préliminaires

Le fichier brut une fois constitué doit d'abord être « nettoyé » pour éliminer erreurs et incohérences. Il comporte alors en général un trop grand nombre de variables. Une exploration des liaisons entre chaque variable X et le critère à prédire Y permet en général d'éliminer les variables non pertinentes. On utilise alors des outils classiques : test du khi-deux de liaison entre variables qualitatives, comparaison des % de bons et de mauvais par catégorie de chaque variable X.

Dans le même temps on procède à des recodages des variables : regroupement de valeurs en classes pour les variables continues (on s'aide d'histogrammes), regroupement de classes pour obtenir la meilleure séparation sur Y. On crée également de nouvelles variables par combinaison de 2 ou plusieurs variables. Par exemple si on s'aperçoit que l'ancienneté dans l'emploi joue différemment selon la profession , sur la probabilité de bon remboursement, on créera une variable croisant les modalités de ces deux variables (cf. exemple plus loin).

Il est couramment admis que toutes ces analyses représentent près de 80% du temps de ce genre d'études.

III La modélisation

Les techniques de « scoring » qui sont les plus utilisées dans le secteur bancaire utilisent des méthodes linéaires pour leur simplicité et leur grande robustesse. Il existe bien d'autres méthodes non-linéaires ou non-parametriques comme les arbres de décision, les réseaux neuronaux etc. dont l'usage se répand (cf. références) mais elles sortent de ce bref exposé.

Un score est une note de risque que l'on calcule comme combinaison linéaire des variables explicatives $S = \sum_{i=1}^{p} a_i X_i$. Les coefficients a_i étant optimisés pour la prédiction de Y.

Pour obtenir le vecteur **a** des coefficients des a_i , il existe diverses techniques d'estimation dont les deux principales sont la fonction linéaire discriminante de Fisher et le modèle logit (encore appelé régression logistique).

III.1 La fonction linéaire discriminante de Fisher.

C'est la plus ancienne (elle remonte à 1936) : c'est la combinaison optimale qui sépare le mieux les moyennes du score dans les deux groupes. Plus précisément si $\overline{s}_1$ et $\overline{s}_2$ sont les scores moyens sur les deux groupes de n_1 et n_2 individus, on maximise $\frac{(\overline{s}_1 - \overline{s}_2)^2}{V(s)}$ où *V(s)* est la moyenne pondérée des variances du score dans chacun des 2 groupes. On montre que **a** est proportionnel à $W^{-1}(g_1-g_2)$ où W est la moyenne pondérée des matrices de variance-covariance des variables explicatives dans chaque groupe et les g les vecteurs des moyennes des variables de chaque groupe. C'est une méthode de moindres carrés.

III.2 La régression logistique ou modèle logit .

On exprime la probabilité *a posteriori* d'appartenance à un des groupes selon :

$$P(G_1 / X) = \frac{\exp(S)}{1+\exp(S)} = \frac{\exp(\sum_{i=1}^{p} a_i X_i)}{1+\exp(\sum_{i=1}^{p} a_i X_i)}$$

et on estime alors les a_i par la méthode du maximum de vraisemblance. X désigne ici le vecteur dont les composantes sont les X_i pour i=1 à p.

Nous avons employé le terme de probabilité *a posteriori* qui renvoie à l'usage de la formule de Bayes. En effet si on connaît les probabilités *a priori* d'appartenance aux deux groupes p_1 et p_2=1- p_1, qui sont en fait les proportions réelles des deux groupes, la probabilité d'appartenir au groupe 1 connaissant les informations fournies par le dossier, c'est à dire les X, est donnée par :

$P(G_1 / X = x) = \frac{p_1 f_1(x)}{p_1 f_1(x) + p_2 f_2(x)}$ où f_k est la densité de probabilité des X dans le groupe k.

Pour de nombreux modèles probabilistes (gaussiens, multinomial etc.) cette probabilité *a posteriori* se met sous la forme logistique précédente : $P(G_1 / X) = \frac{\exp(S)}{1+\exp(S)}$

En particulier si le vecteur aléatoire des X suit une loi normale de même matrice de variance-covariance dans les deux groupes, la règle qui consiste à classer une observation x dans le groupe qui a la plus forte probabilité *a posteriori* est équivalente à la règle qui consiste à classer une observation dans un groupe selon que son score est inférieur ou supérieur à un certain seuil.

Les deux méthodes.(Fisher et logit) ne conduisent pas aux mêmes estimations des coefficients, mais celles-ci sont en général assez proches. Le choix entre les deux ne doit pas être une question d'école : moindres carrés contre maximum de vraisemblance, mais plutôt se faire sur leur capacité prédictive, c'est à dire sur de nouvelles observations.

La règle « naïve » de Bayes qui consiste à prédire le groupe le plus probable, donc ici à choisir le groupe qui a une probabilité *a posteriori* supérieure à 0.5, n'est en général pas adaptée à la prédiction d'un groupe rare. On cherche plutôt à détecter un maximum d'individus à risque, et on choisira le seuil de décision en conséquence (voir plus loin).

III.3 Cas de prédicteurs qualitatifs.

Le cas où les variables explicatives X_i sont qualitatives nécessite un traitement particulier. En effet comment faire une combinaison linéaire de variables qualitatives ? Cela n'a évidemment pas de sens. La solution retenue est basée sur ce que l'on appelle la forme disjonctive d'une variable qualitative X à m modalités (comme une profession). On définit les m variables indicatrices des modalités ($\mathbf{1}_1$, $\mathbf{1}_2$, ….,$\mathbf{1}_m$) telles que $\mathbf{1}_j$ vaut 1 si on appartient à la modalité j, 0 sinon. Seule une des indicatrices vaut 1, celle qui correspond à la modalité prise. Les *m* indicatrices sont donc équivalentes à la variable qualitative. Le score est alors une combinaison linéaire des indicatrices, ce qui revient à donner une note partielle à chaque modalité de chaque variable. Le score final étant la somme des notes partielles (à telle profession correspond telle note). Les variables explicatives qui interviennent dans les formules sont donc les indicatrices de toutes les variables.
Une difficulté intervient cependant : la matrice *W* n'est pas de plein rang et n'est donc pas inversible car la somme des indicatrices des modalités de chaque variable vaut 1. Cela signifie qu'il existe une infinité de solutions équivalentes pour estimer les coefficients :une des solutions couramment utilisée consiste alors à ne prendre que *m-1* indicatrices pour chaque variable qualitative puisque la dernière est redondante.

III.4 Un exemple

Les valeurs suivantes sont fictives (mais réalistes) et ne servent qu'à illustrer la méthode. Considérons le cas d'un établissement financier qui veut prédire la solvabilité d'entreprises pour savoir s'il doit ou non accorder un prêt. On connaît pour chaque entreprise les deux variables suivantes : X_1 part des frais financiers dans le résultat en %, et X_2 délai de crédit fournisseurs (nombre de jours avant de payer les fournisseurs) .

Sur l'échantillon des entreprises solvables la moyenne de X_1 vaut 40, celle de X_2 90. Sur l'échantillon des entreprises non solvables ces moyennes sont respectivement 90 et 100. On admet que les écart-types sont les mêmes d'un groupe à l'autre et sont respectivement s_1=40, s_2=20, et que X_1 et X_2 présentent la même corrélation r=0.8 dans chaque groupe. La covariance entre X_1 et X_2 vaut rs_1s_2=640.

La matrice de variance commune (dite également intra-classe) est alors $W = \begin{pmatrix} 1600 & 640 \\ 640 & 400 \end{pmatrix}$

et le vecteur de différence des moyennes $g_1 - g_2 = \begin{pmatrix} -50 \\ -10 \end{pmatrix}$

Il est facile d'en déduire la fonction de Fisher par la formule $a = W^{-1}(g_1-g_2)$. Les coefficients étant définis à une constante multiplicative près, on peut prendre pour a le vecteur de composantes –1 et 1.2 .

La fonction de score est alors $S = -X_1 + 1.2\ X_2$

On en déduit facilement par transformation linéaire que le score moyen des entreprises solvables vaut 68 tandis que le score moyen des entreprises non solvables vaut 30. Les écart-types des variables étant supposés identiques dans les deux groupes on trouve que $V(S) = V(X_1) + 1.2^2 V(X_2) - 2(1.2)\mathrm{cov}(X_1; X_2) = (25.3)^2$

On supposera pour la simplicité de l'exposé que la distribution du score suit dans chaque groupe une loi normale. Quand il n'en est pas ainsi, les densités de probabilité, les fonctions de répartition, etc. doivent être estimées d'une autre manière.

Un usage classique dans les études de ce type est de recaler le score S pour qu'il prenne la quasi totalité de ses valeurs dans l'intervalle [0 ; 1000]. Cela se fait simplement par transformation affine.

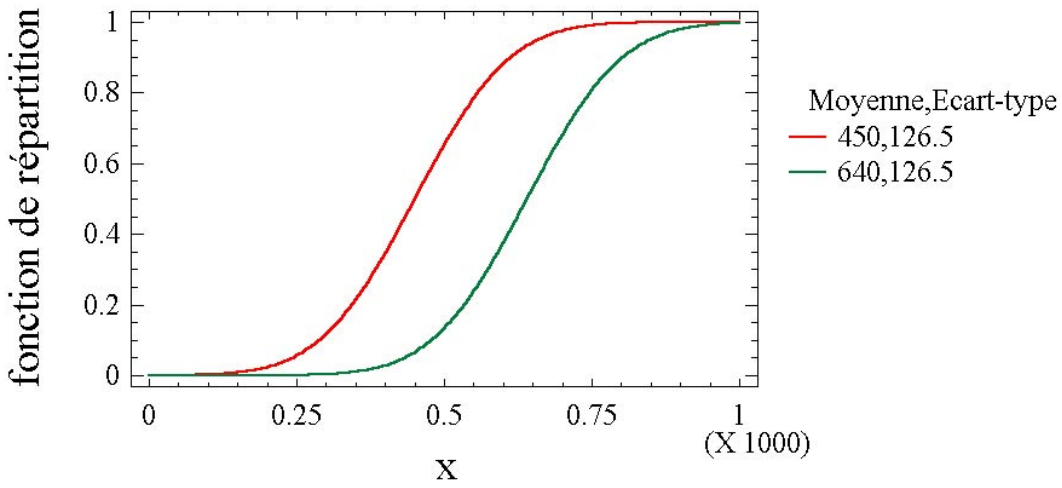
Ceci peut être réalisé approximativement dans notre exemple en multipliant le score par 5 et en ajoutant 300.

La fonction de score vaut donc $S= -5X_1 + 6 X_2 + 300$

V. Qualité et utilisation d'un score

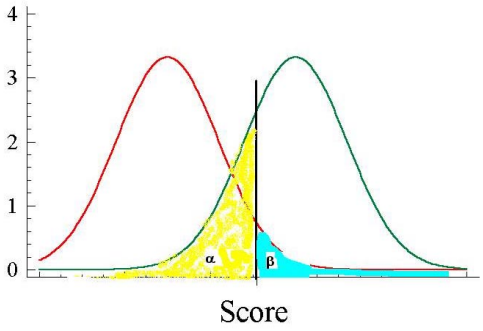
On estime tout d'abord les distributions conditionnelles du score dans chacun des deux groupes. Un score efficace doit conduire à des distributions bien séparées. Dans l'exemple précédent le score suit une loi normale N(640 ;126.5) pour le groupe des entreprises solvables ou une loi N(450 ;126.5) pour les entreprises non-solvables. On vérifiera que le score *S* donne une meilleure séparation que chaque variable prise séparément en calculant l'écart réduit entre moyennes, c'est à dire la différence en valeur absolue entre moyennes divisée par l'écart-type commun.

On considérera également les fonctions de répartition :



L'utilisation est la suivante : si on refusait de prêter de l'argent aux entreprises ayant une note de score inférieure à 556, on éliminerait 80% des entreprises insolvables (les « mauvaises ») mais on refuserait à tort 25% des entreprises solvables (les « bonnes »). Le choix du seuil dépend des risques financiers et est fixé par un raisonnement économique prenant en compte les coûts d'erreur de mauvaise classification : en effet accorder un prêt à une entreprise qui se révélera insolvable a un coût différent de celui de perdre un bon client.

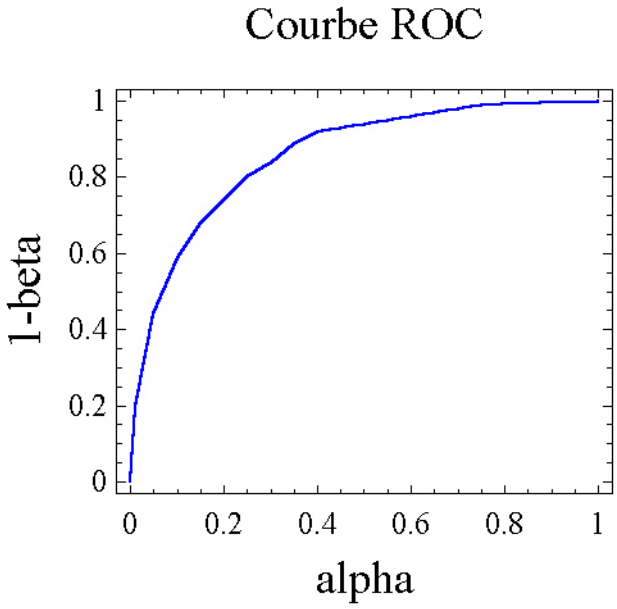
D'une manière similaire à la présentation classique d'un test statistique, la situation peut se décrire à l'aide des deux densités :



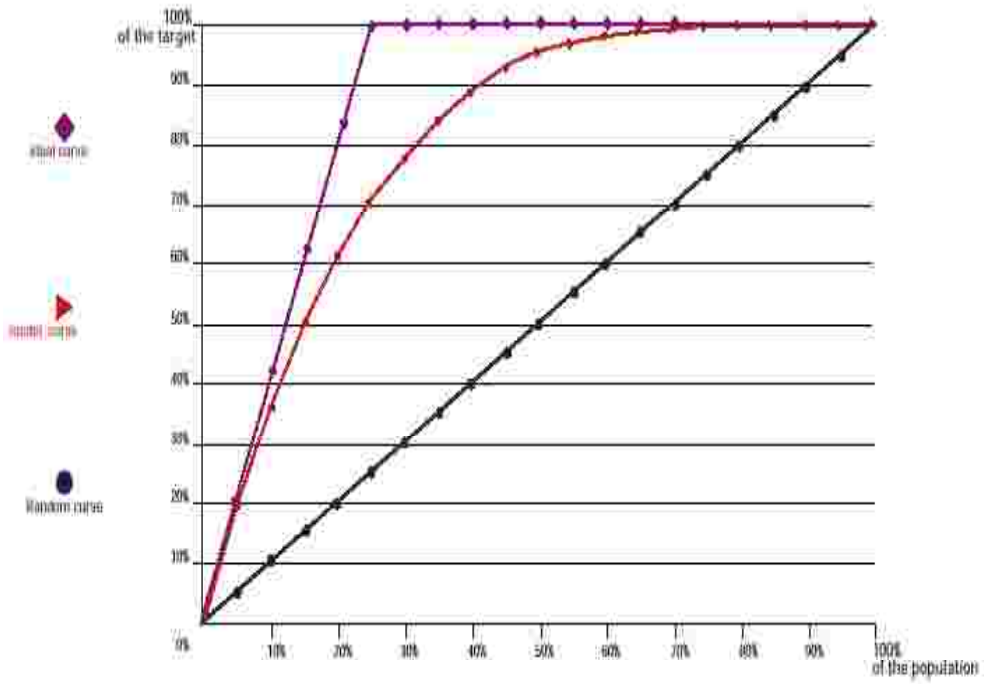
En faisant varier le seuil, on voit qu'en augmentant le pourcentage α de faux mauvais, on augmente aussi le pourcentage $1-\beta$ de vrais mauvais.

La courbe suivante (appelée courbe Roc pour « receiver operating curve ») est souvent utilisée pour mesurer le pouvoir séparateur d'un score. Elle donne $1-\beta(s)$ en fonction de $\alpha(s)$ lorsque l'on fait varier le seuil s du score. Plus elle est proche de la partie supérieure du carré, meilleure est la séparation. Lorsque les deux densités sont identiques, la courbe ROC se confond avec la diagonale du carré.

La surface entre la courbe et l'axe des abscisses, comprise entre 0 et 1, est également parfois utilisée. On peut montrer qu'elle est théoriquement égale à la probabilité que $P(X_1>X_2)$ si X_1 et X_2 sont deux variables tirées indépendamment, l'une dans la distribution des « bons », l'autre dans la distribution des « mauvais ».



Les courbes précédentes ne font pas intervenir les proportions réelles de « bons » et de « mauvais ». Les praticiens utilisent alors la courbe de « lift » ou d'efficacité de la



sélection : en abscisse le % de tous les individus bons et mauvais ayant un score inférieur à s, en ordonnée le % de mauvais ayant un score inférieur à s.
La courbe idéale est le segment brisé qui correspond au cas où la distribution des « mauvais » est entièrement inférieure à la distribution des « bons ».

VI Validité prédictive

Mesurer l'efficacité d'un score, comme d'ailleurs de toute règle de sélection, sur l'échantillon dit « d'apprentissage » , c'est à dire celui qui a servi à estimer les coefficients de la fonction de score, conduit à des résultats trop optimistes : en effet les coefficients ayant été optimisés sur cet échantillon, les taux d'erreur sont des estimations biaisées du vrai taux d'erreur, que l'on aura sur de nouvelles données issues de la même population. On peut en effet obtenir de très bons taux de reconnaissance sur l'échantillon d'apprentissage si le nombre de variables explicatives est très élevé : à la limite avec autant de variables que d'observations on pourrait classer sans erreur toute observation, mais ce résultat est purement artificiel.

La validation du score se fait donc à l'aide d'observations supplémentaires, mises de côté, pour lesquelles on connaît Y , et qui servent à simuler le comportement futur du score.

Conclusion

Les méthodes de score, largement utilisées se perfectionnent sans cesse. Elles sont également appliquées dans d'autres domaines : en assurance automobile pour détecter les conducteurs à risque, en prospection publicitaire pour sélectionner des adresses sur un fichier en vue d'un courrier commercial, pour analyser le risque de perte d'un client etc.

Leur usage basé sur une approche statistique permet de mieux quantifier les risques. Bien sur, comme toute méthode statistique, le scoring commet des erreurs et un individu qui a la malchance d'avoir un profil proche de celui de mauvais payeurs sera considéré comme tel ; mais ce type de méthodes commet moins d'erreur et est plus objectif que les jugements d'expert.

Par ailleurs le score de risque bancaire pour un prêt n'est qu'un élément dans le processus de décision et comme le rappelle la CNIL dans sa Délibération n° 88-083 du 5 juillet 1988 portant adoption d'une recommandation relative à la gestion des crédits ou des prêts consentis à des personnes physiques par les établissements de crédit :
« *conformément à l'article 2 de la loi du 6 janvier 1978, aucune décision accordant ou refusant un crédit ne peut avoir pour seul fondement un traitement automatisé d'informations donnant une définition du profil ou de la personnalité de l'intéressé ».*

Pour en savoir plus :

M Bardos, « Analyse discriminante, application au risque et scoring financier », Dunod, 2001
Ouvrage de niveau 2ème cycle universitaire, écrit par la responsable de l'Observatoire des Entreprises de la Banque de France. Unique en son genre, en français.

T.Hastie, R.Tibshirani, J.Friedman , « The Elements of Statistical Learning Theory », Springer-Verlag, 2001
Le livre de référence pour les années à venir, balayant toutes les techniques de modélisation prédictive. Niveau mathématique : 3 ème cycle.