UNE MÉTHODE DE TRAITEMENT DES REFUSÉS DANS LE PROCESSUS D'OCTROI DE CRÉDIT

Asma Guizani[1] & Salwa Ben Ammou[2] & Gilbert Saporta[3]

[1]Institut Supérieur de Gestion de Sousse, rue Abdlaaziz il Behi . Bp 763. 4000 Sousse Tunisie.
[2]Faculté de Droit & des Sciences Economiques et Politiques de Sousse Cité Erriadh - 4023 Sousse Tunisie.
[3]Laboratoire Cédric - CNAM, 292 rue Saint Martin, 75141 Paris cedex 03, France.

Abstract

The object of our paper is to build a credit scoring model based on a representative sample of the total population (accepted + rejected) to remedy the problem of selection bias, PLS-DA regression (PLS Discriminant Analysis) and canonical discriminant analysis are used for their simplicity and efficacy.

Résumé

On présente deux modèles de scoring construits sur la base d'un échantillon représentatif de la population globale (acceptés + refusés) pour remédier au problème du biais de sélection. Les modèles adoptés dans notre cas c'est la régression PLS-DA et l'analyse factorielle discriminante.

Thèmes : Apprentissage et classification, Modèles pour les assurances et les finances.
Mots clés : crédit scoring, régression PLS-DA, analyse factorielle discriminante, augmentation simple, courbe ROC.

1. Introduction

Le risque de défaillance de l'emprunteur a toujours constitué, pour une banque, le cœur même de la problématique de l'analyse financière qui accompagne toute demande de crédit. Disposer de modèles statistiques pour prédire la défaillance est donc devenue primordial pour une banque, surtout dans le contexte actuel de renforcement du contrôle des risques bancaires et le respect de la nouvelle réglementation prudentielle préconisé par les accords « Bâle II » et « Bâle III ».
Le crédit scoring est un outil fondamental de prévision des risques basé sur les caractéristiques du demandeur de prêt. À partir d'un échantillon de dossiers acceptés dont la qualité est connue, on calcule une note, le score, dont on déduit la probabilité de défaut. Cette probabilité ne peut être estimée que pour les dossiers acceptés. On ne peut donc l'estimer pour les demandeurs rejetés dés le départ, (données incomplètes) ce qui peut conduire à des estimations incorrectes (biais de sélection [2]) et à un éventuel manque à gagner pour la

banque si le client rejeté a priori alors qu'en réalité c'est un bon payeur qui n'aurait pas fait défaut si un crédit lui avait été octroyé.
C'est la problématique de la classification semi-supervisée où l'on dispose à la fois d'un ensemble de données étiquetées (dossiers acceptés) et d'un ensemble de données non étiquetées (dossiers refusés). Le grand intérêt de la classification semi-supervisée est de pouvoir combiner l'information contenue dans des données étiquetées et celle contenue dans les données non étiquetées afin d'atteindre des taux de classification plus élevé.
Le traitement des refusés (reject inference en anglais) tente de remédier à ce problème et de corriger biais de sélection en réintégrant les dossiers refusés à l'échantillon initial et par la suite rendre ce dernier représentatif de la population globale (admis+refusés).
Dans la section 2, nous présentons les techniques classiques de traitement des refusés.
La méthode d'inférence de rejet, adoptée dans notre cas, est présentée dans la section 3. Ensuite, nous mettons en application, dans la section 4, cette méthode pour répondre à notre problématique et nous comparons la performance des modèles de score obtenus. Enfin, la section 5 est consacrée aux conclusions et perspectives de recherches pour la mise en ouvre d'autres méthodes et la définition du modèle le plus performant.

2. Panoramas des techniques de traitement des refusés

Parmi les nombreuses méthodes de traitement des refusés dans la littérature, mentionnons :

2.1 L'augmentation

Cette méthode consiste à construire d'abord un score d'acceptation qui prévoit la probabilité d'être accepté parmi la population globale. On applique ce modèle à la population toute entière et on sépare cette dernière en intervalles ou bandes de score selon le critère de notre choix. Á chaque intervalle de score, on définit un poids, chaque dossier accepté est pondéré par ce poids et un modèle de score de défaut est construit sur les acceptés ainsi pondérés.

Intervalle de score	**Nombre d'acceptés**	**Nombre de refusés**	**Poids**
1	A1	R1	(A1+R1)/A1
…	**…**	**…**	**…**
N	An	Rn	(An+Rn)/An

Tableau n°1 : Table de calcul de poids [7]

2.2 Parceling [7]

On construit un modèle de score sur les acceptés, ensuite on départage la population en intervalles de score puis on calcule, sur chaque intervalle, le nombre de défaut et de non défaut des dossiers acceptés et le nombre total des refusés. On applique une hypothèse de taux de défaut sur les refusés (qui nous donne le nombre de défaut et non défaut dans la population des refusés). On définit ensuite, au hasard, les refusés de chaque intervalle en deux classes défaut/non défaut tout en respectant le nombre de défaut/non défaut calculé dans chaque intervalle. On aboutit ainsi à la constitution du « augmented data set » (on regroupe les acceptés initiaux et les refusés calculés à partir de l'hypothèse de taux de défaut) sur le quel on va construire notre modèle de score.

2.3 L'extrapolation [7]

On construit un modèle de score sur les acceptés et on l'applique à tous les dossiers. L'inconvénient de cette méthode, c'est qu'il y a un biais qui peut être positif ou négatif.

2.4 Reclassification « augmented data set » [7]

On construit un modèle de score de défaut sur les acceptés et on l'applique aux dossiers rejetés. Ces derniers sont classifiés en deux catégories défaut et non défaut.
On construit alors le « augmented data set » qui consiste à ajouter aux dossiers acceptés initiaux (étiquetés défaut/non défaut) les dossiers rejetés étiquetés inférés (défaut/non défaut).
Finalement, on produit le modèle de score sur le « augmented data set ». Cette méthode aboutit donc à l'hypothèse implicite que la distribution des défaut/non défaut est la même dans les populations d'acceptés et de rejetés, ce qui est en réalité faux.

2.5 Groupe de contrôle [7]

Cette méthode consiste à accepter tous les dossiers d'un groupe de contrôle représentatif de la population complète. On construit par la suite notre modèle de score sur cet échantillon départagé en deux catégories défaut et non défaut.

3. La méthode de l'augmentation simple [5]

Afin de résoudre le problème du biais de sélection, nous utilisons la méthode de l'augmentation simple qui se résume en les étapes suivantes :

- Etape 1 : construire un modèle de score sur la base d'un échantillon composé des dossiers acceptés seulement qui sont étiquetés en bon et mauvais payeur (c'est notre échantillon d'apprentissage).
- Etape 2 : appliquer le modèle établi sur les refusés et déterminer le taux de défaut de ces derniers (c'est le principe même du crédit scoring adopté par les banques mais qui est ici appliquer sur des dossiers qui devront normalement être refusés).
- Etape 3 : étiqueter les dossiers refusés par bon ou mauvais selon le taux de défaut.
- Etape 4 : une fois l'échantillon des refusés est défini en bon et mauvais payeur, il sera par la suite réintégrer à l'échantillon d'apprentissage de l'étape 1 pour reconstruire un nouveau modèle non biaisé sur la base cette fois ci d'un échantillon représentatif de la population globale (acceptés et refusés). Ce qui représente un risque considérable à prendre par la banque et qui peut lui coûter cher.

Dans notre cas, nous adoptons deux méthodes pour la construction du modèle du score, la régression PLS-DA (Partial Least Squares Discriminant Analysis) [6] et l'analyse factorielle discriminante (AFD) [3]. La variable à prédire étant binaire (1 ou 0) selon la qualité du dossier. On sait en effet que l'analyse discriminante linéaire est dans ce cas identique à une régression multiple. Le score est alors défini comme la combinaison linéaire des variables obtenues par la régression et on étudie les performances de prédiction en faisant varier le terme constant (ou seuil).

4. Modélisation

Les données utilisées proviennent de la compétition PAKDD 2010 (Pacific-Asia Conference on Knowledge Discovery and Data mining qui s'est déroulée en Inde du 21 au 24 juin 2010)[1]. Ce sont des dossiers d'octroi de crédit d'une banque brésilienne sur une période s'étalant de 2006 à 2009. Notre objectif est de construire un modèle qui sert à étiqueter les clients en bon et mauvais payeur. La variable dépendante est donc une variable binaire qui indique si le client a fait défaut pour une période de 60 jours pendant la première année du crédit (Y = 1), sinon Y = 0. Nous comptons 22 variables explicatives dont 7 sont des variables quantitatives et les 15 autres sont des variables qualitatives. Dans ce cadre de mélange de données qualitatives et quantitatives, on a transformé chaque variable qualitative à r modalités en r variables numériques indicatrices de chaque modalité. Nous disposons en tout de 15000 dossiers dont 10000 dossiers étiquetés en bon et mauvais payeur et ne comportant que les dossiers acceptés (ils seront utilisés pour la construction de notre modèle) et 5000 dossiers non étiquetés comportant les dossiers acceptés et refusés et seront utilisés pour la prévision.
Afin de remédier au problème de traitement des refusés, nous avons appliqué la méthode de l'augmentation simple (évoquée précédemment) qui consiste tout d'abord à construire un modèle de score sur la base des 10000 dossiers (acceptés seulement qui sont étiquetés en bon et mauvais payeurs : c'est notre échantillon d'apprentissage) en utilisant une régression PLS-DA que nous comparons par la suite au modèle obtenu par l'analyse factorielle discriminante. Une fois le modèle établi (quelque soit par une régression PLS-DA ou une analyse factorielle discriminante), il sert à prédire et étiqueter les 5000 dossiers supplémentaires (acceptés+refusés : non étiquetés) en bon et mauvais payeur, la règle de décision bancaire en matière d'étiquetage des dossiers s'appuie sur le score obtenu à partir du modèle. Donc, à un score Z calculé inférieur à Z_c (score limite ou seuil) serait associée l'étiquette bon payeur (Y=0), à un score Z calculé supérieur à Z_c serait associée l'étiquette mauvais payeur (Y=1). Une fois, les dossiers étiquetés (selon le modèle construit initialement), on les réintègre aux 10000 premiers dossiers pour reconstruire un nouveau modèle de score non biaisé car il prend en considération un échantillon représentatif de la population globale.
Pour étudier la performance de nos modèles, nous avons utilisé la courbe ROC (Receiver Operating Characteristics) qui relie la proportion de vrais positifs (bons dossiers classés tels) à la proportion de faux négatifs (mauvais dossiers classés bons) lorsqu'on fait varier le seuil du score d'acceptation. L'aire sous la courbe (ou Area Under the Curve – AUC) est un indice synthétique calculé pour la courbe ROC. L'AUC correspond à la probabilité pour qu'un événement positif ait une probabilité donnée par le modèle plus élevée qu'un événement négatif. L'AUC appartient à l'intervalle [0,1], un modèle est considéré comme idéal si l'AUC est égal à 1. La figure 1 représente les courbes ROC pour les deux modèles (régression PLS-DA et analyse factorielle discriminante) avant réintégration des 5000 dossiers supplémentaires.
Nous remarquons, d'après cette figure, que le modèle construit avec l'analyse factorielle discriminante (AUC=0,603) est plus performant que celui construit avec la régression PLS-DA (AUC=0,595), ce qui n'est cependant pas une très bonne performance.

[1] http://sede.neurotech.com.br/PAKDD2010/

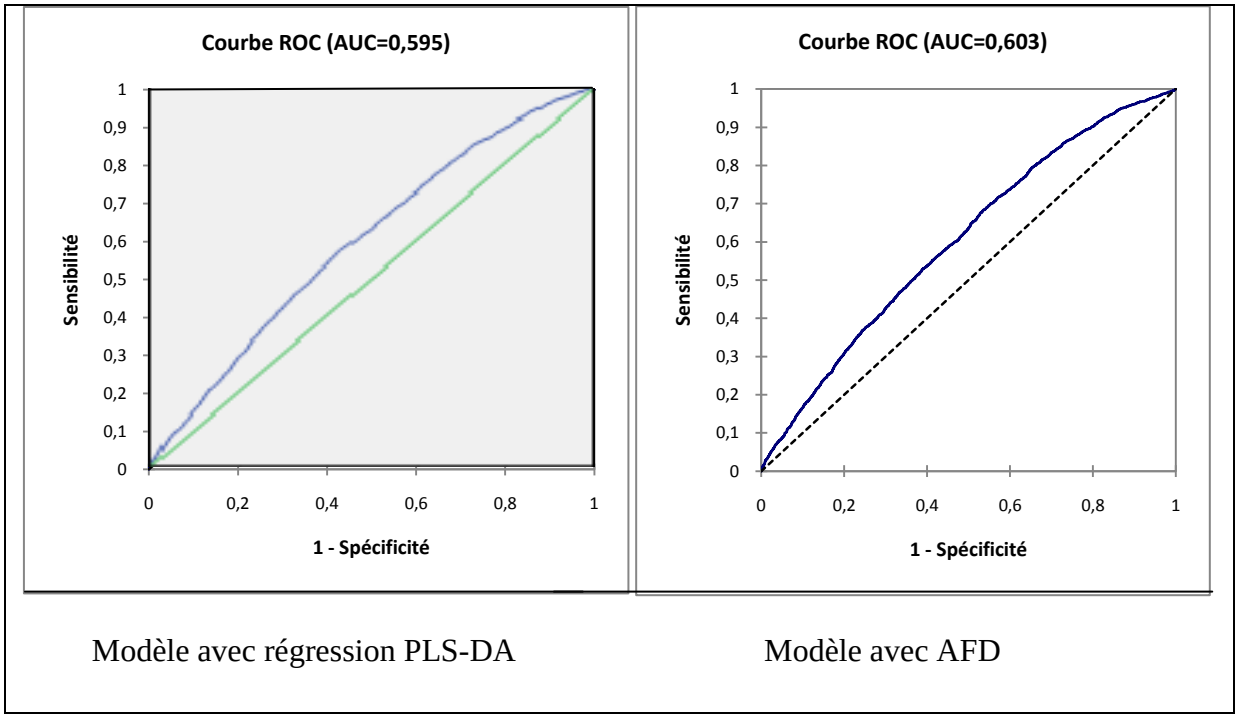

Figure 1 : Courbes ROC avant réintégration des dossiers supplémentaires

La figure 2 représente les courbes ROC des deux modèles une fois les 5000 dossiers réintégrés.

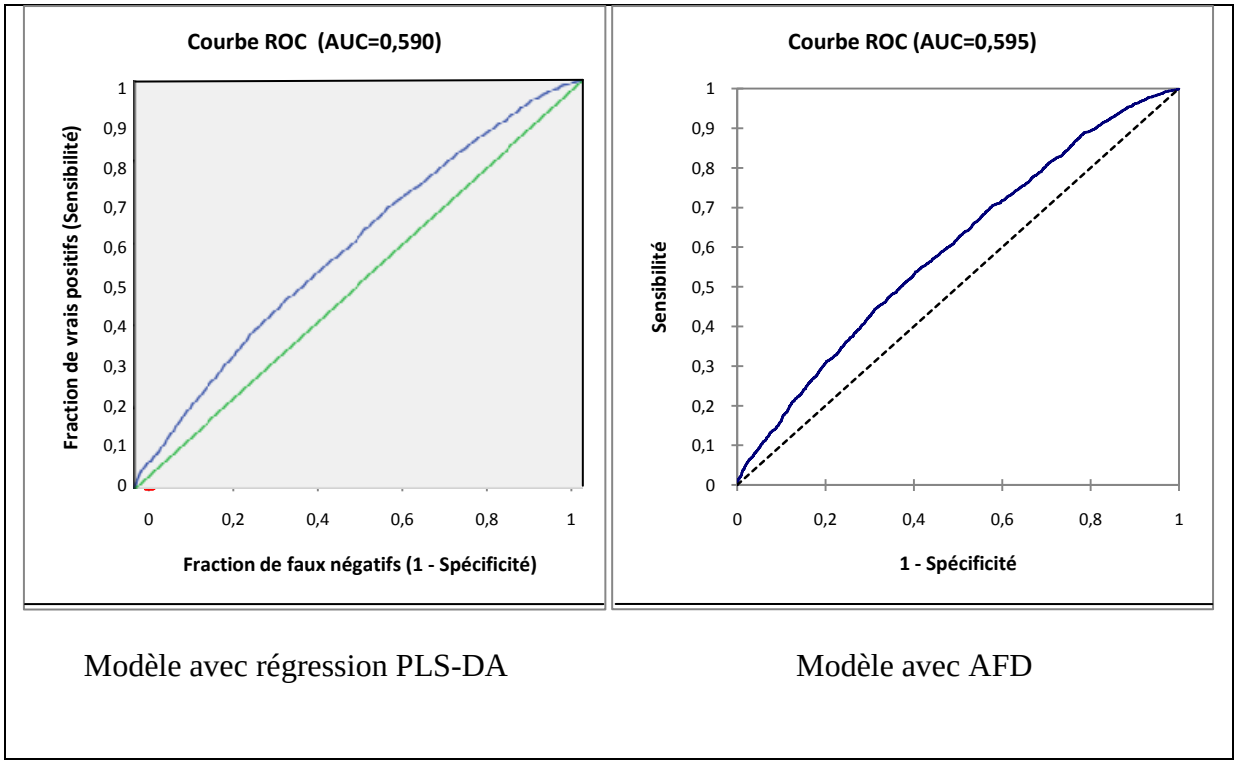

Figure 2 : courbes ROC après réintégration des 5000 dossiers supplémentaires

Nous constatons que la valeur de l'AUC, pour les deux modèles construits avec la régression PLS-DA et l'analyse factorielle discriminante, reste presque la même pour les deux cas de figure (avant et après réintégration des dossiers supplémentaires). Elle passe de 0.595 à 0.590

pour le premier modèle et de 0.603 à 0.595 pour le deuxième modèle, mais il reste toujours plus performant que celui de la régression PLS-DA.
Pour valider nos deux modèles, nous avons calculé, pour l'analyse factorielle discriminante, le taux de bon classement des individus. Ce qui aboutit aux résultats suivants : 55.2% des observations d'origine sont correctement classées et 54.9% des individus ayant subi une validation croisée sont bien classées, ce qui révèle que notre modèle discriminant est moyennement significatif. Quant' à la régression PLS-DA, nous avons appliqué le modèle sur l'échantillon de prévision pour déterminer la classe de chaque individu, la valeur de l'AUC est de 0.588.

5. Conclusions et perspectives

Le traitement des refusés dans le processus d'octroi de crédit est une méthode qui permet de remédier au problème du biais de sélection en réintégrant les dossiers refusés par la banque dans l'échantillon qui sert à construire le modèle de score.
Nous nous sommes basés sur la méthode de l'augmentation simple comme méthode de l'inférence de rejet pour aboutir à un modèle de score construit sur la base d'un échantillon représentatif de la population globale. Nous avons mis en œuvre deux techniques statistiques pour construire notre modèle de score, à savoir la régression PLS-DA et l'analyse factorielle discriminante. Cette dernière s'avère plus performante que la première. L'objectif pour la suite est de mettre en œuvre d'autres méthodes de réintégration des refusés et de comparer les performances de chacune pour déterminer la méthode la plus efficace à corriger le biais de sélection et ainsi d'éviter à la banque le manque à gagner en refusant un dossier qui peut s'avérer par la suite un bon payeur.

Bibliographie

[1] Anderson, R. (2007) *The credit scoring toolkit theory and practice for retail credit risk management and decision automation*. Oxford University Press, New York.

[2] Banasik, J. et Crook, J. (2007*) Reject inference, augmentation, and sample selection*. European Journal of Operational Research, 183, 1582-1594.

[3] Bardos, M. (2001) *Analyse discriminante ; application au risque et scoring financier*. Dunod, Paris.

[4] Bardos, M. (2005) *Les scores de la Banque de France : leur développement, leurs applications, leur maintenance*. Bulletin de la Banque de France, 144, 63-73.

[5] Siddiq, N. (2006) *Credit risk scorecards developing and implementing intelligent credit scoring*. John Wiley & Sons, Inc., New Jersey.

[6] Tenenhaus, M. (1998) *La régression PLS*. Editions Technip, Paris

[7] Viennet, E., et Fogelman Soulié, F. (2007) *Le traitement des refusés dans le risque crédit*. Revue des Nouvelles Technologies de l'Information (RNTI-A-1), 23-45.