Le Traitement des Refusés dans le Risque Crédit

Emmanuel Viennet*, Françoise Fogelman Soulié**

*Université Paris 13, Institut Galilée, LIPN UMR 7030 CNRS
99, Avenue J-B. Clément - 93430 Villetaneuse France
Emmanuel.Viennet@lipn.univ-paris13.fr,
http://www-lipn.univ-paris13.fr

**Kxen, 25 Quai Gallieni, 92 158 Suresnes cedex
Francoise.SoulieFogelman@kxen.com
http://www.kxen.com

Résumé. Nous présentons la problématique du traitement des refusés dans le cadre de l'octroi de crédit. Les modèles d'octroi sont basés sur l'historique de remboursement du crédit par les clients acceptés par l'organisme de crédit, l'information sur les refusés manque donc systématiquement : il en résulte que les modèles sont construits sur des échantillons intrinsèquement biaisés. Dans le cadre législatif actuel, les organismes doivent industrialiser leurs procédures et mettre en place des procédures de traitement documentées, notamment pour les refusés. Nous passons en revue quelques unes des techniques utilisées aujourd'hui par les grands organismes de crédit (reclassification, parcelling, reweighting et groupe de contrôle). Nous montrons sur un cas réel le comportement de ces techniques et évaluons l'impact du traitement sur le risque : en particulier, l'utilisation d'un groupe de contrôle bien choisi est évaluée. La capacité à évaluer rapidement et effectivement les diverses techniques apparaît comme un point clé pour l'industrialisation des techniques de traitement des refusés.

1 Introduction

Le traitement des refusés (ou *Reject Inference*) est un problème rencontré de façon systématique dans les activités bancaires d'octroi de crédit. Ce problème est lié au fait que les scores d'octroi sont construits sur les historiques de données disponibles (clients de la banque ayant obtenu et remboursé un crédit, en faisant ou non défaut sur une ou plusieurs échéances) qui ne comprennent pas, par construction même, d'information sur l'éventuel défaut de remboursement des clients à qui l'on a refusé le crédit (l'échantillon des données utilisées pour le calibrage du modèle est biaisé en recrutement : seuls les Acceptés en font partie). Il existe dans la littérature de très nombreuses techniques qui permettent d'effectuer le traitement des Refusés. Cependant, *aucune* n'est réellement fondée théoriquement, et on peut même se demander (Hand et Henley, 1993) si ces techniques ont une quelconque efficacité pour redresser ce biais. La pratique des établissements de crédit consiste alors, dans cette situation, soit à affiner progressivement une technique choisie plus ou moins a priori, soit à comparer, pour chaque

nouvelle situation (crédit sur un certain produit), un ensemble de techniques dont la meilleure sera dorénavant utilisée pour ce score produit. La difficulté réside alors dans la capacité à produire facilement de nombreux modèles et à les comparer efficacement.
Nous présentons ici une méthodologie systématique de traitement des refusés, basée sur la production de modèles selon différentes techniques classiques (que nous présentons) et leur évaluation. Nous utilisons le logiciel KXEN qui permet, très rapidement de produire les modèles, quelle que soit la volumétrie des données.
Nous montrons que différents critères de performance peuvent être utilisés : les critères globaux (tels que l'AUC ou l'indice Gini présentés plus loin) sont peu appropriés au contexte du score crédit, qui met toujours en œuvre un seuil et donc un point de fonctionnement autour duquel la décision d'Accepter / Refuser va se faire. Nous montrons donc qu'il est préférable d'utiliser des indicateurs locaux (taux d'erreur, revenu par exemple) qui seuls permettent d'évaluer le risque du score d'octroi (Hand, 2005).
De plus, les différents modèles ont en général, sur des données réelles, des performances très voisines : il n'existe pas de technique nettement supérieure aux autres. Il est donc important, pour pouvoir comparer les techniques, de calculer les barres d'erreur sur les indicateurs de performance retenus. Nous présentons donc un certain nombre de moyens de calculer ces barres d'erreur pour les critères globaux et locaux.
Nous présentons dans la section 2 la problématique du score d'octroi et montrons que les techniques classiques sont biaisées ; nous décrivons ensuite (section 3) les différentes techniques de traitement des refusés ; nous introduisons en section 4 les différents critères d'évaluation des performances en indiquant comment calculer les barres d'erreur pour un critère global (AUC). La section 5 illustre la méthodologie présentée en s'appuyant sur un ensemble de données réel, mais de petite taille.

2 Le problème

La plupart des grands organismes de crédit (crédit immobilier ou crédit à la consommation) utilisent des techniques de scoring pour décider de l'opportunité d'accorder leurs prêts et évaluer leurs risques. Le domaine du score crédit est donc très développé et les techniques utilisées variées : le score crédit a pour objectif de fournir un instrument pour mesurer le risque, guider la décision d'octroi ou refus du crédit et gérer le portefeuille des risques crédit. Il fournit une estimation de la probabilité qu'un client qui demande un prêt ou qui a déjà un crédit fera défaut dans le remboursement de son emprunt. On distingue classiquement deux types de problématiques : le *score d'octroi* estime le risque d'un dossier au moment où celui-ci est déposé, avant que le prêt soit accordé ; le *score comportemental* estime le risque tout au long de la durée de l'emprunt, il est calculé pour l'ensemble des clients emprunteurs et est régulièrement recalculé pour prendre en compte les comportements récents du client. Dans toute la suite de cet article, nous nous intéresserons uniquement au calcul du score d'octroi.
Supposons que pour chaque dossier de demande de prêt, on observe k *caractéristiques* $x = (x_1, x_2, \ldots, x_k)$. Ces caractéristiques comprennent des informations clients statiques (par exemple la date de naissance, le nombre d'enfants, la situation de famille, le salaire mensuel, la détention ou non d'une carte bancaire), des données dynamiques concernant l'activité bancaire de ce client (par exemple le niveau d'endettement, la fréquence de découvert du compte courant,...) et enfin des informations concernant le prêt demandé (le montant du prêt, son ob-

jet et sa durée,. . .).
Notons y le résultat de l'emprunt : $y \in \{0, 1\},\ y = 0$ si le prêt est en défaut, $y = 1$ si l'emprunt est remboursé sans problème. Remarquons que la notion de *défaut* peut être définie de nombreuses façons plus ou moins restrictives, par exemple avoir un retard de paiement une fois sur une échéance ou encore avoir un retard de paiement d'au moins 10 jours sur une période de 6 mois. Notons d la décision d'octroi :$d \in \{0, 1\}, d = 1$ si le prêt est accordé, $d = 0$ sinon. Évidemment, le résultat de l'emprunt y est connu uniquement pour les clients à qui on a accordé un crédit dans le passé, c'est-à-dire ceux pour lesquels $d = 1$. Par ailleurs, l'information sur le résultat y est habituellement disponible longtemps après la décision d'octroi (plusieurs années dans le cas de crédit immobilier, plusieurs mois pour le crédit à la consommation). Les données disponibles ont donc la structure suivante (cf. tableau 1) : l'historique clients conserve, pour chaque client ayant demandé un crédit, ses caractéristiques, la décision d'octroi ou de refus du crédit et le résultat, défaut ou non défaut. Pour les clients refusés, l'information du résultat est *manquante* : l'ensemble de données disponible a *un biais de recrutement*.

Caractéristiques	X_1	X_2	X_3	...	X_k	d	y
Client							
110						1	1
125						1	1
175						1	0
305						1	1
472						1	1
525						1	1
792						1	0
1254						1	1
2553						0	?
3201						0	?

TAB. 1 – *Les données clients. Les caractéristiques X_i représentent les données sur le client (données statiques, données dynamiques et données sur le prêt)*

Les données clients sont utilisées pour construire un score, $S(x)$, c'est-à-dire une estimation de la probabilité de non-défaut : $S(x) = \hat{P}(y = 1/x)$.
La décision d'octroi est alors basée sur le score S et sur un seuil s : $d = 1$ si : $S(x) \geq s$ et $d = 0$ sinon. Selon la valeur du seuil s, et pour un même score S, on pourra faire varier la règle de décision d et obtenir des taux d'acceptation différents. Dans la pratique, le processus d'octroi (décrit dans la figure 2) peut être compliqué par le fait que certains clients dont le dossier a été accepté peuvent finalement refuser l'offre, ou que certains dossiers refusés ou acceptés sur la base du score peuvent être retravaillés manuellement et donner lieu à une décision différente en application de règles internes à l'institution. Dans la suite de cet article, et notamment dans l'exemple présenté, nous supposerons que la décision Accepter / Refuser est uniquement basée sur le score. Notons enfin qu'un certain nombre de clients acceptés peuvent finalement refuser l'offre (préférant par exemple un concurrent) : nous ne prenons pas ce problème en compte dans la suite et traitons donc ces cas éventuels comme les refusés.
Quand on construit le score, on utilise l'historique disponible, décrit par le tableau 1. Généralement, les méthodes classiques de scoring (Hand, 2001) utilisent uniquement les données *étiquetées* par l'information de résultat, c'est-à-dire dans notre contexte les données des Acceptés. Des techniques récentes (Chapelle et *al.*, 2000, 2003) visent à utiliser également les données non étiquetées, c'est-à-dire ici les Refusés ; toutefois, ces techniques n'ont pas encore

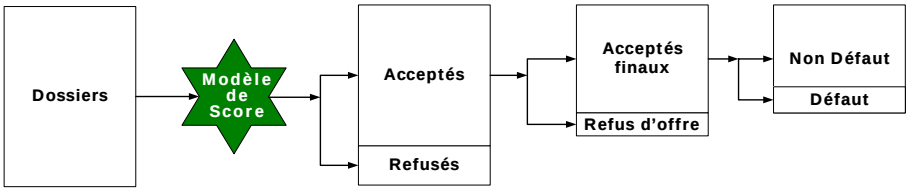

FIG. 1 – *Le processus d'octroi de crédit.*

été appliquées à des problèmes de très grande taille (plusieurs millions de clients, caractérisés par plusieurs centaines d'attributs) et un travail d'investigation systématique reste à faire.

Nous nous plaçons donc ici dans le cadre suivant : étant donné une base de données contenant les données historisées d'un ensemble de clients Acceptés $A = \{i = 1, \ldots, N\}$ et Refusés $R = \{i = N+1, \ldots, N+M\}$, caractérisés par un vecteur $\{x^i = (x_1^i, x_2^i, \ldots, x_k^i), y^i\}$, pour les Acceptés et $\{x^i = (x_1^i, x_2^i, \ldots, x_k^i)\}$ pour les Refusés. On calcule un score de non-défaut estimant la probabilité qu'un dossier accepté ne fasse pas ultérieurement défaut, en utilisant les données des Acceptés : $\{[x^i = (x_1^i, x_2^i, \ldots, x_k^i), y^i], i = 1, \ldots, N\} \rightarrow S$, avec $S(x) = \hat{P}(y = 1/x)$. On verra dans la suite qu'on a parfois aussi besoin d'un score d'acceptation, estimant la probabilité qu'un dossier soit refusé, en utilisant les données des Acceptés et des Refusés :$\{[x^i = (x_1^i, x_2^i, \ldots, x_k^i), d^i], i = 1, \ldots, N\} \rightarrow S$, avec $S(x) = \hat{P}(d = 1/x)$.

Une fois construit, le score de non-défaut est appliqué à tous les dossiers de demande de crédit, qui seront acceptés ou refusés selon que $S(x)$ est supérieur ou inférieur au seuil fixé s. Cependant, les données des Acceptés utilisées pour construire S ne constituant pas un échantillon représentatif de la population totale Acceptés + Refusés, sur laquelle le score est utilisé, une erreur systématique est commise. Cette erreur est appelée dans la littérature *sample selection error ou censoring*. La difficulté, en particulier dans le contexte de Bâle II, est que les établissements de crédit devant fournir une estimation de leur risque souhaitent avoir une estimation de ce biais, qui peut en fait, selon les cas, être positif ou négatif (sur ou sous-estimation du risque de défaut), ce qui introduit un doute dans les estimations nécessaires pour la constitution des fonds propres.
Remarquons que ce problème ne se présente pas pour le score d'acceptation, puisqu'on a là bien l'information Accepté / Refusé pour tous les dossiers.
Le *traitement des refusés* ou *Reject Inference* a pour but de corriger ce biais.

3 Techniques de traitement des refusés - état de l'art

3.1 Contexte

Comme indiqué dans l'introduction, le traitement des refusés comprend toujours la production d'un - ou plusieurs - score(s). Il existe de très nombreuses techniques dans la littérature pour construire un score, depuis les techniques classiques d'analyse discriminante jusqu'aux

techniques les plus modernes (réseaux de neurones, SVM, ...). Cependant, il semble que, dans le domaine du risque crédit et contrairement à d'autres domaines, les techniques modernes n'apportent pas d'avantage significatif : Thomas et *al.* (2005) par exemple considèrent que les différences d'erreur entre les différents modèles sont moins importantes que l'erreur due à la mauvaise qualité des données, ce que semblent confirmer les résultats de comparaison détaillée de Baesens et *al.* (2003) [1]. Par ailleurs, les techniques modernes (réseaux de neurones, SVM) sont en général refusées par le régulateur Bâle II, du fait de la difficulté à documenter et expliquer la décision d'Acceptation finale, notamment auprès du client. Nous ne discuterons donc pas ici du choix de la technique de score utilisée : dans les expériences décrites au §5, nous utiliserons une régression régularisée produite par le module K2R de KXEN (cf. les détails plus bas.)

Les techniques de traitement des refusés sont également très nombreuses (Ash and Meester, 2002, Crook and Banasik 2004, Feelders, 2000, 2003, Hand, 2001) : cependant, la plupart des résultats publiés semblent indiquer des résultats pour le moins mitigés [2]. Ces méthodes sont souvent basées sur des hypothèses *ad-hoc* (par exemple la normalité des distributions de $x_{|y=1}$ et $x_{|y=0}$ pour l'analyse discriminante) qui ne sont pas nécessairement valides sur les données disponibles : le traitement des refusés par une méthode pour laquelle les hypothèses ne seraient pas vérifiées peut alors tout à fait conduire à une dégradation des performances ! L'efficacité de la méthode dépendra donc fortement de la nature des données utilisées.

C'est pourquoi, et en accord avec Ash et Meester (2002), nous pensons que la meilleure méthodologie de traitement des refusés doit être basée sur l'implémentation systématique d'un ensemble de méthodes, leur comparaison et la sélection de la méthode la plus adaptée à l'ensemble de données particulier disponible. En effet, en pratique, la banque dispose d'un ensemble de données fixe et la question pour elle n'est pas de déterminer la meilleure méthode en général (qui n'existe d'ailleurs sans doute pas, en l'état actuel des connaissances), mais bien celle qui est la meilleure pour cet ensemble de données là. Ces deux problèmes («conditionnel aux données» et «inconditionnel»), selon la terminologie discutée dans (Hand, 2005) sont différents et peuvent tout à fait donner des réponses opposées.

Notons finalement que la meilleure méthode de traitement des refusés reste d'obtenir l'information manquante : en octroyant un crédit à un échantillon représentatif de la population globale, quelle que soit leur note de score, on peut construire un modèle applicable à l'ensemble de la population. Bien que cette méthode ait un coût (puisque le crédit est octroyé à des clients qui auraient dû normalement être refusés et dont les perspectives de défaut sont donc certainement plus élevées), ce coût peut sans doute être compensé par une amélioration significative de l'estimation du risque par le modèle obtenu (Hand, 2001) : nous revenons sur ce point au § 5. Les établissements de crédit toutefois sont assez réticents pour utiliser cette approche, à la fois par crainte du risque encouru, mais également parce qu'il est alors difficile d'éviter l'afflux d'une population à haut risque spécifiquement attirée par l'information que les dossiers ne seront pas filtrés.

Nous allons maintenant présenter rapidement quelques méthodes classiques de traitement des refusés, notre but n'étant évidemment pas l'exhaustivité, mais simplement de donner des exem-

[1] « The more sophisticated and flexible modern methods do not generally have much of an advantage over the older more straithforward methods »(Hand, 2001). « The majority of classification techniques yielded classification performances that are quite competitive with each other »(Baesens et *al.* 2003).

[2] «None of these methods is ideal, all having serious weaknesses, and none can be applied in every situation» (Hand, 2001).

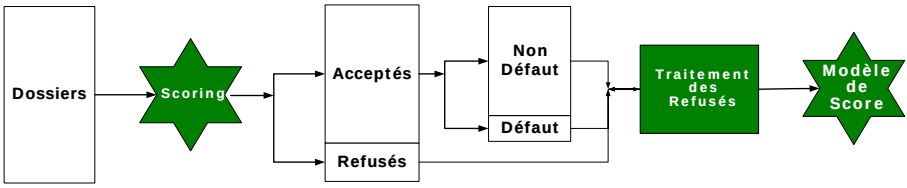

FIG. 2 – *Le traitement des refusés dans le processus d'octroi de crédit.*

ples des techniques les plus utilisées que nous illustrerons dans les expérimentations présentées au §5.
Le *processus de traitement des refusés* est présenté figure 2 : une première méthode de scoring plus ou moins manuelle a abouti à la génération d'une population d'Acceptés (parmi lesquels certains se sont trouvés en défaut par la suite, et d'autres pas) et de Refusés. Cette méthode doit être une méthode systématique (carte de score avec éventuellement forçage manuel par exemple) de façon à ce que la décision Accepter / Refuser soit significative d'une réelle différence entre les dossiers respectifs : c'est la technique de scoring que l'institution a utilisée jusque là et qu'il s'agit d'améliorer en implémentant le traitement des refusés, de façon à corriger le biais d'échantillon : les nouveaux dossiers seront ensuite scorés avec ce nouveau modèle. Ce nouveau modèle de score devra être plus performant et également pouvoir être documenté pour répondre aux exigences des organismes de régulation (Bâle II).

3.2 Techniques classiques de traitement des refusés

3.2.1 Extrapolation

La méthode la plus simple, dite d'*extrapolation* (Ash and Meester, 2002), reste encore de ne rien faire : on construit un modèle de score de défaut sur les Acceptés et on l'applique à l'ensemble de la population, en extrapolant donc sur la population des refusés. On sait qu'il y a un biais, dans un sens ou dans l'autre [3]. Cependant, quand le taux de rejet est très faible, cette technique peut s'appliquer.

3.2.2 Reclassification

Cette méthode, dite aussi d'*ensemble augmenté* («augmented data set») (Ash and Meester, 2002) consiste à itérer la méthode d'extrapolation : on construit un modèle de score de défaut sur les Acceptés et on l'applique à l'ensemble de la population, en extrapolant sur la population des Refusés (qui reçoivent donc une étiquette défaut ou non-défaut). On construit alors un nouveau score sur la population globale - dite ensemble augmenté - Acceptés + Refusés en utilisant les étiquettes d'origine pour les Acceptés et les étiquettes extrapolées pour les Refusés. Cette technique fait implicitement l'hypothèse que $P(y/d = 1) = P(y/d = 0)$ c'est-à-dire que la distribution des défauts / non-défauts est la même dans les populations populations

[3] Kraft et *al.* (2004) fournissent une borne sur l'erreur commise. L'étude systématique de cette borne reste à faire

d'Acceptés et de Rejetés, ce qui est évidemment faux : on n'a donc aucune garantie que la technique de reclassification améliore les performances.

3.2.3 Augmentation

Cette méthode, dite aussi de *re-weighting* (Hsia, 1978) consiste cette fois à construire d'abord un score d'acceptation : on prend l'ensemble de la population Acceptés + Refusés et on cherche à prévoir la probabilité d'être Accepté. On applique ensuite ce modèle à la population et on sépare la population en *intervalles* ou *bandes de score* selon le critère de son choix (nombre donné d'intervalles, de même longueur, de même effectif, ...). Ensuite, on définit, pour chaque intervalle de score, le *poids* de cet intervalle comme l'inverse de la fréquence d'Acceptés dans cet intervalle (tableau 2). Ensuite chaque Accepté est pondéré par ce poids et un score de défaut est construit sur les Acceptés ainsi pondérés.
Remarquons que si le score d'acceptation est presque parfait (c'est à dire que tous les scores des clients faisant défaut sont inférieurs aux scores de ceux ne faisant pas défaut), certains intervalles auront un nombre nul d'acceptés, et que donc le poids ne pourra pas être calculé. Enfin, comme précédemment, cette technique fait implicitement l'hypothèse que $P(y/d = 1) = P(y/d = 0)$. Cette hypothèse est là également non fondée, et la technique d'augmentation n'a donc aucune raison - théorique - de fonctionner.

Intervalle de score	Nombre d'acceptés	Nombre de refusés	Poids
1	A_1	R_1	$(A_1 + R_1)/A_1$
...	...	...	...
n	A_n	R_n	$(A_n + R_n)/A_n$

TAB. 2 – *Calcul des poids dans la méthode d'augmentation*

3.2.4 Parcelling

Cette méthode consiste cette fois à construire un modèle de score de défaut sur les Acceptés. Ensuite, on sépare la population en *intervalles de score* selon le critère de son choix, puis on calcule, sur chaque intervalle de score, le nombre de défaut / non-défaut dans la population des Acceptés et le nombre total de Refusés. On fait ensuite une *hypothèse* de taux de défaut des refusés par intervalle de score : on en déduit le nombre de défaut / non-défaut dans la population des Refusés (tableau 3). On étiquette ensuite les Refusés de chaque intervalle en attribuant au hasard l'étiquette défaut / non-défaut tout en respectant les nombres de défaut / non-défaut calculés dans cet intervalle. On constitue ainsi un ensemble augmenté et, comme dans la méthode de reclassification, on construit un nouveau score sur la population globale Acceptés + Refusés en utilisant les étiquettes d'origine pour les Acceptés et les étiquettes calculées pour les Refusés.

Intervalle de score	Acceptés		Refusés			
	Nb de Non-Défaut	Nb de Défaut	Nb de Refusés	Nb de Non-Défaut	Nb de Défaut	Taux de Défaut
1	1400	13	4	4	0	1%
...	...	...	...	...	...	...
n	117	177	97	37	59	61%

TAB. 3 – *Calcul des étiquettes des refusés dans la méthode parcelling.*

3.2.5 Groupe de contrôle

Cette méthode consiste à accepter tous les dossiers d'un groupe de contrôle constitué de façon à représenter la population complète ou par toute autre méthode (Thomas el *al.* 2005). On construit ensuite un modèle de score sur cet échantillon où tous les éléments, ayant été acceptés, sont étiquetés de l'information défaut / non-défaut. Cette technique est la seule qui soit valide du point de vue statistique et tous les outils disponibles en statistique permettent d'estimer, en fonction de la taille du groupe de contrôle choisi, si l'application à la population globale est valide. C'est la meilleure - et la seule - technique de Traitement des Refusés (Hand, 2005), même si, en pratique, le coût encouru en acceptant des dossiers à fort potentiel de défaut doit être contrôlé, en particulier en regard du bénéfice attendu.

3.3 Méthodologie de traitement des refusés

Nous proposons donc la méthodologie suivante pour le traitement des Refusés pour un ensemble de données $\{[x^i = (x^i_1, x^i_2, \ldots, x^i_k), y^i], i = 1, \ldots, N; [x^i = (x^i_1, x^i_2, \ldots, x^i_k)], i = N+1, \ldots, N+M\}$.
On commence par préparer les données : on collecte toutes les informations disponibles, on effectue les préparations nécessaires (jointures de tables, calcul de variables supplémentaires ...) et on les intègre dans un «Analytic Data Set »de façon à les avoir sous un format accessible (par exemple, un tableau plat), puis on construit plusieurs modèles selon les différentes techniques de Traitement des Refusés présentées au § précédent. On compare ensuite ces différents modèles et on choisit le meilleur modèle qui fournira le score utilisé ensuite. Cette méthodologie (figure 3) est donc basée sur *la production systématique de plusieurs modèles, leur comparaison et la sélection du meilleur modèle*. Nous présentons au §4 différentes techniques de comparaison de performance. Cette méthodologie vise bien à choisir le meilleur modèle conditionnellement aux données, il est tout à fait possible qu'un ensemble de données amène à choisir une technique et un autre ensemble une autre technique.

4 Evaluation de performances

L'évaluation et la comparaison des performances de modèles est un domaine largement développé. Classiquement, les classifieurs sont évalués sur la base de leur taux de classification (PCC : Percentage Correctly Classified : nous définissons ci-dessous de façon plus précise les indicateurs discutés ici), c'est-à-dire le nombre d'observations, dans un ensemble test, correctement identifiées. Cependant, le taux de classification pose des problèmes bien connus. Tout

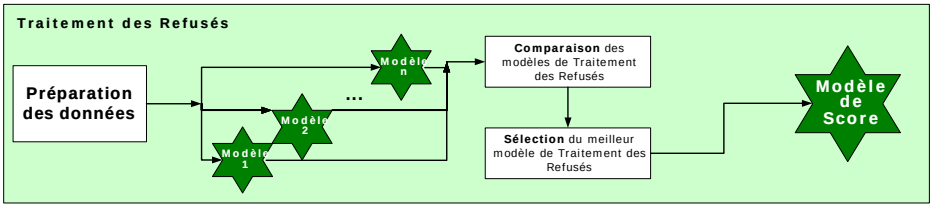

FIG. 3 – *Le processus de traitement des refusés.*

d'abord, si on suppose un problème dans lequel les deux classes (défaut / non défaut) sont très déséquilibrées (par exemple 1 % de défaut et 99 % de non défaut), le classifieur constant qui classe toutes les observations comme non-défaut a un PCC de 99 % ! Il faudrait donc recalibrer (stratifier) la distribution des classes, ce qui est un problème difficile (Provost et *al.* 1998). Ensuite, quand on compare des classifieurs sur des ensembles de données réelles, leurs taux de classification sont en général très similaires et la comparaison revient, pratiquement, à choisir le modèle dont le modélisateur est un expert.
Par ailleurs, les erreurs de classification n'ont pas toutes le même coût : il est sans doute plus coûteux d'accepter un client qui fera défaut que de refuser un client qui n'aurait pas fait défaut. Dans le cas du crédit, la situation est encore plus complexe puisque le coût attaché à un client dépend du montant du crédit, du nombre d'échéances, du taux d'intérêt. Il faudrait donc là plutôt s'intéresser au taux de profit (nous reviendrons sur ce point plus bas.)
D'autres indicateurs ont donc été introduits : AUC (Area Under the ROC Curve), Gini, Kolmogorov Smirnoff (Bradley, 1997, Kraft et *al.* 2004). Dans le §5, nous utilisons également l'indicateur KI introduit par KXEN ($KI = 2AUC - 1$).
Ces indicateurs, très liés, sont intéressants quand on veut comparer globalement des classifieurs. Par contre, quand le classifieur doit travailler en un point de fonctionnement particulier (ou une zone), comme c'est le cas dans le domaine du crédit, les indicateurs globaux peuvent fournir des résultats de comparaison erronés (voir par exemple Hand, 2005). On définit donc des critères en un point de fonctionnement particulier (taux d'erreur, revenu, ...).
Enfin, quand on veut comparer les performances de plusieurs modèles, et ce quel que soit l'indicateur utilisé, il faut être capable de décider si la différence obtenue est significative et donc être capable de calculer des barres d'erreur sur les indicateurs. Quelques travaux ont été consacrés à ce sujet (Bradley, 1997, Provost et *al.* 1998, Maloof, 2002), mais en pratique, ces techniques ne sont pas toujours mises en application dans le cas du *credit scoring*.

Définissons maintenant plus précisément les indicateurs que nous utiliserons dans la suite. Bien que certains de ces indicateurs soient très similaires, nous les décrivons tous, de façon à ce que le lecteur puisse retrouver l'indicateur qu'il a l'habitude d'utiliser. Étant donné un ensemble de N observations $\left\{[x^i = (x^i_1, \ldots, x^i_k), y^i], i = 1, \ldots, N\right\}$, nous supposons que nous avons produit un score $S : x \rightarrow y = S(x)$ estimateur de la probabilité de non-défaut de x (exemple «positif»).
Nous définissons maintenant une règle de décision d qui pour un seuil s est la suivante : $d_s(x) = 1$ si $S(x) \geq s$ et $d_s(x) = 0$ si $S(x) < s$.

		Réel	Réel
		P	N
Classé	P	VP	FP
Classé	N	FN	VN

		Réel	Réel
		P	N
Classé	P	$\alpha(s)$	$1-\beta(s)$
Classé	N	$1-\alpha(s)$	$\beta(s)$

TAB. 4 – *Matrice de confusion.*

Le dossier x est considéré comme «positif» et on décide alors de l'accepter si son score est supérieur au seuil s, «négatif» sinon et donc refusé.
Pour chaque dossier, on pourra ensuite vérifier si le crédit entre en défaut et alors, l'une des quatre situations suivantes peut se produire : x se révèle positif et avait été classé positif (x est un *vrai positif*), x est positif et avait été classé négatif (*faux négatif*), x est négatif et avait été classé positif (*faux positif*), x est négatif et avait été classé négatif (*vrai négatif*). La matrice de confusion résume ceci (table 4). On écrit aussi souvent la matrice de confusion en utilisant les pourcentages au lieu des nombres de dossiers : si VP, VN, FN, VN sont, dans la population considérée, les nombres respectivement de vrais positifs, vrais négatifs, faux négatifs, vrais négatifs, et $nb(P)$ et $nb(N)$ les nombres de dossiers Positifs / Négatifs, les taux de vrais positifs et vrais négatifs au seuil s sont : $\alpha(s) = \frac{\text{VP}}{nb(P)}$ et $\beta(s) = \frac{\text{VP}}{nb(N)}$.
Le *Taux de classification* (correcte) global ou PCC (Percentage Correctly Classified) est alors : $\text{PCC} = \text{CR}(s) = \frac{\text{VP+VN}}{nb(P)+nb(N)}$; le *taux d'erreur* est évidemment $1 - \text{PCC}$.
Si on ne s'intéresse qu'aux observations classifiées positives (comme dans le domaine du crédit, où on refuse les observations classifiées négatif), on définit le *taux d'erreur sur les positifs* (ou *taux de défaut sur les Acceptés* en risque crédit) par :

$$Err_{pos}(s) = \frac{\text{FP}}{nb(P)}$$

La *courbe* ROC (ou Receiver Operating Characteristic curve, figure 4) est la courbe représentant le taux de VP en fonction du taux de FP, *i.e.*, si les observations sont ordonnées dans l'ordre des scores décroissants, la courbe représentant $\alpha(s)$ en fonction de $1 - \beta(s)$.
L'*indicateur* AUC est l'aire sous cette courbe, c'est-à-dire : $\text{AUC} = \int_{-\infty}^{\infty} \alpha(s) d[1-\beta(s)]$.
L'indicateur AUC varie entre 0 et 1 (il vaut $\frac{1}{2}$ pour la décision aléatoire). Ce critère dépend de l'ensemble de la distribution des scores, c'est-à-dire de la distribution des positifs / négatifs dans toute la population.
La courbe de lift (figure 4) est la courbe représentant le taux de VP en fonction du nombre de positifs dans la population, *i.e.*, si les observations sont ordonnées dans l'ordre des scores décroissants, $\alpha(s)$ en fonction de $1 - F(s)$, où $F(s)$ est le taux de Positifs dans la population avec score supérieur à s (*i.e.*, $F(s) = \alpha(s)p_p + [1-\beta(s)]p_N$, avec p_P le taux de positifs et p_N le taux de négatifs : $p_p = \frac{nb(P)}{nb(P)+nb(N)}$ et $p_N = \frac{nb(N)}{nb(P)+nb(N)}$.
L'indicateur KI est le rapport de l'aire sous la courbe de lift au dessus de la diagonale à l'aire sous le modèle parfait («l'oracle») au-dessus de la diagonale (figure 4) : $\text{KI} = \frac{M}{W}$. Comme AUC, KI varie entre 0 et 1 et vaut 0 pour la décision aléatoire. Il est facile de montrer que $\text{KI} = 2\text{AUC} - 1$.
On pourra donc indifféremment utiliser les critères AUC ou KI pour comparer les performances.

La *courbe de Lorenz* (positif) est la courbe (figure 4) représentant le taux de VP en fonction du nombre de positifs dans la population, (*i.e.*, les vrais positifs en fonction des classés positifs). Si les observations sont ordonnées dans l'ordre des scores croissants (*i.e.*, dans le sens contraire des courbes précédentes), c'est la courbe représentant $1 - \alpha(s)$ en fonction de $1 - F(s)$. L'indicateur de *Gini* est alors défini en relation avec l'aire sous la courbe de Lorenz par :

$$\text{GINI} = 1 - 2\int_{-\infty}^{\infty}[1 - \alpha(s)]d[1 - F(s)]$$

On peut montrer que $\text{GINI} = p_N(2\text{AUC} - 1)$ et que le taux d'exactitude (ou Accuracy Ratio : cf Kraft et *al.* 2004) défini par : $\text{AR} = \frac{\text{GINI}}{\text{GINI}_{opt}} = \frac{\text{GINI}}{p_N}$ et que donc KI est identique à AR. On voit donc que les trois indicateurs AUC, KI et GINI sont équivalents. Les praticiens sont, selon les cas, habitués à utiliser l'un ou l'autre.
Enfin, l'indicateur de Kolmogorov-Smirnoff KS mesure la différence entre les distributions des positifs et des négatifs, ou plus précisément : $\text{KS} = \max_s \{\beta(s) - [1 - \alpha(s)]\}$.
La figure 4 présente ces courbes et ces indicateurs. Le logiciel KXEN fournit, pour chaque modèle produit, l'ensemble de ces courbes et indicateurs (ainsi que d'autres que nous n'utiliserons pas dans la suite et que nous ne présentons donc pas).
Il est possible d'estimer un intervalle de confiance sur la valeur de l'AUC (Hanley et McNeil, 1982 ; Provost et *al.*, 1990 ; Bradley, 1997). Nous utilisons ici la technique de Hanley et McNeil : si l'on a p exemples positifs et n exemples négatifs, et si l'on note $\mathbf{P}_p$ la distribution des scores des exemples positifs, et $\mathbf{P}_n$ la distribution des scores des exemples négatifs, la variance de l'AUC s'exprime comme :

$$\sigma_A^2 = \frac{A(1-A) + (p-1)(\mathbf{P}_{ppn} - A^2) + (n-1)(\mathbf{P}_{ppn} - A^2)}{pn}$$

où $\mathbf{P}_{ppn}$ est la probabilité que le modèle fournisse pour deux exemples positifs choisis au hasard un score plus élevé qu'un exemple négatif, et $\mathbf{P}_{ppn}$ la probabilité qu'il place deux exemples négatifs choisis au hasard plus bas qu'un exemple positif. Si l'on fait l'hypothèse (difficile à justifier rigoureusement) que les distributions P_p et P_n sont exponentielles, on montre (Hanley, 1982) que $P_{ppn} = \frac{A}{2-A}$ et $P_{ppn} = \frac{2A^2}{1+A}$. Ce sont ces expressions que nous avons utilisées pour obtenir les résultats présentés dans la section suivante.
Récemment, d'autres expressions (Cortes et *al.*, 2005) ont été obtenues pour calculer des intervalles de confiance indépendants des distributions pour l'*AUC* ; d'autres chercheurs (Macskassy et *al.*, 2005) s'intéressent à la détermination de «bandes de confiance» autour des courbes ROC.

5 Expériences

Nous avons mené quelques expériences permettant d'illustrer notre propos et d'orienter la réflexion sur ce que pourrait être une méthodologie systématique de traitement des refusés. Le but de ces expériences n'est pas d'évaluer la performance finale retenue ou de démontrer la supériorité d'une technique sur une autre, mais bien de démontrer, qu'en l'absence de technique statistiquement fondée (Hand et Henley, 1993), la seule méthode réellement efficace reste de

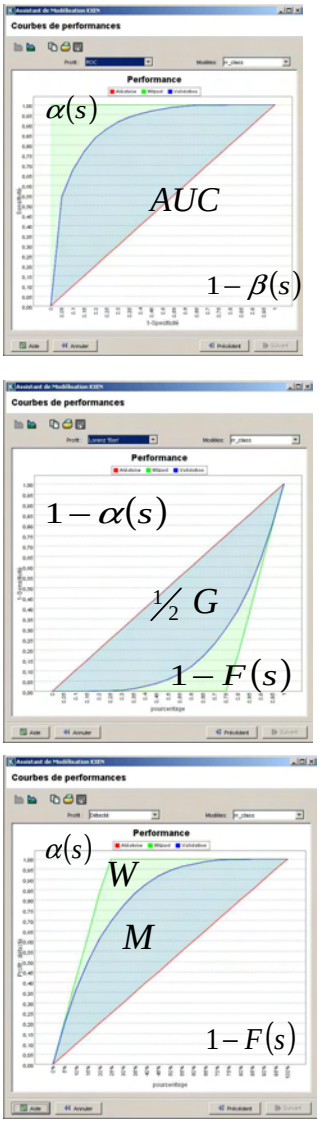

FIG. 4 – *Courbes ROC (en haut), de Lorenz (milieu), de lift (en bas), et indicateurs AUC, Gini et KI.*

produire des modèles et de déterminer la meilleure solution conditionnée aux données disponibles (Hand, 2005).
Nous utilisons pour nos expériences un jeu de données réelles fourni par un client - anonyme - de KXEN. Ces données concernent l'évaluation du risque crédit et comprennent une trentaine de variables pour environ 6 000 clients :

- signalétique client : des indicateurs (niveau de revenu, localisation géographique, profil socioéconomique,. . .), dont certains sont numériques (revenu), d'autres «ordinaux» (valeurs discrètes ordonnées, comme le nombre d'enfants) ou alphanumériques (catégorie socioprofessionnelle) ;
- informations bancaires (activité compte courant, carte bancaire, découvert...) ;
- Informations sur le crédit demandé : montant, durée de remboursement, montant des échéances ;
- issue du crédit : nous savons si le client a fait défaut ou pas avant le terme de son crédit.

Cet ensemble de données réelles est de très petite taille, tant par le nombre réduit de variables que par le faible nombre de clients. Les résultats obtenus ne sont donc pas très significatifs, et la robustesse en est souvent faible (ce qu'indique le critère de robustesse *KR*, écart entre les courbes de lift des ensembles d'estimation et de validation, calculé automatiquement par KXEN). Répétons ici que notre but n'est pas de «résoudre » le problème des Refusés pour cet exemple, mais simplement d'illustrer la démarche. Le lecteur ne devra donc pas s'attacher aux performances obtenues, mais plutôt à la méthodologie suivie pour les obtenir.
Les données disponibles ne comportent donc que des informations sur les Acceptés. Pour réaliser nos expériences, la population des Refusés a été simulée : nous avons pour cela construit un modèle sur l'ensemble de la population, réalisé une segmentation grossière, choisi le segment ayant le plus fort taux de défaut et tiré au hasard $0,9\,r$ clients dans ce segment (si r est le taux de refusés visé) et avons de plus tiré au hasard $0,1\,r$ clients dans les autres segments (pour simuler les imperfections du processus actuel d'attribution de crédit).
A l'issue de ce processus, artificiel mais vraisemblable dans le cadre de cette simulation, nous avons 5% de clients refusés, soit 291 clients dont 93 ont fait défaut (table 5).

Data set	Nombre	Nb positifs	Nb négatifs	%	% Positifs	% Négatifs
Acceptés	5493	5080	413	94.97%	92.48%	7.52%
Refusés	291	198	93	5.03%	68.04%	31.96%
Total	5784	5278	506	100%		

TAB. 5 – *Données d'expériences*

Afin de pouvoir estimer simplement la performance et la robustesse des modèles, nous avons séparé l'ensemble des clients en deux groupes : l'ensemble d'estimation, clients utilisés pour la détermination des paramètres des modèles, et l'ensemble de validation, qui nous permettra de mesurer la performance des modèles sur de nouveaux clients. L'ensemble d'estimation comporte 4 257 clients (dont 5% de refusés) et celui de validation les 1 527 clients restants.

Nous avons ensuite construit des modèles permettant d'ordonner les clients en fonction de leur risque de défaut estimé. Les techniques présentées plus haut (section 3) ont été mises en œuvre à l'aide du logiciel KXEN. Les deux modules utilisés sont les modules d'encodage automa-

tique (K2C) et de classification / régression (K2R).
Le module K2C prend l'ensemble des données disponibles (continues, ordinales ou nominales), analyse ces données en fonction de la cible à calculer (ici la probabilité de défaut), détecte les données manquantes et les *outliers*, regroupe les catégories et finalement code ces données, le tout automatiquement sans intervention de l'utilisateur.
Le module K2R réalise une régression polynomiale régularisée sur les données codées par K2C. L'ensemble de ces deux modules s'appuie sur les techniques de SRM (Structural Risk Minimization) de Vapnik (Vapnik, 1995), ce qui garantit un compromis optimal entre précision et robustesse des modèles calculés. Pour chacun des modèles présentés ici, la manipulation des données a pris quelques heures et la production du modèle lui-même quelques minutes (temps de calcul).

Nous avons donc six modèles :

- $\mathbf{M}_0$: apprentissage sur toutes les données, y compris les refusés ; ce modèle donne la borne supérieure du gain que l'on pourrait attendre d'un traitement optimal des refusés ;
- $\mathbf{M}_1$: extrapolation : pas de traitement des refusés ;
- $\mathbf{M}_2$: reclassification ;
- $\mathbf{M}_3$: augmentation : on a défini 20 bandes de score, correspondant chacune à 5 % de la population, calculé les poids pour chaque bande et pondéré les Acceptés de chaque bande par le poids correspondant ;
- $\mathbf{M}_4$: parcelling : on a défini 5 bandes de score, correspondant chacune à 20 % de la population et fait les hypothèses d'augmentation des taux de défaut par bande de score comme indiqué table 6 ;
- $\mathbf{M}_5$: groupe de contrôle. Le groupe de contrôle a été constitué en sélectionnant 15% des clients les moins risqués, choisis en appliquant le modèle $\mathbf{M}_1$ et en prenant dans les bandes de score les refusés ayant le score de défaut le plus faible.

	Acceptés					Refusés			
Bande de score	Nombre	Nb Positifs	Nb Négatifs	% Positifs	% Négatifs	Nombre	Nb Positifs	Nb Négatifs	Supplément de % Négatifs
< -0.08	1413	1400	13	99.08%	0.92%	4	4	0	0%
[-0.08,0.06]	2701	2646	55	97.96%	2.04%	16	16	0	0%
[0.06,0.21]	1049	914	135	87.13%	12.87%	33	29	4	1%
[0.21,0.36]	249	117	117	39.80%	60.20%	97	37	60	2%
> 0.36	36	3	33	8.33%	91.67%	141	8	133	3%
Total	5493	5080	413	92.48%	7.52%	291	93	198	

TAB. 6 – *Calcul des étiquettes pour la méthode de parcelling*

5.1 Comparaison globale des performances

Dans un premier temps, nous comparons globalement les performances des différentes approches. Comme nous l'avons dit, les indicateurs globaux mesurent la «performance» d'un modèle, globalement pour tous les points de fonctionnement (seuils de décision). La figure 5 présente les courbes ROC, et la table 7 les valeurs des indicateurs couramment utilisés.

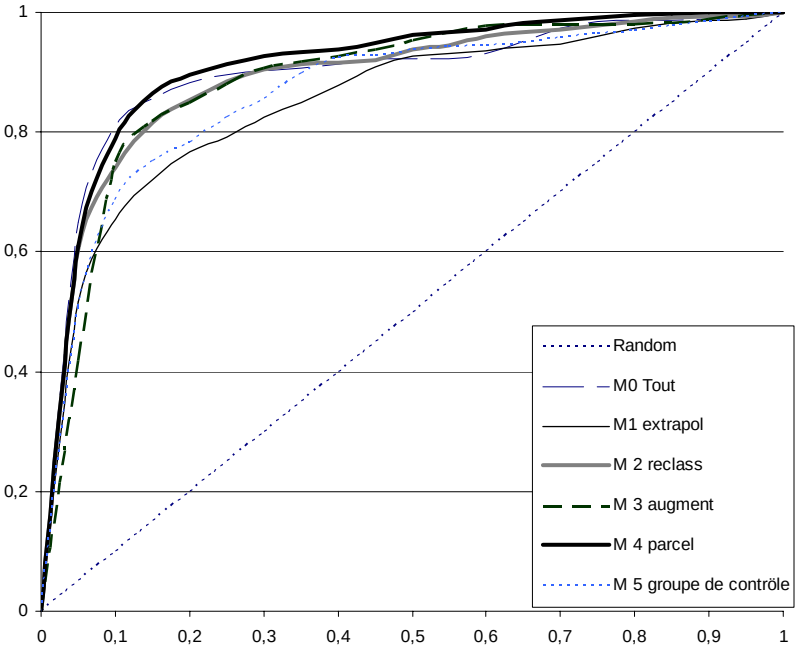

FIG. 5 – *Courbes ROC sur l'ensemble de validation.*

	AUC		KS		GINI		KI	
	Estim.	Valid.	Estim.	Valid.	Estim.	Valid.	Estim.	Valid.
M_0 tout	0,927	0,904	0,800	0,726	0,753	0,724	0,854	0,808
M_1 extrapol.	0,905	0,856	0,654	0,582	0,750	0,659	0,810	0,713
M_2 reclass.	0,896	0,897	0,700	0,668	0,722	0,715	0,792	0,794
M_3 augment.	0,942	0,887	0,761	0,697	0,717	0,652	0,884	0,775
M_4 parcel.	0,905	0,915	0,744	0,734	0,724	0,746	0,810	0,831
M_5 control.	0,900	0,870	0,641	0,598	0,738	0,678	0,800	0,740

TAB. 7 – *AUC, KS, GINI et KI calculés sur les 6 modèles testés. Mesures sur les ensembles d'estimation et de validation*

Les courbes ROC montrent que non seulement les modèles testés ont des performances très semblables, mais aussi que, les courbes se croisant, le «*meilleur» modèle peut dépendre du point de fonctionnement* (voir section suivante).
La table 7 permet de se convaincre que, comme nous l'avions indiqué, les critères AUC, KS, $Gini$ et KI sont bien équivalents pour la sélection de modèle, comme démontré dans la section 4 de cet article. Les résultats sur l'ensemble d'estimation (4 257 points) diffèrent peu de ceux mesurés sur l'ensemble de validation (1 527 points), ce qui indique que les modèles construits sont relativement *robustes* (ils sont capables de généraliser à de nouvelles données) : la modélisation produite par KXEN assure d'ailleurs cette propriété. L'utilisation de l'un quelconque de ces critères conduirait ici à choisir le modèle $\mathbf{M}_4$ (parcelling.)
Par contre, l'utilisation du critère du taux d'erreur global $(1 - \text{PCC})$ conduirait (table 8) à choisir le modèle $\mathbf{M}_1$ (extrapolation.)
Si l'on estime un intervalle de confiance pour l'AUC (figure 6), on constate que les différentes méthodes de traitement des refusés donnent des résultats très proches. Elles apportent toutes un gain de performances : le plus mauvais modèle est bien $\mathbf{M}_1$, dans lequel on ne fait aucun traitement des Refusés. Cependant, pour cette application, les résultats de la méthode «optimale» (modèle $\mathbf{M}_0$, apprentissage sur tous les clients, dont on connaît ici les étiquettes) sont comparables à ceux des autres techniques !

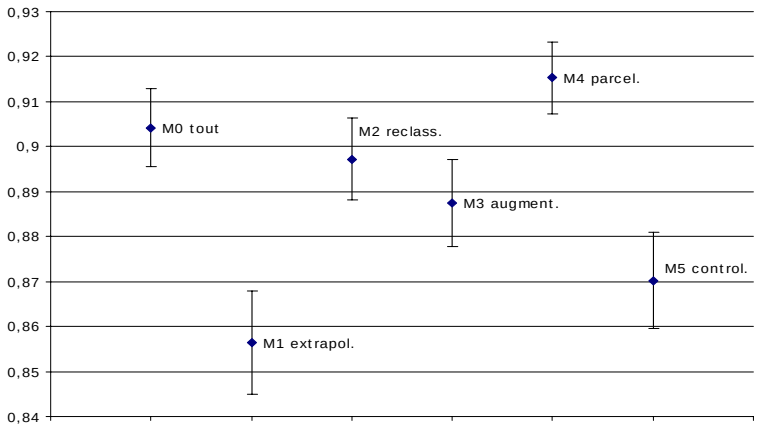

FIG. 6 – *Intervalles de confiance sur l'AUC, calculé selon (Hanley,1982).*

5.2 Comparaison à seuil fixé

Notre exemple concerne l'évaluation du risque d'octroi de crédit : après avoir déterminé un modèle de score, on fixe un seuil s (par exemple pour assurer un certain taux $1 - r$ d'Acceptés) et on accepte ensuite les dossiers dont le score $S(x)$ est supérieur à s. Dans ce cas, la

mesure de performance pertinente n'est pas un critère global comme l'AUC ou PCC, mais bien le *taux de défaut sur les Acceptés* $Err_{pos}(s)$ défini plus haut, c'est-à-dire les clients qui vont faire défaut parmi ceux que l'on a acceptés, puisque ce sont ceux qui sont générateurs de risque (Hand 2005). La décision d'acceptation étant prise en comparant au seuil fixe s le score calculé par le modèle, les performances ne dépendent pas de la distribution particulière des Non Défaut dans les différentes bandes de score.
La table 8 et les figures 7 et 8 présentent le taux d'erreur (taux de défaut sur les acceptés) en fonction du taux d'acceptation (proportion globale d'acceptés). Les intervalles de confiance (donnés en faisant une hypothèse de loi binômiale) sont du même ordre de grandeur pour toutes les courbes (seuls ceux de M5 sont tracés par souci de lisibilité). Le modèle «oracle» représente la décision optimale. Le «meilleur» modèle n'est pas le même selon le seuil choisi (et donc le taux d'acceptation qui en résulte) : si on accepte 25% des clients, les deux meilleurs modèles au sens du taux d'erreur sont (si l'on excepte $\mathbf{M}_0$ qui utilise les refusés durant l'apprentissage et nous sert de référence) le *parcelling* ($\mathbf{M}_4$) et l'*extrapolation* $\mathbf{M}_1$ (autrement dit aucun traitement des refusés). Si on en accepte 85%, c'est l'*extrapolation* qui est le meilleur ; et si l'on en accepte 95%, c'est le *groupe de contrôle* ($\mathbf{M}_5$). Si nous avions utilisé un critère global comme l'AUC, nous aurions choisi le modèle $\mathbf{M}_4$ et avec le critère d'erreur global le modèle $\mathbf{M}_1$!

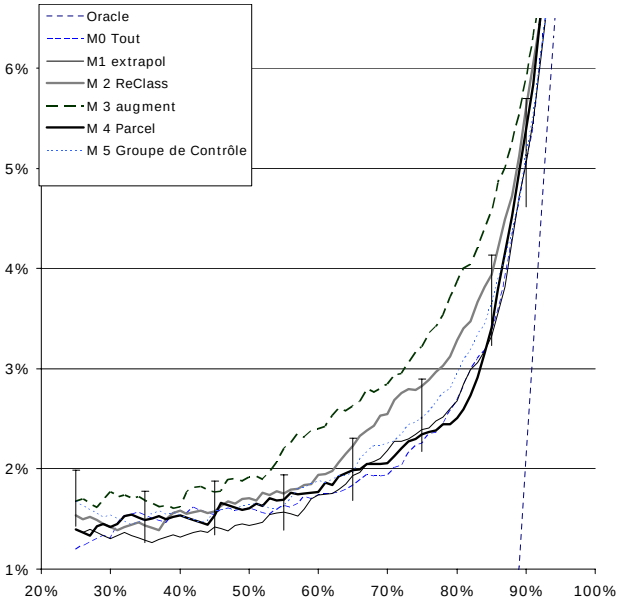

FIG. 7 – *Taux d'erreur pour les différents modèles en fonction du taux d'acceptation global.*

Modèle	Taux d'erreur global	Taux de défaut à 25% d'acceptés	Taux de défaut à 85% d'acceptés	Taux de défaut à90% d'acceptés	Taux de défaut à 95% d'acceptés
$\mathbf{M}_0$ Tout	7,71 %	0,69 %	2,87 %	4,55 %	7,35%
$\mathbf{M}_1$ extrapol	7,45 %	0,90 %	2,83 %	4,55 %	7,55%
$\mathbf{M}_2$ reclass	8,18 %	1,04 %	3,44 %	5,11 %	7,72%
$\mathbf{M}_3$ augment	11,58 %	1,18 %	4,07%	5,42%	7,70%
$\mathbf{M}_4$ parcel	8,01 %	0,90 %	2,91%	4, 90 %	8, 01%
$\mathbf{M}_5$ groupe de contrôle	8,58 %	1,18 %	3,15 %	4,63 %	7,42%

TAB. 8 – *Taux d'erreur global et taux de défaut sur les Acceptés pour différents niveaux de taux d'acceptation, calculés sur les 6 modèles testés. Mesures sur l'ensemble de validation.*

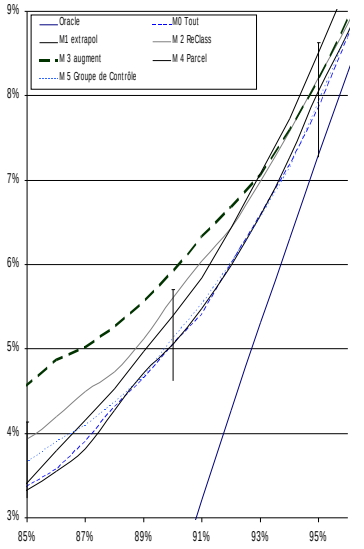

FIG. 8 – *Agrandissement de la figure 7 sur la zone 85-97.*

5.3 Calcul du revenu

En définitive, le critère le plus pertinent pour sélectionner un modèle serait, dans le contexte du score d'octroi, de calculer le revenu global attendu. Le calcul du revenu n'est jamais simple : il faut être capable de calculer exactement le revenu généré par un client, ce qui nécessite sans doute de développer de nouveaux modèles (un client rapporte en général plus que les simples revenus liés au crédit considéré : domiciliation de compte, achat d'autres services et il faut pouvoir estimer ces revenus), mais aussi de réaffecter tous les coûts individuellement sur chaque client, ce qui est beaucoup plus complexe.
Nous nous plaçons ici dans un cadre *très simplifié*, voire caricatural, qui doit simplement servir d'illustration de notre propos : si le client ne fait pas défaut, nous supposons que le gain est

simplement le montant des intérêts perçus par la banque (le montant du crédit multiplié par le taux d'intérêt, fixé à 10% dans les simulations présentées). Si le client fait défaut, le coût est le capital emprunté restant dû, et nous négligeons tous les autres coûts : gestion de dossier, coûts généraux... Notre ensemble de données ne contient en effet pas de données nous permettant d'obtenir une évaluation plus fine du revenu.
Le revenu global dépend donc uniquement du nombre de clients faisant ou non défaut parmi ceux que l'on accepte. Ce critère donnerait des résultats équivalents au calcul de taux d'erreur de la section précédente si tous les clients avaient contracté le même emprunt au même instant. Dans la réalité, le comportement du client est corrélé au type d'emprunt, et le modèle qui minimise le taux d'erreur n'est pas nécessairement celui qui maximise le revenu. Si la méthode utilise un groupe de contrôle, nous évaluons de plus le coût lié à la constitution de ce groupe de contrôle de la même façon : les «mauvais» clients acceptés parce qu'ils font partie du groupe de contrôle entraînent des pertes additionnelles.
La figure 9 présente le revenu obtenu en utilisant les différents modèles en fonction du taux d'Acceptés sur la population. On constate là encore que les différentes méthodes donnent des résultats très semblables, et qu'aucune n'est meilleure dans tous les cas. Il est intéressant de constater que l'utilisation d'un groupe de contrôle peut dans certains cas apporter un gain de revenu, *même si l'on prend en compte les pertes* générées par la constitution de ce groupe : le modèle $\mathbf{M}_5$ est tracé (en pointillés, ligne fine) sans prendre en compte le coût de constitution du groupe de contrôle, puis en prenant en compte ce coût (ligne en dessous) ; il est dans les deux cas très compétitif par rapport aux autres modèles. Le meilleur modèle possible est évidemment «l' oracle». Le modèle M4 (parcelling) est autour de 90% le meilleur modèle (aux barres d'erreur près !)

6 Conclusion

Nous avons présenté une méthodologie de traitement des refusés ainsi que les principaux modèles utilisés classiquement pour le traitement des refusés. Les simulations effectuées permettent de comparer ces modèles dans le cadre d'une expérience plausible, et montrent qu'aucun d'entre eux n'est significativement supérieur aux autres, conclusion qui avait déjà été suggérée par d'autres auteurs dans d'autres contextes (Hand, 2005). Insistons sur l'importance de la méthodologie d'évaluation et de comparaison des modèles. Il est important de choisir un critère de comparaison (AUC, taux d'erreur ou revenu) adapté à l'application, et d'être capable d'évaluer des barres d'erreurs ou d'effectuer des tests d'écart significatif. En effet, la méthode à choisir dépend souvent du point de fonctionnement. Pour mettre en œuvre cette méthodologie, il est indispensable de disposer de la capacité de traiter facilement les données et de construire et évaluer de nombreux modèles statistiques : en effet, les limitations des techniques actuelles de traitement des refusés ne permettent pas d'établir la supériorité absolue d'une technique par rapport aux autres, et donc nécessitent, pour chaque ensemble de données, de rechercher la technique optimale conditionnellement aux données.
Des outils tels que ceux proposés par KXEN sont dans ce cas très précieux. Lorsque c'est possible, nous préconisons l'utilisation d'un groupe de contrôle soigneusement choisi. Les simulations présentées dans cet article montrent que le surcoût lié à ce groupe de contrôle est facilement compensé par les gains de performances apportés.

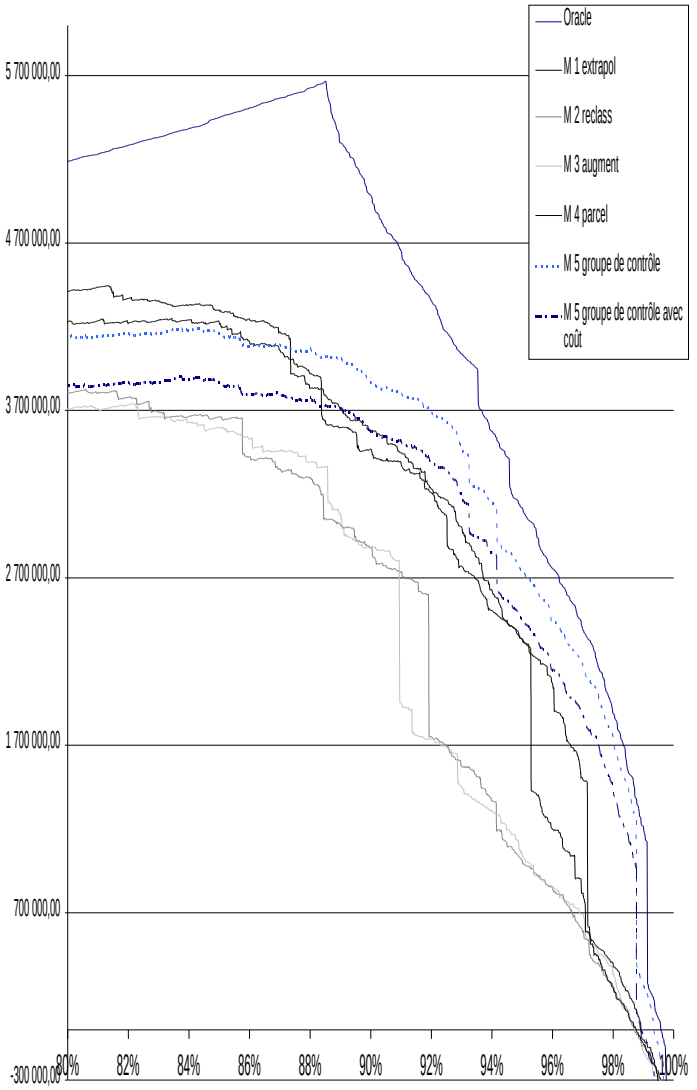

FIG. 9 – *Revenu en fonction du taux d'Acceptés pour différents modèles.*

Nous étudions actuellement comment appliquer à la problématique du traitement des refusés les développements récents de la théorie de l'apprentissage statistique, tels que les approches «transductives» et l'apprentissage semi-supervisé. Des travaux complémentaires seraient sans

doute nécessaires pour affiner les calculs de barres d'erreur sur les indicateurs de performance retenus.

Remerciements

L'un des auteurs (F. Fogelman Soulié) remercie vivement David Hand pour les stimulantes conversations lors de la conférence de Niort-2005 et ses pointeurs vers la littérature, notamment l'article (Hand, 2005). Khosrow Hassibi, Stuart Clarke, Benoît Rognier, de KXEN, ont très largement contribué aux expériences présentées dans cet article : qu'ils en soient ici remerciés. Enfin, les auteurs souhaitent également remercier un des rapporteurs anonymes pour ses commentaires constructifs.

Références

Ash, D., S. Meester (2002) Best practices in reject inferencing. Conference presentation. *Credit Risk Modelling and Decisioning Conference*. Wharton Financial Institutions Center, Philadelphia. http://fic.wharton.upenn.edu/fic/ash.pdf.

Baesens, B., T. VanGestel, S. Viaene, M. Stepanova, J. Suykens, J. Vanthienen (2003) Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society* 0, 1-9.

Bradley, A.P. (1997). The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition* 30 (7). 1145-1159.

Chapelle, O., J. Weston, L. Bottou and V. Vapnik (2000) Vicinal Risk Minimization. *Advances in Neural Information Processing Systems* 13.

Chapelle, O., B. Schölkopf and J. Weston (2003) Semi-Supervised Learning through Principal Directions Estimation. *ICML Workshop, The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*. http://www.kyb.mpg.de/publication.html?user=chapelle.

Cortes, C., Mehryar, M. (2005) Confidence Interval for the Area under the ROC Curve. *Advances in Neural Information Processing Systems (NIPS 2004)*, volume 17, Vancouver, Canada 2005, MIT Press.

Crook, J. , J. Banasik (2004) Does Reject Inference Really Improve the Performance of Application Scoring Models ? *Journal of Banking and Finance* 28, 857-874.
http://www.crc.man.ed.ac.uk/workingpapers/workingpaper02-3.pdf.

Feelders A.J. (2000) Credit scoring and reject inference with mixture models. *International Journal of Intelligent Systems in Accounting, Finance and Management* 9, 1-8.

Feelders, A.J. (2003) An Overview of Model Based Reject Inference for Credit Scoring. *Banff Credit Risk Conference 2003*.
http://www.pims.math.ca/birs/workshops/2003/03w5023/Feelders presentation.pdf.

Hand, D.J. (2001). Reject inference in credit operations : theory and methods. *In The Handbook of Credit Scoring*. E. Mays ed., Glenlake Publishing Company, 225-240.

Hand, D.J. (2005) Good practice in retail credit scorecard assessment. *Journal of the Operational Research Society*, 56, 1109-1117.

Hand, D.J., W.E. Henley (1993). Can reject inference ever work ? *IMA Journal of Mathematics Applied in Business and Industry* 5, 45-55.

Hanley J.A., B.J. McNeil (1982). The meaning and use of the Area under a Receiver Operating Characteristic Curve (ROC). *Radiology*, 143(1), 29-36.

Hsia, D. C., (1978). Credit scoring and the equal credit opportunity act,*The Hastings Law Journal*, 30, November, 371-448

Kraft, H., G. Kroisandt, M. Müller (2004) Redesigning Ratings : Assessing the discriminatory power of credit scores under censoring. http://www.itwm.fhg.de/fm/projects/rating/uKKM.pdf.

Macskassy, S.A., F. Provost, S.Rosset (2005). ROC Confidence Bands : An Empirical Evaluation. *Proceedings of the 22th International Conference on Machine Learning (ICML)*, Bonn, Allemagne, 2005.

Maloof, M.A. (2002). On Machine learning, ROC Analysis, and Statistical Tests of Significance. Proc. 16th Int. Conf. on Pattern Recognition, ICPR02.

Provost F., T. Fawcett, R. Kohavi (1998). The Case Against Accuracy Estimation for Comparing Induction Algorithms. *Proc. 15th Int. Conf. on Machine Learning.* 445-553.

Thomas, L.C., R.W. Oliver, D.J. Hand (2005). A survey of the issues in consumer credit modeling research. *Journal of the Operational Research Society*, 56, 1006-1015.

Vapnik, V.N. (1995). The Nature of Statistical Learning Theory. *Springer Verlag*.

Summary

We present the problem of «Reject Inferencing »for credit acceptance. Models for credit acceptance are based upon credit payments historic data for customers accepted and granted credit. Information on repayment history for rejected customers is thus systematically lacking. As a result, models are built on intrinsically biased samples. In the current legal framework, institutions need to industrialize their processes and reject inferencing is one of those processes.

In this paper, we present a brief review of the techniques currently used today in major institutions (reclassification, parcelling and re-weighting). We show on a specific example that it is possible to rapidly produce results for these various techniques and discuss ways for comparing results and evaluating risk. In particular, we demonstrate the benefit of using a control group. The ability to rapidly produce and compare various techniques thus appears key in the industrialization of the Reject Inferencing process.