Traitement des dossiers refusés dans le processus d'octroi de crédit aux particuliers.

Asma Guizani

L'INSTITUT SUPÉRIEUR DE GESTION (SOUSSE)

LE CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS (PARIS)

ÉCOLE DOCTORALE PARIS "Abbé Grégoire"

ÉCOLE DOCTORALE SOUSSE "École doctorale en sciences de gestion"

Laboratoire Interdisciplinaire de Recherche en Sciences de l'Action (Paris)

Computational Mathematics Laboratory (Monastir)

THÈSE DE DOCTORAT

présentée par : **Asma GUIZANI**

soutenue publiquement à Paris le : **19 Mars 2014**

Pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Discipline/Spécialité : **Théorie des Systèmes Économiques**

et le grade de : **Docteur en Sciences de Gestion**

TRAITEMENT DES DOSSIERS REFUSÉS DANS LE PROCESSUS D'OCTROI DE CRÉDIT AUX PARTICULIERS

THÈSE CO-DIRIGÉE PAR:

Mme. Salwa BENAMMOU	*Professeur, Faculté des Sciences Économiques et de Gestion (Sousse)*
M. Gilbert LAFFOND	*Professeur, Conservatoire National des Arts et Métiers (Paris)*
M. Gilbert SAPORTA	*Professeur, Conservatoire National des Arts et Métiers (Paris)*

RAPPORTEURS :

M. Taher HAMZA	*Professeur, Institut Supérieur de Gestion (Sousse)*
M. Cristian PREDA	*Professeur, École Polytechnique Universitaire de Lille (Lille)*

EXAMINATEURS :

M. Pierre CAZES	*Professeur émérite, Université Paris-Dauphine (Paris)*
Mme. Meglena JELEVA	*Professeur, Economix, Université Paris Ouest-Nanterre-La Défense (Paris)*

À mon très cher papa, à ma très chère maman.

À mon très cher Lamine, à mon petit ange Slim.

À ma très chère tata Hayet.

À toute ma famille...

Remerciements

Ce travail a été réalisé en cotutelle au sein du laboratoire Computational Mathematics Laboratory (CML) à la faculté des sciences de Monastir du côté tunisien et au laboratoire Interdisciplinaire de Recherche en Sciences de l'Action (LIRSA) du côté français.

Je tiens tout d'abord à remercier mon directeur de thèse le professeur Gilbert LAFFOND, pour m'avoir accueillie au sein du laboratoire (LIRSA) et pour m'avoir ainsi permis de réaliser cette thèse.

Je remercie très chaleureusement mon co-directeur de thèse le professeur Gilbert SAPORTA pour m'avoir confié ce sujet de thèse et pour sa gentillesse, sa disponibilité et ses judicieux conseils. Ce fut très sincèrement un réel plaisir de travailler avec lui pendant ces années.

Je tiens à exprimer ma profonde gratitude et ma sincère reconnaissance à mon directeur de thèse le professeur Salwa BENAMMOU pour sa disponibilité, la confiance qu'elle m'a accordée et pour m'avoir souvent donné le courage d'avancer dans mes recherches, en me motivant lorsque j'en éprouvais le besoin et sans qui cette thèse n'aurait jamais pu voir le jour.

Mes vifs remerciements s'adressent aussi au professeur Samir BENAMMOU, directeur du laboratoire Computational Mathematics Laboratory (CML) m'avoir accueillie dans son laboratoire.

Je remercie les professeurs Taher HAMZA de l'Institut Supérieur de Gestion de Sousse et Cristian PREDA de l'École Polytechnique Universitaire de Lille pour avoir consacré du temps à examiner mon travail en tant que rapporteurs.

Je remercie enfin les professeurs Pierre CAZES de l' Université Paris IX Dauphine (Paris) et Meglena JELEVA de l'Université Paris Ouest de Nanterre pour l'intérêt qu'ils ont porté à cette thèse en acceptant de siéger dans ce jury.

Je tiens à remercier M. Sylvain CHAMLEY de nous avoir fourni la base de données sur laquelle nous avons pu effectuer nos applications.

Je remercie vivement Mme Françoise FOGELMAN, M. Abbas MOHSENI, M. Stéphane TUFFÉRY et Mme Mireille BARDOS hélas décédée, pour leurs précieux conseils.

Je tiens à remercier toute l'équipe du CNAM qui a rendu mes journées de travail à Paris plus agréables en m'acceptant parmi eux : Ndeye NIANG, Sabine GLODKOWSKI, Giorgio RUSSOLILLO, Christiane MOREL...

Je remercie, particulièrement, Madame Sonia GHANNOUCHI d'avoir toujours cru en moi et en mes capacités.

Un grand merci à mon amie Sirine et son mari pour leur accueil. Merci à toi Sirine pour les bons moments qu'on a passé ensemble pendant mes séjours à Paris et d'avoir supporté mes états d'âme.

Je tiens aussi à remercier Haifa pour avoir partagé avec moi sa chambre à Paris et pour son accueil.

Je remercie, également, mon amie Besma pour son aide, ses conseils et son soutien moral.

Merci aussi à mon oncle Sadok pour son immense aide, ses conseils et sa patience lors de la relecture des versions intermédiaires de cette thèse..

Merci à toutes mes chères amies pour tous les moments très agréable partagés avec elles : Raoudha , Nabiha, Asma, Olfa, Aida, Samia, Sina, Mariem, Nesrine, Rabia.

J'adresse toute ma gratitude, à mes parents, pour tout ce qu'ils ont fait pour moi. Ma reconnaissance envers eux est inexprimable.

Je tiens aussi à exprimer ma profonde affection à ma belle mère Hayet pour son soutien et pour toute sa patience dont je serai à jamais redevable.

Je remercie énormément mes sœurs Hajer et Chiraz, mon frère Fouad, mes belles sœurs Dalenda et Ibtissem, mes beaux frères Houcem, Fehmi, Jemil, , Zied, Becem, mes neveux et mes nièces, qui ont toujours été là pour moi.

Je ne saurai finir sans remercier mon tendre époux Lamine, pour toute sa patience, son amour et pour tout le soutien qu'il m' a apporté. je voudrai aussi demander pardon à mon petit ange Slim, pour tout ce temps qui aurai du être le sien.

Merci à toutes les personnes qui m’ont permis de mener à bien cette thèse…

Résumé

Le credit scoring est généralement considéré comme une méthode d'évaluation du niveau du risque associé à un dossier de crédit potentiel. Cette méthode implique l'utilisation de différentes techniques statistiques pour aboutir à un modèle de scoring basé sur les caractéristiques du client.

Le modèle de scoring estime le risque de crédit en prévoyant la solvabilité du demandeur de crédit. Les institutions financières utilisent ce modèle pour estimer la probabilité de défaut qui va être utilisée pour affecter chaque client à la catégorie qui lui correspond le mieux: bon payeur ou mauvais payeur. Les seules données disponibles pour construire le modèle de scoring sont les dossiers acceptés dont la variable à prédire est connue. Ce modèle ne tient pas compte des demandeurs de crédit rejetés dès le départ ce qui implique qu'on ne pourra pas estimer leurs probabilités de défaut, ce qui engendre un biais de sélection causé par la non-représentativité de l'échantillon. Nous essayons dans ce travail en utilisant l'inférence des refusés de remédier à ce biais, par la réintégration des dossiers refusés dans le processus d'octroi de crédit. Nous utilisons et comparons différentes méthodes de traitement des refusés classiques et semi supervisées, nous adaptons certaines à notre problème et montrons sur un jeu de données réel, en utilisant les courbes ROC confirmé par simulation, que les méthodes semi-supervisé donnent de bons résultats qui sont meilleurs que ceux des méthodes classiques.

Mots-clés

Réglementation prudentielle, crédit scoring, méthodes d'inférence des refusés, méthodes de l'apprentissage semi-supervisé.

Abstract

Credit scoring is generally considered as a method of evaluation of a risk associated with a potential loan applicant. This method involves the use of different statistical techniques to determine a scoring model. Like any statistical model, scoring model is based on historical data to help predict the creditworthiness of applicants.

Financial institutions use this model to assign each applicant to the appropriate category : Good payer or Bad payer. The only data used to build the scoring model are related to the accepted applicants in which the predicted variable is known. The method has the drawback of not estimating the probability of default for refused applicants which means that the results are biased when the model is build on only the accepted data set. We try, in this work using the reject inference, to solve the problem of selection bias, by reintegrate reject applicants in the process of granting credit. We use and compare different methods of reject inference, classical methods and semi supervised methods, we adapt some of them to our problem and show, on a real dataset, using ROC curves, that the semi-supervised methods give good results and are better than classical methods. We confirmed our results by simulation.

Key words

prudential regulation, credit scoring, reject inference methods, semi-supervised learning methods

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

ACM : Analyse des Correspondances Multiples

AFD : Analyse Factorielle Discriminante

AUC : Area Under the Curve

AUL : Area Under the Lift curve

BM : Variable à prédire Bon / Mauvais

CAH : Classification Ascendente Hierarchique

CCC : Cubic Clustering Criterion

CCCP : Concave Convex Procedure

CET1 : Common Equity Tier 1

CNIL : Commision Nationale Informatique et Libertés

EAD : Exposure At Default

IRB : International Rating Based Approche

KS : Kolmogorov-Smirnov

LCR : Liquidity Coverage Ratio

LGD : Loss Given Default

M : Maturity

MAR : Missing At Random

MCAR : Missing Completely At Random

MNAR : Missing Not At Random

NI : Notation Interne

NIPS : The Neural Information Processing Systems

NSFR : Net Stable Funding Ratio

PA : Perte Attendue

PD : Probability of Default

PI : Perte Inattendue

PLS : Partial Least Squares Regression

PLS-DA : Partial Least Squares Regression Discriminant Analysis

PSCR : Somme Prédite des Carrés Résiduels

RMSSTD : Root Mean Square Standard Deviation

ROC : Receiver Operating Characteristic

RSQ : R-Square

SCR : Somme des Carrés Résiduels

S_e : Sensibilité

S_p : Spécificité

SPRSQ : Semi-Partial R-Square

SVM : Support Vector Machine

TFN : Taux de Faux Négatifs

TFP : Taux de Faux Positifs

TSVM : Transductive Support Vector Machines

TVN : Taux des Vrais Négatifs

TVP : Taux de Vrais Positifs

VA : Valeur Ajoutée

Introduction

La crise financière que connaît le monde actuellement, notamment les faillites successives de grandes banques internationales et les difficultés économiques de certains pays comme la Grèce, l'Espagne, l'Italie ou encore Chypre, a mis en évidence les différentes failles et lacunes de la réglementation prudentielle de Bâle I. Initialement, cette règlementation avait pour mission d'organiser la coopération interbancaire et d'harmoniser internationalement le contrôle prudentiel pour une meilleure pratique bancaire visant à éviter les crises.

À ce titre, la crise des crédits dite des "subprimes" a montré la complexité et la difficulté de la mise en œuvre de cette réglementation et a constitué le phénomène initiateur de la crise financière de 2007 aux USA.

Parmi les failles de la règlementation prudentielle, on peut citer la perte de contrôle du risque dû à sa propagation vers d'autres entités non régulées par la titrisation, ou encore le manque d'intérêt accordé au risque d'illiquidité qui aggrave la situation. La faille majeur des accords de Bâle II consiste dans la manière d'évaluer le risque de crédit qui n'était pas considérée comme règlementation principale.

Face à toutes ces lacunes et à l'ampleur de la crise, les régulateurs se sont trouvés dans l'obligation de faire une révision de l'accord de Bâle II et de proposer des plans de relance pour améliorer la qualité et le niveau des fonds propres des banques dans le but de surmonter les périodes difficiles. D'autres directives ont été prises comme la maitrise de l'effet de levier, l'amélioration de la gestion des liquidités et celle de la gestion et de la surveillance des risques, notamment celle du risque de crédit.

Par ailleurs, et vu la lourde conséquence du risque du crédit, les établissements bancaires considèrent la minimisation de ce risque comme l'une de leurs principales préoccupations. Pour se faire, l'étape d'évaluation de ce risque s'avère cruciale.

Pour l'octroi de crédit, une banque calcule généralement le score de chaque dossier demandeur de crédit et elle décide de refuser ou d'accorder le crédit. Les dossiers acceptés sont suivis par la banque contrairement aux refusés qui sont rejetés initialement.

Paradoxalement, un client accepté par une banque, estimé comme "bon payeur", peut avoir des défaillances et devenir "mauvais payeur" et un client refusé car estimé comme "mauvais payeur" peut être un "bon payeur".

Dans ce cadre, différentes techniques statistiques sont utilisées pour définir un modèle de scoring. Le modèle est calibré à partir des défauts des acceptés et non des refusés, par défaut de suivi, pour ce type de dossiers. Les résultats obtenus sont biaisés du fait que l'échantillon d'apprentissage ne tient pas compte des dossiers refusés, ce qui donne un biais de sélection.

Pour remédier à ce biais, plusieurs méthodes d'inférence des refusés ont été développées, par des chercheurs, tels que "l'augmentation" (Crook & Banasik, 2004), "la reclassification itérative" (Joanes, 1993), le "parceling" (Dempster et al., 1977), "l'extrapolation" (Meester, 1997), etc.

Nous proposons ici d'utiliser des méthodes semi-supervisées comme nouvelles méthodes de réintégration des refusés dans l'échantillon servant au calibrage des modèles de scoring en plus des méthodes classiques de l'inférence des refusés, et ces méthodes sont l'algorithme AdaBoost, l'algorithme Gentle AdaBoost, l'algorithme Logit Boost, le Co-training et l'auto-apprentissage par les SVMs. C'est ce qui diffère notre étude des précédentes recherches menées dans le domaine de l'inférence des refusés.

Nous avons subdivisé notre travail en 3 chapitres. Dans le premier, nous présentons les trois accords de Bâle, nous rappelons le credit scoring, et donnons un panorama des différentes techniques de scoring, ainsi que les mesures de performance, les avantages et les inconvénients de ces techniques de scoring.

Dans le deuxième chapitre, nous donnons des rappels sur le problème du biais de sélection, l'inférence des refusés, les différents types de données manquantes ainsi que les différentes méthodes classiques d'inférence des refusés et l'apport de l'apprentissage semi supervisé dans le traitement des refusés dans le processus d'octroi de crédit .

Dans le dernier chapitre, nous mettons en application les techniques les plus utilisées d'inférence des refusés ainsi que les méthodes de l'apprentissage semi supervisé que nous avons adapté au problème de réintégration des refusés dans le processus d'octroi de crédit sur une base de données réelles que nous a fournie une agence de notation externe concernant 9892 dossiers de crédit à la consommation observés entre les années 2000 et 2001. Ces

dossiers contiennent des dossiers acceptés et refusés. Une discrimination acceptés-refusés, par une analyse discriminante, nous a permis de déterminer la probabilité de refus que nous avons appliqué sur les acceptés uniquement en simulant le processus de refus.

L'échantillon simulé est utilisé pour comparer la performance des différentes méthodes que nous avons présenté et adapté au problème de réintégration des refusés dans le processus d'octroi de crédit et de proposer aux banques une solution au biais de sélection afin de mieux cibler leurs clients potentiels. Nous montrons que les méthodes les plus performantes sont généralement issues de l'apprentissage semi-supervisé.

CHAPITRE 1

Règlementation bancaire et panorama des techniques de scoring

La réglementation prudentielle a pour mission d'assurer la fiabilité et la sécurité du système financier en assurant la diffusion et la promotion de meilleures pratiques bancaires et de surveillance pour éviter les crises. C'est un exercice qui s'avère difficile et complexe à mettre en œuvre car les innovations et les fluctuations conjoncturelles dépassent toujours les normes prudentielles mises en vigueur.

La crise des Subprimes a mis en évidence les différentes failles et lacunes de la réglementation prudentielle de Bâle I et II. Par ailleurs, les régulateurs n'ont pas donné une grande importance quant à la manière d'évaluer le risque de crédit dans l'accord de Bâle II. En conséquence de quoi, cette simple crise de contrepartie, restreinte au niveau de l'Etat américain qui a dégénéré en une crise économique locale, s'est généralisée dans la plupart des pays du monde entier et ne s'est pas à ce jour encore estompée. Plusieurs pays se trouvent actuellement en difficulté, à l'instar de la Grèce, l'Espagne, l'Italie, Chypre, en sus de certains pays du tiers monde.

Face à cette crise, la révision de l'accord de Bâle II qui a donné naissance à Bâle III a retenu les objectifs de l'amélioration de la qualité des fonds propres des banques afin d'augmenter leur capacité à couvrir les pertes, de l'amélioration du niveau des fonds propres, de la maitrise de l'effet de levier, de l'amélioration de la gestion du risque de liquidité, de l'amélioration de la gestion et de la surveillance des risques et surtout le risque de crédit. La maîtrise de ce type de risque constitue l'une des principales préoccupations de la plupart des organismes bancaires. Dès lors, plusieurs banques sont aujourd'hui amenées à intégrer le risque de crédit dans leur gestion afin de le minimiser. Pour cela, toutes les banques doivent être dotées d'un mécanisme d'évaluation du risque de crédit. Ce risque est alors évalué à partir des techniques de credit scoring dont nous allons présenter, dans ce chapitre, les plus utilisées.

I. Le comité de Bâle et ses différents accords

La chute de la banque allemande « Herstatt », en 1974, est considérée, par certains, comme une crise financière en elle même. Elle a poussé les principales puissances économiques mondiales à procéder à une révision profonde du système bancaire et à fonder le "Comité de Bâle"[1] au cours de la même année.

Appelé initialement "Comité Cooke", (du nom de Peter Cooke, directeur de la Banque d'Angleterre qui était le premier à proposer la création de ce Comité et son président), ce comité avait pour objectif d'organiser la coopération et de veiller à l'harmonisation internationale en terme de contrôle prudentiel bancaire. Erigé en institution, le "Comité Cooke" regroupe les représentants des banques centrales et des autorités de régulation financière des principales autorités économiques mondiales. Le Comité a pour rôle de formuler des normes de surveillance et des directives, dans l'espoir de voir les différentes autorités tenir compte des mesures proposées, et de les mettre en application dans leurs propres systèmes nationaux, cependant il n'a aucune autorité directe sur les banques et ses conclusions n'ont pas force de loi.

Selon le rapport[2] de la Banque des règlements internationaux, les missions principales du comité de Bâle sont :

- Le renforcement de la sécurité et de la fiabilité du système financier ;
- L'établissement de standards minimaux en matière de contrôle prudentiel ;
- La diffusion et la promotion de meilleures pratiques bancaires et de surveillance ;
- La promotion de la coopération internationale en matière de contrôle prudentiel.

En 1988, le premier "accord de Bâle I" a été mis en place. Il est basé sur un ensemble de recommandations dont le pivot est la mise en place d'un ratio appelé "ratio Cooke". Ce ratio établit un minimum d'exigence de couverture des risques de crédit par des fonds propres.

[1] Le Comité se compose des représentants des banques centrales et autorités de contrôle de l'Afrique du Sud, l'Allemagne, l'Arabie Saoudite, l'Argentine, l'Australie, la Belgique, le Brésil, le Canada, la Chine, la Corée du sud, l'Espagne, les États-Unis, la France, Hong-Kong, l'Inde, l'Indonésie, l'Italie, le Japon, le Luxembourg, le Mexique, les Pays-Bas, le Royaume-Uni, la Russie, Singapour, la Suède, la Suisse et la Turquie. Le Secrétariat du Comité est sis à la Banque des Règlements Internationaux, à Bâle (Suisse). (Source : Rapport BRI, octobre 2010).

[2] BIS, fiche d'information, "Comité de Bâle sur le contrôle bancaire", 2010.

1. Le premier accord de Bâle I

L'accord de Bâle de 1988 a mis en application le "ratio Cooke", ratio de solvabilité bancaire, qui impose aux banques un niveau d'exigence en fonds propres au moins égal à 8% des risques pondérés (Vo thi, 2004).

L'application de ce ratio a poussé les banques, devenues plus conscientes du risque de crédit, à renforcer leurs exigences en fonds propres. Cette réglementation, qui au départ ne concernait que les pays membres du Comité de Bâle, a été appliquée dans plus de 100 pays.

Malgré l'efficacité de ce ratio qui s'explique par la diminution des faillites bancaires, cette réglementation s'avère imprécise et présente certaines limites.

Les deux failles majeures de ce ratio sont les suivantes (Dumontier et al. 2008) :

- le ratio ne tient compte que du risque de crédit et aucune exigence de fonds propres n'est proposée pour couvrir le risque de marché.
- "*les capitaux propres minimaux sont déterminés par la nature des emprunteurs, et non pas par leurs risques de défaut effectifs*" (Dumontier et al. 2008).

À cause des limites du "ratio Cooke", Bâle I ne parvenait plus à remplir son objectif initial, qui était de consolider la stabilité du système financier. Le Comité de Bâle a donc décidé, en 1998, d'apporter des modifications à la réglementation pour la rendre cohérente avec la pratique réelle des affaires bancaires (internationales) et pour améliorer particulièrement l'ajustement aux risques des exigences en matière de fonds propres, ce qui a donné, plus tard, la réglementation "Bâle II".

2. Le deuxième accord de Bâle II

Le Comité de Bâle a proposé en 2004[3] un nouvel ensemble de recommandations, au terme duquel était définie une mesure plus pertinente du risque de crédit, qui avait été adoptée par les gouverneurs des banques centrales et les superviseurs des pays qui font partie du Comité de Bâle.

[3] Source : Groupe de travail Bâle II : La réforme de Bâle II décembre 2004

Le ratio de solvabilité des établissements bancaires a fait l'objet d'une réforme importante à la fin de l'année 2006. La réglementation de "Bâle II" s'appuie sur les trois piliers (voir annexe 1) décrits par le tableau n°1 :

TABLEAU 1 *: Les trois piliers de Bâle II*

Pilier 1 L'exigence de fonds propres	Pilier 2 Le processus de surveillance prudentielle	Pilier 3 La discipline de marché
Définition des modalités de calcul des exigences en fonds propres nécessaires pour couvrir chacune des catégories de risque : - Le risque de crédit ; - Le risque de marché ; - Le risque opérationnel	Détermination des modalités de surveillance exercée par les autorités de contrôle sur les établissements de crédit : - Contrôler le respect des exigences minimales de fonds propres ; - Contrôler les méthodes d'évaluation et de gestion des risques	Renforcement de la communication à publier en matière de : - Dotation en fonds propres ; - Risques de crédit ; - Risques de marché ; - Risques opérationnels ; - Opérations de titrisations ; - Méthodes d'évaluation et de gestion des risques.

(Source :Dumontier et al. 2008)

- **Les faiblesses de l'accord de Bâle II**

La réglementation prudentielle mise en vigueur par les accords de Bâle II n'a pas permis d'éviter la crise financière de 2007, dite crise des subprimes (Artus, et al. 2008) et en conséquences a présenté plusieurs faiblesses parmi lesquelles nous pouvons citer :

- La nécessité de mise en place d'une surveillance macroprudentielle, alors que l'objectif initial était limité et focalisé sur le niveau microprudentiel;
- Le transfert d'une large part des risques des banques sous forme de titrisation vers d'autres agents, qui ne sont pas forcement soumis aux mêmes exigences réglementaires;
- Les agences de notations externes auxquelles recourent les banques pour évaluer leurs risques, ne sont pas soumises aux exigences règlementaires de Bâle II ce qui peut biaiser leurs estimations du risque;
- Les mesures de risque sur lesquelles se base le comité de Bâle II ne sont pas très fiables et il existe de meilleures indicateurs de risque;
- Le phénomène de procyclicité qui s'explique par un comportement instinctif caractérisé par une exubérance en phase haussière du cycle économique et par une contraction de crédit en phase de ralentissement. Cette contraction conduit à une dégradation majeure de l'activité économique.

La règlementation présente aussi des faiblesses de leur mise en application tels que:

- La mise en place des accords de Bâle II est très complexe et onéreuse;
- Le coût de gestion des risques revient très cher à la banque ;
- L'efficacité des procédures de notation est altéré à long terme par des changements conjoncturels ;
- Les risques opérationnels sont difficiles à évaluer contrairement aux risques de crédit et de marché ;
- La communication d'informations pour une meilleure transparence sur le marché engendre des coûts supplémentaires pour les banques.

Après la crise des Subprimes, quelques mesures ont été prises, comme le renforcement des réserves des banques en liquidités. Ces réserves sont évaluées par des tests de résistance « stress tests » et des exercices de prévention de risque. L'introduction de coussins contracycliques de capital dans les dispositifs de fonds propres et les pratiques de provisionnement conduira la reconstitution de réserves en période d'expansion qui seront utilisées en période de récession (Wellink, 2009).

3. Le troisième accord de Bâle III

La réforme de Bâle II, conçue pour surmonter les lacunes de Bâle I, a nécessité elle-même une mise à jour, après la crise des Subprimes.

En avril 2008, le Comité de Bâle a commencé à prendre des directives pour l'amélioration de certains axes de l'accord de Bâle II, tels que le traitement des dérivés de crédit, la prise en compte de leurs risques de défaut et le traitement du risque de liquidité durable. Les régulateurs ont considérablement amélioré leur manière d'évaluer le risque de crédit, la réglementation imposée aux banques est devenue plus stricte pour l'accord de crédit aux personnes physiques. Les nouvelles recommandations concernent aussi le risque opérationnel, elles tiennent compte de l'intégration des nouvelles technologies et de la technicité croissante des opérations financières(Bouslama et al.2009).

En juillet 2009, le Comité de Bâle s'est intéressé au problème de la titrisation et des activités de marché. En novembre 2010 et après le sommet du G20 à Séoul, de nouveaux accords ont été définis et un nouveau cadre global "Bâle III" a vu le jour (Bouslama et al.2009).

Ce n'est qu'en décembre 2010, que le texte définitif de l'accord de Bâle III a été publié et il n'a été mis en vigueur qu'au cours de l'année 2011 (Bouslama et al.2009).

Pour cette nouvelle réforme, le Comité de Bâle a visé à éviter d'autres crises telle que celle de 2007 et à mettre en œuvre des mesures qui aideront les banques à faire face aux difficultés rencontrées sur le marché. Parmi les objectifs de Bâle III, on peut citer les cinq objectifs décrit, dans le rapport de KPMG intitulé " Bâle III les impacts à anticiper" (Mars 2011)

i. Amélioration de la qualité des fonds propres

Le but des banques est d'augmenter leur capacité à couvrir leurs pertes en améliorant la qualité des fonds propres.

- Augmentation de la part du Common Equity dans le Tier 1[4]: le Common Equity du Tier 1 (CET1) sera exclusivement composé d'actions ordinaires, de réserves et de report à nouveau.
- Harmonisation et simplification de la composition du Tier 2[5] en une seule catégorie au lieu de deux.
- Déductions d'intérêt minoritaires au niveau du CET1.
- Simplification de la composition des fonds propres en excluant progressivement les produits hybrides.
- Disparition petit à petit des fonds surcomplémentaires ou « Tier 3 », qui sont utilisés pour la couverture du risque de marché[6]. ont tendance à disparaitre.

ii. Amélioration du niveau des fonds propres

La crise financière a montré qu'il fallait augmenter le niveau des fonds propres pour faire face à une phase de récession par :

- Augmentation du ratio sur les fonds propres de base durs "Core Tier One " de 2% à 4,5%.

[4] Le Tier 1 est décomposé en deux catégories le Core Tier one (appelé common equity ou fonds propres durs incluant le capital et les réserves) et le Tier one (appelé aussi fonds propres de base incluant le capital, les réserves et certains titres hybrides).

[5] Le Tier 2 désigne les fonds propres complémentaires (plus values latentes, provisions, titres participatifs), il comprend deux catégories : "Upper Tier 2" et "Lower Tier 2".

[6] Source: Banque de France, rapport annuel 2010.

- Prise en considération de matelas de sécurité (coussin de conservation) de 2,5% en 2019.
- Augmentation du ratio sur les fonds propres de base (Tier one) à 7% en 2019.
- Le ratio de solvabilité va passer de 8% à 10,5% (incluant les coussins de conservation).

La mise en place de ces directives se fera progressivement entre 2013 et 2019.

iii. Maitrise de l'effet de levier

En complément au ratio de solvabilité, Bâle III a introduit un nouveau ratio de capital visant à maitriser la croissance des bilans, c'est le ratio de levier qui s'exprime par le rapport des fonds propres de base Tier one au total de bilan et des éléments du hors bilan.

Le Comité de Bâle commence par tester un ratio de levier de 3% du Tier 1 durant une période d'évaluation allant du 1er janvier 2013 au 1er janvier 2018 (Eurogroup Consulting. Avril 2011).

iv. Amélioration de la gestion du risque de liquidité

Pendant la crise, les banques se sont trouvées face à une pénurie de liquidité, ce qui a causé la faillite de plusieurs d'entre elles.

Le communiqué de Bâle III (rapport BIS. octobre 2010) présente les différentes décisions prises par le comité de Bâle pour instaurer des exigences minimales de liquidités à fin de renforcer la capacité des banques à surmonter d'éventuelles situations d'illiquidité.

Ces normes proposent deux ratios à deux horizons temporels différents, le ratio de liquidité à court terme (*Liquidity Coverage Ratio* "LCR") [7]et le ratio structurel de liquidité à long terme *(Net Stable Funding Ratio "*NSFR")[8].

[7] le ratio de liquidité à court terme (Liquidity Coverage Ratio "LCR") qui mesure la situation de liquidité à un horizon de 30 jours. Il a pour but de s'assurer que les banques auront les moyens de surmonter une période de récession brusque pendant 30 jours en mobilisant des actifs liquides qui pourront, au moment de la crise, générer de l'espèce.

[8] Le ratio structurel de liquidité à long terme *(Net Stable Funding Ratio "*NSFR") qui vient compléter le premier. Il est utilisé sur un horizon d'un an et il est conçu pour remédier aux problèmes d'asymétrie de financement et inciter ainsi les banques à s'approvisionner en ressources stables pour financer leurs activités.

v. Gestion et surveillance des risques

L'amélioration de la gestion et de la surveillance des risques vient s'ajouter à ces normes prudentielles exigées par le comité de Bâle.

En juillet 2009, le comité de Bâle (rapport BIS. octobre 2010) a décidé de réviser le second pilier de l'accord de Bâle II pour combler plusieurs insuffisances surtout celles se rapportant à la gestion des risques des établissements bancaires. Parmi ces révisions, nous pouvons citer :

- la gouvernance et la gestion des risques au sein des établissements;
- la prise en compte des risques liés aux expositions du hors-bilan et aux opérations de titrisation ;
- la gestion des concentrations de risque ;
- l'incitations pour les banques à mieux gérer les risques et les rendements sur le long terme; .

En plus de ces nouvelles reformes se rapportant au pilier 2, le comité de Bâle a renforcé ses directives prudentielles dans les autres domaines telles que :

- la gestion du risque de liquidité : l'augmentation des actifs liquides au bilan des banques, à court ou à long terme, est nécessaire pour améliorer l'absorption des pertes face à des situations de difficultés financières.
- la discipline de marché : communication financière plus détaillée concernant les composantes des fonds propres réglementaires et leur rapprochement avec les comptes publiés;
- la titrisation : exigence, pour les banques, d'analyser plus rigoureusement la qualité de crédit de leurs expositions de titrisation notées par un organisme externe, etc.

Face à ces nouvelles réformes de Bâle III, les institutions bancaires se voient attribuer une grande importance à la maitrise du risque de crédit qui constitue l'une de leurs préoccupations majeures. Pour cela, toutes les banques doivent être dotées d'un mécanisme d'évaluation du risque de crédit afin de le minimiser. Ce risque est alors évalué à partir des techniques de credit scoring.

Le credit scoring a donc pour but de prédire la probabilité de défaillance d'un client et en conséquence de réduire le risque d'avoir des engagements non honorés. Comme toute méthode de prévision, le credit scoring s'appuie sur l'historique des résultats de remboursement, des caractéristiques des clients et du type du crédit pour construire une fonction de score qui sera utilisée pour la discrimination (bon payeur / mauvais payeur) des nouveaux demandeurs de crédits.

II. Panorama des techniques de scoring appliquées au "crédit aux particuliers"

Nous présentons ici les techniques de scoring les plus utilisées.

1. Histoire du credit scoring

Le credit scoring consiste à départager au mieux une population donnée en classes différentes. Ce principe de discrimination a été introduit dans les modèles statistiques par Fisher (1936) pour différencier trois types de variétés d'iris. Durand (1941) était le premier à utiliser ces techniques de discrimination pour départager les bons et les mauvais demandeurs de crédit en utilisant certaines caractéristiques de ces derniers. En 1958, le cabinet "Fair et Isaac" fût le pionnier dans l'automatisation des décisions d'accord de crédit, et ce n'est qu'à partir des années 1960 que le traitement de masse des dossiers des demandeurs de crédit est devenu possible.

Le credit scoring fût ensuite appliqué aux entreprises pour prévoir leur défaillance à partir de ratios financiers. Beaver (1968) a utilisé une technique de classification supervisée (l'analyse discriminante univariée) pour prévoir, a priori, les entreprises en difficulté. Altman (1968) s'est aussi intéressé à l'activité d'octroi de crédit aux entreprises. Il a développé une technique d'analyse discriminante linéaire que nous présentons plus loin. Feldman (1997) a expliqué l'intérêt d'utiliser le credit scoring dans le domaine des crédits aux petites entreprises.

À partir des années 1980, le credit scoring a connu un grand essor dans le domaine du crédit à la consommation. L'application des techniques de scoring pour ce type de produits n'a cessé d'augmenter (Malhotra & Malhotra, 2003, Sustersic et al. 2009, etc.).

Nous présentons ci-après trois des modèles les plus courants, parmi les séries de travaux relatifs au credit scoring aux entreprises: le modèle d'Altman, le modèle de Conan et Holder et le modèle établi par la Banque de France.

- **Le modèle d'Altman**

Altman (1968) a développé un modèle de score établi sur la base d'un échantillon de 66 entreprises, dont 33 sont considérées comme défaillantes et 33 comme saines. Nous remarquons que la taille de l'échantillon utilisée dans le modèle d'Altman est faible ce qui peut altérer la qualité des résultats.

La technique statistique adoptée dans ce modèle est celle de l'analyse discriminante multivariée. Elle repose sur une fonction de score combinaison linéaire des cinq ratios financiers jugés les plus pertinents pour départager au mieux les deux groupes d'entreprises (saines ou défaillantes).

La fonction de score d'Altman, couramment nommée Z-score, s'exprime par la relation (1.1) (Voir par exemple Sadi, 2009) :

$$Z = 1.2\,R_1 + 1.4\,R_2 + 3.3\,R_3 + 0.6\,R_4 + 0.9\,R_5 \qquad (1.1)$$

Avec

$$R_1 = \frac{\text{Fond de Roulement Net}}{\text{Actif Total}}$$

$$R_2 = \frac{\text{Bénéfice non réparti}}{\text{Actif Total}}$$

$$R_3 = \frac{\textit{Bénéfice avant intérêts et impôts}}{\text{Actif Total}}$$

$$R_4 = \frac{\text{Capitaux Propres}}{\text{Dettes Totales}}$$

$$R5 = \frac{\text{Chiffre d'Affaire Hors Taxe}}{\text{Actif Total}}$$

Le risque encouru par la banque varie dans le sens contraire de Z, avec 3 comme valeur critique tel que résumé par la figure n°1 :

Figure 1: *Règles de décision pour le modèle d'Altman*

Si le Z- score est supérieur à 3, l'entreprise a peu de risque de faire défaut et s'il se situe entre 2,7 et 3, l'entreprise est à risque. Pour un score compris entre 1,8 et 2,7, la probabilité de faire défaut est importante et l'entreprise est jugée à haut risque. Enfin pour un score inférieur à 1,8 la probabilité d'un problème financier est très élevée (Hull et al. 2007).

- **Le modèle de Conan et Holder**

Ce modèle est basé sur un échantillon de 190 petites et moyennes entreprises industrielles dont 95 considérées comme saines et 95 comme défaillantes. Les deux auteurs ont observé les valeurs de 31 ratios financiers se rapportant à toutes les entreprises de l'échantillon. Ils ont conclu que parmi ces 31 ratios seuls 5 ratios sont les plus significatifs et ont abouti à la formalisation de la fonction score (notée aussi Z) par l'équation (1.2) (Voir par exemple Sadi, 2009) :

$$Z = 0.24\,R_1 + 0.22\,R_2 + 0.16\,R_3 - 0.87\,R_4 - 0.10\,R_5 \qquad (1.2)$$

Où ici $R_1 = \frac{\text{Excédent Brut d'Exploitation}}{\text{Total des Dettes}}$

$R_2 = \frac{\text{Capitaux Permanents}}{\text{Total de l'Actif}}$

$R_3 = \frac{\text{Réalisables et Disponibles}}{\text{Total de l'Actif}}$

$R_4 = \frac{\text{Charges Financières}}{\text{Chiffre d'Affaire Hors Taxe}}$

$R_5 = \frac{\text{Charges du Personnel}}{\text{Valeur Ajoutée}}$

Pour affiner le Z-score et perfectionner la règle de décision des banquiers, Conan et Holder (1979) proposent une probabilité de défaillance selon la valeur du score Z. En fonction de ce score l'entreprise est considérée comme saine ou défaillante.

Le calcul de cette loi de probabilité de défaillance, en plus du score Z, permet de classer les entreprises selon leurs niveaux de risque présumé comme le montre la figure n°2.

(Source : Sadi, 2009)

Figure 2 : *Règles de décision pour le modèle Conan et Holder*

- Si le score Z > 0.10 : l'entreprise a une très bonne situation financière avec un risque de défaillance inférieur à 30%.
- Si 0.04 < Z < 0.10 : l'entreprise est dans une zone d'alerte avec une probabilité de défaillance de 30% à 65%.
- Si -0.05 < Z < 0.04: l'entreprise est dans une zone de danger avec une probabilité de défaillance de 65% à 90%.
- Et finalement, Si Z < -0.05 : l'entreprise est classée comme défaillante avec une probabilité de défaillance qui dépasse 90%.

- **La fonction de score utilisée par la Banque de France**

La Banque de France a élaboré son modèle, en 1985, à partir des documents comptables et financiers de 3000 entreprises industrielles de moins de 500 salariés, qui ont été départagées en trois classes : saines, défaillantes et vulnérables. Les données sont formées de ratios (présentés ci-après). Les entreprises défaillantes ont été observées pendant les trois années précédant leur dépôt de bilan. L'étude a porté sur la période 1975-1980.

La Banque de France a établi une fonction de score (notée Z) qui est donnée par la relation (1.3) (Sadi, 2009) :

$$100\,Z = -85{,}544 - 1{,}255\,R_1 + 2{,}003\,R_2 - 0{,}824\,R_3 + 5{,}221\,R_4 - 0{,}689\,R_5 - 1{,}164\,R_6 + 0{,}706\,R_7 + 1{,}408\,R_8 \quad (1.3)$$

Avec $R_1 = \frac{\text{Frais Financiers}}{\text{Résultat Économique Brut}}$

$R_2 = \frac{\text{Ressources Stables}}{\text{Capitaux Investis}}$

$R_3 = \frac{\text{Capacité d'Autofinancement}}{\text{Endettement Global}}$

$R_4 = \frac{\text{Résultat Économique Brut}}{\text{Chiffre d'Affaire Hors Taxe}}$

$R_5 = \frac{\text{Dettes Commerciales}}{\text{Achats TTC}}$

R_6 : Taux de croissance de la Valeur Ajoutée (VA) (%): $\frac{VA_t - \text{VA}_{t-1}}{\text{VA}_{t-1}}$

$R_7 : délai\ de\ crédit\ clients = \frac{\text{[Stocks de travaux en cours – Avances clients + Créances d'Exploitation]}}{\text{Production (jour)}}$

$R_8 = \frac{\text{Investissements Physiques moyens}}{\text{Valeur Ajoutée}}$

A partir de la valeur de Z calculée, l'exposition au risque de l'entreprise est ainsi déduite :

- Si Z > 0,125, l'entreprise est considérée comme saine
- Si Z < -0,25, l'entreprise risque un dépôt de bilan au cours des exercices prochains.
- Si -0,25 < Z < 0,125, la valeur de Z est insuffisante pour jugersur l'état financier de l'entreprise : c'est une zone d'incertitude.

2. Application du credit scoring aux particuliers

Le credit scoring fût appliqué aux particuliers à partir des années 1980. Actuellement, la plupart des institutions financières sont dotées d'un système de notation interne ou font recours à des organismes de notation externes pour évaluer et prédire si un emprunteur sera bon ou mauvais payeur.

Ces systèmes de notation utilisent des modèles statistiques qui aboutissent à un score d'octroi de crédit à partir d'un seuil fixé par l'institution financière, selon sa tolérance au risque. Ce score est déterminé pour chaque demandeur de crédit et il sera utilisé par les banquiers pour l'aide à la décision.

Nous présentons ci-après les différentes étapes pour déterminer un score de crédit (Saporta, 2006).

2.1. Préparation des données

Le processus de crédit, commence d'abord par une phase de collecte d'informations auprès du client, et auprès de sources externes, afin de former le dossier de crédit. Ces informations portent sur la forme (la"qualité" de l'emprunteur, l'équilibre du montage financier[9], le respect de la réglementation, etc.), le fond (la capacité de remboursement des emprunteurs, les éléments d'appréciation du risque, etc.) et les garanties (cautions, hypothèques, etc.).

2.2. La quantification et la détermination des variables du crédit

Le banquier étudie le dossier du client et prend sa décision d'accorder ou non ce crédit en se basant sur le score calculé. Avant la phase de modélisation, les données collectées sont analysées pour déterminer les variables les plus pertinentes et discriminantes des clients. Notons que la plupart des variables du modèle qui se dégagent du dossier de prêt sont de type qualitatif et ne peuvent être intégrés directement dans un modèle statistique. Pour ce faire, certaines transformations sont nécessaires (en variables dichotomiques, quantification par Analyse des Correspondances Multiples (ACM), etc.).

2.3. La phase de modélisation

Une fois la phase de préparation des données achevée, commence la phase de modélisation qui permet le calcul des scores. Différentes techniques de scoring , qui sont explicitées dans la section 3, pourront être appliquées pour la construction de la fonction de score.

Dans la pratique, le principe le plus simple du scoring, que les banques adoptent dans leur processus d'octroi de crédit, est le suivant : une fois l'étape de collecte de données achevée, des scores partiels sont attribués à chaque variable du dossier et le score final est obtenu par leur sommation. La dernière étape du processus consiste à comparer ce score au seuil fixé par la banque. Si le score dépasse ce seuil alors la banque accorde le crédit sinon la demande est rejetée : c'est le principe de la grille de score, que nous illustrons par l'exemple ci-après:

[9] Le montage financier comprend des éléments essentiels à la réalisation d'un projet comme le plan de financement, le plan de trésorerie et le compte de résultat prévisionnel.

- **Exemple de calcul du score d'un demandeur de crédit immobilier**

Nous supposons que pour un demandeur de crédit immobilier, sept variables sont considérées pour calculer son score (l'âge, l'ancienneté dans la profession, la situation matrimoniale, le nombre d'enfants, l'endettement, l'apport personnel et l'état du bien).

Le tableau n°2 représente la grille de score avec les différents points attribués à chaque variable. A titre d'exemple, le demandeur de crédit est âgé de 55 ans, marié, ayant 4 enfants à charge et qui travaille depuis 6 ans comme professeur universitaire. Il désire acquérir un appartement neuf pour 200 000 € avec une capacité d'autofinancement de 15%, et il n'a jamais contracté de crédit.

TABLEAU 2 : *Un exemple simple de grille de score*

1.Age du client	
Attribut	**Note**
< 20 ans	0
[20, 30 [	0.5
[30, 50 [	5
[50, 60 [	**3**
Plus de 60 ans	1

2.Ancienneté professionnelle	
Attribut	**Note**
≤ 1 an	0
2 à 3 ans	1
3 à 5 ans	3
5 à 7ans	**4**
Plus de 7 ans	5

3.Situation matrimoniale	
Attribut	**Note**
Divorcé	1
Célibataire	2
Marié	**4**
Veuf	3

4.Nombre d'enfants	
Attribut	**Note**
≥ 3 enfants	**1**
2 enfants	3
≤ 1 enfants	5

5.Endettement	
Attribut	**Note**
< 20%	**0**
[20%, 30% [	1
≥ 30%	3

6.Apport personnel	
Attribut	**Note**
< 15%	1
[15%, 40% [	**3**
≥ 40%	5

7.Etat du bien	
Attribut	**Note**
Nécessitant des travaux	1
Ancien en bonne état	3
Neuf	**5**

Le score attribué au dossier de crédit est égal à 23, calculé comme suit :

Score = 3 + 4 + 4 + 1 + 3 + 3 + 5 = 23.

Avec l'hypothèse que, selon la banque, le seuil d'admission d'un dossier est de 20 points, le crédit sera donc accordé au client.

Il est à noter que ce score est utilisé comme aide à la décision d'octroi, et que le banquier pourra ne pas le prendre en considération.

3. Les différentes techniques de discrimination et de prédiction appliquées au scoring

Les méthodes de scoring les plus utilisées par les banques, à cause de leur simplicité d'interprétation et leur grande fiabilité, sont généralement de type linéaire telles que l'analyse discriminante linéaire ou encore la régression logistique.

Il existe d'autres méthodes non linéaires et non paramétriques comme les réseaux de neurones et les arbres de décision qui sont également utilisés dans le domaine du crédit scoring.

On peut aussi citer les systèmes experts qui sont basés sur les règles de décision d'octroi de crédit déduites des caractéristiques du demandeur par les responsables du crédit. Ces règles vont permettre d'identifier et de mesurer le risque de défaut des emprunteurs et elles vont être intégrées dans le système opérationnel de décision.

On peut définir deux types de méthodes de scoring, celles aboutissant à une fonction de score selon laquelle la décision sera prise ou encore les méthodes de décision et de classement directes.

3.1. Les méthodes aboutissant à une fonction de score

3.1.1. L'analyse discriminante

Sur une population de n individus, on observe une variable qualitative Y à k modalités et p variables quantitatives X_i, $i = 1, \dots, p$. La variable Y permet de diviser la population en k groupes disjoints.

L'analyse discriminante permet de mettre en évidence la différence entre classes et de trouver une règle de décision basée sur la connaissances de Y et des X_i , permettant d'affecter un nouvel individu (pour qui Y est inconnue) dans le groupe qui lui est le plus proche. Pour une présentation détaillée de l'analyse discriminante voir par exemple Saporta (2011).

L'analyse discriminante est utilisée dans l'octroi de crédit par les banques en prenant pour Y la variable qualitative ayant pour modalités : bon payeur et mauvais payeur.

3.1.2. La régression PLS (Partial Least Squares Regression)

Tenenhaus (1998) a expliqué que la régression PLS a été proposée par Svante Wold (1984) qui s'est inspiré des premiers travaux de son père Herman Wold (1966) sur l'approche PLS dans les modèles à équations structurelles.

C'est une technique de régression qui consiste à remplacer une matrice de m variables prédictrices X_i (problème de multicolinéarité des variables) à n lignes et m colonnes, par une matrice, tirée de X_i, notée T comprenant les mêmes observations que X_i mais avec un nombre de colonnes (k) inférieur à m. Les colonnes de la matrice T sont des combinaisons linéaires des variables initiales X_i.

$$X_{(n,m)} = \begin{bmatrix} X_{11} & \cdots & X_{1m} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nm} \end{bmatrix} \qquad T_{(n,k)} = \begin{bmatrix} T_{11} & \cdots & T_{1k} \\ \vdots & \ddots & \vdots \\ T_{n1} & \cdots & T_{nk} \end{bmatrix}$$

matrice initiale matrice des composantes T

$$T = \sum_{i=1}^{k} \lambda_i X_i$$

Les X_i correspondent donc aux variables prédictrices, i= {1,…,*p*} et *p* le nombre de variables qui maximisent à la fois la variance de T et la corrélation entre T et Y. Ce qui revient à maximiser le carré de la covariance : c'est le critère d'optimalité de Tucker (voir annexe 2).

$$\text{Max } cov^2(T\ ,Y) = cor^2(\ T\ ,\ Y)\ .\ var\ (T\ )\ .\ var\ (Y) \qquad (1.4)$$

La régression PLS peut être utilisée aussi bien pour expliquer une seule variable et elle est alors notée PLS1 ou bien pour expliquer plusieurs variables auquel cas elle est notée PLS2 (Tenenhaus, 1998).

Pour la discrimination, un cas particulier de régression PLS noté régression PLS discriminante (PLS-DA) est utilisée (voir Saporta, 2009). Dans ce cas la variable dépendante Y est un ensemble de variables binaires décrivant la classe à laquelle appartient une observation à partir d'un ensemble de variables explicatives X. La variable qualitative Y est remplacée par l'ensemble des variables indicatrices de ses modalités. La régression PLS-DA est alors définie comme la régression PLS 2 de Y sur X (des variables numériques ou indicatrices de modalités des variables qualitatives).

3.1.3. La régression logistique

La régression logistique (Tenenhaus, 2007) est un modèle multivarié qui permet d'expliquer, sous forme de probabilité, la relation entre une variable dépendante Y qualitative le plus souvent binaire, $Y \in \{0, 1\}$, et une ou plusieurs variables indépendantes X qui peuvent être quantitatives ou qualitatives.

Le modèle fournit la probabilité qu'un événement se produise ou non (dans notre cas défaut ou non défaut) et les variables indépendantes X sont celles susceptibles d'influencer la survenue ou non de l'événement.

La fonction logistique est la suivante :

$$Y = P(Y = 1) = \beta_0 + \beta_1 . x_1 + \beta_2 . x_2 + \cdots + \beta_n . x_n + \cdots + \beta_k . x_k \quad (1.5)$$

avec

$$Y = Log \left[\frac{p}{1-p}\right] \quad (1.6)$$

et

$$p = \frac{e^{\beta_0 + \sum_{i=1}^{k} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{k} \beta_i x_i}} \quad (1.7)$$

Pour ajuster ce modèle de régression logistique, il suffit de déterminer les coefficients β de la fonction Y, de l'équation (1.5), avec k est nombre de variables exogènes. Pour cela, il suffit d'utiliser la méthode du maximum de vraisemblance qui vise à fournir une estimation, des paramètres, qui maximise la probabilité d'obtenir les valeurs réellement observées sur l'échantillon.

3.2. Les méthodes de décision et de classement directs

3.2.1. Les réseaux de neurones

Ce sont les travaux de Mc Culloch et al. (1943) qui ont montré, pour la première fois, que les réseaux de neurones pouvaient être appliqués dans la résolution des fonctions logiques, arithmétiques et symboliques complexes.

Les réseaux de neurones sont formés d'une couche d'entrées (variables input), d'une couche de sortie (variables output) et d'une ou plusieurs couches cachées de traitements qui forment l'ensemble des nœuds cachés connectés entre eux. Chaque couche prend ses entrées sur les sorties de la précédente. A ce titre, si une couche (i) est composée de *N*(i) neurones, celles-ci prennent leurs entrées sur les neurones de la couche précédente de rang (i-1). Chaque neurone

(ou processus élémentaire) reçoit un nombre variable d'entrée X_i en provenance de neurones en amont.

Le processus de traitement se présente comme suit:

- **Phase 1: Les inputs**

  À chacune des entrées est associé un poids w_i (qu'on appelle poids synaptique) représentatif de la force de connexion (figure n°3). Le neurone ne traite pas chaque information reçue unilatéralement, mais effectue une somme pondérée de toutes les entrées. Cette somme représente la fonction de combinaison suivante :

$$a = \sum_{i=1}^{R} w_i X_i - b \tag{1.8}$$

  avec :

  b : le biais de neurone ou seuil d'activation du neurone.

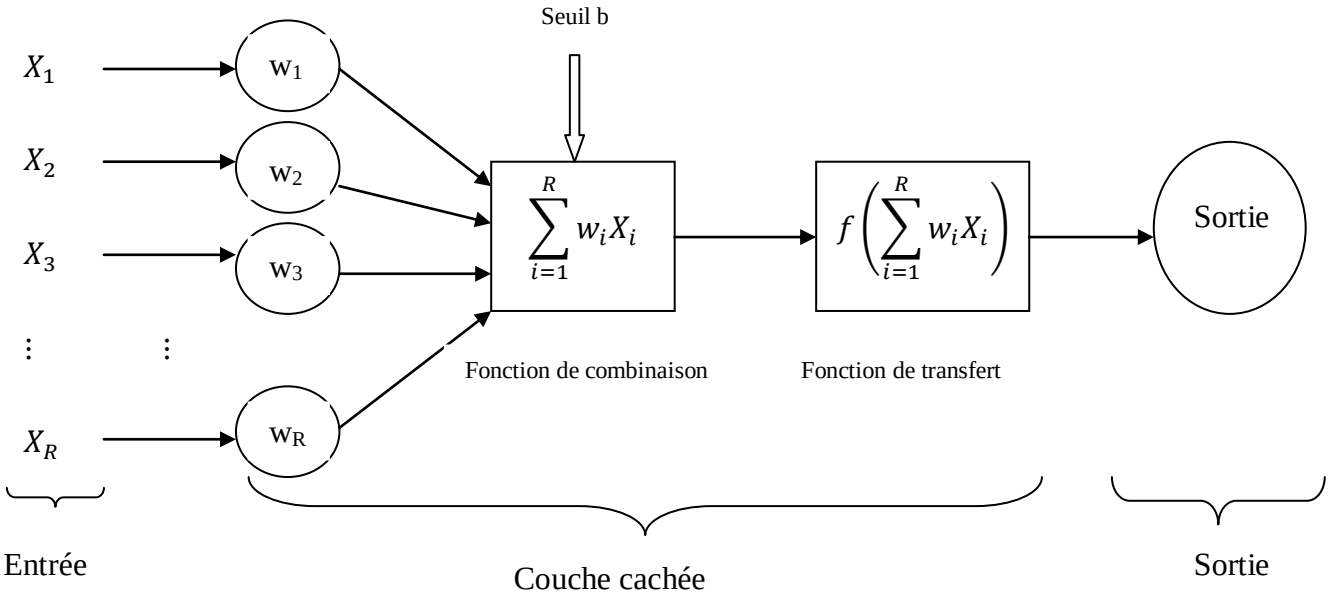
  X_i : l'information qui parvient aux neurones de rang i de la couche d'entrée

  R : le nombre d'informations

  w_i: la pondération du signal émis par le neurone de la couche d'entrée vers la neurone de la couche cachée.

  a: le niveau d'activation du neurone, qui est le signal total reçu par le neurone de la couche cachée.

Un réseau de neurone peut être schématisé par la figure n°3:



(Adapté de : Tufféry, 2012)

Figure 3 : *Conception d'un réseau de neurones*

- **Phase 2 : Les fonctions de transfert**

Afin de déterminer une valeur de sortie, une seconde fonction, Y, appelée fonction de transfert ou d'activation, est appliquée à la valeur a.

$$Y = f(w^t X - b)$$

$$Y = f\left(\sum_{i=1}^{R} w_i X_i - b\right) \qquad (1.9)$$

La fonction de transfert la plus utilisée est la fonction sigmoïde qui est définie par la formule suivante :

$$Y = \frac{1}{1 + \exp(-a)} \qquad (1.10)$$

Avec Y comprise entre 0 et 1 dans ce cas.

- **Phase 3 : L'apprentissage**

L'apprentissage est l'une des propriétés les plus importantes des réseaux de neurones. Elle consiste à développer le réseau de neurones jusqu'à atteindre le comportement désiré. Cette phase s'effectue à partir d'un échantillon de la population étudiée, les entrées X_i de l'échantillon lui permettent d'ajuster le poids w_i des connexions entre les nœuds de façon à améliorer la prédiction par un processus itératif, car il est souvent impossible de fixer à priori la valeur des poids des connexions d'un réseau pour une application donnée. Une fois les poids fixés, la phase d'apprentissage s'achève et commence alors la phase d'utilisation du réseau de neurones.

Les réseaux de neurones, dans le cadre du credit scoring, permettent de mettre en relation les inputs (la base de données qui est composée des dossiers de crédits) et les outputs (le résultat du crédit : bons payeurs ou mauvais payeurs) sans supposer que cette relation est linéaire.

La figure n°4 illustre le traitement des dossiers de crédit par les réseaux de neurones pour étudier la présence ou non du risque de crédit :

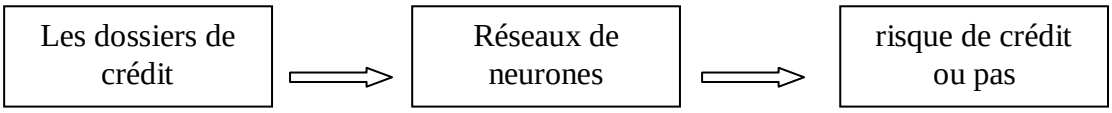


Figure 4: *Application des réseaux de neurones au risque de crédit*

3.2.2. Les arbres de décision

Les arbres de décision ont été développés dans les années 1960 (AID[10] de Morgan et Sonquist) et sont très utilisées en marketing. Au départ, ces méthodes n'étaient pas utilisées par les statisticiens, ce n'est qu'à partir de 1984, qu'ils connaissent un grand essor avec les travaux de Breiman et al. À partir de là, les arbres de décision sont devenus un des outils les plus importants du Data Mining et ceci à cause de la lisibilité des résultats et de la simplicité des interprétations.

Le principe des arbres de décision est de prédire une variable Y quantitative (arbre de régression) ou qualitative, dans notre cas bon ou mauvais payeur, (arbre de décision, de classification, de segmentation) à l'aide de variables explicatives quantitatives ou qualitatives. C'est le principe de l'arbre de décision binaire qui commence par construire l'arbre maximal,

[10] AID : Automatic Interaction Detection

puis, à partir de la phase d'élagage, détermine l'arbre optimal qui sera testé à l'aide d'un échantillon test.

i. La construction de l'arbre[11]

Cette phase consiste à répartir les individus d'une population (échantillon d'apprentissage) en n classes prédéfinies (le plus souvent n=2). Pour ce faire, il faut choisir la variable explicative qui sépare le mieux les individus de chaque classe. La règle de division d'un nœud (segment) dépend de la nature statistique de la variable explicative:

- Si la variable est binaire "B" prenant pou valeurs 0 ou 1 : une seule division est possible
- Si la variable est nominale "N" à "k" modalités : $(2^{k-1}-1)$ divisions sont possibles
- Si la variable est ordinale "O" à "k" modalités : (k-1) divisions sont possibles
- Si la variable est quantitatives "Q" ("q" valeurs distinctes): (q-1) divisions sont possibles

Une fois la règle de division déterminée, elle va être appliquée à l'échantillon d'apprentissage pour une première division en plusieurs segments (nœuds). Cette opération est répétée sur chaque nœud pour augmenter la discrimination, jusqu'à atteindre l'arbre maximal. L'arbre maximal est atteint si :

- la division n'est plus possible.
- il ne reste plus qu'un seul individu dans chaque nœud.
- les individus sont identiques et ne peuvent plus être subdivisés.
- un critère d'arrêt de division de l'arbre est satisfait.

Pour choisir la meilleure séparation d'un nœud, on dispose de plusieurs critères tels que:

- Le critère du Chi-deux, lorsque les variables explicatives sont qualitatives ou discrètes.
- Le critère de Gini, pour tout type de variables explicatives.

[11] Cette partie qui suit est inspirée du papier de SAPORTA, G. Introduction au Data Mining et à l'apprentissage statistique. http://cedric.cnam.fr/~saporta/DM.pdf

- Le critère Twoing, pour tout type de variables explicatives avec une variable dépendante à plusieurs modalités.
- L'entropie ou information, pour tout type de variables explicatives.

ii. La phase d'élagage

Une fois l'arbre maximale obtenu, on procède à une phase d'élagage. Cette phase consiste à tester chaque sous arbre de l'arbre maximal en utilisant un échantillon test distinct de l'échantillon d'apprentissage. On considère comme meilleur élagage, le sous arbre dit "optimal" qui minimise le taux d'erreur déterminé en utilisant l'échantillon test.

Grâce à cette phase, les segments les moins informatifs seront supprimés pour ne garder que les "purs" et qui contiennent des observations toutes identiques. Pour sélectionner l'arbre dit "optimal", on détermine l'ensemble optimal de sous arbres emboîtés { A_{max} -1, ..., A_h, ..., A_1} pour $1 \leq h < max$.

Cet ensemble vérifie l'hypothèse:

$$TEA(A_h) = min_{A \in S_h}\{TEA\ (A)\} \tag{1.11}$$

avec $TEA(A_h)$ est le taux d'erreur d'apprentissage du sous arbre (A_h) et S_h est l'ensemble des sous arbres de A_{max} ayant "h" segments terminaux.

iii. La phase de test

Elle consiste à sélectionner le meilleur sous arbre A* tel que l'erreur de classement associé à A* (ETC) sur l'échantillon test soit la plus petite possible parmi tous les sous arbres et donc vérifie l'hypothèse suivante:

$$ETC\ (A^*) = min_{1 \leq h \leq max}\{ETC\ (A_h)\} \tag{1.12}$$

3.2.3. Les Support Vector Machine (SVM)

Vapnik (1995) fût le pionnier a avoir suggéré la méthode des SVMs pour différentes applications. La technique (SVM) fait partie des techniques de data mining. Elle fait partie des méthodes d'apprentissage qui ont réalisé des performances meilleures que les méthodes statistiques traditionnelles en matière de classification.

Récemment les SVMs sont utilisés dans plusieurs domaines, y compris celui de la reconnaissance des formes (Pontil et al. 1998), la bioinformatique (Yu, et al. 2003), la recherche de l'information, la finance, etc.

L'expérience de l'utilisation de la méthode des SVMs dans le domaine du credit scoring est assez récente et performante comme démontré dans plusieurs articles et travaux (Baesens et al., 2003).

Les SVM peuvent être utilisés dans les cas de séparation linéaire ou non linéaire entre classes.

i. Cas de classification linéaire

Lorsqu'il n'y a que deux classes et qu'on se trouve dans le cas ou les observations sont linéairement séparées, la règle de séparation des SVMs est équivalente à une frontière linéaire. Dans ce cas , les SVMs tentent de départager en deux classes les individus par un hyperplan optimal qui garantie une grande marge de séparation entre ces deux classes.

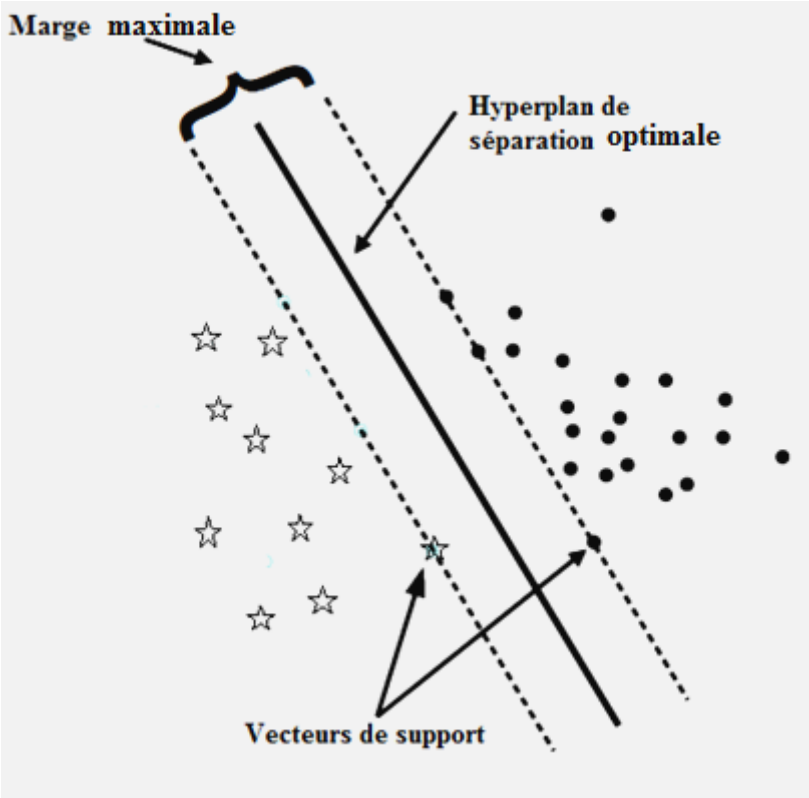


Figure 5 : *Exemple de deux classes linéairement séparables par SVM*

- **Cas linéairement séparable**

L'approche peut être modélisée comme suit : on cherche l'hyperplan d'équation $y_t = \mathbf{w}^T\mathbf{x} + b$ qui maximise la marge entre les deux classes et où $y_t \in \{-1, +1\}$ est l'étiquette de la classe associée à une donnée x_t (avec t=1, ..., N), $\mathbf{x} = (x_1, \dots, x_N)$ est le vecteur de données et $\mathbf{w}$ est le vecteur des poids associé à **x.**

Ainsi, il suffit de trouver $\boldsymbol{w}$ solutions du problème d'optimisation convexe en suivant le procédé ci-après:

La distance d'un point au plan est donnée par :

$$d(x) = \frac{|\mathbf{w}.x+b|}{\|\mathbf{w}\|} \tag{1.13}$$

Maximiser la marge de séparation revient à minimiser $\|w\|$ sous contraintes :

$$\text{Min } \left(\frac{\|\mathbf{w}\|^2}{2}\right), \text{ sous} \tag{1.14}$$

$$y_t(w.\mathbf{x}_t + b) - 1 \geq 0 \; ; \; \forall \, t = 1, \dots, N \tag{1.15}$$

Où $x_t \in R^N$ représentant les N données d'apprentissage et $y_t \in \{-1, +1\}$. L'optimisation se résout par la méthode de Lagrange.

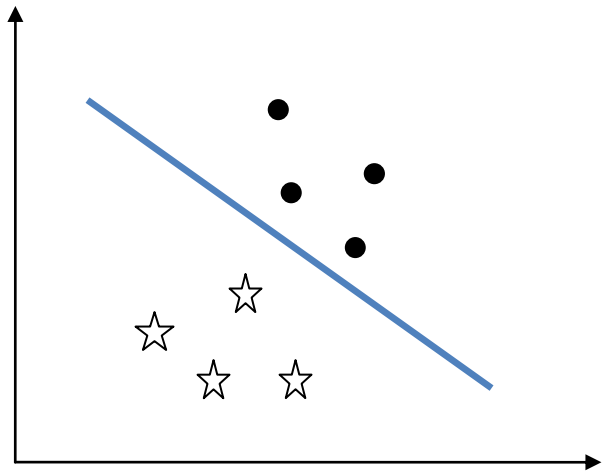


Figure 6 : *Classification linéaire séparable*

La règle de classification se fait selon le signe de $(w.\mathbf{x}_t + b)$ alors :

$$\begin{cases} \text{si } (w.\mathbf{x}_t + b) \geq 0 \text{ alors } y_t = +1 \\ \text{si } (w.\mathbf{x}_t + b) < 0 \text{ } alors \text{ } y_t = -1 \end{cases}$$

- **Cas non linéairement séparable**

Dans le cas où les données ne sont pas linéairement séparables, c'est à dire que les deux groupes à discriminer ne sont pas linéairement séparables, on introduit des variables «ressort », ξ_t, pour assouplir les contraintes :

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{t=1}^{l} \xi_t \qquad (1.16)$$

$$\forall t, y_t(\mathbf{w}.\mathbf{x}_t + b) \geq 1 - \xi_t$$

$$\xi_t \geq 0$$

Où *C* est la constante choisie par l'utilisateur, le problème d'optimisation se résout de la même manière que celui où les données sont linéairement séparables.

Lorsqu'un nouvel élément **x** se présente, il suffit d'étudier la fonction de décision suivante :

$$f(\mathbf{x}) = \sum_{t \in SV} \alpha_t^* y_t \, \mathbf{x}_t . \mathbf{x} + b \qquad (1.17)$$

Où $\alpha_t \geq 0$ sont les multiplicateurs de Lagrange et $\alpha_t^* > 0$ pour le cas linéairement séparable et $0 < \alpha_t^* < C$ pour le cas non séparable.

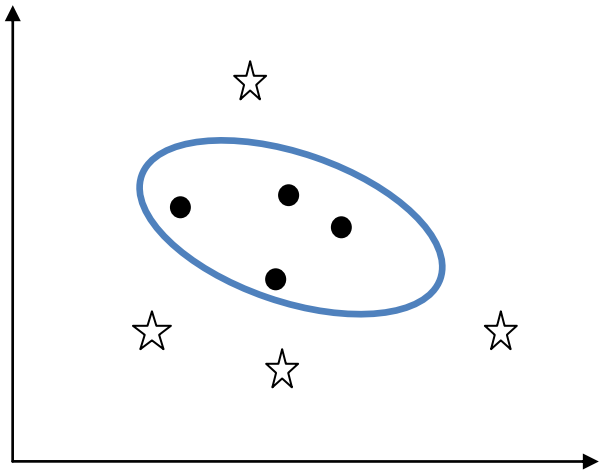


Figure 7 : *Classification non linéairement séparable*

ii. Cas de classification non linéaire

Dans le cas non linéaire, le principe de SVM est de projeter, par une fonction noyau, les données de départ dans un espace de grande dimension (éventuellement infinie), on considère alors : $\Phi: x \to \Phi(x)$, on peut remplacer alors les x par les $\Phi(x)$, ainsi la classification d'un nouvel élément est donnée par la fonction de décision:

$$f(\mathbf{x}) = \sum_{t \in SV} \alpha_t^* y_t \Phi(\mathbf{x}_t) . \Phi(\mathbf{x}) + b$$

$$f(\mathbf{x}) = \sum_{t \in SV} \alpha_t^* y_t K(\mathbf{x}, \mathbf{x}_t) + b \qquad (1.18)$$

Où K c'est la fonction noyau, les plus utilisées sont le noyau polynomial,

$K(x, x_t) = (x . x_t + 1)^d, d \in \mathbb{N}$, et le noyau Gaussien $K(x, x_t) = e^{-\gamma \|\mathbf{x} - \mathbf{x}_t\|^2}, \gamma \in \mathbb{R}^+$.

Le choix du noyau et l'optimisation de ses paramètres dépendent de l'application (Laanaya et al. 2008).

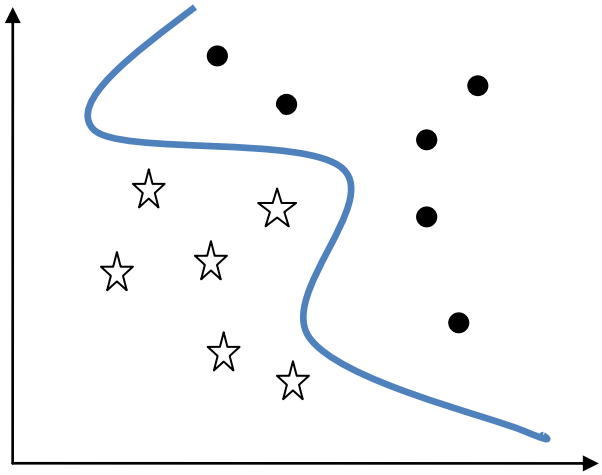


Figure 8 : *Classification non linéaire*

Finalement, on peut dire que les SVMs offrent plusieurs avantages tel que leur capacité à modéliser les phénomènes non linéaires, la précision dans certains cas de leurs prédictions et le fait que l'hyperplan optimal ne soit déterminé que par les points les plus proches (non pas ceux qui sont éloignés), ce qui explique la robustesse de cette technique.

Les SVMs retournent des scores bruts pour chaque observation. Cependant, l'estimation de la probabilité, $p(\mathbf{x})$, a posteriori d'appartenance à une classe à partir des outputs des SVMs est déterminée en utilisant la fonction logistique proposée par Platt (1999). La fonction est la suivante :

$$\mathrm{P(classe/input)} = \mathrm{P(y = 1/\mathbf{x})} = \mathrm{p(\mathbf{x})} = \frac{1}{1+\exp(\mathrm{A.f(\mathbf{x})+B})} \qquad (1.19)$$

où $f(\mathbf{x})$ est la sortie brute des SVMs et A et B des paramètres estimés par le maximum de vraisemblance. Cette transformation est appliquée sur les scores bruts pour chaque observation pour obtenir des scores normalisés comparables à d'autres méthodes de scoring.

3.2.4. Les algorithmes génétiques

En 1970, l'approche de l'algorithme génétique a été développée par l'école de John Holland (Mitchell, 1995). Cette méthode est inspirée de la théorie de l'évolution selon laquelle les individus les mieux adaptés à leur environnement transmettent leurs gènes aux descendants. La méthode essaye de reproduire ce principe, en cherchant les règles les plus appropriées pour avoir la meilleure résolution du problème de prévision. La méthode fait des croisements et des

mutations si nécessaire sur les règles pour aboutir au final à un modèle ayant de meilleur pouvoir prédictif.

Un algorithme génétique se déroule en trois phases :

Phase A. Génération aléatoire des règles initiales

La création des règles initiales se fait de façon aléatoire sous la contrainte d'être différentes et distinctes.

Supposons que nous disposons par exemple des trois variables suivantes:

- L'âge avec les quatre modalités suivantes : [20 - 35 ans], [36 - 50 ans], [51 - 65 ans] et plus de 65 ans.
- Le nombre d'enfant avec les trois modalités suivantes : [0 - 2], [3 - 4] et plus de 4 enfants.
- Le revenu mensuel avec les trois modalités suivantes : [400 - 600 euro[, [600 - 1500 euro] et plus de 1500 euro.

Une règle initiale pourra être alors :

- individu appartenant à la tranche d'âge [36 - 50 ans] et avec un revenu appartenant à [400 - 600 euro[.

Phase B. Sélection des meilleures règles

Une fonction d'adaptation est appliquée, à ces premières règles, dont l'objectif est d'évaluer ces dernières. Cette fonction permet de retenir les meilleures règles qui la maximisent sous la contrainte que chaque règle retenue soit satisfaite par un nombre minimum d'individus.

Si nous voulons évaluer la règle précédente, dans le cadre du scoring, nous allons déterminer, parmi les clients d'une banque âgés entre 36 ans et 50 ans et avec des revenus dans la tranche [400 - 600 euro[, le pourcentage de client à qui la banque à accorder le crédit. Ce pourcentage représente la fonction d'adaptation.

Phase C. Génération de nouvelles règles par les mutations ou les croisement

Les règles retenues de la phase "B" vont ensuite être aléatoirement mutées ou croisées. La mutation consiste à remplacer une variable ou la modalité d'une variable de la règle d'origine par une autre. Si nous revenons à la règle initiale de notre exemple qui était:

- Individu appartenant à la tranche d'âge [36 - 50 ans] et avec un revenu mensuel appartenant à [400 - 600 euro[.

Une mutation de cette règle donnera alors une nouvelle règle :

- Individu appartenant à la tranche d'âge [36 - 50 ans] et ayant 2 enfants.

Où nous avons remplacé le revenu par le nombre d'enfants.

Le croisement de deux règles initiales est l'échange de certaines de leurs variables ou modalités pour donner de nouvelles règles. Si nous prenons comme règles initiales les deux suivantes :

- Individu âgé de plus de 20 ans avec un revenu appartenant à [600 - 1500 euro].
- Individu appartenant à la tranche d'âge [36 - 50 ans], ayant 2 enfants et avec un revenu mensuel de plus de 1500 euro.

Leur croisement pourra donner les deux nouvelles règles suivantes :

- Individu appartenant à la tranche d'âge [20 - 35 ans] et avec un revenu appartenant à [600 - 1500 euro] et ayant 2 enfants.
- Individu appartenant à la tranche d'âge [36 - 50 ans] et avec un revenu mensuel de plus de 1500 euro.

Où nous avons ajouté à la première règle la modalité 2 enfants que nous avons enlevé de la deuxième.

Une nouvelle règle obtenue après mutation ou croisement est appelée règle « fille ».

L'algorithme prend fin lorsque :

- soit un nombre fixé à priori d'itérations a été atteint.
- soit à partir d'une génération donnée, les règles de cette génération et des générations précédentes sont presque identiques.

4. Comparaison des différentes techniques de scoring

Nous présentons (tableau n°3) les avantages et les inconvénients de chaque technique de scoring :

TABLEAU 3 : *Tableau comparatif des techniques de scoring*

Technique de scoring	**Avantages**	**Inconvénients**	**Règle classification**
Analyse discriminante	- Des prédictions explicites. - Un résultat analytique direct. - Des calculs très rapides. - Ne nécessite pas un échantillon de grande taille pour l'apprentissage. - Tient compte des variables qualitatives (procédure DISQUAL)	- Variables explicatives continues et sans valeurs manquantes. - Sensible aux individus hors norme. - Absence de tests statistiques de significativités des coefficients.	Score d'appartenance à une classe.
Régression PLS	- Utilisable en présence de multicolinéarité entre les variables. - Le nombre de variables peut être supérieur au nombre d'observations. - L'algorithme de la régression PLS est rapide. - L'algorithme de la régression PLS est une suite de régressions simples sans inversion, ni diagonalisation de matrices. - La prédiction est meilleure avec une régression PLS sur *p* composantes qu'une régression sur les *p* premières composantes principales. - Efficace sur un grand volume de données. - Possibilité de présence de valeurs manquantes.	- nécessité d'adapter au cas d'une réponse binaire -PLS-DA ou logistique PLS; - Calculs supplémentaires (bootstrap, validation croisée) pour obtenir des erreurs standards sur les coefficients.	Score d'appartenance à une classe
Régression logistique	- Variables explicatives discrètes, qualitatives ou continues. - Variables à expliquer ordinale ou nominale. - Pas d'hypothèses de multinormalités, ni d'homoscédasticités pour les variables explicatives. - possibilité de prise en compte les interactions entre variables.	- Les variables explicatives doivent être non colinéaires. - Calcul itératif plus long qu'une analyse discriminante de Fisher. - La précision est moindre que celle de l'analyse discriminante. - La régression logistique ne converge pas toujours vers une solution optimale.	Probabilité que l'évènement de défaut se produise

	- Résultats faciles à interpréter.	- Ne traite pas les valeurs manquantes. - Sensible aux valeurs hors norme.	
Réseaux de neurones	- Modéliser des relations non linéaires entre les données. - Modéliser des problèmes de différents types. - Résiste aux données défectueuses.	- Les résultats ne sont pas explicites et sont difficile à comprendre par les utilisateurs. - Le risque de sur-apprentissage. - Ne traite pas un grand nombre de variables. - la convergence vers la meilleure solution globale n'est pas toujours garantie.	Affecter l'appartenance des individus aux classes définies.
Arbres de décision	- Les résultats sont exprimés sous forme de condition explicites sur les variables d'origine. - Compréhensibilité des résultats pour les utilisateurs. - Les variables explicatives peuvent ne pas suivre des lois probabilistes particulières. - Les arbres ne sont pas affectés par les individus hors norme. - Traite les données manquantes. - Tous types de variables : continues, discrètes et qualitatives. - Simple à utiliser.	- La détermination des nœuds du niveau (n+1) dépend fortement du nœud précédent (n). - L'apprentissage d'un arbre de décision nécessite un nombre assez grand d'individus. - Le score d'un individu dépend de la feuille à laquelle le conduisent les valeurs de ses prédicteurs.	Associer une observation à l'attribut attaché à la feuille à la quelle il appartient.
Support Vector Machine	- Capacités à modéliser les phénomènes non linéaires. - Précision de prédictions dans certains cas.	- Résultats non explicites. - Difficulté des choix des paramètres. - Temps de calcul longs. - Risque de sur-apprentissage. - Programmable sur peu de logiciels.	Estimation d'une frontière de classification et l'affectation d'un individu à une classe se fait par rapport à sa position à cette frontière.
Algorithmes génétiques	- Améliorer la performance de certains outils de prédiction comme les réseaux de neurones.	- Algorithme assez lent. - La complexité de cet algorithme augmente de manière exponentielle en fonction du nombre de règles utilisées. - Utilisable sur un volume de données assez faibles. - Réglage délicat. - Peu répondu dans les logiciels.	Trouver des règles de prédictions intéressantes pour les appliquées aux individus et trouver l'attribut qui leur correspond.

(Adapté de : Tufféry, 2012)

5. Les mesures de performance

La performance d'un modèle est mesurée par son pouvoir discriminant et sa capacité d'estimer la probabilité de défaut. Plusieurs mesures sont proposées dans la littérature, on peut en distinguer deux types :

- Les mesures adaptées à des scores
- Les mesures adaptées à une décision

5.1. Les mesures adaptées à des scores

Ce sont des méthodes qui mesurent le pouvoir prédictif du score établi. Nous pouvons citer l'exemple de la statistique de "Kolmogorov-Smirnov" (KS) (Anderson, 2007).
La statistique"KS" est un test non paramétrique, basé sur la comparaison de la fonction de distribution cumulée des emprunteurs qui ont fait défaut (mauvais payeur F(M)) celle des emprunteurs qui n'ont pas fait défaut (bon payeur F(B)) . Plus la distance est grande, en valeur absolue, entre les deux distributions de score, meilleure est la capacité du modèle à différencier les deux catégories d'emprunteurs.

La figure n°9 illustre le test de Kolmogorov-Smirnov:

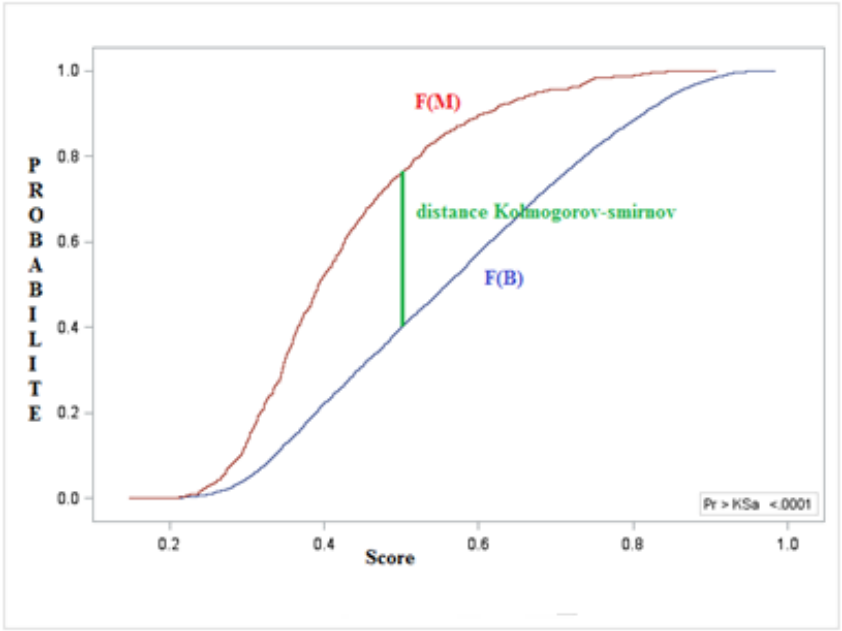


Figure 9 : *La statistique de Kolmogorov-Smirnov*

5.2. Les mesures adaptées à une décision

Ces mesures ont aussi pour objectif de tester le pouvoir discriminant d'un modèle comme la statistique "KS", sauf qu'au lieu de s'intéresser aux scores, elles évaluent la qualité de la règle de décision établie par le modèle et sont donc préférables.

5.2.1. Le taux de bons classements

L'évaluation de la règle de décision , pour cette méthode, repose souvent sur l'estimation du taux d'erreur. Le tableau n°4, nommé table d'affectation ou matrice de confusion, présente les différentes possibilités d'affectation d'une observation à une classe par un modèle, ce qui permet d'estimer le taux de bons classements et le taux d'erreur.

TABLEAU 4 : *Table d'affectation*

	Affectation	
Situation réelle	Mauvais payeur	Bon payeur
Mauvais payeur	*TVN*	*TFP*
Bon payeur	*TFN*	*TVP*

avec :

- *TVP* : Taux de vrais positifs (bons dossiers classés comme tels).
- *TFP* : Taux de faux positifs (mauvais dossiers classés comme bons).
- *TVN* : Taux de vrais négatifs (mauvais dossiers classés comme tels).
- *TFN* : Taux de faux négatifs (bons dossiers classés comme mauvais).

La figure n°10 présente les différents cas qui forment la matrice de confusion en fixant un seuil de score. Habituellement le seuil selon le quel le dossier de crédit est affecté à l'une de ces parties, est égal à 0.5.

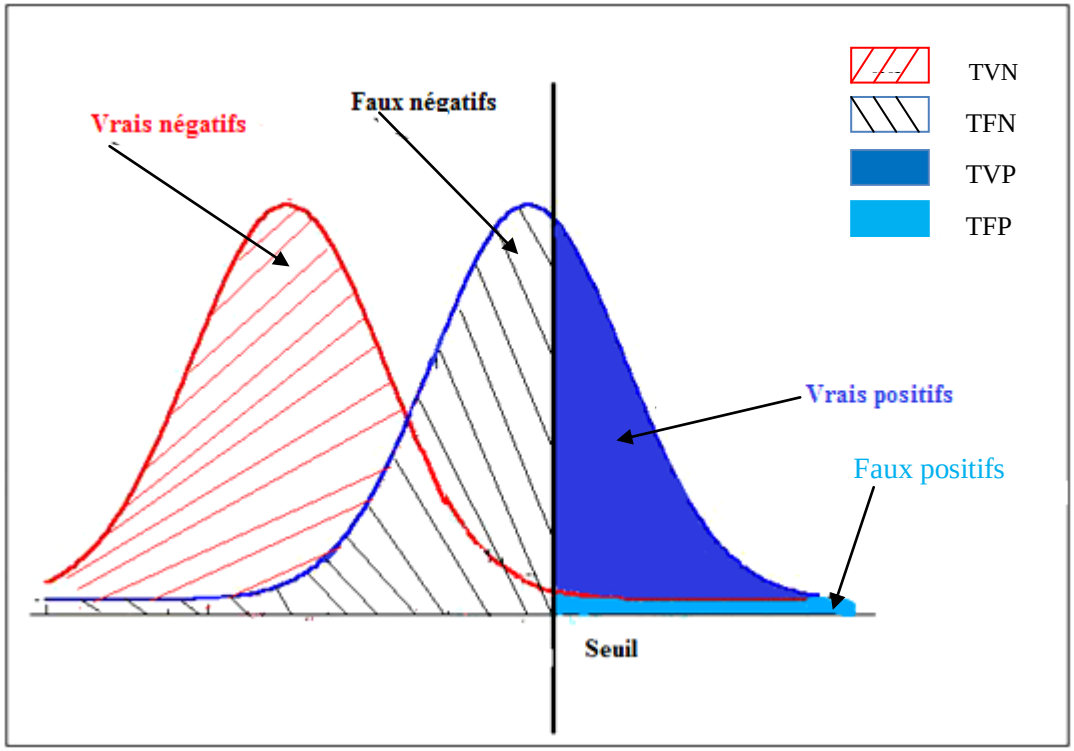


Figure 10 : *Les différents cas d'affectation d'un dossier de crédit*

À toute observation correspond deux groupes : celui auquel elle appartient réellement et celui auquel elle est affectée. Le décompte des affectations correctes, c'est-à-dire les observations affectées à leur groupe réel, fournit une estimation des taux réels de bon classement.

- le taux de bon classement réel estimé pour les Mauvais payeurs, noté t_D, est donné par la relation (1.20) :

$$t_D = \frac{TVN}{TVN + TFP} \tag{1.20}$$

- le taux de bon classement réel estimé pour les Bons payeurs, noté t_{ND}, est donné par la relation (1.21) :

$$t_{ND} = \frac{TVP}{TFN + TVP} \tag{1.21}$$

- le taux de bon classement réel général estimé sur l'échantillon, noté t, est donné par la relation (1.22):

$$t = \frac{TVP + TVN}{TVP + TVN + TFN + TFP} \tag{1.22}$$

On note que :

- **l'erreur de type 1** : représente les dossiers de crédit réellement mauvais, déclarés par le modèle comme Bons (risque de non remboursement du crédit).
- **l'erreur de type 2** : représente les dossiers de crédit réellement bons, qualifiés par le modèle de mauvais payeurs (manque à gagner).

Chaque erreur de classement engendre un coût supplémentaire relatif à la mauvaise décision que l'on aurait prise. Il s'agit alors de minimiser ce coût causé par le classement incorrect des éléments.

Nous pouvons conclure que la matrice de confusion permet de déterminer le pourcentage d'observations bien classées pour un seuil de classement donné. Dans certains cas, on dispose d'un échantillon avec des classes très déséquilibrées (par exemple plus de bons payeurs que de mauvais), donc la matrice de confusion peut donner une idée erronée sur le pouvoir prédictif du modèle. De plus la matrice de confusion donne directement la classe à laquelle une observation est affectée pour un seuil donné donc pour comparer deux classifieurs, il faut se fixer le même seuil pour les deux cas afin de pouvoir trancher. On peut dire qu'il y a une perte d'information car la comparaison donnera un résultat bien déterminé soit que l'un est plus performant que l'autre soit que les deux ont la même performance.

La courbe Receiver Operating Characteristic (ROC) qui donne la représentation graphique de toutes les matrices de confusion pour différents seuils vient remédier à ce problème. La comparaison, de deux classifieurs dans ce cas, donnera un résultat de la forme parfois meilleur parfois moins bonne. Donc la courbe ROC compense le problème de perte d'information et elle est bien adaptée à notre cas. L'idée générale de la courbe ROC est d'évaluer la performance du modèle pour tous les seuils de classement possibles (le seuil peut aller de 0 à 1 si le score est une probabilité sinon il faut faire varier le score de son maximum à son minimum). pour chaque seuil de classement, il faut calculer le taux de vrai positifs (TVP : Bons dossiers classés comme tels) et le taux de faux positifs (TFP : mauvais dossiers classés comme bons) que l'on reporte sur un graphique ayant en abscisse le TFP et en ordonnée le TVP.

5.2.2. La Courbe Receiver Operating Characteristic (ROC)

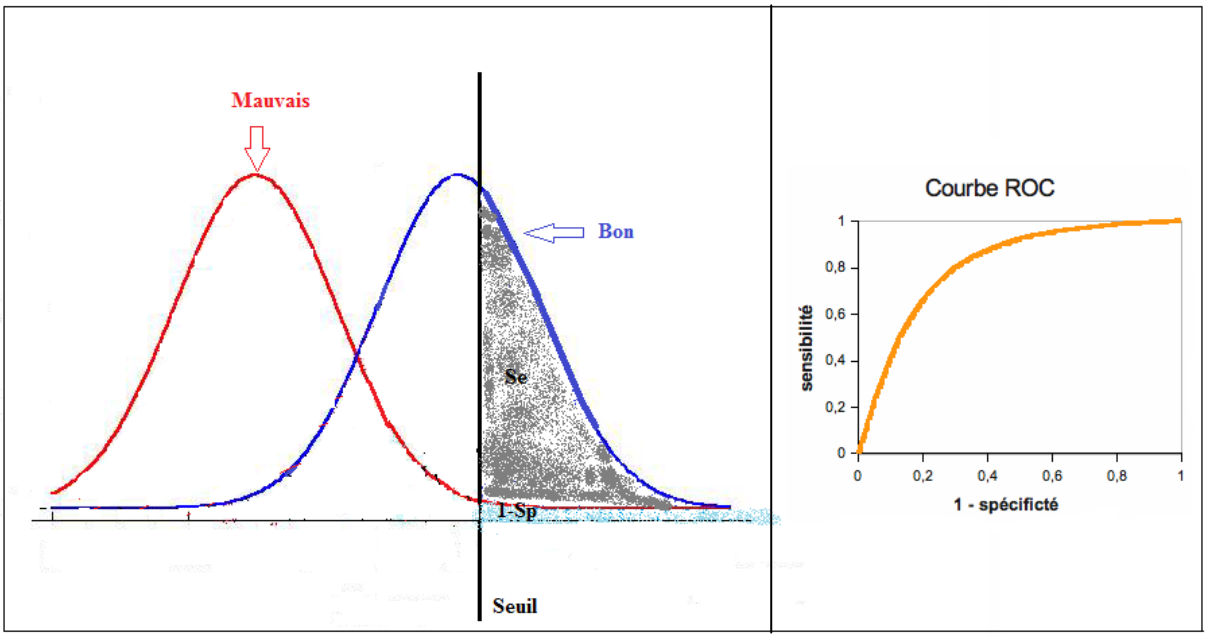
La courbe ROC est une représentation graphique des performances des classifieurs à deux classes. Elle a longtemps été utilisée dans la théorie de détection du signal (Swets et al. 2000), dans l'analyse du comportement du système de diagnostic (Swets, 1988), dans la médecine (Zou, 2002 et Swets et al. 2000). Parmi les pionniers qui l'ont utilisé dans le domaine du "machine learning" on peut cité Spackman en 1989 et ce pour l'évaluation et la comparaison des différents algorithmes dans ce domaine. L'utilisation des courbes ROC est toujours d'actualité. En plus du domaine, du "machine learning", cette méthode a été utilisée pour des travaux d'analyse de la performance, Metz 1978, Provost et Fawcett en 1997, Hand en 1997.

En matière de crédit bancaire, cette courbe relie la proportion de vrais positifs (bon dossier classés comme tels) à la proportion de faux positifs (mauvais dossiers classés comme bons) lorsqu'on fait varier le seuil du score d'acceptation du dossier.

La proportion de vrais positifs, appelée aussi sensibilité et notée S_e est définie comme étant le pourcentage des cas positifs correctement identifiés alors que la proportion des vrais négatifs appelée aussi spécificité et notée S_p est définie comme les cas négatifs correctement identifiés. Compte tenu des ces définitions les expressions de S_e et S_p se présentent comme suit:

$$S_p = \frac{TVN}{TFP + TVN} \text{ et } S_e = \frac{TVP}{TVP + TFN}$$

La figure n°11 montre les parties correspondantes à la sensibilité et la spécificité:

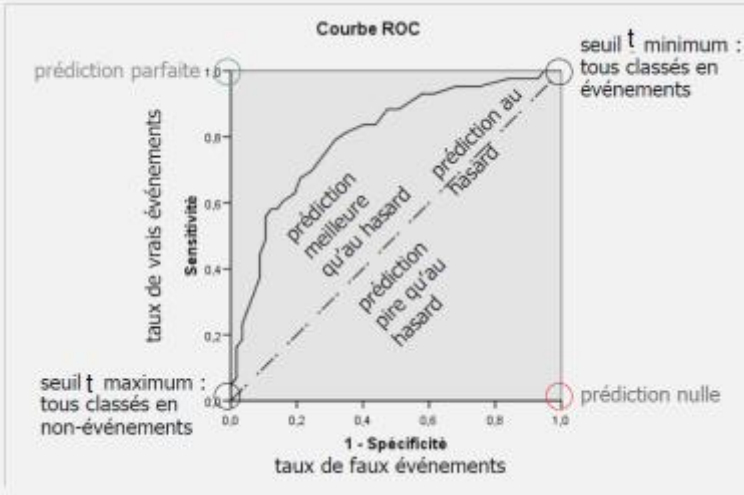


*(*Adapté de*: Saporta, 2009)*

Figure 11 : *Présentation de la sensibilité et la spécificité*

La règle de classification consiste à affecter, pour un seuil de classification donné t, un individu avec un score s > t, à la classe 1, et à la classe 0 si s < t.

La courbe ROC permet de définir la valeur seuil en représentant la sensibilité S_e en ordonné en fonction de (1- S_p) en abscisse. Si on note par classe 0 les vrais positifs et par classe 1 les faux positifs, la courbe ROc est alors représentée dans un plan dont l'axe vertical noté $F_0(t)$ correspond à la proportion des individus de la classe 0 qui sont bien classés. L'axe horizontal noté $F_1(t)$ correspond à la proportion des individus de la classe 1 qui sont mal classés. La courbe passe par le point (0,0) au point (1,1) quand t décroît, la courbe ROC peut être présentée par la figure n°12 :



(Source: Tufféry, 2012)

Figure 12 : *Lecture d'une courbe ROC*

La représentation graphique se fait dans un carré de côté égal. La courbe ROC passe par (0,0) et (1,1). Plus cette courbe s'éloigne de la diagonale par le haut meilleure est la règle de classification. Si cette courbe s'approche trop ou coïncide avec la diagonale la règle de classification n'est pas meilleure qu'une classification aléatoire. Si la courbe se trouve sous la diagonale la règle de classification est jugée mauvaise.

L'étude de la performance des modèles se fait en comparant leurs indices AUC, les aires sous leurs courbes ROC (Area Under the curve : AUC). Un modèle est d'autant plus performant que son AUC est plus proche de 1, et il est d'autant moins performant que son AUC est proche de zéro. Dans le cas où l'AUC est égale à 0,5, le modèle est considéré comme inintéressant.

Parmi les inconvénients des courbes ROC est que pour pouvoir comparer différents modèles et pour que l'utilisation de l'indice AUC soit intéressante, il faut que les courbes ROC soient estimées sur un échantillon test non pas sur l'échantillon d'apprentissage car cela risque de donner des résultats trop optimistes. De plus, l'AUC se trouve dans l'incapacité de trancher entre différents classifieurs, concernant leur performance de prédiction, si les courbes ROC de ces derniers se croisent.

Au final, l'AUC est considéré comme l'indice approprié pour évaluer la performance du pouvoir discriminant de plusieurs modèles lorsque les courbes ne sont pas de formes équivalentes et ne présentent pas des zones de croisement.

L'indice de Gini, mesure classique de l'inégalité, est une simple transformation de l'AUC donnée par la relation (1.23) :

$$Gini = 2 * AUC - 1 \tag{1.23}$$

Cet indice représente le double de la surface entre la courbe ROC et la diagonale, donc plus cet indice est élevé, meilleur est le modèle.

Bien que l'indice AUC soit considéré comme un outil très intéressant pour la comparaison entre différentes règles de classification et l'étude de la performance des classifieurs, il présente un problème majeur faisant partie intégrante de sa définition. Ce problème a été explicité par Hand (Hand, 2009 et Hand, 2010) qui a supposé que pour les deux types d'erreur de classification (classer des objets de classe 0 en classe 1,ou bien classer des objets de classe 1 en classe 0) il existe deux types de coûts de mauvais classement. Pour une paire de coûts d'erreur de classement donnée, on cherche à choisir, le seuil de classement t, de façon à minimiser la perte globale.

Hand a démontré que l'indice AUC est équivalent à l'espérance de la perte due à l'erreur de classement. Cette perte est calculée sur la base d'une distribution des coûts d'erreurs de classement qui dépendent des distributions empiriques des scores. Ces distributions dépendent du classifieur utilisé. Donc pour évaluer la performance d'un classifieur, l'indice AUC utilise différentes métriques qui dépendent du classifieur qu'on est en train d'évaluer.

On peut donner un exemple qui explique très bien ce problème : dans le cadre du credit scoring, le fait de classer un client comme mauvais payeur au lieu de bon payeur est dix fois plus risqué si l'outil de classification est la régression logistique, alors qu'il sera cent fois plus risqué si on utilise un arbre de décision. Hand (2010) déclare que" *la gravité relative des mauvais classements ne peut dépendre des outils choisis pour construire la discrimination*." On peut conclure alors que la dépendance, de la distribution, des coûts d'erreurs à la distribution empirique des scores implique des choix sous optimaux de règles de discrimination qui conduiront à des erreurs de classement qui auraient pu être évitées. Les conséquences peuvent être très graves et coûter très cher dans certains domaines.

Flach et al. (2011) trouvent que les critiques de Hand (2009 et 2010) concernant l'AUC sont dévastatrices. Il ont donné une nouvelle alternative pour l'interprétation de l'AUC comme étant une association de la performance de classement pour tous les seuils de classification

possibles et les différents paramètres de coûts. Une fois cette interprétation est mise en place, Flach et al. démontrent qu'on peut dès lors comparer facilement les performances de différents outils de discrimination.

6. Les avantages et les inconvénients du credit scoring

Comme cité par Schreiner (2003), le credit scoring possède certains avantages et certains inconvénients que nous essayons de résumer ci-après:

6.1. Les avantages du credit scoring

6.1.1. Cohérence de l'évaluation statistique

Toutes les techniques de scoring, sans exception, traitent de façon identique toutes les candidatures potentielles. Deux personnes présentant les mêmes caractéristiques auront le même niveau de risque estimé.

Les pondérations des caractéristiques d'un client, dans une fonction score, sont basées sur des données archivées sans l'influence de jugements subjectifs. Dans un système d'évaluation subjective, les variations des jugements des agents responsables du crédit sont possibles et parfois même dépendantes de leur humeur du moment. Cette évaluation subjective dépend d'un procédé flou que même les agents de crédit auront du mal à justifier. Alors que ce problème ne se pose pas dans l'évaluation statistique, des risques, qui est connue et qui peut être communiquée.

6.1.2. Amélioration de la productivité et de la rentabilité de la banque

Le rejet, dès le départ, de certains clients jugés à risque permet de ne pas perdre de temps sur le recouvrement de créances des clients en difficulté et libère du temps pour étendre les prestations de la banque auprès d'un plus grand nombre de clients jugés bons payeurs.

D'autres parts, les techniques de scoring permettent une meilleure gestion du risque, traitent davantage de prêts toutes choses égales par ailleurs et donnent une couverture plus vaste, ce qui permet aux banques de bénéficier d'une plus grande part de marché et d'augmenter leurs gains.

6.1.3. Quantification du risque de crédit

Les responsables de prêts ont besoin d'énormément de temps et d'information pour être en mesure d'utiliser des paramètres et des politiques de notation subjectives et pouvoir estimer le risque. Par contre, la notation statistique de crédit prévoit le risque sur la base des caractéristiques quantifiées enregistrées dans une base de données. Les relations entre le risque et les caractéristiques du dossier de crédit de la clientèle sont exprimées sous la forme d'ensembles de règles, dans une formule mathématique estimant le risque comme une probabilité, et permettant une gestion plus aisée du risque.

6.1.4. Facilité de gestion des portefeuilles

La notation constitue également un outil puissant, susceptible de permettre aux responsables d'ajuster la composition et les orientations des portefeuilles en fonction de l'activité économique, compte tenu d'un niveau souhaité de tolérance au risque.

6.2. Les inconvénients du credit scoring

Les techniques du credit scoring présentent un certain nombre de faiblesses. En effet, l'avenir peut être différent du passé, alors que les prédictions fondées sur la notation s'appuient uniquement sur les événements passés. Les changements pouvant survenir sont susceptibles d'affecter la qualité de prédiction du modèle.

6.2.1. Les problèmes pratiques

i. Coût onéreux de mise en place

La collecte et le traitement des données nécessaires au calcul d'un score entraînent des surcoûts qui peuvent être importants pour les institutions financières les moins pourvues en moyens humains et matériels. Cette charge provient des coûts engendrés par la collecte d'un grand nombre de données de bonne qualité, des coûts de traitement automatisé, de stockage et de la maintenance du système d'informations nécessaires, et enfin des coûts relatifs à la formation des responsables de prêts, aux techniques de collecte des données supplémentaires, qui s'avèrent nécessaires.

À ce titre le projet de notation induit, lui-même, un coût important lié au développement du score, à l'intégration au système d'information, à l'ajustement du système d'information, à la formation des utilisateurs finaux et au suivi. En particulier, l'ajustement du système d'information, pour le calcul et l'édition automatique des

prévisions du risque, peut s'avérer extrêmement longue et difficile et peut représenter une part considérable du budget prévu.

iii. Complexité de mise en place

La mise en place d'un système de scoring est souvent techniquement complexe. L'évaluation statistique nécessite du personnel spécialiste et expérimenté. De plus, ce personnel doit suivre des formations concernant ces nouvelles techniques pour se familiariser avec les résultats et pouvoir les interpréter.

iv. L'accès aux données pertinentes

La procédure d'archivage des dossiers acceptés et refusés est rarement informatisée, souvent les dossiers refusés, jugés inutiles, ne sont même pas conservés. Il est indispensable de prévoir une phase d'inventaire, de codification et de saisie de tous les dossiers. Il faut aussi mettre en place une phase de mise à jour des données pour l'alimentation de la base de données qui pourra servir pour des futures études de scoring.

v. Aspect règlementaire

Les réglementations prudentielles mises en vigueur par le comité de Bâle n'avaient pas force de loi, donc certains établissements bancaires et agences de notation ne tiennent pas compte des mesures proposées.

La crise des Subprimes illustre bel est fort ce problème. En effet, la sous évaluation des risques des crédits des « Subprimes », sortaient des bilans des banques via la titrisation, vers d'autres établissements financiers qui n'étaient forcement pas soumis à la réglementation prudentielle. Le modèle « originate and distribute » (octroi puis cession des crédits) n'incite pas les établissements financiers à donner une grande importance à l'évaluation du risque.

vi. Autorisation de la Commission Nationale Informatique et Libertés "CNIL"

Le traitement informatisé des données relatives aux des personnes physiques doivent obéir en France à un certains nombre de règles. Dans le scoring, le CNIL déclare qu' "*aucune décision accordant ou refusant un crédit ne peut avoir pour seul fondement un*

traitement automatisé d'informations donnant une définition du profil de la personnalité de l'intéressé" (article 2 de la loi du 6 janvier 1978), ce qui est le cas pour le crédit à la consommation où le banquier doit prendre sa décision d'accorder ou pas le crédit rapidement. Dans ce cas, le scoring est considéré par le banquier, non pas comme un outil d'aide à la décision qu'il va prendre, mais plutôt, la décision elle même.

Le CNIL exige des banques qu'elles déclarent les variables utilisés pour le calcul du score, ainsi que les paramètres de ce dernier et les grilles de pondérations. Le CNIL déclare aussi que toute personne qui se voit refusé un crédit, a le droit d'accéder à son dossier de crédit et peut même exiger des rectifications (Tufféry, 2012).

6.2.2. Les problèmes de méthodologie

i. La population et l'échantillonnage

Les seules données disponibles pour construire un modèle de scoring sont relatives à l'historique des dossiers acceptés dont la variable à expliquer est connue (bon payeur ou mauvais payeur). La probabilité de défaut n'est ainsi estimée que pour les dossiers acceptés.

Le modèle de scoring ne tient donc pas compte des demandeurs rejetés dès le départ, donc il ne pourra pas estimer la probabilité de défaut de ces derniers. Ses résultats sont donc biaisés car les dossiers refusés n'entrent pas dans l'étape d'apprentissage du modèle de score établit. Le problème de la mise à jour ou d'actualisation de l'échantillon doit aussi être effectué.

En conclusion, deux types de redressements sont obligatoires :

- Le premier a pour but de corriger la structure de l'échantillon pour remédier au problème des refusés.

- Le deuxième doit envisager de permettre de corriger les effets de vieillissement de l'échantillon pour lui redonner la structure de la demande.

ii. Le choix des variables

Le choix des variables à introduire dans le modèle n'est pas évident. Il est nécessaire de mettre au point un algorithme de sélection afin de déterminer les variables les plus

discriminantes et les plus significatives, pour la classification des dossiers de crédit. Une fois la sélection des variables faite, le modèle pourra alors être aisément construit.

iii. Ecarts entre prévisions et réalisations

Une fois le modèle élaboré, il sera utilisé, en premier lieu, pour la sélection des dossiers basée sur la grille de score. D'autre part, il sera utilisé pour réaliser des prévisions d'impayés, compte tenu des politiques de sélection. Ses écarts entre prévisions et réalisations peuvent apparaitre et doivent être expliqués puis maitrisés. Ces écarts peuvent être dus à une évolution de l'environnement ou une évolution de la demande ou encore à des biais non pris en compte lors de la construction de l'échantillon. Pour déceler à temps les phénomènes initiateurs des écarts, il est fondamental de mettre en place des outils permanents de contrôle.

III. Conclusion

Les réglementations de l'accord de Bâle III ont été mise en place pour faire face aux crises, surtout la crise des Subprimes qui était la cause principale de la révision de l'accord de Bâle II. Cette crise est particulièrement reliée au risque de crédit qui peut être lourd de conséquence pour la banque si cette dernière voit ses clients ne pas honorer leurs engagements envers elle.

En effet, le non remboursement d'une dette représente une perte sèche pour les banques qui exigent toujours des recouvrements qui dépendent du montant du crédit et du risque encouru. La décision de notation des créances a pour objectif de renforcer le contrôle du risque de crédit. Les principaux scores utilisés par une banque sont au nombre de cinq :

- le **score d'appétence** qui estime l'intérêt qu'un client porte à un produit bancaire.
- le **score comportemental** qui estime le risque de non remboursement tout au long de la durée d'emprunt.
- le **score d'octroi** qui estime le risque d'un nouveau dossier de crédit.
- le **score de recouvrement** qui estime le montant susceptible d'être récupéré dans un cas de non remboursement.
- le **score d'attrition** qui estime la probabilité qu'un client quitte la banque.

Dans le chapitre suivant, nous nous intéressons qu'à deux types de scores qui sont le score d'octroi (accepter ou refuser le dossier de crédit) et le score comportemental (les dossiers de crédits acceptés feront ou non défaut) qui seront déterminés par l'utilisation de différentes techniques statistiques.

Basé sur les caractéristiques du dossier du client accepté seulement, le score comportemental estime le risque de crédit en prévoyant la solvabilité du demandeur de crédit. Le problème qui peut être rencontré dans ce cas de figure est le biais de sélection que nous expliquons dans le chapitre 2 et dont une solution serait la prise en considération des dossiers refusés dans le calcul du score comportemental, c'est le principe de l'inférence des refusés ("reject inference"). Nous présentons aussi dans le chapitre 2 les différentes techniques classiques de réintégration des refusés dans le processus d'octroi de crédit ainsi que des méthodes de l'apprentissage semi-supervisé que nous adaptons au problème de traitement des refusés.

CHAPITRE 2

Les méthodes de traitement des refusés dans le processus d'octroi de crédit

La probabilité de défaut dans le processus d'octroi de crédit est la probabilité qu'un client n'arrive pas à honorer ses engagements envers la banque.

Pour déterminer cette probabilité de défaut, plusieurs méthodes ont été proposées telle que l'analyse discriminante, la régression linéaire, la régression logistique, les arbres de décision, etc.

Les modèles statistiques utilisés pour l'évaluation de la probabilité de défaut ne considère que les dossiers acceptés ce qui peut biaiser les résultats de prédiction (c'est le biais de sélection). L'inférence des refusés consiste à réduire cette erreur de sélection en réintégrant les dossiers refusés à l'échantillon initial.

Dans ce chapitre nous décrivons, en premier lieu le processus de crédit et le biais de sélection et nous présentons par la suite les méthodes classiques de l'inférence des refusés ainsi que de nouvelles méthodes semi supervisées adaptées au problème de traitement des refusés pour remédier à ce biais.

I. L'inférence des refusés et le traitement des données manquantes

Nous expliquons ici le problème de l'inférence des refusés et nous présentons les méthodes de traitement des refusés basées sur les modèles de données manquantes.

1. Le processus de crédit

Quand le client d'une banque sollicite un prêt, la demande peut être acceptée ou rejetée par le banquier (on parle alors du score d'octroi), le client dont la demande est acceptée recevra un prêt. Selon son comportement de remboursement après une certaine période, il sera classé

dans la catégorie bon payeur ou dans la catégorie mauvais payeur (on parle alors du score comportemental).

On distingue donc deux types de score dans le processus d'octroi de crédit (Viennet et al. 2007) :

- Le score d'octroi (score d'acceptation): C'est le score qui estime le risque que présente un dossier au moment où celui-ci est déposé et qui aidera à la décision d'accorder ou non le crédit.

- Le score comportemental : Il estime le risque tout au long de la durée d'emprunt. Il permet d'étudier le comportement de remboursement des clients acceptés et les classe selon deux catégories : défaut si le client ne respecte pas l'échéance de remboursement ou non défaut dans le cas contraire.

Le processus d'octroi de crédit est illustré dans la figure n°13:

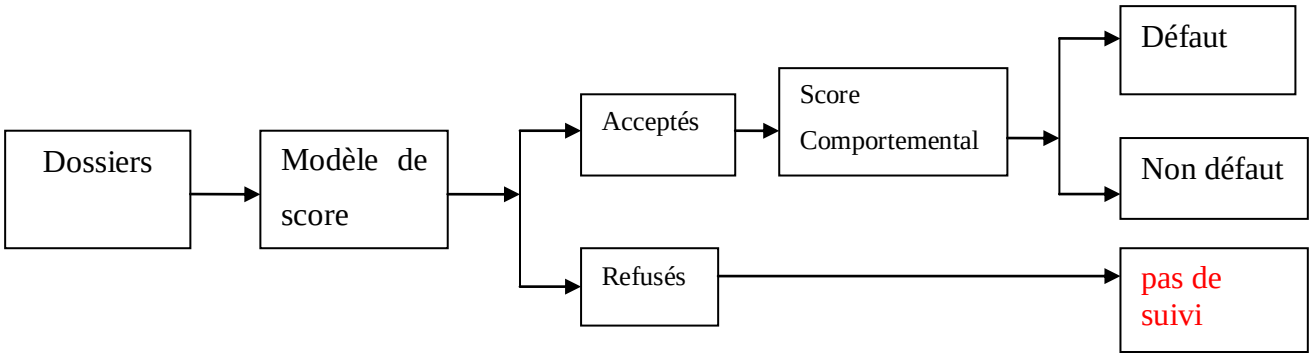


Figure 13 : *Les étapes d'octroi de crédit*

Notons ici que le risque de défaut des demandeurs rejetés est inconnu puisqu'il n'y a pas eu d'octroi de crédit et donc pas de comportement de remboursement à étudier.

On suppose que les données relatives à un client sont représentées par un ensemble de k variables noté $\mathbf{X} = (x_1, x_2, \dots, x_n, \dots, x_k)$, qui corresponde aux caractéristiques de chaque dossier de crédit. Ces caractéristiques comprennent des informations statiques (par exemple le salaire, l'âge, la profession, nombre d'enfants, etc.), des informations dynamiques concernant l'activité bancaire du client (le montant d'endettement, la fréquence de découvert, etc.) et des informations concernant le prêt demandé (le montant de crédit, la nature de crédit, la durée de remboursement, etc.). Ces données sont utilisées pour construire un premier score $S(\mathbf{X})$ qui est un score d'acceptation selon lequel le crédit sera accordé ou non selon un seuil s. Soit :

a est la décision d'octroi de crédit. a peut prendre deux valeurs :

$$a = \begin{cases} 0 \text{ si le crédit est non accordé} & S(\mathbf{X}) < s \\ 1 \text{ sinon} & S(\mathbf{X}) \geq s \end{cases}$$

Une fois le crédit accordé (a = 1), un score comportemental p est calculé pour déterminer si le client fera défaut ou non.

$p = P(y = 1 / \mathbf{X})$ avec y est le résultat de l'emprunt :

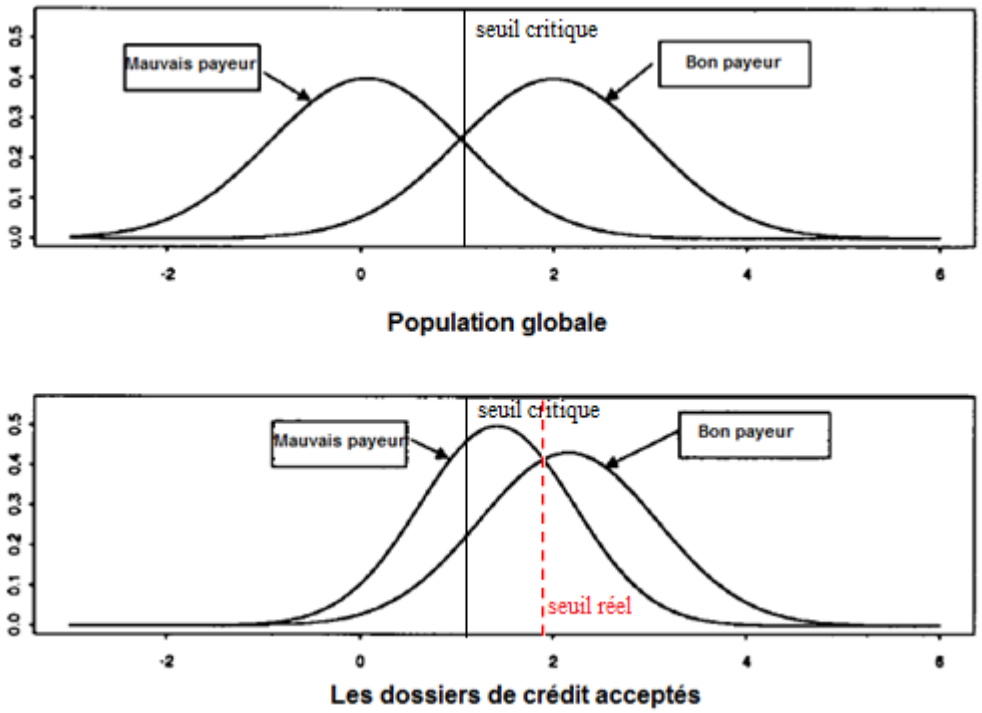
$$y = \begin{cases} 0 & s'il\ y\ a\ défault \\ 1 & si\ l'emprunt\ est\ remboursé\ sans\ défaut \end{cases}$$

2. Le biais de sélection

La probabilité de défaillance ne peut être estimée que pour les demandeurs acceptés par la banque sur la base des caractéristiques figurant dans les dossiers de crédit. Pour les demandeurs de crédit qui sont rejetés par la banque à cause de leurs profils à haut risque se traduisant par un faible score d'acceptation, la probabilité de défaut ne pourra être estimée par défaut de suivi pour ce type de dossiers.

Dans certains cas, les dossiers de crédit rejetés auraient pu être de bons payeurs s'ils avaient eu le crédit. C'est ce qui pourra être considéré comme un manque à gagner pour la banque, cette perte pourra être encore plus importante pour la banque, si elle accordait aléatoirement des crédits.

Étant construit sur la base d'un échantillon non représentatif de la population totale (les dossiers acceptés seulement), le modèle de score donnera des résultats biaisés. Ce biais est appelé "biais de sélection" ou "sample selection bias". La figure n°14 illustre bien le problème du biais de sélection:



(Adapté de : Feelders,2000)

Figure 14: *Biais de sélection*

Le premier graphique représente la distribution des bons et des mauvais payeurs pour la population entière (acceptés et refusés). Les distributions des deux populations (acceptés et refusés) sont bien séparées et le banquier pourra facilement prendre sa décision pour l'octroi du crédit.

Le deuxième graphique, représente la distribution des bons et des mauvais payeurs pour les demandeurs de prêt acceptés seulement. Sachant que le seuil critique est placé au milieu des bons et mauvais payeurs, on remarque, pour le deuxième graphique, que la partie des bons payeurs est sur-estimée et que l'écart entre les bons et les mauvais payeurs (qu'on appelle zone de doute) est sous-estimé. Ce qui rend plus difficile la décision du banquier vu que les deux courbes sont presque confondues.

Le biais de sélection peut être évité en incluant les caractéristiques des dossiers refusés dans l'estimation de cette probabilité ce qui devrait améliorer significativement les résultats obtenus et le modèle sera plus performant et répondra mieux aux réglementations prudentielles en vigueur.

Les dossiers refusés sont dans ce cas considérés comme des données manquantes. Il existe différents types de données manquantes comme démontrer ci après.

3. Les méthodes de traitement des refusés basées sur les modèles de données manquantes

3.1. Types de données manquantes

Selon Little & Rubin (1987), on peut distinguer trois types de données manquantes :

- Les données manquantes complètement aléatoirement " missing completely at random MCAR" ;
- Les données manquantes aléatoirement "missing at random MAR" ;
- Les données manquantes non aléatoirement " missing not at random MNAR" .

3.1.1. Les données manquantes complètement aléatoirement "MCAR"

Soient :

- **X** : le vecteur des variables indépendantes qui sont toujours observées et
- y : la variable cible comportant des valeurs manquantes (dans notre cas le résultat de l'emprunt).

On suppose ici que la probabilité d'avoir une valeur manquante (noté $y_{manquant}$) associée à la variable cible y est indépendante de X et de y.

$$P(y_{manquant} / y, \mathbf{X}) = P(y_{manquant})$$

Dans ce cas, la probabilité d'acceptation ne dépend ni des caractéristiques des clients ni de la valeur de la variable réponse y.

Il n'y a, donc, aucun problème d'inférence des refusés car les institutions financières accordent des crédits à tous les clients qui se présentent.

La probabilité d'acceptation (a = 1 crédit accordé) de crédit est alors formulée par la relation (2.1) (Feelders, 2000):

$$P(a = 1 / y, \mathbf{X}) = P(a = 1) \qquad (2.1)$$

3.1.2. Les données manquantes aléatoirement "MAR"

On suppose ici que la probabilité d'avoir une valeur manquante associée à la variable cible y est dépendante de X et non de y.

$$P(y_{manquant} / y, \mathbf{X}) = P(y_{manquant} / \mathbf{X})$$

Dans ce cas la valeur de la variable y manque de façon aléatoire, lorsque la probabilité d'acceptation de crédit dépend seulement des caractéristiques du client.

La probabilité d'acceptation est alors donnée par la relation (2.2) :

$$P(a = 1 / y, \mathbf{X}) = P(a = 1 / \mathbf{X}) \quad (2.2)$$

Cette situation est très fréquente dans la pratique puisque plusieurs institutions financières utilisent des modèles statistiques pour la sélection des clients potentiels pour un crédit (Feelders, 2000). Dans ce cas de figure, la variable y est observée seulement pour les dossiers de crédit acceptés.

On démontre que la probabilité de défaut (y = 1, pas de défaut) dépend aussi des caractéristiques des dossiers des clients et non pas du modèle de sélection (le score d'acceptation) ce qui implique la formule (2.3) (Feelders, 2000) :

$$P(y = 1 / \mathbf{X} ; a = 1) = P(y = 1/ \mathbf{X} ; a = 0) = P(y = 1/ \mathbf{X}) \quad (2.3)$$

Donc la probabilité de défaut, sachant toutes les variables indépendantes, reste la même aussi bien pour les dossiers acceptés que refusés.

3.1.3. Les données manquantes non aléatoirement "MNAR"

La probabilité d'avoir une valeur manquante associée à la variable cible y, sachant **X**, dépend de y. Dans ce cas, on doit réaliser des modèles pour les données manquantes pour obtenir des estimations non biaisées.

On peut citer comme exemple une personne qui a un revenu important et qui refuse de le dévoiler.

Les demandes de crédit rejetées ne manquent pas aléatoirement, ceci se produit typiquement quand le modèle de sélection est en partie basé sur des caractéristiques du dossier de crédit qui ne sont pas observables et ne font pas partie de l'ensemble de variables **X**, comme par

exemple, l'impression générale que le responsable du crédit a sur le client. Il peut également se produire quand les résultats du modèle de sélection ne sont pas pris en compte par le banquier ou lorsque certaines caractéristiques ne sont pas enregistrées dans l'ensemble de variables **X** (Feelders, 2000) .

La probabilité est donnée par (2.4) (Feelders, 2000) :

$$P (a = 1/ \mathbf{X}, y) \neq P (a = 1/ \mathbf{X}) \quad (2.4)$$

On peut déduire que la probabilité de ne pas faire défaut (y = 1) dépend entre autre du modèle de sélection quand on tient compte des caractéristiques des demandeurs de crédit acceptés. Ce qui donne à la formule (2.5) (Feelders, 2000) :

$$P (y = 1/ \mathbf{X} ; a = 1) \neq P (y = 1/ \mathbf{X} ; a = 0) \quad (2.5)$$

3.2. Quelques solutions pour le traitement des données manquantes

Les approches de traitement des refusés, différent selon le fait que le modèle de sélection est ignoré ou pris en compte par le banquier.

- Si le mécanisme de sélection est ignorable par le banquier alors les demandes refusées manquent aléatoirement (MAR) : les modèles les plus utilisés sont la régression logistique, l'analyse discriminante et l'approche mélange (un mélange de distribution).

- Si le mécanisme de sélection est non ignorable par la banquier ce dernier devrait le prendre en considération pour l'élaboration du modèle final afin d'obtenir des estimations fiables et pertinentes de la variable, y, et ainsi prédire correctement si le demandeur fera défaut ou pas (le cas de MNAR) : le modèle le plus approprié ici est celui de Heckman.

3.2.1. Le mécanisme de sélection ignorable

Dans le cas du mécanisme de sélection ignorable, Feelders, (2003) utilise trois méthodes : la régression logistique, l'analyse discriminante et l'approche mélange.

3.2.1.1. La régression logistique

Quand les données sont MAR, le résultat de l'emprunt, y, est la probabilité de ne pas faire défaut sachant les caractéristiques du dossier du client donnée par (2.6).

$$p = P(y = 1 / \mathbf{X}) = 1 - P(y = 0 / \mathbf{X}) \quad (2.6)$$

Dans ce cas, la relation entre les k variables explicatives ($\mathbf{x}_1$, …, $\mathbf{x}_n$, …, $\mathbf{x}_k$) = $\mathbf{X}$ et la variable à prédire binaire y qui est bon payeur ou mauvais payeur peut être établie par la régression logistique (Tenenhaus, 2007).

La catégorie (défaut/ non défaut) à laquelle appartient un nouveau client demandeur de crédit est déterminée après avoir estimé la probabilité $\pi(\mathbf{X})$ d'appartenance à l'un des groupes.

Quand $\pi(\mathbf{X}) \geq c$ avec c le seuil critique fixé par le banquier, le client est classé bon payeur. Alors que lorsque $\pi(\mathbf{X}) < c$ le client fera défaut et il sera classé mauvais payeur.

3.2.1.2. L'analyse discriminante

Dans l'analyse discriminante décisionnelle la détermination de la probabilité d'appartenance à un groupe, soit la probabilité a posteriori, peut être calculée, en utilisant le théorème de Bayes, à partir de la probabilité conditionnelle d'appartenance à un groupe j, $p_j(\mathbf{X}) = P(\mathbf{X} / y = j)$ avec $j \in \{0,1\}$ et de la probabilité a priori π_j ; $\pi_j = P(y = j)$.

Selon le théorème de Bayes, la probabilité a posteriori peut être exprimée par la relation (2.7):

$$P(y = j/\mathbf{X}) = \frac{\pi_j p_j(\mathbf{X})}{\sum_{j=0}^{1} \pi_j p_j(\mathbf{X})} \quad (2.7)$$

avec $j \in \{0,1\}$.

Une fois ce calcul effectué pour chacun des deux groupes, la règle de décision sera alors d'affecter un individu i au groupe pour lequel la probabilité à posteriori sera maximale. Donc la règle de décision est donnée par la relation (2.8) :

$$\begin{aligned} P(y = 0 / \mathbf{X}) > P(y = 1 / \mathbf{X}) &\Longrightarrow \mathbf{X} \in \Omega_0 \\ P(y = 0 / \mathbf{X}) < P(y = 1/\mathbf{X}) &\Longrightarrow \mathbf{X} \in \Omega_1 \end{aligned} \tag{2.8}$$

avec Ω_0 et Ω_1 les deux classes distinctes des bons payeurs ou des mauvais payeurs.

La fonction discriminante étant basée sur les caractéristiques des demandeurs de prêt, donc dans le cas où l'estimateur $p_j(\mathbf{X})$ de la probabilité conditionnelle d'appartenance à un groupe j, est basée seulement sur les dossiers acceptés, cette estimation sera biaisée à cause de la règle d'acceptation. Ce biais peut être évité en prenant en compte les caractéristiques des dossiers refusés dans l'estimation de la distribution $p_j(\mathbf{X})$. La solution à ce problème est d'appliquer l'approche mélange qui est la plus approprié dans ce cas de figure.

3.2.1.3. L'approche mélange

Dans un mélange de distributions (bons et mauvais payeurs), des hypothèses concernant le paramètre de densité de la matrice $\mathbf{X}$ sont mises en place. La fonction de probabilité $P(\mathbf{X})$ est donnée par la relation (2.9) :

$$\begin{aligned} p(\mathbf{X}) &= p(\mathbf{X} / y = 0) + p(\mathbf{X} / y = 1) \\ &= \pi_0 p_0(\mathbf{X}, \vartheta_0) + \pi_1 p_1(\mathbf{X}, \vartheta_1) \end{aligned} \tag{2.9}$$

où la probabilité a priori π_j, avec $0 \leq \pi_j \leq 1$, est considérée comme étant une proportion $p_j(\mathbf{X}, \vartheta_j)$ dans un modèle mixte et ϑ_j un paramètre inconnu avec $j \in \{0,1\}$.

On note que $\pi_0 + \pi_1 = 1$. En réalité, le résultat de l'emprunt est observé seulement pour les dossiers acceptés et non pas pour ceux refusés. Lorsqu'on tient compte du problème des données manquantes, la fonction de vraisemblance l_i sera définie par la relation (2.10):

$$l_i = \begin{cases} \pi_0 p_0(\mathbf{x}_i, \vartheta_0) + \pi_1 p_1(\mathbf{x}_i, \vartheta_1) & \text{si y est manquante} \\ \pi_j p_j(\mathbf{x}_i, \vartheta_j) & \text{si } y = j \text{ avec } j \in \{0,1\}; i \in \{1, \dots, n\} \end{cases} \tag{2.10}$$

Si on a m dossiers refusés et n dossiers acceptés, la fonction de vraisemblance pour les données manquantes ll_{man} est formulée par l'équation (2.11) le logarithme de vraisemblance :

$$ll_{man} = \sum_{i=1}^{m} \log\{\pi_0 p_0(\mathbf{x}_i, \vartheta_0) + \pi_1 p_1(\mathbf{x}_i, \vartheta_1)\} + \sum_{i=m+1}^{m+n} (1 - y_i) \log\{\pi_0 p_0(\mathbf{x}_i, \vartheta_0) + y_i \log(\pi_1 p_1(\mathbf{x}_i, \vartheta_1))\} \tag{2.11}$$

Le vecteur de paramètres $\boldsymbol{\varphi} = (\pi_0, \pi_1, \vartheta_0, \vartheta_1)$ est estimé par la méthode du maximum de vraisemblance. Pour trouver l'estimateur du maximum de vraisemblance $\hat{\boldsymbol{\varphi}}$, on doit recourir à

un algorithme spécial, adapté pour l'objet, connu sous le nom de l'algorithme EM (Espérance-Maximisation) (Feelders, 2000).

3.2.2. Le mécanisme de sélection non ignorable

3.2.2.1. Le modèle de Heckman

Pour le cas où les données ne sont pas MAR et où le mécanisme de sélection est non ignorable (Feelders, 2003), la méthode la plus appropriée est celle du modèle de Heckman (Puhani, 2000) donnée par :

$$a_i = \begin{cases} 0 & \text{si le crédit est non accordé} \quad S(\mathbf{X}) < s \\ 1 & \text{si le crédit est accordé} \quad S(\mathbf{X}) \geq s \end{cases}$$

$$y_i = \begin{cases} 0 & \text{s'il y a défaut} \\ 1 & \text{sinon} \end{cases}$$

avec

a_i : mécanisme de sélection pour l'observation i.

y_i : le résultat de l'emprunt de l'observation i est observable si a_i=1.

$$\begin{cases} a_i = \mathbf{x}_i\beta + d_i \\ y_i = \mathbf{x}_i\gamma + e_i \end{cases}$$

Où β et γ sont les paramètres à estimer, et (d_i, e_i) est un couple de termes d'erreur qui suivent une loi normale centrée et la matrice de variance-covariance **Σ**.

avec : $\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ et $(d_i, e_i) \rightsquigarrow \mathcal{N}(0, \mathbf{\Sigma})$,

Le coefficient de corrélation ρ entre les termes d'erreurs d_i et e_i est donné par la formule classique : $\rho = \frac{\mathrm{cov}(d_i, e_i)}{\sigma_{d_i}\sigma_{e_i}}$

Nous donnons le détail de l'adaptation du modèle de Heckman au problème d'inférence des refusés en l'annexe 3.

II. Les méthodes classiques de l'inférence des refusés

De nombreuses méthodes de traitement des refusés peuvent être utilisées, nous présentons ici les plus connues.

1. L'augmentation simple

L'augmentation simple (Siddiqi, 2006) se déroule en quatre étapes :

Étape 1: Construction d'un modèle de score sur la base d'un échantillon composé des dossiers acceptés seulement, étiquetés en bons et mauvais payeurs.

Étape 2: Détermination du taux de défaut des refusés par l'application du modèle établi dans l'étape 1 sur ces derniers.

Étape 3: Étiquetage des dossiers refusés par bon ou mauvais selon le taux de défaut.

Étape 4: Réintégration des refusés étiquetés dans l'échantillon initial des acceptés et détermination d'un nouveau modèle de score.

L'augmentation simple est facile à appliquer mais elle présente certains inconvénients : l'algorithme ne tient pas compte du score d'acceptation, en plus il suppose que le taux de défaut pour les dossiers acceptés est le même que pour les dossiers refusés ce qui n'est sans doute pas le cas dans la réalité.

2. L'augmentation

Cette technique dite aussi "repondération" ou "reweighting" (Banasik et al., 2007) est la technique d'inférence des refusés la plus utilisée, elle consiste à utiliser, en premier lieu, un modèle pour prédire la probabilité d'être accepté. Ce modèle est appliqué sur la population totale (acceptés et refusés) et un score d'acceptation est attribué à chaque demandeur de prêt. Ensuite la population est segmentée en intervalles dit "bandes de score". Ces bandes sont déterminées selon le nombre d'intervalles nécessaires, ou encore selon l'effectif des demandeurs de prêt de cet intervalle. Ces intervalles représentent tous des scores équivalents et ils sont constitués par les deux types de demandeurs de crédit, les acceptés et les refusés. Parmi les dossiers acceptés, on peut distinguer ceux qui ont fait un défaut de remboursement dans les 12 premiers mois du crédit (noté comme mauvais payeurs) et ceux dont le comportement de remboursement est parfait (noté comme bons payeurs).

On définit pour chaque intervalle de score un poids, ensuite chaque demandeur accepté sera pondéré par le poids de l'intervalle auquel il appartient et un score de défaut est construit sur la base des acceptés ainsi pondérés. Le calcul des poids des intervalles est illustré dans le tableau n°5:

TABLEAU 5 : *Calcul des poids dans la méthode d'augmentation*

Numéro de l'intervalle	Nombre de Bons	Nombre de Mauvais	Nombre d'acceptés	Nombre de refusés	Poids
1	n_{b1}	n_{m1}	$n_{A1} = n_{b1} + n_{m1}$	n_{R1}	$\frac{(n_{R1} + n_{A1})}{n_{A1}}$
2	n_{b2}	n_{m2}	$n_{A2} = n_{b2} + n_{m2}$	n_{R2}	$\frac{(n_{R2} + n_{A2})}{n_{A2}}$
⋮	⋮	⋮	⋮	⋮	⋮
j	n_{bj}	n_{mj}	$n_{Aj} = n_{bj} + n_{mj}$	n_{Rj}	$\frac{(n_{Rj} + n_{Aj})}{n_{Aj}}$
⋮	⋮	⋮	⋮	⋮	⋮
N	n_{bn}	n_{mn}	$n_{An} = n_{bn} + n_{mn}$	n_{Rn}	$\frac{(n_{Rn} + n_{An})}{n_{An}}$

(Adapté de:Viennet, 2007)

Nous reprenons, à titre illustratif, l'exemple donné par Siddiqi dans le tableau n°6:

TABLEAU 6 : *L'inférence des refusés*

Score	Mauvais	Bons	n_{Aj}	% de Mauvais	% de Bons	Refusés n_{Rj}	$\left[\frac{(n_{Rj} + n_{Aj})}{n_{Aj}}\right]$	Acceptés repond[12]	Bon repond[12]	Mauvais repond[12]
de 0 à 169	290	971	1261	23%	77%	1646	2.3	2900	2233	667
de 170 à 179	530	2414	2944	18%	82%	1732	1.58	4652	3815	837
de 180 à 189	365	2242	2607	14%	86%	3719	2.42	6309	5426	883
de 190 à 199	131	1179	1310	10%	90%	7334	6.59	8633	7770	863
de 200 à 209	211	2427	2638	8%	92%	1176	1.44	3794	3490	304
de 210 à 219	213	4047	4260	5%	95%	3518	1.82	7753	7365	388
de 220 à 229	122	2928	3050	4%	96%	7211	3.36	10248	9838	410
de 230 à 239	139	6811	6950	2%	98%	3871	1.55	10773	10558	215
de 240 à 249	88	10912	11000	0,8%	99,2%	4773	1.43	15730	15604	126
250 et plus	94	18706	18800	0,5%	99,5%	8982	1.47	27636	27498	138

(Adapté de : Siddiqi, 2006)

Prenons l'exemple d'un client qui se situe dans l'intervalle [170, 179] (donc dans le deuxième intervalle), le poids qui lui sera attribué est : (2944+1732) / 2944 = 1.58 donc le nombre

[12] repond : repondérés

d'acceptés sera multiplié par ce poids pour atteindre 4652 qui seront répartis en $4652 \times 0{,}18 = 837$ mauvais payeurs et $4652 \times 0{,}82 = 3815$ bons payeurs. Un score de défaut est ensuite construit sur les acceptés pondérés.

Crook et Banasik (2004) ont expliqué la théorie de l'augmentation et ont démontré qu'elle exige en premier lieu l'estimation du modèle d'origine (le modèle d'acceptés et refusés) pour déterminer la probabilité d'acceptation. La technique de repondération attribue un poids n_{A_j} aux acceptés du $j^{ème}$ intervalle notés A_j et aux refusés notés R_j. Chaque accepté A_j est pondéré par le poids, $\left[\frac{(R_j+A_j)}{A_j}\right]$, l'inverse de la probabilité d'acceptation dans le $j^{ème}$ intervalle. Ils ont aussi démontré que l'augmentation fourni une performance prédictive nettement inférieure à celle appliquée sur l'échantillon composé uniquement de candidats qui pourraient normalement être acceptés.

Le poids d'un demandeur de crédit accepté est d'autant plus élevé que sa probabilité d'acceptation est faible. Par la suite, on attribue à chaque client son poids et le modèle (bon ou mauvais) sera estimé sur la base de l'échantillon pondéré.

3. L'extrapolation

Cette technique consiste à considérer d'abord un modèle de score de défaut sur les acceptés seulement, à l'appliquer ensuite à l'ensemble de la population, en extrapolant la probabilité de faire défaut sur les cas refusés, et un seuil limite de probabilité est défini, par le banquier selon sa tolérance au risque, pour classer les refusés en deux catégories (bon et mauvais). Enfin, on détermine un nouveau modèle (bon/mauvais) pour toute la population.

Cette technique repose sur l'hypothèse qu'il n'y a pas de différence pour la détermination de bon ou mauvais payeur pour le cas des acceptés ou des refusés. Les banques se trouvent confrontées à un grand risque, celui de voir leur clients faire défaut.

Meester (1997) donne des résultats montrant des possibilités modestes d'amélioration de la prédiction de ce modèle.

4. La reclassification itérative

Cette technique dite aussi "l'ensemble augmenté" est une varinte de la méthode d'extrapolation. Elle consiste, en premier lieu, à construire un modèle de score de défaut sur

les acceptés et par la suite, comme c'est le cas pour la technique d'extrapolation, on applique ce modèle sur la population totale, en extrapolant la probabilité de faire défaut sur le cas des refusés. Une fois les refusés étiquetés en bons et mauvais payeurs par le modèle précédent, on construit, un nouveau modèle de score sur la population totale. C'est ce qui explique le nom de cette technique car la population est composée des acceptés et des refusés, sachant que les acceptés sont dés le départ partagés en deux catégories (bon et mauvais). On applique alors ce nouveau modèle aux refusés pour les classer en « bon » ou « mauvais ». On réitère le processus jusqu'à stabilisation des scores obtenus.

On peut remarquer que comme dans l'extrapolation, cette méthode pose comme hypothèse P(y/ a= 1) = P(y/ a= 0) ce qui signifie qu'il n'y a pas de différence entre la distribution des défauts et celle des non défauts dans le cas des acceptés et des refusés.

Joanes (1993) a utilisé la méthode de la reclassification itérative en appliquant la régression logistique sur la base d'un échantillon composé de dossiers acceptés pour construire le modèle de score et l'appliquer sur les dossiers refusés pour les étiqueter.

5. Le parceling

Cette méthode, qui a été présentée par Dempster et al. (Dempster et al., 1977), est considérée comme une amélioration de la reclassification itérative. La classification des dossiers refusés est faite proportionnellement à la probabilité de défaut correspondante à l'intervalle de score auquel le dossier appartient.

Elle consiste à construire un modèle de score de défaut sur les acceptés. Ensuite, on sépare la population en intervalles de score. Puis, on détermine pour chaque intervalle le nombre de bons et mauvais payeurs pour les acceptés ainsi que le nombre total de refusés.

Pour déterminer le taux de défaut des refusés par intervalle de score, quatre méthodes différentes sont utilisées :

5.1. « Polarised parceling » pour la population entière des refusés

Cette méthode consiste à utiliser un score pour classer les refusés et les attribuer, selon un seuil de probabilité, à la bonne catégorie :

- Défaut : si le score est inférieur au seuil de probabilité,
- Non défaut : si le score est supérieur au seuil de probabilité.

Le seuil de probabilité est déterminé par les experts de la banque selon leur niveau de tolérance au risque encouru. Par exemple, si le seuil est fixé à 200 (selon le tableau n°6) les refusés qui auront un score inférieur à 200 seront considérés comme mauvais payeurs, dans le cas contraire ils seront considérés comme bons payeurs.

5.2. « Polarised parceling » sur une base stratifiée

La différence entre cette méthode et la précédente est que le seuil de classification n'est pas appliqué sur la population entière des refusés, mais plutôt sur chaque intervalle séparément. Un seuil, fixé de la même façon que pour le cas de la population entière des refusés, est déterminé pour chaque intervalle de score pour séparer les cas qui ont fait défaut de ceux qui n'ont pas fait défaut dans la population des refusés. Au final, on obtient des seuils différents d'un intervalle à l'autre. Par exemple pour les deux intervalles de score [0-169] et [170, 179] (d'après le tableau n°6), si le seuil est fixé à 50 pour le premier intervalle, les 1646 refusés qui font partie de cet intervalle seront départagés selon ce seuil (d'après la figure n°15), il y a 1000 clients dont le score dépasse 50 et ils sont considérés comme bons, les 646 autres sont considérés comme mauvais. Pour le deuxième intervalle, si le seuil est fixé à 175, les 1732 refusés qui font partie de cet intervalle seront départagés selon ce seuil (d'après la figure n°15), il y a 532 clients dont le score dépasse 175 et ils sont considérés comme bons, les 1200 autres sont considérés comme mauvais.

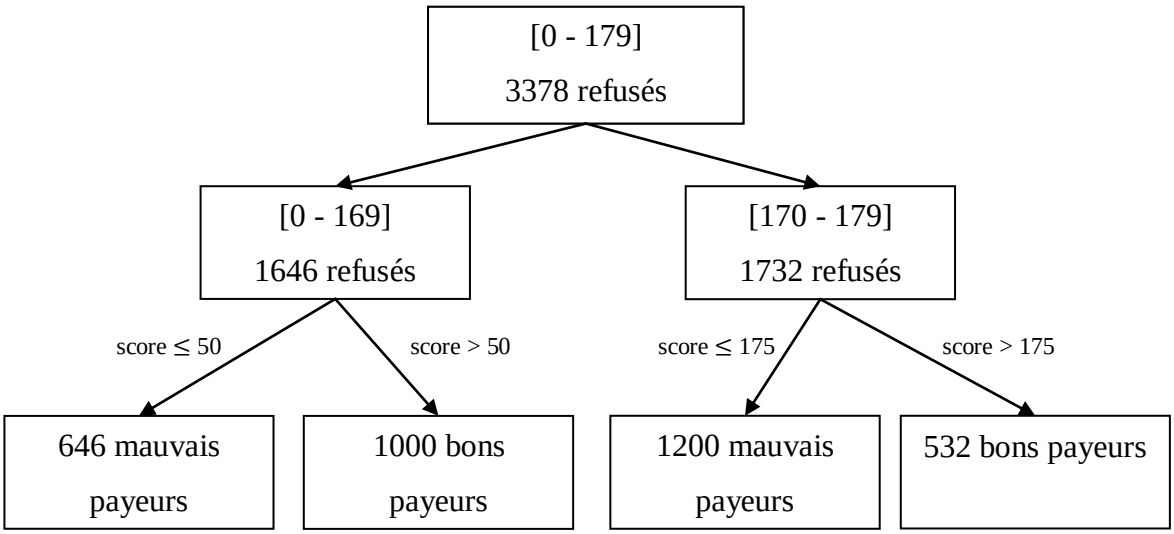


Figure 15 : *Exemple de classification des refusés*

5.3. Parceling aléatoire

La détermination des cas qui ont fait défaut et ceux qui n'ont pas fait défaut dans la population des refusés est faite aléatoirement, que ce soit pour la région entière des refusés ou pour chaque intervalle séparément.

- Pour la région entière des refusés :

  Après avoir déterminé le nombre de défaut et de non défaut sur toute la population des refusés, ces derniers sont assignés, aléatoirement, à l'une des catégories. Si l'on dispose de 10 000 dossiers de crédit, dont 6000 acceptés et 4000 refusés, et si parmi les 6000 dossiers acceptés, il y a 4200 (70%) dossiers qui ne feront pas défaut et 1800 (30%) qui feront défaut, en appliquant ce taux de défaut à la population des refusés, il y aura, parmi les 4000 dossiers refusés, 2800 (70%) bons payeurs et 1200 (30%) mauvais payeurs. Les clients rejetés seront affectés à ces deux catégories aléatoirement.

- Pour une base stratifiée :

  Cette approche diffère de la première décrite ci-dessus par le fait que les refusés sont assignés à l'une des catégories (défaut ou non défaut) pour chaque intervalle de score séparément et non pour toute la population des refusés.

  Elle consiste à prendre une hypothèse sur les taux de défaut des refusés par intervalle de score. On en déduit le nombre de ceux qui ont fait défaut et ceux qui n'ont pas fait de

défaut dans la population des refusés, qui sont par la suite assignés au hasard à l'une des catégories (défaut ou non défaut), en respectant le nombre de défaut et de non défaut calculé dès le départ dans l'intervalle considéré.

À titre d'exemple, si nous reprenons l'exemple de l'intervalle [0-169] du tableau n°6. Il y a 23% de "mauvais payeurs" et 77% de "bons payeurs" dans l'intervalle de score [0-169] parmi les admis. Les 1646 dossiers refusés devront être départagés, selon ces probabilités, en 379 Mauvais payeurs (23%) et 1267 Bons payeurs (77%). Par la suite l'affectation des dossiers rejetés dans chaque intervalle de score sera faite aléatoirement.

A la fin, on procède à un nouveau modèle de score sur la population totale en considérant les étiquettes d'origine pour les acceptés et les étiquettes déterminées par le parceling aléatoire stratifié pour les refusés.

5.4. « Fuzzy parceling » ou Duplication

Cette méthode est aussi nommée "duplication" ou "reclassification partielle". Elle consiste à créer deux copies pour chaque cas refusé, une copie est assignée à la catégorie de non défaut et l'autre à la catégorie de défaut, en considérant la probabilité de faire défaut et celle de ne pas faire de défaut comme poids attribués respectivement à chaque copie.

Par la suite, on inclut chacune des deux copies à la catégorie (défaut/ non défaut) initiale correspondante pour la construction du modèle de score. Si par exemple, le cas refusé initial a un poids de l'ordre de 5, avec une probabilité de non défaut de 77% (on considère l'intervalle de score [0-169]) alors les deux copies défaut ou non défaut auront respectivement un poids de 1.15($= 5 \times 0.23$) et 3.85 ($= 5 \times 0.77$) et seront réaffectées à l'échantillon d'apprentissage pour la construction de la fonction de score avec le poids qui leur correspond.

6. Le groupe de contrôle

Pour cette technique (proposée par Rosenberg et Gleit (1994)), l'échantillon est composé des acceptés et des refusés sans sélection. En fait, tous les dossiers qui se présentent pour la demande de crédit sont acceptés. On construit ensuite un modèle de score sur cet échantillon, qu'on appelle groupe de contrôle, en assignant chaque dossier à la bonne catégorie (défaut ou non défaut).

C'est la technique, jugée par les spécialistes, comme la plus fiable en tant que technique de traitement des refusés mais très rarement appliquée à cause du haut risque de non remboursement qu'elle fait courir à la banque.

7. La classification mixte

C'est une méthode qui relève du domaine de l'apprentissage non supervisé, le but étant de répartir un ensemble de données hétérogènes en des groupes de données homogènes. L'algorithme associé à cette méthode se déroule en deux étapes.

L'ensemble des observations à classer passe par un premier partitionnement utilisant la méthode des k-means[13] de façon à réduire la dimension initiale de cet ensemble de données en des groupes bien homogènes.

Les groupes obtenus subissent une deuxième classification à savoir la Classification Ascendante Hiérarchique (CAH). Le nombre final de classes à retenir sera fixé par l'un des indicateurs statistiques de la méthode.

D'après Lebart et al. (1995), la classification mixte présente l'avantage de pouvoir traiter un nombre volumineux de données à faible coût, en revanche, elle a l'inconvénient de produire des groupes qui dépendent des centres de gravités choisis à la première étape de partitionnement et de fixer le nombre de ces groupes a priori.

Les étapes de la méthode de classification mixte sont les suivantes:

Étape 1: Classification par la méthode des k-means : elle consiste à utiliser la méthode des k-means pour un partitionnement initial sur l'ensemble des observations (acceptés et refusés) de façon à obtenir k groupes homogènes avec «k» déterminé par le critère de Wong qui suggère que $k \geq n^{0.3}$ (Tufféry, 2009), n étant la taille de la population étudiée. Chaque classe contient une forte proportion de bons ou de mauvais dossiers en plus des dossiers refusés.

Étape 2: Les k classes ainsi obtenues seront, dans un second temps, réduites à q classes ($q < k$) par une Classification Ascendante Hiérarchique (CAH) appliquée aux centres de gravités des groupes trouvés par la première classification. Le nombre q de

[13] La méthode des k-means consiste à diviser la population en k classes de façon arbitraire. On calcule par la suite le centre de gravité de chacune des classes et les distances de chaque individu aux k centres obtenus et on attribue un individu à la classe qui lui est la plus proche et on réitère le processus jusqu'à ce qu' aucun individu ne change de classe.

classes est déterminé par analyse des indicateurs statistiques comme le R-Square (RSQ) qui détermine la proportion de la variance expliquée par les classes, le Semi-Partial R-Square (SPRSQ) qui mesure la perte d'inertie interclasse provoquée en regroupant deux classes (le but étant d'avoir une inertie interclasse maximale), le Root-Mean-Square Standard Deviation (RMSSTD), le Pseudo F qui mesure la séparation entre toutes les classes, le Cubic Clustering Criterion (CCC)[14] qui mesure un caractère plus au moins significatif d'une partition en référence à une hypothèse nulle d'absence de classes, ou encore par le dendrogramme qui suggérera le nombre final de classes par l'éventuelle coupure de la plus haute branche de l'arbre de classification.

Une fois les q classes définies, les observations non étiquetées, appartenant à chacune de ces classes, seront affectées à l'une des catégories (par exemple dans notre cas les dossiers refusés seront affectés à l'une des catégories "bon payeur" ou "mauvais payeur") en fonction de la catégorie dominante de la classe. À l'issue de cette méthode, toutes les observations sans étiquettes se voient attribuer une étiquette, un modèle (dans notre cas un modèle de score) est alors construit sur la base des observations d'origine étiquetées (dossiers acceptés avec des étiquettes bons ou mauvais) et les observations ainsi étiquetés (les dossiers refusés ainsi étiquetés).

La pertinence des résultats et la fiabilité de cette méthode reposent sur des hypothèses sous-jacentes : il serait mieux de standardiser les variables surtout si elles présentent des variances sensiblement inégales. Il faut tester l'adéquation des résultats à la réalité par simulation ou par des jeux de données test homologués pour en déduire le nombre de classes stables.

En plus, pour la mise en œuvre d'une classification mixte, il faut choisir une mesure d'éloignement ou de ressemblance entre individus, choisir le critère d'homogénéité des classes à optimiser, il faudrait aussi mesurer la qualité de la classification et choisir un nombre de classes optimales et interpréter ces classes.
Nous présentons dans le tableau n°7 les avantages et les inconvénients des différentes méthodes classiques de l'inférence des refusés :

[14] si CCC > 2 : laclassification est jugée bonne; si CCC est compris entre 0 et 2 alors la classification est peut être bonne mais c'est à vérifier mais CCC < 0 reflète la présence de valeurs extrêmes qui peuvent biaiser la classification.

TABLEAU 7 : *Avantages et inconvénients des méthodes classiques de l'inférence des refusés*

Méthodes classiques de l'inférence des refusés	**Avantages**	**Inconvénients**	**Règle d'affectation des dossiers refusés**
Augmentation simple	- L'Application simple.	- Ne pas tenir compte du score d'acceptation. - La probabilité de défaut est la même pour les dossiers acceptés et refusés.	Extrapolation des probabilités de défaut des acceptés sur les cas refusés
Augmentation	- Tenir compte du score d'acceptation - La pondération des dossiers.	- Utiliser la distribution des dossiers refusés pour corriger le bais de sélection. - Perte d'information pour les cas refusés. - Les dossiers refusés ne contribuent pas à l'apprentissage du modèle final.	Utilisation des dossiers refusés pour calculer les poids
Reclassification itérative	- L'amélioration de la performance du classifieur d'apprentissage d'une itération à l'autre.	- Ne pas tenir compte du score d'acceptation. - La probabilité de défaut est la même pour les dossiers acceptés et refusés.	Extrapolation des probabilités de défaut des acceptés sur les cas refusés
Parceling	- Simple et rapide à mettre en oeuvre. - L'affectation des dossiers refusés est proportionnelle à la probabilité de défaut de chaque intervalle. - définir un taux de défaut pour les refusés non sous-estimé.	- La probabilité de défaut est la même pour les dossiers acceptés et refusés.	Extrapolation des probabilités de défaut des acceptés sur les cas refusés proportionnellement à chaque intervalle
Groupe de contrôle	- Échantillon représentatif de la population entière. - Technique, jugée par les spécialistes, comme la plus fiable en tant que technique de traitement des refusés.	- Méthodes trop risquée dans la pratique. - Accord de crédit sans étude, a priori, des dossiers. - un grand risque de voir ses clients faire défaut.	Accorder le crédit sans exception à toute demande de crédit
Classification mixte	- La méthode considère les dossiers acceptés et refusés au même temps. - Pas de perte d'information appartenant aux dossiers refusés. - Combinaison des avantages de la méthode de k-means et de la CAH. - traitement d'ensemble de données volumineux.	- Le nombre de classes est fixé a priori pour le premier type de partitionnement : k-means. - Les classes finales dépendent des premiers centres de gravités déterminés dès le départ par les k-means.	Affectation des dossiers refusés à la catégorie la plus dominante de la classe auxquels ils appartiennent

8. Comparaison des différentes techniques classiques d'inférence des refusés

Comme présenté précédemment, la méthode de l'augmentation présente plusieurs inconvénients (voir Ash & Meester (2002)). Cette méthode repose sur l'hypothèse que la probabilité de défaut pour les acceptés est égale à celle des refusés, partant du fait qu'il peut exister quelques cas d'acceptés qui ont un profil similaire aux refusés.

Il se peut que l'affectation à la catégorie de bon ou mauvais payeur, d'un nombre élevé de refusés, selon un taux de défaut relatif aux acceptés soit liée à un nombre réduit d'acceptés (prenons l'exemple du tableau n°6 précédent), il y a seulement 1310 acceptés pour l'intervalle [190-199] contre 7334 refusés, ces derniers vont être assignés à l'une des catégories (bon ou mauvais payeur) selon le taux de défaut des acceptés c'est à dire 733 (10%) des refusés feront défaut contre 6601 (90%) qui ne feront pas de défaut.

Ces hypothèses sont irréalistes et ne peuvent être mises en pratique dans la réalité, car elles supposent toutes que les acceptés doivent avoir le même profil que les refusés.

Tufféry (2012) expose quelques inconvénients de la méthode de l'augmentation. Le premier est que l'hypothèse, que les "acceptés bons" appartenant à un intervalle de score donné ont le même profil que les "refusés bons" de ce même intervalle, reste à vérifier. Certaines variables qui font parties du profil de ces clients n'ont pas été utilisées dans le calcul du score d'acceptation (respectivement pour les cas des mauvais).

La prise en considération du poids dans la méthode de l'augmentation peut être une solution pour rétablir une quantité correcte de bons payeurs et de mauvais payeurs mais cette correction n'est forcément pas basée sur les véritables profils des clients d'où l'obtention de résultats biaisés.

Le deuxième point faible de la méthode est qu'elle repose sur l'hypothèse qu'il n'y a pas de différence pour la détermination de bons ou mauvais payeurs pour le cas des acceptés ou des refusés. Or si on se trouve dans le cas de figure où les refusés et les acceptés n'ont aucun point commun (une discrimination parfaite), la méthode de l'augmentation ne sera plus applicable. L'hypothèse précédente ne sera jamais vérifiée.

La troisième critique, présentée par Tufféry, est que les acceptés appartenant à un intervalle de score avec une probabilité d'acceptation faible (proche de zéro), se verront attribuer une pondération très forte vu que le poids est égal à l'inverse de la probabilité d'acceptation. Le problème se pose alors lorsque le nombre de ces acceptés, fortement pondérées, est faible. Dans ce cas, les résultats d'estimation du modèle seront influencés par cette minorité de clientèle et risque clairement de biaiser la modélisation. Il est clair que la méthode de l'augmentation est déconseillée dans ce cas.

Une autre critique, faite par Tufféry (2012), est que les dossiers refusés ne participent pas directement à l'étape d'apprentissage du modèle, mais seulement par l'intermédiaire du calcul de la pondération appliquée aux acceptés. Or les autres méthodes d'inférence des refusés exposées dans ce chapitre intègrent les dossiers refusés directement dans la modélisation (imputation des dossiers refusés), après les avoir affectés aux classes leurs correspondant (bon ou mauvais).

On peut remarquer que même si la pondération permet en quelque sorte de « compenser » l'absence des cas refusés dans la modélisation, la méthode de l'augmentation conduit à une perte d'information contenue dans ces derniers. Cette perte d'information s'explique par le fait que la méthode ne cherche pas à estimer l'étiquette des dossiers refusés pour que par la suite les combiner aux dossiers acceptés et construire le modèle de score final.

La méthode d'extrapolation et de reclassification itérative ainsi que celle du parceling reposent sur la même hypothèse que celle de la technique de l'augmentation qui considère qu'il n'y a pas de différence pour la détermination des bons ou mauvais payeurs pour le cas des acceptés ou des refusés.

Cette hypothèse est très contradictoire avec la réalité, et expose les institutions financières à un risque, très élevé, de voir leurs clients faire défaut. Dans la pratique on cherche à diminuer ce risque. Donc la proportion de bons et mauvais payeurs parmi les refusés ne doit pas être la même que celle parmi les acceptés. Logiquement, la proportion de mauvais payeurs doit être plus élevée pour les refusés et en général le taux de défaut attribué pour cette catégorie de clientèle devrait être supérieur à celui des acceptés.

La méthode de parceling est simple et rapide à mettre en œuvre. Pour cette méthode, le modèle de score de défaut doit être bien défini car c'est sur la base de ce modèle qu'on va

départager la population en bandes de score et qu'on va placer les dossiers refusés dans les bandes de score qui leur correspondent.

Pour perfectionner cette méthode, il faut tenir compte de l'hypothèse d'affectation d'une grande proportion de dossiers refusés à la catégorie de mauvais payeurs. Cette méthode permettra de définir un taux de défaut pour les refusés non sous-estimé.

Dans la pratique la technique de groupe de contrôle est considéré comme la plus risquée puisque le crédit est accordé sans étudier les dossiers a priori et la banque est ainsi exposée à un grand risque de voir ses clients faire défaut et c'est pour cette raison que cette technique n'est pas mise en œuvre.

Hand & Henley (1993) ont évoqué, dans leur article "can reject inference ever work?", différentes approches pour résoudre le problème d'inférence des refusés qui sont classées en trois catégories : les méthodes d'extrapolation, les méthodes qui utilisent la distribution des refusés et les méthodes basés sur les informations supplémentaires.

Ils ont démontré qu'il fallait distinguer entre deux méthodes. La première est basée sur l'estimation directe de la probabilité d'appartenance à un groupe, comme la régression logistique. La deuxième méthode passe par l'estimation de la probabilité conditionnelles d'appartenance à un groupe (bon ou mauvais payeur) en utilisant le théorème de Bayes, c'est le cas de l'analyse discriminante. Dans le cas où le modèle est construit en se basant uniquement sur les dossiers acceptés, les estimations des probabilités conditionnelle d'appartenance à un groupe seront biaisées contrairement aux méthodes qui estiment directement la probabilité de défaut. Ce biais est expliqué par la règle d'acceptation qui est déterminée sur la base des dossiers acceptés uniquement.

Plusieurs travaux de recherches empiriques (Eisenbeis, 1978) ont été faits pour expliquer ce problème.

Hand & Henley (1993) ont démontré aussi que lorsqu'on construit un premier modèle de score sur un ensemble d'observations X et que par la suite un nouveau modèle est construit sur un ensemble d'observations Y, avec Y un sous ensemble de X, cela impliquera des résultats biaisés. En effet, les modèles ne sont pas établis sur le même ensemble d'observations. Pour éviter ce biais, il faut impérativement inclure toutes les observations utilisées par le premier modèle de score, dans la construction du nouveau modèle.

Selon Reichert et al. (1983) le plus important est de savoir si on doit accorder ou non le crédit (discrimination entre acceptés et refusés) au lieu de distinguer entre les bons et les mauvais payeurs une fois le crédit accordé (discrimination entre bons et mauvais payeurs), donc ils proposent de distinguer trois classes (bons acceptés, mauvais acceptés et refusés) par un modèle de score construit sur la base de la population entière. En effet, ils jugent cette idée plus intéressante que de construire un modèle de score sur les acceptés pour distinguer les bons des mauvais payeurs. Or le but est de diviser la population entière en deux groupes (bon et mauvais), alors que l'identification de la catégorie des refusés peut très bien être déterminée par n'importe quelle méthode.

Le problème majeur, pour l'inférence des refusés, est qu'on se trouve dans le cas de données manquantes "missing at random" car la variable à prédire n'est connue que pour les dossiers acceptés et manque au hasard pour les dossiers refusés. Pour avoir un modèle de score avec un bon pouvoir prédictif, il faut considérer dans la modélisation une partie des dossiers refusés. Ce qui représente un grand risque pour le banquier. Pour que le banquier prend ce risque, il faut que ces données améliorent nettement la prédiction du modèle.

Hand & Henley (1993) ont conclu que les techniques d'inférence des refusés ne sont pas fiables et même si les modèles de score s'améliorent, après avoir réintégré les dossiers refusés, ce n'est dû qu'au hasard.

Or, nous savons très bien que, pour chaque modèle statistique, l'échantillon doit impérativement être représentatif de la population qu'on souhaite étudiée. Dans notre cas, la population est composée de dossiers acceptés et de dossiers refusés, donc il ne faut pas tenir compte des dossiers acceptés uniquement. En plus, pour pouvoir comparer les qualités prédites et réelles de la variable à prédire afin d'étudier la performance des modèles de score des différentes techniques d'inférence des refusés, il faut savoir si les dossiers refusés ont fait défaut ou non. Dans la pratique, nous ne savons pas si les refusés feront défaut ou non, et c'est ce qui explique le nombre très réduit d'études empiriques dédiées à la comparaison des techniques d'inférence des refusés. L'une des solutions les plus utiles est la simulation du processus de refus pour pouvoir évaluer les techniques d'inférence des refusés.

Dans le cadre de l'inférence des refusés, nous pouvons procéder, en plus des méthodes évoquées précédemment, à un apprentissage semi supervisé, car le problème qui se pose ici relève de la classification où l'on dispose à la fois d'un ensemble de données étiquetées (les

dossiers acceptés) et d'un ensemble de données non-étiquetées (les dossiers refusés). Le but sera d'atteindre un taux de classification élevé en combinant l'information contenue dans les données étiquetées et celle contenue dans les données non-étiquetées.

III. L'apport de la théorie de l'apprentissage au traitement des refusés

L'objectif de l'apprentissage statistique est d'améliorer la qualité de prédiction des modèles réalisés à partir d'observations, avec un échantillon test et de généraliser ces modèles à de nouvelles situations.

On distingue trois types d'apprentissage (Chapelle et al., 2006) :

- **L'apprentissage supervisé** : l'apprentissage supervisé nécessite la présence d'un ensemble d'observations déjà classées (étiquetées) pour les utiliser dans le processus d'apprentissage du modèle qui va permettre par la suite d'associer à toute nouvelle observation la classe qui lui correspond aux mieux. Nous pouvons citer par exemple l'analyse discriminante, la régression logistique, la méthode du plus proche voisin, les réseaux de neurones, les SVM, CART, etc.

- **L'apprentissage non-supervisé** : il s'agit de répartir un ensemble hétérogène de données en sous groupes homogènes regroupant chacun des données similaires. On parle de clustering. L'ensemble des données utilisées est non étiqueté et on ne dispose d'aucune information sur les classes. Les méthodes connues dans ce cas sont : la classification hiérarchique, la carte de Kohonen, les méthodes des K-means, extractions de règles, etc.

- **L'apprentissage semi-supervisé** : Ces méthodes se situent à mi chemin entre l'apprentissage supervisé et le non-supervisé. Comme dans le cas de l'apprentissage supervisé, ces méthodes ont pour objectif de déterminer un modèle qui prédit l'appartenance des données à différentes classes prédéfinies, tout en tenant compte des données non étiquetées qui peuvent apporter des informations supplémentaires dans un but d'améliorer la performance du processus d'apprentissage.

Dans cette partie, nous nous intéressons à l'apprentissage semi supervisé.

Il existe deux types d'apprentissage semi supervisé :

- **l'apprentissage inductif** dont le but est de déterminer une fonction capable de classer chaque observation de l'ensemble des données.

- **l'apprentissage transductif** dont le but est de chercher à classer les données non étiquetées en se référant aux autres données déjà classées (données étiquetées).

Dans notre cas, nous nous intéressons plutôt aux techniques transductives qui sont les mieux adaptées au problème d'inférence des refusés.

1. Analyse des algorithmes transductifs

L'objectif principal de l'apprentissage semi-supervisé est d'améliorer la performance de classification d'un modèle en combinant les données étiquetées et non étiquetées. Dans la pratique, le mécanisme d'étiquetage s'avère très couteux en argent et en temps puisqu'il nécessite du personnel qualifié et bien formé.

L'apprentissage semi-supervisé a l'avantage, en prenant en considération les données non étiquetées et les données étiquetées, de construire un modèle représentatif de la population entière, ce qui évitera le problème du biais de sélection. En effet au début des années soixante-dix, l'apprentissage semi-supervisé a représenté une solution au problème de traitement des données non étiquetées qui vient combler la faiblesse de l'analyse discriminante de Fisher.

Plusieurs techniques ont été développées dans le domaine de l'apprentissage semi-supervisé ; dans cette section, nous présentons certaines catégories de ces techniques.

1.1. L'auto-apprentissage (*self training*)

L'auto-apprentissage est considéré comme l'une des premières méthodes à être utilisée en apprentissage semi-supervisé. Elle est apparue dans les années 1960. Cette méthode consiste à entraîner le processus de classification par les données étiquetées et par la suite le résultat trouvé sera utilisé pour classer les données non étiquetées. En d'autres termes, le classifieur utilise ses propres prédictions pour s'auto-former d'où le nom d'auto-apprentissage. L'algorithme de l'auto-apprentissage est très simple, il consiste tout d'abord à apprendre la règle de classement (ou fonction de décision) notée « f » à partir de l'échantillon des données étiquetées. À chaque itération, et tant qu'il y a des données non étiquetées, le classifieur

sélectionne un sous ensemble de ces données sur le quel on applique la fonction de décision, déduite à l'étape précédente, pour l'étiquetage. Les données non étiquetées ainsi classées sont ajoutées aux données étiquetées pour former l'ensemble d'apprentissage. Finalement, une nouvelle fonction de décision est établie avec les nouvelles données étiquetées. L'algorithme est répété jusqu'à satisfaction d'une condition d'arrêt telle que par exemple l'épuisement des données non étiquetées. Nous pouvons dire que la méthode de la reclassification itérative est une variante de l'auto-apprentissage. En effet, les étapes de la reclassification itérative sont identiques à celles expliquées ici à une différence que les données non étiquetées au lieu d'être traitées en des sous ensembles, sont étiquetées toutes en même temps par le classifieur.

L'auto-apprentissage présente certains inconvénients : si par exemple, une erreur est commise lors du déroulement de l'algorithme d'auto-apprentissage. Cette erreur va générer des données incorrectement étiquetées. En utilisant ces résultats incorrects, l'itération suivante va être erronée et ainsi de suite,..., l'erreur va se propager. Donc pour éviter ce type d'erreur, il vaudrait mieux avoir le plus grand nombre de données étiquetées.

Maldonado & Paredes (2010) utilisent l'algorithme de l'auto-apprentissage avec quelques ajustements dont l'objectif de l'adapter au problème de l'inférence des refusés.

Le principe est alors d'entrainer un classifieur SVM en se basant uniquement sur les dossiers acceptés et de calculer la probabilité de faire défaut pour les dossiers refusés en utilisant la fonction logistique proposée par Platt (1999). Une fois cette probabilité déterminée et en se basant sur un seuil donné, les dossiers refusés sont affectés à l'une des deux classes (bon ou mauvais) et puis réintégrés aux dossiers acceptés.

Vapnik[15] a aussi proposé l'approche des SVM semi-supervisés, dont l'objectif est de combiner en même temps les données étiquetées et non étiquetées pour construire un modèle sur la base des deux types de données.

Le modèle final tient, alors, compte à la fois des dossiers acceptés et des dossiers refusés ainsi étiquetés et permet d'avoir des prédictions non biaisées pour un nouveau dossier de crédit.

[15] Cette référence a été citée par : BENNETT, K.P. and DEMIRIZ, A. *Semi Supervised Support Vector Machines. Proceedings of Neural Information Processing Systems, Denver, 1998.*

Pour ajuster le classifieur SVM, tout dossier refusé qui est classé comme bon payeur par l'algorithme d'auto-apprentissage, mais avec un taux de confiance faible (proche de l'hyperplan) sera pénalisé et réaffecté à la classe des mauvais payeurs.

Soit un ensemble de données réunion de l'ensemble des données étiquetées et des données non étiquetées, $N = L \cup U$.

où $L = \{(x_1^l, y_1^l), \quad \dots, \quad (x_l^l, y_l^l)\} \subset X \times Y$ est l'ensemble des données étiquetées,

$U = \{x_{l+1}^u, \quad x_{l+2}^u, \dots, \quad x_n^u\} \subset X$ est l'ensemble des données non étiquetées.

"y" la classe à laquelle appartient une observation et prend ses valeurs dans $\{+1, -1\}$ (dans notre cas bon ou mauvais payeur)

Pour une classification binaire, le modèle SVM fournit l'hyperplan optimal $< w, b >$ qui sépare au mieux les bons payeurs des mauvais payeurs, le modèle est définit par la relation (2.12) avec w est le vecteur des poids associé à l'ensemble des données étiquetées :

$$f(x) = w^T \cdot x_i^l + b \tag{2.12}$$

Dans le cas où les classes sont linéairement séparables, l'hyperplan tente de départager au mieux et en deux classes (+1) et (-1) les individus en cherchant l'hyperplan optimal qui sépare les deux groupes, en garantissant une grande marge de séparation entre ces deux classes.

Dans ce contexte, la procédure SVM vise à résoudre le problème d'optimisation suivant :

$$\underset{w,b,\xi}{Min} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{L} \xi_i \tag{2.13}$$

Avec C un paramètre de pénalisation des erreurs de classification dans l'échantillon d'apprentissage et ξ_i est le terme d'erreur, soient :

$$\begin{aligned} & y_i^l \cdot \left(w^T \cdot x_i^l + b\right) \geq 1 - \xi_i \qquad i = 1, \dots, L \\ & \xi_i \geq 0 \end{aligned} \tag{2.14}$$

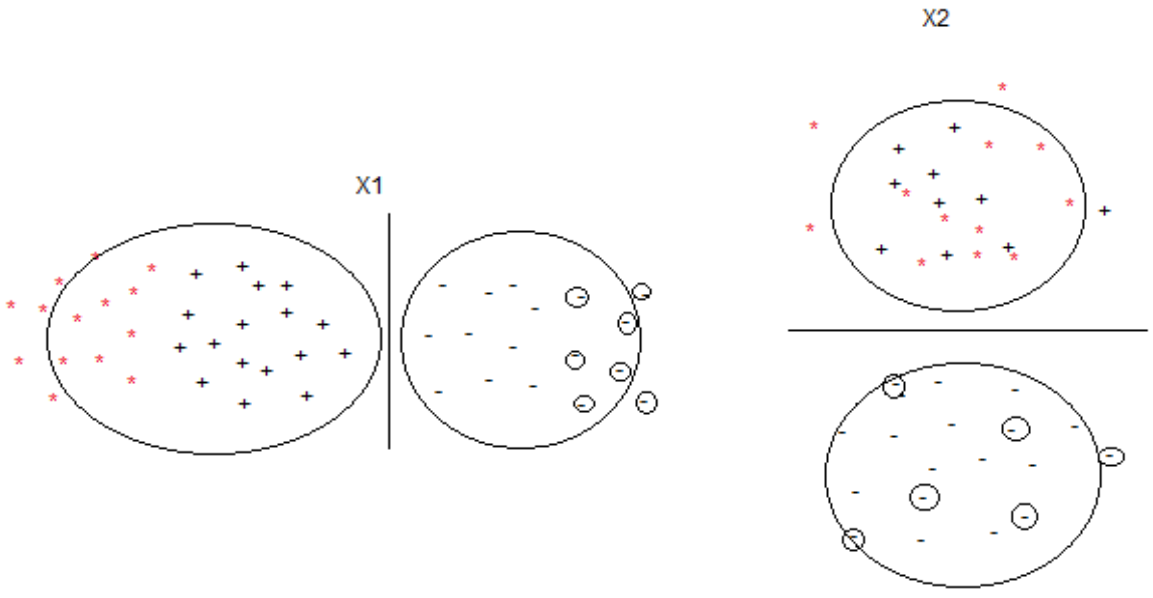
Les dossiers qui sont représentés par les points les plus éloignés dans l'hyperplan de séparation, auront un taux de confiance élevé et seront plus facilement affectés aux classes qui leur correspondent

1.2. Le Co-apprentissage (*Co-training*)

Pour mettre en œuvre cette méthode, il faut disposer d'un ensemble de données avec suffisamment d'attributs pour les départager en deux sous ensembles statistiquement indépendants. Chaque attribut doit respecter les deux hypothèses suivantes :

- Un attribut est suffisant pour effectuer une classification efficiente (engendrer suffisamment de données étiquetées)
- Les attributs sont indépendants conditionnellement à la classe Y à laquelle appartient l'observation.

L'idée est alors de partager un même ensemble de données étiquetées en deux projections indépendantes (Voir figure n°16), par la suite entraîner deux classifieurs selon ces deux projections.



Première projection X1 Deuxième projection X2

Figure 16 : *Répartition de l'ensemble de données en deux projections X1 et X2*

(Source : Divvala, S.K.[16])

Les deux classifieurs sont utilisés pour étiqueter les données qui ne le sont pas. L'algorithme de Co-training (Blum & Mitchell, 1998) s'illustre comme suit :

[16] DIVVALA, S.K. *Co-training and its applications in vision*. http://homes.cs.washington.edu/~santosh/presentations/coTraining_miscRead_web.pdf

Algorithme : Co-training

Données :

Soit un ensemble de données $N = L \cup U$ où $L = \{(x_1^l, y_1^l), \ \dots, \ (x_l^l, y_l^l)\} \subset X \times Y$ est l'ensemble des données étiquetées, et $U = \{x_{l+1}^u, \ x_{l+2}^u, \dots, \ x_n^u\} \subset X$ est l'ensemble des données non étiquetées avec "y" la classe à laquelle appartient une observation et prend ses valeurs dans l'ensemble $\{+1, -1\}$.

Faire :

i. L'ensemble d'attributs des données étiquetées X^l (données d'apprentissage) est divisé en deux sous ensembles indépendants X_1^l et X_2^l avec $X^l = X_1^l \times X_2^l$.

ii. Deux classifieurs f^1 et f^2 sont entrainés en utilisant les deux ensembles respectivement X_1^l et X_2^l .

iii. Appliquer les deux classifieurs aux données non étiquetées pour leur affecter des étiquettes

iv. Les données étiquetées avec une bonne confiance par l'un des classifieurs sont ajoutées aux données d'apprentissage de l'autre classifieur.

v. La phase d'apprentissage des classifieurs est réitérée sur le nouvel ensemble d'apprentissage.

vi. Retirer ces données ainsi étiquetées avec une bonne confiance des données non étiquetées restantes

Fin :

Refaire la procédure jusqu'à satisfaire l'une des conditions d'arrêt : soit il n'y a plus de données non étiquetées, soit les deux classifieurs ne peuvent plus classer avec succès ces données.

L'objectif principal de la technique de Co-training est qu'au final, on obtienne un ensemble d'apprentissage beaucoup plus grand et cela en combinant les prédictions des deux classifieurs.

On peut résume cet algorithme par un schéma simple (Figure n°17):

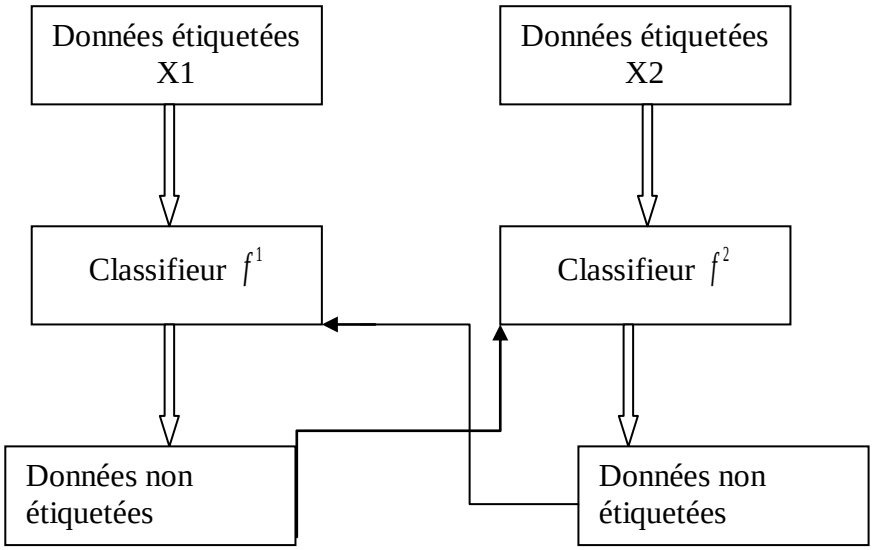


Figure 17 : *Le processus de Co-training*

(Source: Divvala, S.K.[17])

2. Les algorithmes de Boosting

Le boosting introduit par Freund & Schapire (1999), est une famille d'algorithmes d'apprentissage automatique dont le but est de combiner un ensemble de classifieurs dits « faibles » pour aboutir à un seul classifieur dit « fort » d'où l'amélioration de la performance de la classification.

2.1. L'algorithme ADABOOST

On dispose d'un ensemble de données d'apprentissage: $\{(x_i, y_i)\}_{i=1,\dots,N}$ où x_i est le vecteur des valeurs caractéristiques et y_i= -1 ou 1 est la classe à laquelle appartient un individu. On définit $F(x) = \sum_1^T \alpha_t f_t(x_i)$ où $f_t(x)$ est le classifieur qui aboutit à une valeur égale à (+1) ou (-1), α_t est une constante et T est le nombre maximal de classifieurs. L'affectation d'une nouvelle observation se fait selon le signe de $F(x)$.

Le principe de cet algorithme est d'attribuer un poids à chaque observation de l'échantillon d'apprentissage, ce poids correspond au niveau de difficulté rencontré pour prédire la classe de cet individu.

L'algorithme commence par construire un premier classifieur sur la totalité des données d'apprentissage.

[17] DIVVALA, S.K. *Co-training and its applications in vision*. http://homes.cs.washington.edu/~santosh/presentations/coTraining_miscRead_web.pdf

Initialement, tous les poids sont identiques et sont utilisés pour construire le premier classifieur, mais aux cours des itérations, les données qui se trouvent dans une classe qui ne leur corresponde pas vont avoir des poids croissants et ceux qui sont bien placés dans leurs classes auront des poids décroissants.

L'algorithme est alors le suivant (Friedman et al. 2000):

Algorithme : AdaBoost

Initialisation: Poids initiaux $w_0(i) = \frac{1}{N}$ avec $i = 1, \dots, N$

Faire pour t = 1, 2, ..., T:

i. Construire un classifieur $f_t(x_i) \in \{-1, 1\}$ en utilisant les poids w_i et les données d'apprentissage.

ii. Calculer le risque empirique pondéré :

$$\varepsilon_t(i) = \sum_i w_t(i) \mathbb{1}_{[y_i \neq f_t(x_i)]}, i = 1, \dots, N \quad (2.15)$$

Si $\varepsilon_t > 0{,}5$ **ou si** toutes les observations sont bien classées **alors** on arrête le processus

Sinon:

i. Calculer $\alpha_t = \frac{1}{2} \log\left(\frac{(1-\varepsilon_t(i))}{\varepsilon_t(i)}\right)$ (2.16)

ii. Mettre à jour les poids

$$w_{t+1}(i) = \frac{w_t(i)}{z_t} \times \begin{cases} e^{-\alpha_t} & \text{si } f_t(x_i) = y \quad \text{bien classé} \\ e^{\alpha_t} & \text{si } f_t(x_i) \neq y \quad \text{mal classé} \end{cases} \qquad i = 1, \dots, N$$

$$w_{t+1}(i) = \frac{w_t(i) \exp(-\alpha_t y_i f_t(x_i))}{z_t} \quad (2.17)$$

Avec z_t est un facteur de normalisation

Fin:

Le classifieur final affecte une nouvelle observation à l'une des classes selon le signe du classifieur

$$\text{sign}[F(x)] = \text{sign}\left(\sum_{t=1}^{T} \alpha_t f_t(x_i)\right) \begin{cases} \text{si } F(x) > 0 \text{ alors } y = 1 \\ \text{si } F(x) < 0 \text{ alors } y = -1 \end{cases} \quad (2.18)$$

2.2. L'algorithme « LogitBoost »

Les fondateurs de l'approche « AdaBoost », Freund & Schapire en 1995, ont toujours essayé d'améliorer cet algorithme, ce qui explique l'apparition de plusieurs variantes de cette méthode qui optimisent différemment la pondération (w_i). On s'intéresse ici à l'algorithme « LogitBoost » qui répond au mieux au problème de classification qui est basé, selon la règle de Bayes, sur la détermination de la probabilité a posteriori P(y = j / x) ou j représente la classe à laquelle une observation x appartient.

Les algorithmes « AdaBoost » sont considérés comme des procédures d'estimation pour la conception d'un modèle de régression logistique additive $F(x) = \sum_{t=1}^{T} \alpha_t f_t(x_i)$ conçue pour minimiser l'espérance de l'exponentielle de la fonction de perte $J(F) = E(e^{-yF(x)})$. L'exponentielle de la fonction de perte évolue de façon exponentielle avec l'erreur de classification ce qui implique la vulnérabilité et l'hyper sensibilité de l'algorithme « adaboost ». Pour remédier à ce problème, Friedman et al. (2000) proposent d'utiliser une fonction de perte binomiale qui est la log-vraisemblance de la fonction de perte $L(J(F)) = E\left[-\log\left(1 + e^{-y(F(x))}\right)\right]$ qui évolue linéairement avec l'erreur de classification et rend ainsi l'algorithme « LogitBoost » plus robuste et plus stable face aux données bruitées et aberrantes.

« LogitBoost » minimise la log-vraisemblance de la fonction de perte binomiale en utilisant les étapes de Newton.

L'algorithme « LogitBoost » se déroule comme suit :

On considère la probabilité pondérée $p(x) \in [0,1]$ donnée par le relation (2.19).

$$p(x) = P(y = 1 / x) = \frac{e^{F(x)}}{e^{F(x)} + e^{-F(x)}} \qquad (2.19)$$

Algorithme : LogitBoost

Données: L'ensemble des données d'apprentissage $S = \{(x_1, y_1), \cdots, (x_N, y_N)\}$ avec $x_i \in X$ et $y_i \in Y = \{-1, 1\}$, (on a en tout T itérations).

Initialisation: On commence dans la première itération par fixer le poids $w_0(i) = \frac{1}{N}$ avec $i = 1, \ldots, N$; $F_0(x_i) = 0$ et la probabilité $p(x) = \frac{1}{2}$

Faire pour t = 1, 2, ..., T:

i. Calculer des poids $w_t(i)$ et des valeurs attendues $z_t(x_i)$ du classifieur faible pour chaque observation:

$$z_t(x_i) = \frac{y_t^*(x_i) - p(x_i)}{w_t(i)} = \frac{y_t^*(x_i) - p(x_i)}{p(x_i)\big(1 - p(x_i)\big)} \tag{2.20}$$

avec

$$y_t^*(x_i) = \frac{y_t(x_i)+1}{2} \tag{2.21}$$

$$\text{et} \quad w_t(i) = p(x_i)\big(1 - p(x_i)\big) \tag{2.22}$$

ii. Construire la fonction $f_t(x_i)$ en utilisant les Moindres Carrées Pondérés de $z_t(x_i)$ sur x_i en se basant sur les poids $w_t(i)$

iii. Mise à jour de F(x) : $F_{t+1}(x_i) \leftarrow F_t(x_i) + \frac{1}{2} f_t(x_i)$ (2.23)

$$\text{et } p_{t+1}(x_i) \leftarrow \frac{e^{F_t(x_i)}}{e^{F_t(x_i)} + e^{-F_t(x_i)}} \tag{2.24}$$

Fin: Les sorties de cet algorithme aboutiront au classifieur final

$$F(x) = \text{signe}\,[F(x)] = \text{signe}\,[\textstyle\sum_{t=1}^{T} f_t(x_i)\,] \tag{2.25}$$

2.3. L'algorithme «Gentle AdaBoost»

L'algorithme « Gentle AdaBoost» (Friedman et al. 2000) est considéré comme une amélioration de l'algorithme «AdaBoost» et il donne un ensemble de classifieurs plus efficients et plus robustes. L'algorithme « Gentle AdaBoost» ne tient pas compte dans ses

différentes étapes du calcul de ratios en fonction logarithmique qui peuvent être instable numériquement. L'algorithme utilise les Moindres Carrés Pondérés pour minimiser la fonction $E(\exp(-y_i F(x)))$ (Zhang, 2012).

Algorithme : Gentle AdaBoost

Données: L'ensemble des données d'apprentissage $S = \{(x_1, y_1), \cdots, (x_N, y_N)\}$ avec

$x_i \in X$ et $y_i \in Y = \{-1, 1\}$, (on a en tout T itérations).

Initialisation: Poids initiaux $w_0(i) = \frac{1}{N}$ avec $i = 1, \ldots, N$ et $F_0(x_i) = 0$

Faire pour t = 1, 2, ..., T:

i. construire la fonction de régression $f_t(x_i)$ à l'aide des Moindres Carrés Pondérés de $y_t(x_i)$ sur x_i avec les poids $w_t(i)$.
ii. Mise à jour F(x) : $F_{t+1}(x_i) \leftarrow F_t(x_i) + f_t(x_i)$
iii. Mise à jour des poids : $w_{t+1}(i) \leftarrow w_t(i) \exp(-y_t(x_i) f_t(x_i))$ et normalization

Fin: L'affectation finale d'une nouvelle observation se fait par le signe du classifieur final :

$$\text{sign}[F(x)] = \text{sign}\left(\sum_{t=1}^{T} f_t(x_i)\right) \begin{cases} \text{si } F(x) > 0 \text{ alors } y = 1 \\ \text{si } F(x) < 0 \text{ alors } y = -1 \end{cases} \quad (2.26)$$

2.4. Application des algorithmes de Boosting pour le traitement des dossiers refusés

Toutes les méthodes de Boosting, que nous avons présenté, s'appliquent dans le cas où on dispose de données étiquetées seulement, or dans notre cas, nous disposons à la fois de données étiquetées (dossiers acceptés) et de données non étiquetées (dossiers refusés). Dans ce cadre d'apprentissage semi-supervisé, une nouvelle approche a été proposée par Bennett et al. (2002) nommée l'algorithme « ASSEMBLE ». Cette approche, qui est basée sur l'arbre de décision comme classifieur, a réussi à donner de bons résultats et a gagné le premier prix dans la compétition NIPS[18] en 2001 parmi 34 autres algorithmes. Les auteurs ont présenté un algorithme bien spécifique aux cas de données étiquetées et non étiquetées, "l'algorithme

[18] The Neural Information Processing Systems

ASSEMBLE.ADABOOST" (*adaptative boosting*) (Chawla & Karakoulas 2005). Le principe de cet algorithme est de commencer par affecter les données non étiquetées à des pseudo classes avant de commencer les autres étapes de l'algorithme. Il suffit alors d'affecter une donnée non étiquetée à la classe de son plus proche voisin. Ainsi, on aura des données qui sont toutes étiquetées. Nous utilisons ce principe d'affectation des données non étiquetées à des pseudo classes, dans notre étude, pour le cas des algorithmes AdaBoost, LogitBoost et Gentle Adaboost pour qu'ils soient tout les trois adaptés aux problèmes de traitement des dossiers refusés que nous considérons comme des données non étiquetées. En effet, nous nous sommes basés sur le principe de l'extrapolation pour affecter les dossiers refusés à des pseudo classes. Il suffit alors de construire un modèle de score de défaut sur les dossiers acceptés en se basant sur l'Analyse Factorielle Discriminante (AFD). Par la suite, nous extrapolons la probabilité de faire défaut, établie par le modèle, sur les cas refusés et nous définissons alors un seuil limite de probabilité. Ce seuil nous permettra de classer les refusés en deux catégories dites pseudo classes (bon et mauvais). L'affectation des dossiers refusés aux pseudo classes se fait selon le score obtenu comme nous l'avons indiqué précédemment. Si le score est supérieur au seuil fixé alors le dossier est considéré comme bon payeur sinon le dossier est considéré comme mauvais payeur.

L'avantage avec ces algorithmes de Boosting est qu'ils combinent un ensemble de classifieurs "faibles" (telles que l'analyse discriminante avec l'algorithme AdaBoost et la régression linéaire avec l'algorithme Logit Boost ou Gentle AdaBoost) pour aboutir à un classifieur final robuste et "fort". Les prédictions issues de ce classifieur final seront plus précises et plus proches de la réalité, donc la prédiction des classes d'appartenance des dossiers refusés sera proche de la réalité.

L'objectif de ces algorithmes de Boosting est d'entrainer le classifieur d'apprentissage, d'une itération à l'autre, afin d'obtenir à un classifieur final robuste. En effet, le classifieur d'apprentissage va prédire, à chaque itération, les classes des dossiers refusés et des dossiers acceptés et ces prédictions seront comparées aux pseudo classes. L'algorithme s'arrête alors lorsque les classes prédites seront stables. L'avantage avec ces méthodes itératives par rapport aux autres méthodes classiques de l'inférence des refusés (à l'exception de celle de la reclassification itérative) est que la robustesse du classifieur s'améliore d'une itération à l'autre.

IV. Conclusion

Nous avons commencé par expliquer le principe de l'inférence des refusés qui se base sur le biais de sélection qui s'explique par des modèles de score basés sur les dossiers acceptés uniquement (échantillon non représentatif de la population totale (dossiers acceptés et dossiers refusés)). Le problème du biais de sélection est alors à l'origine de résultats biaisés de ces modèles de score.

Pour remédier à ce problème, il est impératif de réintégrer les dossiers refusés à l'échantillon d'origine c'est à dire les dossiers acceptés. Dans le cadre d'une modélisation des dossiers refusés comme données aléatoirement manquantes, nous avons présenté différentes méthodes statistiques traitant ce type de données dans le contexte du scoring. Parmi ces méthodes, nous avons exposé des méthodes classiques de l'inférence des refusés et de nouvelles méthodes de l'apprentissage semi supervisé que nous avons adapté pour résoudre le problème de traitement des dossiers refusés.

Nous nous proposons de comparer la performance de ces deux types de méthodes pour voir si les méthodes semi supervisées donnent de meilleurs résultats et aident à remédier au problème de biais de sélection. Pour cela, dans le troisième et dernier chapitre, nous mettons en application certaines techniques d'inférence des refusés proposées dans la littérature et d'autres méthodes de l'apprentissage semi-supervisé sur données réelles. Nous confirmons nos résultats par simulation.

CHAPITRE 3

Étude empirique des performances de certaines méthodes de réintégration des refusés

Afin de comparer les performances des diverses méthodes de réintégration des refusés (voir chapitre 2 (§ II) pour les méthodes classiques et (§ III-1 et § III-2) pour les méthodes semi supervisées), nous utilisons des données bancaires françaises.

Il s'agit d'étudier les résultats obtenus par les différentes méthodes et de déterminer celle qui donne le modèle le plus performant. Pour cela, nous calculons les aires sous les courbes ROC de chaque méthode, celle qui possède l'aire la plus proche de 1 sera jugée la meilleure.

I. Présentation des données

Nous disposons d'une base de données réelles qui nous a été fournie par l'agence française de notation externe "Experian". Ses données proviennent initialement de la société « Financo » qui est un organisme de crédit à la consommation dans l'automobile, la moto, les véhicules de loisirs, l'habitat et l'équipement général des ménages.

Généralement une banque classe ses clients en deux : les acceptés et les refusés.

Les refusés sont subdivisés en deux catégories les refusés systématiques qui sont jugés par la banque ne jamais pouvoir rembourser un emprunt et ceux qui pourraient sous certaines conditions rembourser l'emprunt (refusés non systématiques). Dans ce travail, nous nous intéressons aux clients du second sous groupe de refusés.

Notre échantillon est formé de 13 319 dossiers de crédit collectés sur une période de deux années entre 2000 et 2001. Le comportement de remboursement des clients est observé sur une durée minimale de 18 mois. Les données relatives aux dossiers conduisant à un refus systématique ne relevant pas de la procédure de réintégration ne nous ont pas été fournies et ont déjà été exclus de l'échantillon par l'agence "Experian". La question de l'inférence des refusés ne se pose que pour les dossiers refusés non systématiques susceptibles d'être acceptés.

Nous notons par « BM » la variable qui définie le comportement de remboursement du client et elle est à prédire. Cette variable est classée en cinq modalités (voir tableau n°8).

TABLEAU 8 : *Description des modalités de la variable « BM »*

Code (BM)	**Étiquette**	**Description**	**Effectifs**
1	Bon	Pas de défaut (bon payeur)	7084
2	Intermédiaire	1 ou 2 défauts de paiement	1831
3	Mauvais	Plus de 3 défauts de paiement	902
98	Sans suite	Dossiers classés sans suite	1596
99	Refusé	Dossiers refusés	1906

(Source : Experian)

Les dossiers intermédiaires, codés par BM = 2, ainsi que ceux classés sans suite, codés par BM = 98, sont exclus de l'échantillon. Au final, nous ne gardons que 9 892 dossiers de crédit composés de 7 986 dossiers acceptés dont la variable BM est connue (bon payeurs ou mauvais payeurs) et 1 906 dossiers refusés non systématiques dont la variable BM n'est pas connue, et que nous cherchons à estimer.

La variable dépendante est une variable binaire qui indique si le client a fait défaut (BM=3) ou bien s'il est classé comme bon payeur (BM=1).

Les variables indépendantes dont nous disposons donnent des renseignements concernant la situation financière, personnelle actuelle et future du client.

Nous comptons 15 variables indépendantes dont 12 variables quantitatives et 3 qualitatives. (voir tableau n°9).

TABLEAU 9 *: Description des variables*

Variables	Description
Variables Qualitatives	Profession
	Situation Familiale
	Type de logement
Variables Quantitatives	Ancienneté bancaire
	Montant d'achat
	Montant du capital emprunt
	Montant de la charge immobilière mensuelle
	Montant de l'échéance
	Revenu mensuel du ménage
	Nombre d'échéances
	Autres revenus mensuels
	Age du demandeur de crédit
	Ancienneté dans le logement
	Ancienneté professionnelle
	Nombre d'enfants

(Source : Experian)

Comme traitement préliminaire des données, nous avons transformé les variables qualitatives en variables dichotomiques : une variable qualitative à r modalités est remplacée en r variables dichotomiques.

L'objectif commun pour les onze méthodes est de réintégrer les dossiers refusés dans le processus d'octroi de crédit et ceci en les affectant à l'une des catégories « bon » ou « mauvais » et avoir ainsi à la fin un modèle de score construit sur la base de la population "entière".

II. Méthodologies de comparaison entre les méthodes de traitement des refusés

Ne pouvant pas apprécier le comportement de remboursement des refusés (ces derniers n'ont pas eu de crédits), nous avons procédé par simulation.

1. Simulation du processus de refus

Pour pouvoir comparer les qualités prédites et réelles de la variable à prédire, nous avons simulé un processus de refus sur les 7986 dossiers acceptés. La variable BM n'est connue que pour les dossiers acceptés et manque au hasard pour les dossiers refusés, nous nous trouvons donc dans le cas de données aléatoirement manquantes dites "missing at random".

Le processus de simulation est le suivant :

À l'aide d'un échantillon suivant la loi uniforme [0, 1] et pour chaque observation, nous comparons la variable uniforme U_i à la probabilité de refus *Pr(i)* obtenue selon la discrimination acceptés-refusés par une analyse discriminante.

Si $U_i < Pr(i)$ alors l'observation *i* est considérée comme un dossier rejeté sinon elle est considérée comme un dossier accepté.

Au final, à partir de l'échantillon de 7986 observations, nous avons simulé 6686 dossiers acceptés composés de 89,64% de bons payeurs et 10,36% de mauvais payeurs et 1300 dossiers refusés simulés comprenant 83,92% de bons payeurs et 16,08% de mauvais payeurs.

Pour pouvoir tester la performance des différentes méthodes, nous avons divisé notre échantillon en deux : échantillon d'apprentissage et échantillon test. L'échantillon d'apprentissage est formé de 70% de l'échantillon total soit 5590 dossiers, l'échantillon test représente 30%, soit 2396 dossiers. La subdivision est similaire à celle de NIANG & SAPORTA (2009). La figure n°18 présente les différentes étapes de répartition de notre échantillon de 7986 dossiers de crédit :

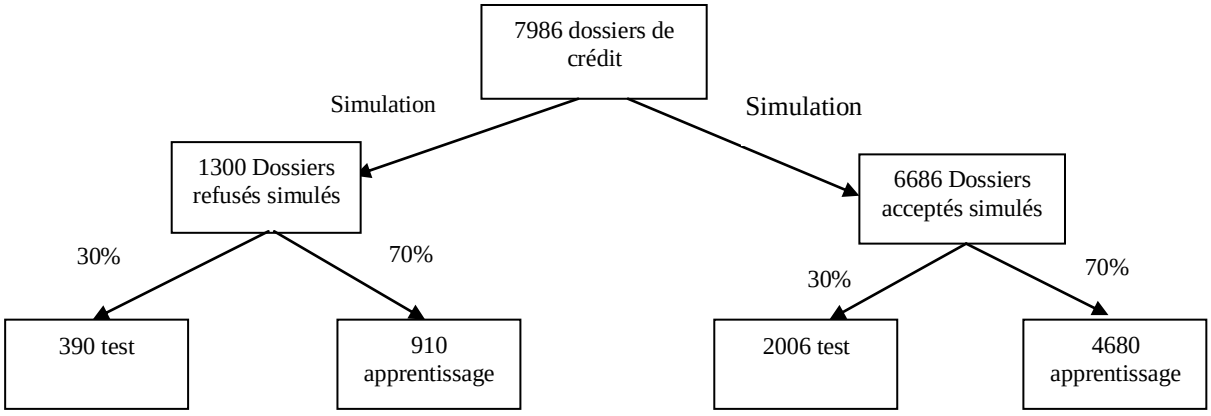


Figure 18 *: La répartition de l'échantillon de dossiers de crédit*

Nous utilisons l'échantillon d'apprentissage pour construire nos modèles de score pour les différentes méthodes.

Pour les méthodes où les dossiers refusés ne participent pas à l'apprentissage du modèle, nous utilisons la totalité (1300 dossiers) des dossiers refusés au lieu des 910 dossiers d'apprentissage.

L'échantillon test est utilisé pour comparer entre ces modèles obtenus en estimant l'erreur de prévision à partir de la matrice de confusion.

Les méthodes de réintégration des refusés peuvent être subdivisées en deux :

- celles qui cherchent à déterminer la catégorie à laquelle appartient les dossiers refusés (bon ou mauvais payeur) et qui les remettent ensuite avec les dossiers acceptés afin de construire le modèle de score à l'aide de l'une des méthodes classiques ou semi supervisées.
- celles où les dossiers refusés ne sont pas intégrés dans l'échantillon final, ces dossiers servent uniquement non pas pour la détermination du modèle de score mais comme étape intermédiaire : calcul des poids pour la méthode de l'augmentation.

2. Les techniques d'inférence des refusés appliquées au problème de réintégration des refusés

Nous donnons ici l'application des techniques d'inférence au problème de réintégration des refusés.

2.1. L'augmentation simple

Nous utilisons l'augmentation simple décrite dans le chapitre 2 (§ II-1). Elle se déroule en quatre étapes :

Étape 1 : L'utilisation de la régression logistique pour la construction du modèle sur la base des 4680 dossiers d'apprentissage acceptés qui sont déjà étiquetés en bons et mauvais payeurs.

Étape 2 : Application du modèle établi sur les 910 dossiers refusés et détermination du taux de défaut de ces derniers.

Étape 3 : Pour chacun des 910 dossiers refusés, si sa probabilité de défaut est supérieure à 0.5, il est considéré comme bon payeur sinon il est considéré comme mauvais payeur..

Étape 4 : Une fois l'échantillon des 910 refusés séparé en bons et mauvais payeurs, il sera mis avec les 4680 dossiers acceptés soit un total de 5590 dossiers. Un nouveau modèle de score est déterminé sur la base du nouvel échantillon (acceptés et refusés)[19].

[19] Cette application a fait l'objet d'une communication (GUIZANI et al. 2011).

2.2. L'augmentation

Nous utilisons ici une variante de la méthode de l'augmentation décrite dans le chapitre 2 (§ II-2), où nous ne départageons pas la population en intervalles mais nous utilisons l'ensemble de la population (acceptés et refusés) (SIDDIQI, 2006).

La méthode se déroule en 3 étapes :

Étape 1 : Construction d'un score d'acceptation à l'aide de la régression logistique pour obtenir la probabilité d'acceptation pour chacun des 5590 individus de l'échantillon d'apprentissage (acceptés et refusés).

Étape 2 : Pondération de chacun des 4680 dossiers acceptés par l'inverse de sa probabilité d'acceptation.

Étape 3 : Construction du modèle de score de défaut "bon ou mauvais" sur les 4680 dossiers acceptés ainsi pondérés.

2.3. La reclassification itérative

Dans la méthode de reclassification itérative (§ II-4 du chapitre 2), nous reprenons les mêmes étapes que celles de l'augmentation simple à chaque itération jusqu'à la stabilisation des scores : Nous commençons par appliquer toutes les étapes de l'augmentation simple jusqu'à obtenir le modèle de régression logistique construit sur la base des 5590 dossiers étiquetés. Nous réappliquons l'augmentation simple avec comme échantillon d'apprentissage les 5590 dossiers et nous déterminons de nouveau le taux de défaut des 910 dossiers refusés. Nous recommençons l'étiquetage des 910 dossiers refusés en bons et mauvais payeurs et nous le comparons avec l'ancien étiquetage des 910 dossiers. La procédure s'arrête lorsque les 910 dossiers gardent la même étiquette d'une itération à l'autre.

2.4. Le parceling

Cette méthode est considérée comme une amélioration de la reclassification itérative. La classification des dossiers refusés se fait proportionnellement à la probabilité de défaut correspondant à l'intervalle de score auquel le dossier appartient (c'est le principe de la méthode du parceling aléatoire pour une base stratifiée présentée dans le chapitre 2 § II-5.3).

On construit les modèles de score par la régression logistique où la variable à expliquer est une probabilité.

Donc pour pouvoir départager la population en intervalles de score, nous avons transformé la probabilité de défaut qui appartient à l'intervalle [0, 1] en un score qui appartient à l'intervalle [0, 1000].

Nous avons départagé notre échantillon en quatre intervalles de score : [0-300[, [300-500[, [500-700[et [700-1000[, le score minimal obtenu étant égal à 149 et le score maximal obtenu étant égal à 998, comme l'indique le tableau n°10:

TABLEAU 10 : *La répartition de l'échantillon des acceptés*

	[0 - 300[	**[300- 500[**	**[500-700[**	**[700-1000[**
Bons acceptés	33	205	698	3264
Mauvais acceptés	14	74	188	204
Pourcentage de Bons acceptés	70.21%	73.48%	78.78%	94.12%
Pourcentage de Mauvais acceptés	29.79%	26.52%	21.22%	5.88%

Nous utilisons les pourcentages obtenus dans le tableau n°10 pour départager la population des refusés en bons et mauvais payeurs comme l'indique le tableau n°11.

TABLEAU 11 : *La répartition de l'échantillon des refusés*

	[0 - 300[	**[300- 500[**	**[500-700[**	**[700-1000[**	**Total**
Refusés	32	138	328	412	910
Mauvais refusés	22	101	258	388	769
Bons refusés	10	37	70	24	141

Une fois les 910 dossiers étiquetés (en 769 mauvais et 141 bons), nous les rajoutons aux 4680 dossiers pour construire le modèle final de score.

2.5. La classification mixte

Pour appliquer la classification mixte (voir § II-7) pour tous les dossiers y compris les dossiers refusés, nous avons commencé par construire une vingtaine de classes par la méthode des k-means. Le nombre de classes est fixé par le critère de Wong et correspond dans notre cas à $5590^{0,3} = 13{,}31$ que nous arrondissons (TUFFÉRY, 2009) à une dizaine près pour arriver à 20 classes homogènes.

Nous calculons par la suite les centres de gravité des 20 classes obtenues. Ensuite nous effectuons une Classification Hiérarchique Ascendante (CAH) (l'indice d'agrégation que nous

utilisons est le critère du saut maximum de Ward dont l'objectif est d'avoir la plus forte inertie interclasse possible.

Le choix du nombre de classe à retenir (c'est à dire le niveau auquel nous choisissons d'arrêter le processus de fusion des classes) se fait selon plusieurs indicateurs statistiques. Ces indicateurs sont :

En premier lieu, nous utilisons le graphique du "R²" qui nous permet de choisir le nombre de classe finale selon le dernier saut important. Dans notre cas, nous avons choisi de nous arrêter à 3 classes selon la figure n°19.

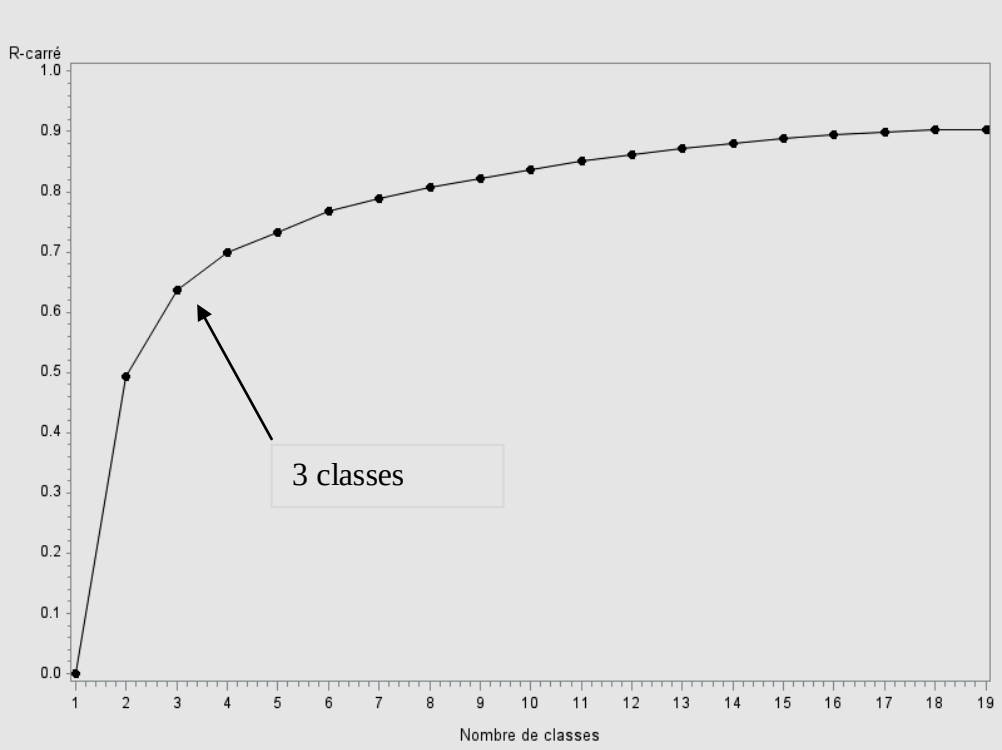


Figure 19 : *Le graphique du R²*

Le second graphique (figure n°20) est celui du "R² semi-partiel". Dans ce cas, la règle du choix de classes est la suivante (TUFFÉRY, 2009) :"*un creux pour k classes et un pic pour k-1 classes, indique une bonne classification en k classes*". Ainsi, nous choisissons 3 classes car la perte d'inertie interclasse est près de 0.06 alors que pour 4 classes, elle est près de 0.035. Le choix de 3 classes serait plus convenable.

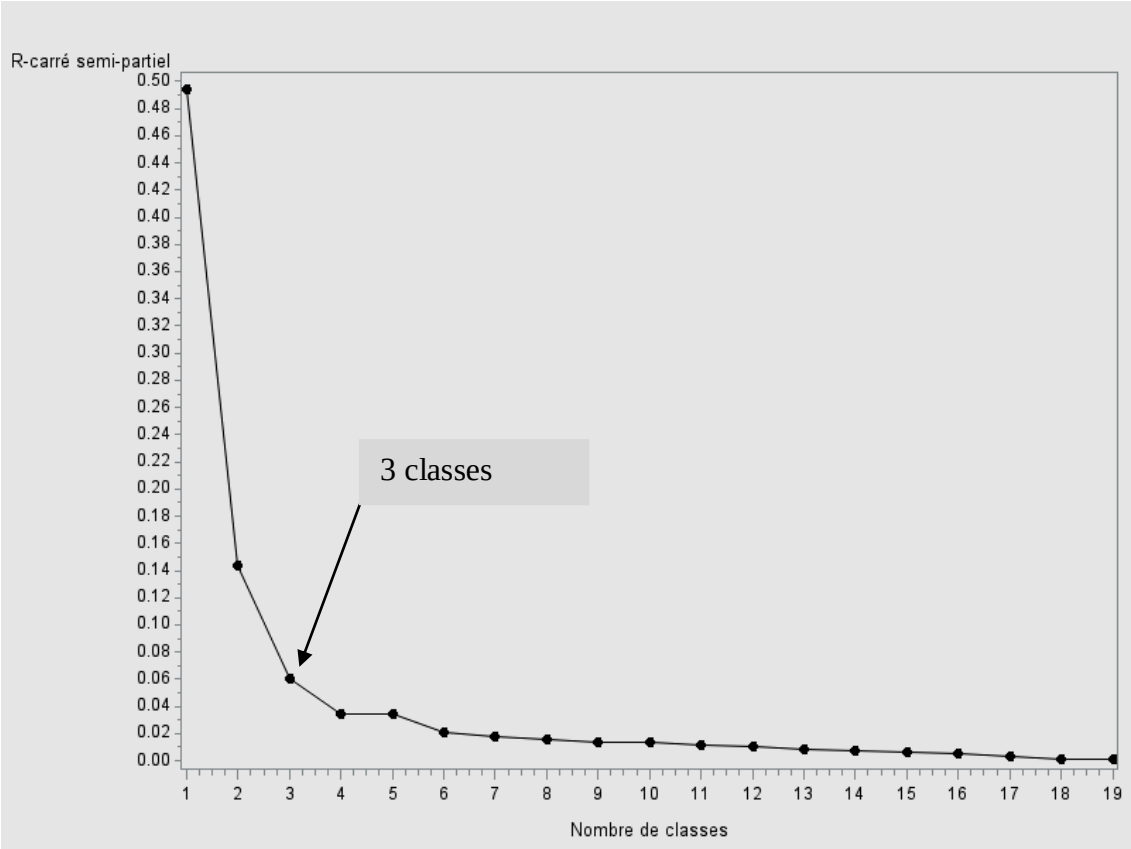


Figure 20 : *Le graphique du R² semi-partiel*

Pour la figure n°21, qui représente le Critère de Classification Cubique (CCC), la règle de décision (TUFFÉRY, 2009) est la suivante :"*un creux pour k classes suivi d'un pic pour k+1 classes indique une bonne classification en k+1 classes*". Donc nous pouvons retenir, selon cette règle, 3 classes.

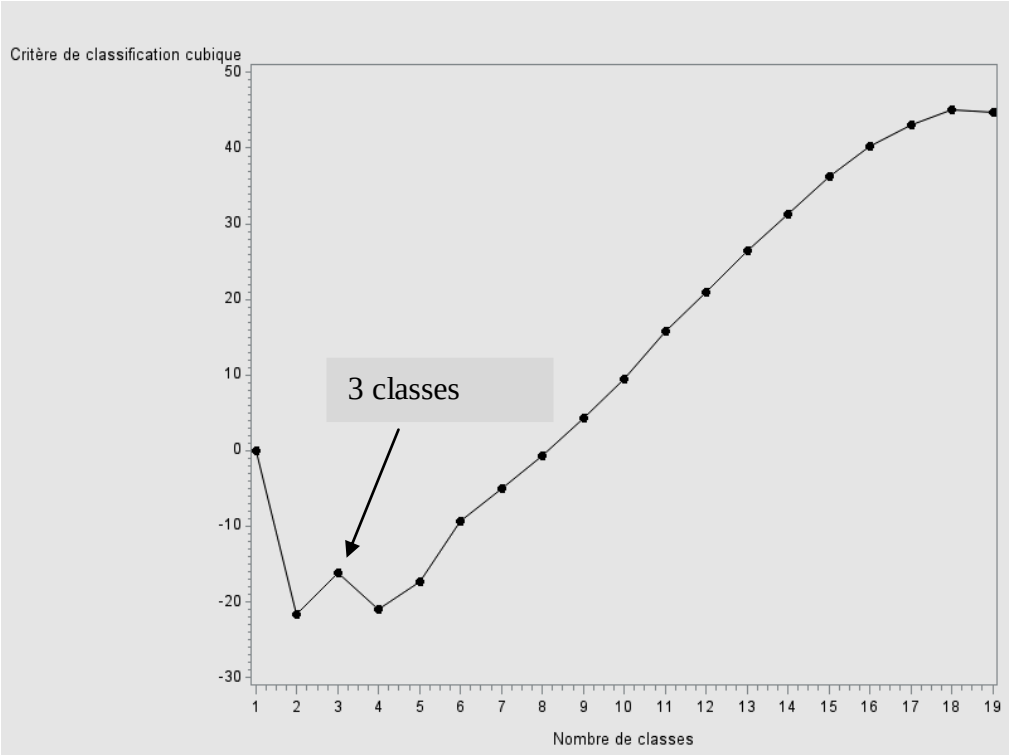


Figure 21 : *Le graphique du Critère de Classification Cubique*

Selon les graphiques précédents, le nombre de classes retenu est de valeurs 3 (la perte d'inertie interclasse est forte voir le graphique du dendrogramme). Donc le nombre optimal de classes à retenir, dans notre cas et selon les indicateurs statistiques explicités plus haut, est fixé à 3 classes finales contenant des dossiers acceptés et des dossiers refusés. Nous pouvons, aussi, afficher le dendrogramme de la CAH (Figure n°22) :

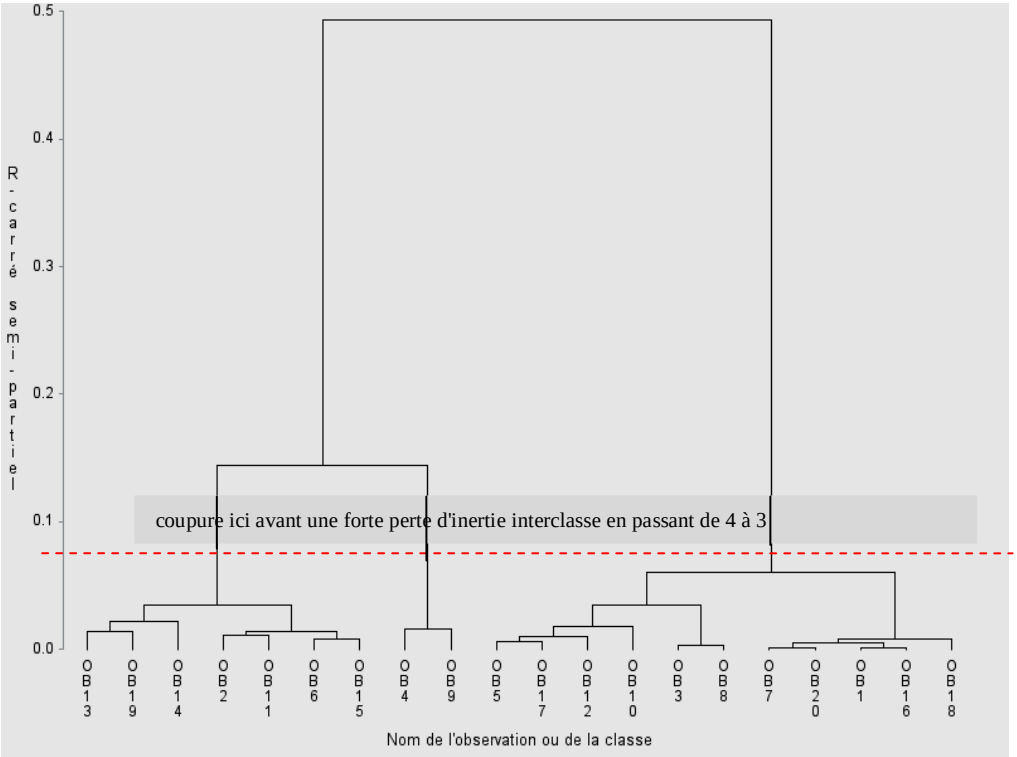


Figure 22 : *Le dendrogramme de la CAH*

Le dendrogramme de la CAH nous permet de visualiser la composition des différentes classes, ainsi que l'ordre de leur formation. Il nous indique aussi la valeur du R^2 semi-partiel de deux classes qui ont été fusionnées à une étape donnée. Dans notre cas, nous choisissons la partition en 3 classes.

Chaque dossier refusé est alors affecté à la catégorie «bon» ou « mauvais » en fonction de la catégorie dominante de sa classe.

2.6. Le modèle Heckman

Dans notre échantillon, les données sont considérées comme tronquées. En effet, ce n'est que pour les personnes ayant octroyé un crédit (ayant un score d'acceptation supérieur à un seuil donné) qu'on peut étudier s'il y a défaut ou non.

En utilisant le modèle de Heckman à deux étapes (chapitre 2 § I-3.2.2.1) (HECKMAN, 1979), notre modèle peut se formaliser comme suit :

Pour chaque individu "i" :

Étape 1 : Définir le mécanisme de sélection des dossiers ayant octroyé un crédit : nous utilisons ici le modèle probit sur la population entière (acceptés et refusés : soit 5590 dossiers pour l'échantillon d'apprentissage) et nous estimons ce modèle par la méthode du maximum de vraisemblance qui permet détiqueter les dossiers en refusés et acceptés.

Étape 2 : Estimation du résultat de l'emprunt (bon ou mauvais payeurs) : cette estimation n'est possible que pour les dossiers acceptés. nous utilisons dans ce cas les moindres carrés ordinaires pour déterminer les coefficients du modèle du résultat de l'emprunt.

2.7. L'auto-apprentissage par les SVMs

Nous nous trouvons face au problème de classification binaire (bon payeur ou mauvais payeur), nous cherchons alors à trouver un hyperplan qui sépare au mieux les bons payeurs des mauvais payeurs (Voir chapitre 2 § III-1.1). Nous disposons, à la fois, des dossiers acceptés dont la variable à prédire est connue et des dossiers refusés dont la variable à prédire est inconnue et que nous cherchons à déterminer. Nous commençons par entrainer un classifieur SVMs en se basant sur les 4680 dossiers acceptés de l'échantillon d'apprentissage. Nous appliquons par la suite le modèle obtenu sur les 1300 dossiers refusés de l'échantillon d'apprentissage pour calculer leur probabilité de défaut en utilisant la fonction logistique de Platt.

Une fois déterminées, ces probabilités de défaut des dossiers refusés, sont affectés à la classe qui leur correspond (bon payeur ou mauvais payeur).

2.8. Le Co-training

L'avantage avec cette méthode (présentée au chapitre 2 § III-1.2) est qu'elle tient compte à la fois des données étiquetées et des données non étiquetées.

Nous rappelons ici que nous travaillons sur les 910 dossiers refusés (par simulation) et que nous savons dès le départ que ce sont des dossiers acceptés et que nous connaissons leur véritable étiquette.

Nous commençons par départager aléatoirement les 4680 dossiers acceptés en deux sous populations égales. Pour chaque sous population, nous effectuons une analyse discriminante. Le classifieur obtenu pour chacune des deux sous populations est appliqué aux 910 dossiers refusés pour les étiqueter en bons ou mauvais payeurs.

Nous comparons alors les véritables étiquettes avec celles que nous avons obtenues et ceci pour chacun des deux classifieurs séparément. Les dossiers refusés qui sont bien classés par l'un des classifieurs sont ajoutés aux 2340 données d'apprentissage de l'autre classifieurs. Nous obtenons 372 dossiers biens classés et 538 dossiers mal classés pour le premier classifieur et 367 dossiers biens classés et 543 mal classés par le second classifieur.

Les 372 biens classés viennent s'ajouter aux 2340 dossiers du second classifieur. Les 367 biens classés du second classifieur viennent s'ajouter aux 2340 dossiers du premier classifieur. On répète le processus avec les nouvelles données pour obtenir deux nouveaux classifieurs que nous appliquons respectivement sur les 538 et les 543 dossiers mal classés. Le processus s'arrête lorsqu'on n'a plus de dossiers mal classés par les deux classifieurs ou redonne les mêmes résultats que l'étape précédente. Dans notre cas au bout de 4 itérations les deux classifieurs n'arrivent plus à classer avec succès les dossiers refusés restants avec un seuil égal à 0.5 pour les probabilités de défaut.

3. Les algorithmes de Boosting appliqués au problème de réintégration des refusés

Nous adaptons ici les algorithmes de Boosting que nous avons présenté au chapitre 2 (§ III-2) pour la réintégration des refusés dans le processus d'octroi de crédit.

En effet les algorithmes de Boosting s'appliquent en général sur des données étiquetés. Dans notre cas, notre échantillon est composé, à la fois, de dossiers acceptés (4680 données étiquetées) et de dossiers refusés (910 données non étiquetées). Pour pouvoir appliquer ces algorithmes de Boosting, nous commençons par une étape préliminaire dont l'objectif est d'étiqueter les dossiers refusés. Nous affectons par la suite les données non étiquetées (dossiers refusés) à des pseudo-classes déterminées par une Analyse Factorielle Discriminante (AFD) (TUFFÉRY, 2012).

La procédure de détermination des pseudo-classes est la suivante : nous déterminons un modèle de score sur la base des dossiers acceptés à l'aide de l'AFD. Ensuite, ce modèle est

appliqué aux dossiers refusés pour déterminer leur probabilité de défaut. Selon cette probabilité chaque dossier refusé est affecté à la classe correspondante avec un seuil égal à 0.5 (bon ou mauvais payeur) d'où le nom de pseudo-classe. À la fin de cette étape tous les dossiers refusés ont leurs pseudo-étiquettes et ainsi l'ensemble des données est étiqueté et les algorithmes de Boosting sont alors applicables.

Les algorithmes de Boosting commencent par pondérer toutes les observations. Généralement quand toutes les données sont étiquetées, le poids initial appliqué pour ce genre de méthode est égal à l'inverse du nombre total d'observations (étiquetés). Dans notre cas, nous allons plutôt appliquer comme poids de départ les proportions réelles des catégories étiquetées et non étiquetées.
Les poids sont alors présentés par la relation (3.1):

$$p_i = \begin{cases} \frac{l}{N} & i \in L \qquad L : \text{données étiquetés} \\ \frac{u}{N} & i \in U \qquad U : \text{données non étiquetées} \end{cases} \tag{3.1}$$

Nous normalisons les p_i pour obtenir les poids w_i avec $(\sum_{i=1}^{N} w_i = 1)$, où l et u sont respectivement les tailles de l'échantillon des données étiquetées (dossiers acceptés) et non étiquetées (dossiers refusés) et $N = u + l$.

Les algorithmes de Boosting commencent tous par prendre des classifieurs dits faibles pour les "entrainer" pour qu'à la fin ils aboutissent à un classifieur dit fort qui n'est qu'une combinaison linéaire des différents classifieurs faibles de chaque itération.

Pour le cas de l'algorithme AdaBoost (présenté dans le chapitre 2 § (III-2.1)), nous avons utilisé comme classifieur faible l'analyse discriminante sur l'ensemble des données (4680 acceptés et 910 refusés étiquetés), soit 5590 dossiers. L'algorithme s'arrête lorsque l'erreur de classification, calculée pour l'ensemble de données mal classées par le modèle, est supérieur à 0,5.

Le classifieur final est une combinaison linéaire des différents classifieurs pondérés α_t. α_t mesure l'importance accordée à chaque classifieur de chaque itération. Lorsque $\varepsilon_t \leq 0{,}5$, alors α_t prend une valeur positive ou nulle et elle augmente quand ε_t diminue (t est la $t^{éme}$ itération et ε_t est l'erreur de classification.

Nous continuons à appliquer par la suite l'algorithme logit Boost ou l'algorithme Gentle AdaBoost de la manière présentée dans le chapitre 2 respectivement § (III-2.2) et § (III-2.3), une fois que les dossiers refusés sont affectés à leurs pseudo-classes comme présenté plus haut. Dans ces deux cas, le classifieur faible utilisé est la régression linéaire qui nous a permis de séparer les valeurs prédites en deux.

4. Critère de comparaison

Pour étudier la performance des modèles que nous avons appliqué, nous utilisons la courbe Receiver Operating Characteristics (ROC) qui relie la proportion de vrais positifs (bons dossiers classés comme tels) à la proportion de faux négatifs (mauvais dossiers classés bons) lorsqu'on fait varier le seuil du score d'acceptation.

Nous avons comparé l'aire sous la courbe ROC (Area Under the Curve : AUC), qui est un indice synthétique de performance, pour chaque méthode. Le modèle qui est jugé le plus performant est celui dont l'AUC est la plus grande.

Nous avons aussi comparé l'air sous la courbe ROC pour chaque méthode appliquée sur les 1300 dossiers refusés.

Nous avons testé le pouvoir prédictif des différentes méthodes sur les 2396 dossiers de l'échantillon test en calculant le taux de bon classement.

Et finalement, pour tester la stabilité des modèles établis pour les différentes méthodes exposées dans ce chapitre, nous avons répété 50 fois la simulation du processus de refus pour obtenir 50 échantillons simulés différents. Pour chaque échantillon nous avons calculé l'indice AUC des différentes courbes ROC de chaque méthode et ceci pour comparer les différentes valeurs de l'AUC obtenues et voir si l'ordre de performance reste le même pour les 50 échantillons.

Les résultats obtenus sont reportés dans des "box-plots" ou "boîte à moustaches" qui représente un graphique permettant la visualisation des valeurs des AUC de chacune des onze méthodes à part pour les 50 échantillons simulés.

III. Résultats et interprétations

1. Performance des modèles après réintégration des dossiers refusés

La figure n°23 (a) représente les différentes courbes ROC relatives aux techniques de réintégration des refusés pour les méthodes classiques. La méthode de la classification mixte (AUC = 0.7397) est considérée comme la plus performante de toute les méthodes classiques.

L'augmentation simple, la reclassification itérative, l'augmentation et le modèle Heckman donnent des modèles de score performants avec des AUCs de l'ordre de 0.73 (respectivement 0.7370, 0.7349, 0.7304 et 0.7283).

Le processus de la méthode de la reclassification itérative s'arrête après 10 itérations lorsque les scores se stabilisent.

En dernière position vient la méthode du parceling qui possède un AUC égale à 0.6543.

Nous donnons par la figure 23 (a) les courbes ROC des différentes méthodes classiques que nous avons utilisé.

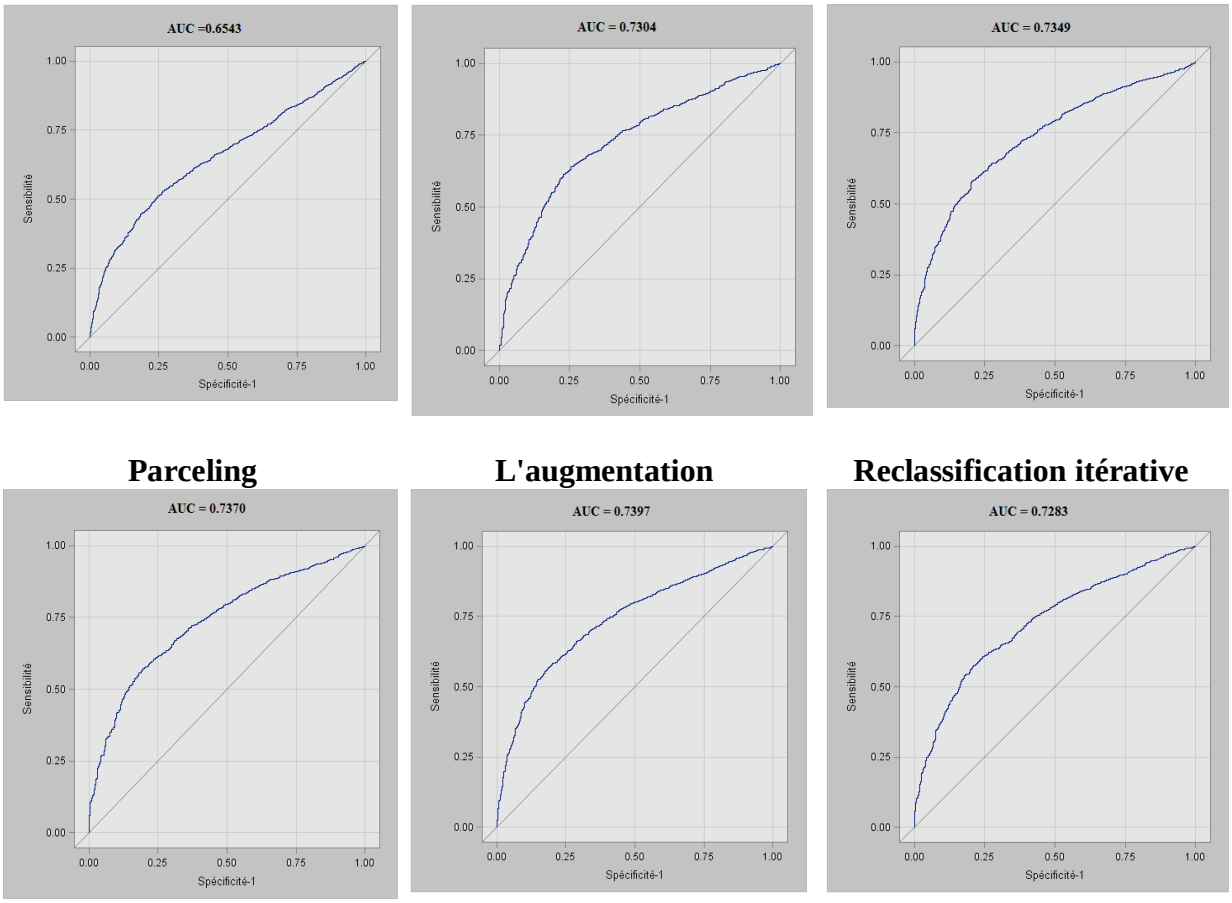


Figure 23 (a) *: Courbes ROC après réintégration des refusés pour les méthodes classiques*

La figure n°23 (b) représente les différentes courbes ROC relatives aux techniques de réintégration des refusés pour les méthodes semi supervisées. La méthode de l'auto-apprentissage par les SVMs (AUC = 0.9141) est considérée comme la plus performante de toute les méthodes semi supervisées.

Les trois méthodes de Boosting à savoir l'AdaBoost (AUC = 0.6945), le Gentle AdaBoost (AUC = 0.7412) et le Logit Boost (AUC = 0.7435) donnent des résultats plutôt performants. La méthode de Co-training donne aussi des bons résultats avec un AUC égale à 0.7383.

Pour le Co-training, les deux classifieurs issus de cette méthodes ne classent plus avec succès les dossiers refusés restants après 4 itérations.

L'algorithme AdaBoost s'arrête au bout de 10 itérations lorsque l'erreur de classification atteint 0.5. Pour l'algorithme Logit Boost, le processus s'arrête à la deuxième itération lorsque les poids "$w_t(x_i)$" des observations et les valeurs attendues du classifieur faible pour chaque

observation "$z_t(x_i)$ " deviennent constants. Pour l'algorithme Gentle AdaBoost, les valeurs prédites restent constantes et ne s'améliorent plus après 12 itérations.

Nous donnons par la figure 23 (b) les courbes ROC des différentes méthodes semi supervisées que nous avons adapté.

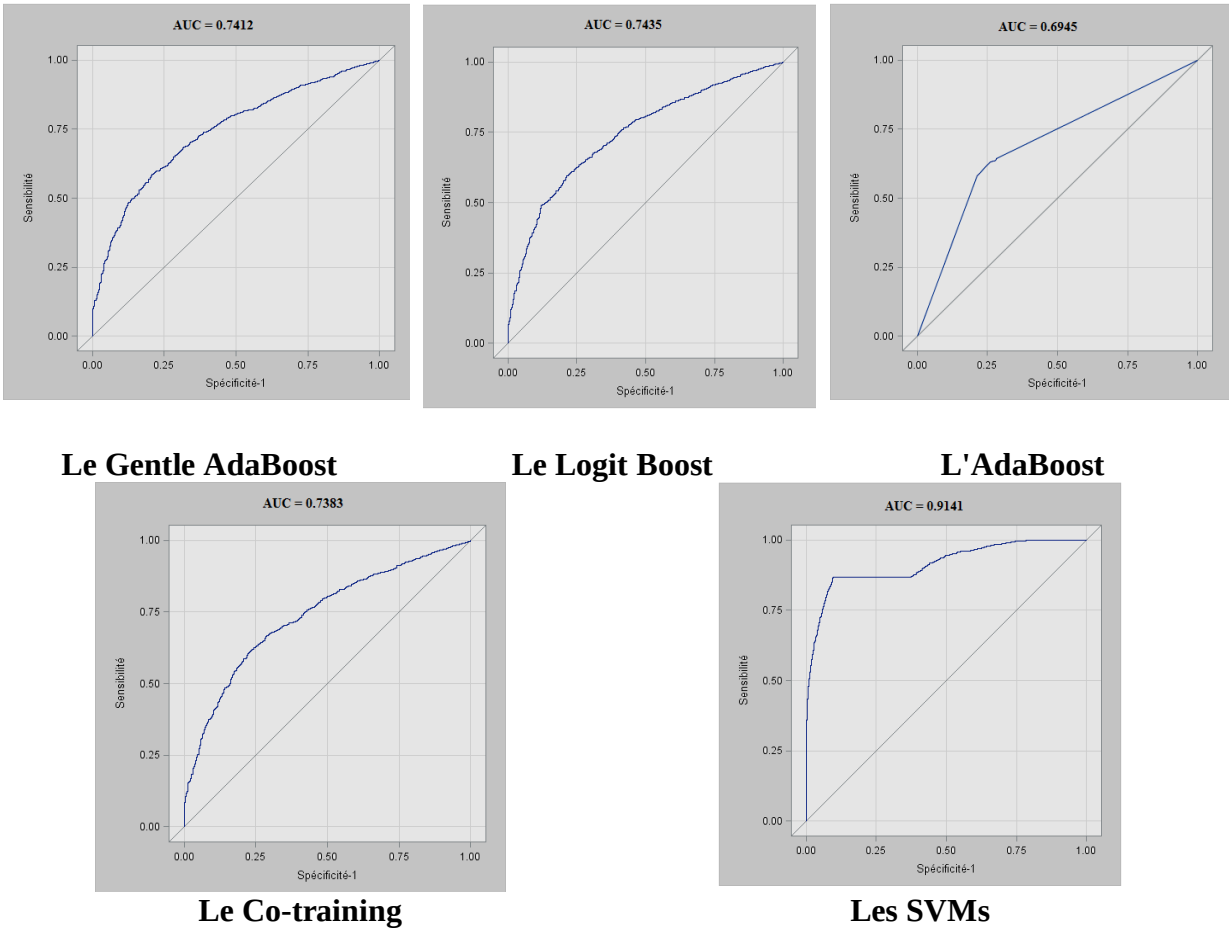


Figure 23 (b) : *Courbes ROC après réintégration des refusés pour les méthodes semi-supervisées*

Nous pouvons conclure que la meilleure méthode est celle de l'auto-apprentissage par les SVMs. Les algorithmes de boosting (logitBoost et Gentle AdaBoost) viennent après et ont un pouvoir prédictif presque identique et meilleur que toutes les autres méthodes. Le Co-training donne un modèle de score dont le pouvoir prédictif est presque équivalent à ceux des méthodes classiques (classification mixte, augmentation simple, reclassification itérative, augmentation) et il reste meilleur que le modèle de Heckman et le parceling.

Les méthodes de boosting, le Co-training et la reclassification itérative donnent des bons résultats. Ces résultats s'expliquent par le fait que ces méthodes ont tendance à améliorer leurs classifieurs d'une itération à l'autre.

Seule l'algorithme AdaBoost est considéré comme le moins performant parmi les méthodes semi supervisée et les méthodes classiques (classification mixte, augmentation simple, reclassification itérative, augmentation, modèle de Heckman) à l'exception de la méthode de parceling.

2. Performance des modèles pour les dossiers refusés

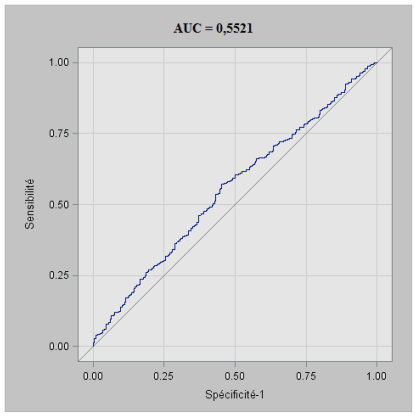
La figure 24 (a) représente les différentes courbes ROC des 1300 dossiers refusés relatives aux techniques de réintégration des refusés pour les méthodes classiques.

Les meilleures méthodes (qui prédisent assez bien les classes des dossiers refusés) sont l'augmentation simple et la reclassification itérative qui donnent presque les mêmes résultats (respectivement AUC = 0.6755 et AUC = 0.6754).

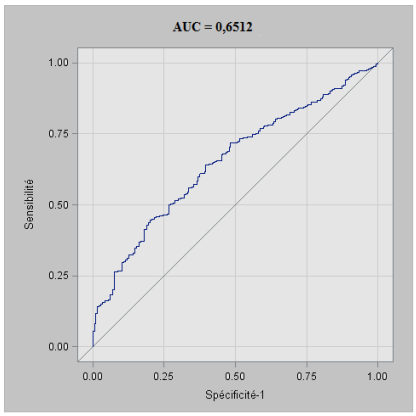
Viennent ensuite la méthode de la classification mixte avec un AUC égal à 0.6631 et l'augmentation (AUC = 0.6512).

Pour la méthode du parceling qui possède un AUC égale à 0.5521 et celle du modèle de Heckman avec un AUC égal à 0.5898. Nous pouvons conclure que leur pouvoir prédictif de ces deux dernières n'est pas meilleur qu'une affectation aléatoire des dossiers refusés dans les deux catégories de dossiers (bons ou mauvais payeurs) et que les deux modèles de score ne sont pas du tout discriminants.

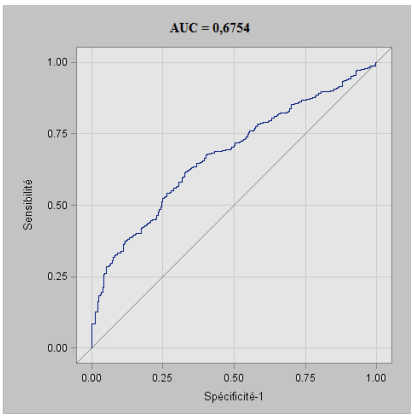
Nous donnons par la figure 24 (a) les courbes ROC des six méthodes classiques étudiées pour les dossiers refusés pour comparer le pouvoir prédictif des différents modèles de score sur les dossiers refusés.



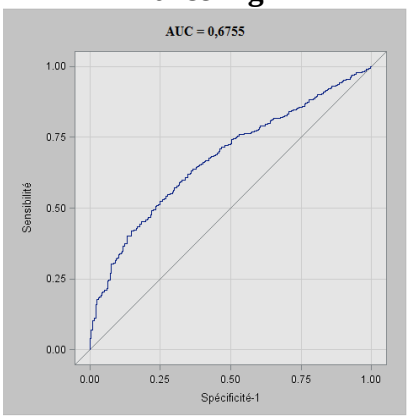
Parceling



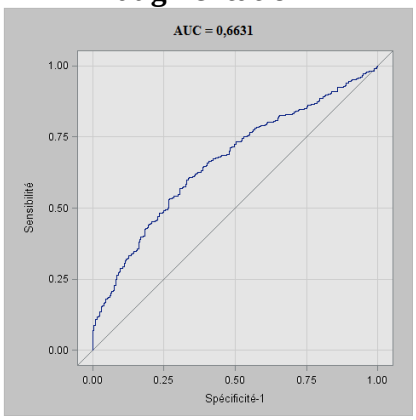
L'augmentation



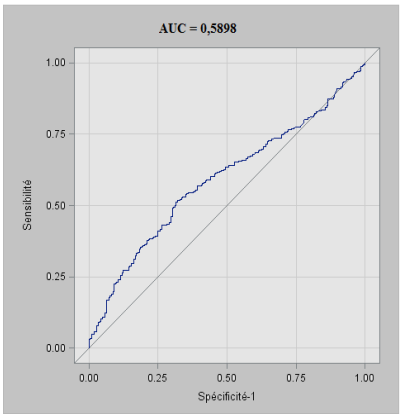
Reclassification itérative



Augmentation simple



Classification mixte



Heckman

Figure 24 (a) : *Courbes ROC des 1300 dossiers refusés des méthodes classiques*

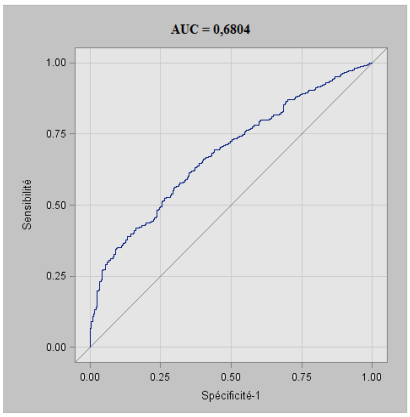
La figure 24 (b) représente les différentes courbes ROC des 1300 dossiers refusés relatives aux techniques de réintégration des refusés pour les méthodes semi-supervisées.

Les meilleures méthodes (donnant des modèles de score qui prédisent assez bien les classes des dossiers refusés) sont l'algorithme LogitBoost (AUC = 0.6860) et l'algorithme Gentle AdaBoost (AUC = 0.6804).

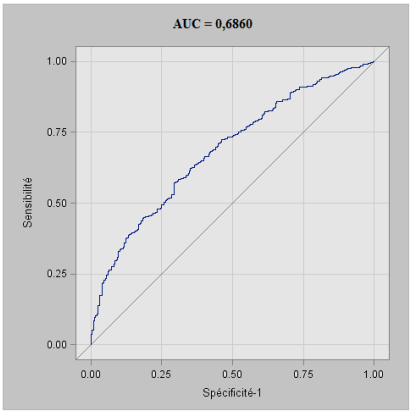
La méthode de Co-training (AUC = 0.6730) et l'algorithme AdaBoost avec un AUC égal à 0.6379 sont moins performantes que les deux premières.

La méthode qui prédit moins bien les dossiers refusés est celle de l'auto-apprentissage par les SVMs avec un AUC égal à 0.6121 malgré qu'elle donne la meilleure valeur de L'AUC sur toute la population (dossiers acceptés et dossiers refusés).

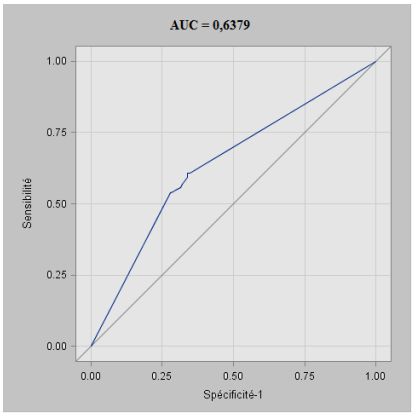
Nous donnons par la figure 24 (b) les courbes ROC des cinq méthodes semi-supervisées étudiées pour les dossiers refusés pour comparer le pouvoir prédictif des différents modèles de score sur les dossiers refusés.



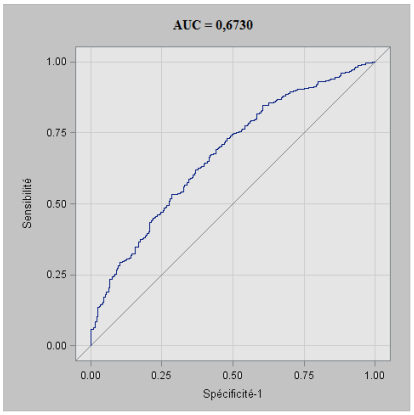
Le Gentle AdaBoost



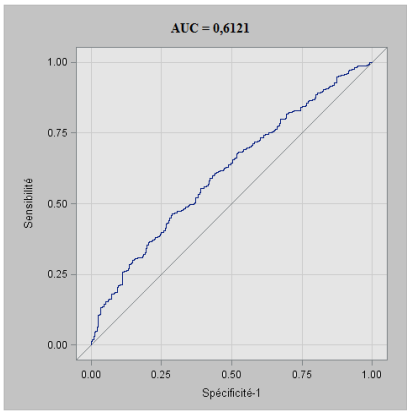
Le Logit Boost



L'AdaBoost



Le Co-training



Les SVMs

Figure 24 (b) : *Courbes ROC des 1300 dossiers refusés des méthodes semi-supervisées*

Nous pouvons conclure que les meilleures méthodes qui prédisent assez bien les classes des dossiers refusés sont celles du boosting à savoir l'algorithme LogitBoost et l'algorithme Gentle Boost.

L'augmentation simple, la reclassification itérative et la méthode de Co-training donnent presque les mêmes résultats avec un AUC de l'ordre de 0.67.

Viennent ensuite la classification mixte avec un AUC égal à 0.6631, l'augmentation (AUC = 0.6512) et l'algorithme AdaBoost avec un AUC égal à 0.6379.

La méthode qui prédit moins bien les dossiers refusés est celle de l'auto-apprentissage par les SVMs ce qui est contradictoire aux résultats trouvés pour la population entière.

Les plus mauvaises méthodes et qui donnent des modèles de score qui ne sont pas du tout discriminants sont la méthode du parceling et celle du modèle de Heckman.

3. Calcul des taux de biens classés dans l'échantillon test

Pour le calcul des taux de biens classés, nous utilisons les 390 dossiers refusés et les 2006 dossiers acceptés soit 2396 dossiers (voir figure n°18).

Nous rappelons que la véritable étiquette est connue et que ces dossiers sont classés comme tels par simulation. Les 2396 dossiers sont utilisés comme échantillon test pour la détermination des taux de biens classés.

Nous donnons dans le tableau n°12 les taux que nous avons obtenu pour un seuil fixé à 0.5 :

TABLEAU 12 : *Les taux de bien classé pour les 11 méthodes*

	Taux de bien classé	
	Bon payeur	**Mauvais payeur**
Augmentation	77.82%	4.29%
Parceling	80.68%	10.73%
Augmentation simple	88.19%	8.89%
Classification mixte	88.19%	8.89%
Classification itérative	80.68%	2.42%
Heckman	74.83%	2.92%
Gentle AdaBoost	88.15%	0.08%
AdaBoost	54.97%	8.47%
LogitBoost	83.72%	2%
Cotraining	52%	8.89%
SVMs	87.40%	5.84%

Nous pouvons remarquer que pour toutes les méthodes, sauf celle de l'AdaBoost et le Co-training, les modèles prédisent bien la catégorie de bon payeur. Alors que pour la catégorie de mauvais payeurs, seules le parceling, l'augmentation simple, la classification mixte, l'AdaBoost et le Co-training donnent des bons résultats. Les taux de bien classé pour la catégorie des mauvais payeurs sont assez faibles, cela s'explique par le faible effectif de cette catégorie : 265 dossiers seulement soit 11,07% contre 2131 dossiers classés bons payeurs soit 88,93%.

4. Comparaison des différentes méthodes sur les échantillons simulés

Pour comparer la dispersion des valeurs de l'AUC pour les onze méthodes, que nous appliquons dans ce chapitre, après les 50 simulations, nous utilisons les boîtes à moustaches (voir figure n°25).

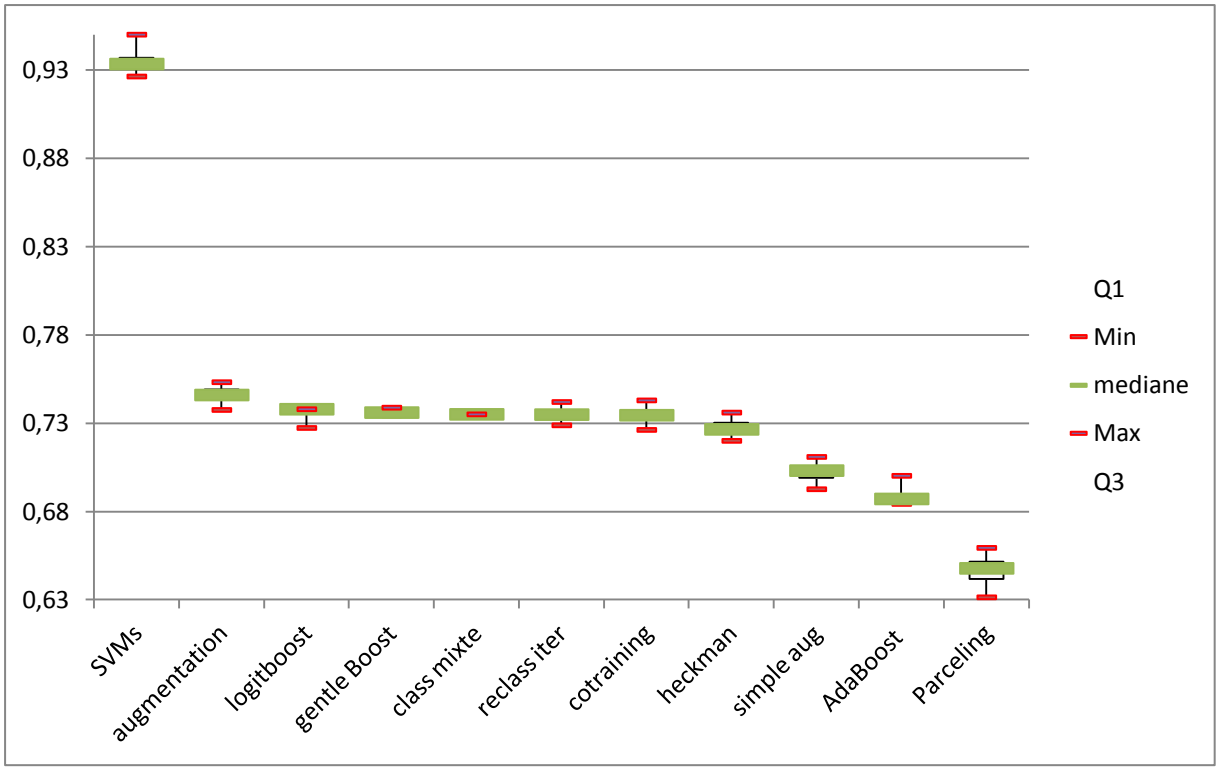


Figure 25 : *Boîtes à moustaches des 11 méthodes*

Nous remarquons que la méthode d'auto-apprentissage par les SVMs donne le modèle de score avec le pouvoir prédictif le plus élevé.

Les autres méthodes semi supervisées (Logitboost, Gentle Adaboost, Cotraining) semblent donner des modèles dont le pouvoir prédictif est équivalent ou presque à ceux des méthodes classiques (augmentation, classification mixte, reclassification itérative, Heckman) et elles restent meilleures que l'augmentation simple et le parceling.

Seule la méthode AdaBoost est moins performante que les méthodes classiques (l'augmentation et l'augmentation simple).

Nous remarquons aussi que les valeurs de l'indice AUC restent presque constantes pour les onze méthodes.

IV. Conclusion

Nous avons appliqué six méthodes classiques du traitement des dossiers refusés (augmentation, augmentation simple, reclassification itérative, parceling, classification mixte et le modèle de Heckman). Nous avons aussi utilisé cinq méthodes issues de l'apprentissage semi-supervisé, l'AdaBoost, le LogitBoost et le Gentle AdaBoost, le Co-training et l'auto-apprentissage par SVMs que nous avons adapté au problème de traitement des refusés. Nous avons étudié la performance de ces différentes méthodes en comparant les différentes valeurs de l'indice AUC des courbes ROC établies pour les onze méthodes.

La meilleure méthode est celle de l'auto-apprentissage par les SVMs. Les modèles de score associés aux méthodes LogitBoost et Gentle AdaBoost ont un pouvoir prédictif presque identique et meilleur que toutes les autres méthodes. Le Co-training donne un modèle de score dont le pouvoir prédictif est presque équivalent à ceux des méthodes classiques (classification mixte, augmentation simple, reclassification itérative, augmentation) et il reste meilleur que le modèle de Heckman et le parceling. La performance du modèle de score établi par la méthode AdaBoost est meilleure que celle du modèle de score du parceling mais moindre que toutes les autres méthodes.

Les résultats obtenus sur les données simulés confirment l'ordre de performance des méthodes, qui est resté presque le même. En effet, la méthode la plus performante reste l'auto-apprentissage par les SVMs. Les modèles de score établis par le LogitBoost et le Gentle AdaBoost sont légèrement moins performants que celui de l'augmentation mais ils sont presque identiques à ceux du Co-training et des méthodes classiques (augmentation, classification mixte, reclassification itérative, modèle de Heckman) et elles restent meilleures que l'augmentation simple et le parceling. Le parceling et l'AdaBoost gardent toujours la moins bonne performance prédictive.

Les valeurs de L'AUC sont restées presque constantes pour les 50 échantillons simulés, donc nous pouvons conclure que les modèles de score obtenus pour les différentes méthodes sont stables.

Conclusion

Dans le processus d'octroi de crédit, les banques ne tiennent pas compte des dossiers refusés pour le calcul de leurs scores, nous nous intéressons dans cette thèse à la réintégration des dossiers refusés dans le calcul de ces scores.

Notre approche n'est pas basée uniquement sur l'étude des méthodes classiques de l'inférence des refusés développées dans la littérature, mais aussi sur des méthodes issues de l'apprentissage semi-supervisé que nous adaptons au problème de traitement des refusés.

Les accords de Bâle ont été mis en place pour éviter de nouvelles crises financières dues à la mauvaise évaluation du risque de crédit par les institutions financières.

Pour combler le manque à gagner des banques du au biais de sélection causé par le rejet des dossiers potentiellement bons. Nous nous proposons ici de réintégrer ces dossiers refusés dans le processus d'octroi de crédit. Ce biais provient du fait que les banques calculent leurs scores sur la base des dossiers acceptés uniquement et n'ont pas de suivi pour les dossiers refusés.

Nous avons appliqué six méthodes classiques du traitement des dossiers refusés (augmentation, augmentation simple, reclassification itérative, parceling, classification mixte et le modèle de Heckman). Nous avons aussi utilisé cinq méthodes issues de l'apprentissage semi-supervisé, l'AdaBoost, le LogitBoost et le Gentle AdaBoost, le Co-training et l'auto-apprentissage par SVMs, que nous avons adapté au problème de traitement des refusés.

L'analyse discriminante sur un échantillon réel a permis de déterminer les probabilités de refus que nous avons appliqué sur les dossiers acceptés uniquement en simulant le processus de refus.

Une comparaison des courbes ROC de ces méthodes nous a permis de conclure que la méthode la plus performante est celle de l'auto-apprentissage par les SVMs.

Les autres méthodes semi supervisées (Logitboost, Gentle Adaboost, Co-training) semblent donner des modèles dont le pouvoir prédictif est équivalent ou presque à ceux des méthodes

classiques (augmentation, classification mixte, reclassification itérative, modèle de Heckman) et elles restent meilleures que l'augmentation simple et le parceling.

Seule la méthode AdaBoost est moins performante que les méthodes classiques (sauf pour le parceling).

En vue de valider nos résultats, nous avons calculé, pour chaque méthode, le taux de bon classement sur un échantillon test. Les résultats de cette étude ont confirmé nos conclusions.

Nous avons testé la stabilité des méthodes utilisées par simulation.

L'application des méthodes adaptées au traitement des dossiers refusés nous a permis de mettre en évidence la nécessité de traiter le problème du biais de sélection. En effet, le fait de travailler sur un échantillon représentatif de la population entière rend les estimations des modèles de score plus performantes.

Nous incitons également les banquiers à se lancer dans le traitement des dossiers refusés pour optimiser leurs gains en optimisant le nombre de clients potentiels.

Sur le plan empirique, nous avons procédé à plusieurs études pour tester la performance des différentes méthodes que nous avons appliqué. L'objectif de ces études est de pouvoir comparer ces méthodes, ce qui constitue la principale contribution de ce travail. Cette partie a démontré que l'application des méthodes issues de l'apprentissage semi-supervisé donnent des modèles de score meilleures que ceux des méthodes classiques.

À l'issue des recherches effectuées, de nombreuses perspectives peuvent être étudiées pour approfondir notre étude : l'intérêt de déterminer les variables qui expliquent aux mieux le rejet de certains dossiers de crédit c'est à dire qu'il faut faire des statistiques descriptives sur l'ensemble des variables utilisé pour notre étude pour pouvoir déterminer celles les plus discriminantes et pertinentes.

La simulation d'autres processus de refus pour tester la robustesse et la stabilité des modèles établis dans cette thèse dans le but d'atteindre un réalisme accru.

Le développement d'autres méthodes semi-supervisées et leur adaptation au problème de réintégration des refusés. En plus des méthodes de Boosting, de nouvelles tendances sont apparues pour la classification comme celles du Bagging..

L'intérêt de répondre à la question "quand" la défaillance aura lieu, mais pas seulement chercher à déterminer "si" il y aura défaut. En effet, la discrimination entre les deux catégories bons ou mauvais payeurs n'est plus le seul but à atteindre pour les banques et ceci surtout pour les prêts à long terme. Ces dernières s'intéressent aussi à la déterminer de la période où le client fera défaut. il est intéressant dans ce cas de développer des modèles de survie pour les données censurées qui permettra de résoudre le problème des données incomplètes comme l'a expliqué Saporta (2006)[20]. Dans notre cas, nous pourrons utiliser les techniques issues de l'analyse de survie en considérant le défaut de payement comme l'événement cible et les dossiers refusés comme des données censurées.

[20] http://cedric.cnam.fr/~saporta/scoring.pdf

Références bibliographiques

ALTMAN, E.I. Financial ratios discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 1968, 589-609.

ANDERSON, R. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. New York, Oxford University Press, 2007

ARTUS, P., BETBEZE, J.P., DE BOISSIEU, C. & CAPELLE-BLANCARD, G. *La crise des subprimes, Conseil d'Analyse Économique*. Paris, 2008.

ASH, D. & MEESTER, S. Best practices in reject inferencing. *Conference credit risk modeling and decisioning*, Wharton FIC, University of Pennsylvania, 2002.

BAESENS, B. VAN GESTEL, T. VIAENE, S. STEPANOVA, M. SUYKENS, J. & VANTHIENEN, J. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 2003, 54, 6, 627–635

BANASIK, J. & CROOK, J. Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, 2007, 183, 1582-1594.

BEAVER, W.H. Alternative accounting measures as predictors of failure. *The Accounting Review*, 1968, 113-122.

BENNETT, K. P., DEMIRIZ, A. & MACLIN, R. Exploiting Unlabelled Data in Ensemble Methods. *Proceeding of the Eighth ACM International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, 2002.

BENNETT, K.P. & DEMIRIZ, A. Semi Supervised Support Vector Machines. *Proceedings of Neural Information Processing Systems*, Denver, 1998.

BLUM, A. & MITCHELL, T. Combining Labeled and Unlabeled Data with Co-training. COLT: *Proceedings of the Workshop on computational Learning Theory*, New York, 1998, 92-100

BOUSLAMA, G., BOUTEILLER, C., LE GUIRRIEC-MILNER, G. & VERNIER, E. *La crise financière en 40 concepts clés*. Paris : REVUE BANQUE Édition, 2009.

BREIMAN, L. FRIEDMAN, J.H. OLSHEN, R.A. & STONE, C.J. *Classification and Regression Trees*. Wadsworth, Canada, 1984.

CHAPELLE, O., ZIEN, A. & SCHOLKOPF, B. Semi Supervised Learning. *The MIT Press*, 2006.

CHAWLA, N.V. & KARAKOULAS, G. Learning from Labeled and Unlabeled Data: An

Empirical Study Across Techniques and Domains. *Journal of Artificial Intelligence*, 2005, 23, 331-366.

CONAN, J. & HOLDER, M. *Variables explicatives de performances et contrôle de gestion dans les PMI*. Thèse de Doctorat d'Etat, Université Paris Dauphine,1979.

CROOK, C. & BANASIK, J. Does Rejct Inference Really Improve the Performance of Application Scoring Models?. *Journal of Banking and Finance*. 2004, 28, 857-874.

DE SERVIGNY, A. & ZELENKO, I. *Le risque de crédit face à la crise*. Paris : DUNOD, 2010.

DEMPSTER, A.P., LAIRD, N.M. & RUBIN, D.B. Maximum Likelihood from incomplete data. *Journal of the Royal Statistical Society*, 1977, 39, 1-38.

DUMONTIER, P. DUPRE, D. & MARTIN, C. Gestion et contrôle des risques bancaires : L'apport des IFRS et de Bâle II. Paris : *Revue Banque Edition*, 2008.

DURAND, D. *Risk Elements in Consumer Instalment Financing*. National Bureau of Economic Research, 1941.

EISENBEIS, R.A. Pitfalls in the Application of Discriminant Analysis in Business. Finance and Economics, *Journal of Finance*, 1978, 32, 3, 875-900.

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

FEELDERS, A.J. An overview of model based reject inference for credit scoring. *BANFF Credit Risk Conference*, Canada, 2003.

FELDMAN, R. Small business loans, small Banks and Big Change in Technology called credit scoring. *Federal Reserve Bank of Minneapolis, The Region*, 1997, 19-25.

FISHER, R.A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 1936, 7, 179-188.

FLACH, P.A., HERNÁNDEZ-ORALLO J., & FERRI, C. A coherent interpretation of AUC as a measure of aggregated classification performance. *In Proceedings of the 28th International Conference on Machine Learning*, 2011.

FREUND, Y. & SCHAPIRE, R. E. A Decision Theoretic generalization of on Line Learning and an Application to Boosting. *Second European Conference on Computational Learning Theory*, 1995.

FREUND, Y. & SCHAPIRE, R.E. A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, 1999, 14, 5, 771-780.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. Additive Logistic Regression: A Statistical View of Boosting. *The Annals of Statistics*, 2000, 28, 2, 337-407.

GUIZANI, A., BENAMMOU, S. & SAPORTA, G. Une méthode de traitement des refusés

dans le processus d'octroi de crédit. *43éme édition des Journées de Statistique, Tunisie, 2011*.

HAND, D. J. & HENLEY W. E. Can reject inference ever work?. *IMA Journal of Mathematics Applied in Business and Industry*, 1993, 4, 5, 45-55.

HAND, D.J. *Construction and Assessment of Classification Rules*. New York, Wiley, 1997.

HAND, D.J. Measuring Classifier Performance: a Coherent Alternative to the Area Under the ROC Curve. *Machine Learning*. 2009, 77, 103-123.

HAND, D.J. L'incohérence de l'aire sous la courbe ROC, que faire à ce propos ? *Revue Modulad*, 2010, 42, 74-80.

HECKMAN, J.J. Sample Selection Bias as a Specification Error. *Econometrica*, 1979, 47, 1, 153-161.

HULL, J., GODLEWSKI, C. & MERLI, M. *Gestion des risques et institutions financières*. Edition Pearson Education, 2007.

JOANES, D.N. Reject Inference Applied to Logistic Regression for Credit Scoring. *IMA Journal of Mathematics Applied in Business and Industry*, 1993/4, 5, 35-43.

LAANAYA, H., MARTIN, A., ABOUTAJDINE, A. & KHENCHAF, A. *Régression floue et crédibiliste par SVM pour la classification des images sonar*. Traitement du signal, 2008, 25, 1-2.

LEBART, L., MORINEAU, A. & PIRON, M. *Statistique exploratoire multidimentionnelle*. Dunod, Paris, 1995.

LITTLE, R. J. A. & RUBIN, D. B. *Statistical analysis with missing data*. Hoboken, John Wiley, New York, 1987.

MALDONADO, S. & PAREDES, G. A Semi Supervised Approach for Reject Inference in Credit Scoring using SVMs. *Lecture notes in computer science, advances in Data Mining Applications and Theoretical Aspects*, 2010, 6171, 558-571.

MALHOTRA, R. & MALHOTRA, D. K. Evaluating consumer loans using Neural Networks. *Omega the International Journal of Management Science*, 2003, 31, 2, 83-96.

MC CULLOCH, W. S. & PITTS, W. H. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 1943,5, 115-133.

MEESTER, L.J. *what's the point of credit scoring?.* Business Review, Federal Reserve Bank of Philadelphia, 1997, 3-16.

METZ, C. E. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 1978, 8, 283-298.

MITCHELL, M. Genetic Algorithms: An Overview. *Complexity*, 1995, 1, 1, 31.

NIANG, N. & SAPORTA, G. Supervised Classification and AUC. Workshop Franco-Brésilien sur la fouille de données, Brésil, 2009.

NOYER, C. Bâle II : Genèse et enjeux. *Conférence- débat, association d'économie financière*, 2004.

PLATT, J. Probability Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers, MIT Press, Cambridge*, 1999, 61-74.

PONTIL, M. & VERRI, A. Support Vector Machines for 3D Object Recognition. *Browse Journals and Magazines, Pattern Analysis and Machine*, 1998, 20, 6, 637-646.

PROVOST, F. & FAWCETT, T. *Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In: Proc. Third International Conference on Knowledge Discovery and DataMining (KDD-97)*, California, 1997, 43–48.

PUHANI, P. A. The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 2000, 14, 1, 53-68.

REICHERT, A.K., CHO, C.C. & WAGNER, G.M. An Examination of the Conceptual Issues Involved in Developing Credit Scoring Models. *Journal of Business and Economic Statistics*, 1983, 1, 101-114.

ROSENBERG, E. & GLEIT, A. Quantitative methods in credit management: A survey. *Operations Research*, 1994, 42, 589–613.

SADI, N.E. *Analyse financière d'entreprise : méthodes et outils d'analyse et de diagnostic en normes françaises et internationales IAS – IFRS*. Editions L'Harmattan, 2009.

SAPORTA, G. *La notation statistique des emprunteurs ou SCORING*. Document réalisé pour la commission "Kahane" de réflexion sur l'enseignement des mathématiques, 2006.

SAPORTA, G. *Probabilités, analyse des données et statistique*. Edition TECHNIP, 2011.

SCHREINER, M. *Les vertus et les faiblesses de l'évaluation statistique (crédit scoring) en microfinance*. Center for Social development, Washington University, 2003.

SIDDIQI, N. *Credit risk scorecards developing and implementing intelligent credit scoring*. John Wiley & Sons, Inc., New Jersey, 2006.

SONQUIST, J.A. & MORGAN, J.N. *The detection of interaction effects*. Institut of Social Research Ann. Arbor, University of Michigan, 1964.

SPACKMAN, K.A. Signal detection theory: Valuable tools for evaluating inductive learning. *In: Proc. Sixth International Workshop on Machine Learning*, 1989, California, 160–163.

SUSTERSIC, M., MRAMOR, D. & ZUPAN J. Consumer credit scoring models with limited data. *Expert Systems with Applications*, 2009, 36, 3, 4736-4744.

SWETS, J. Measuring the accuracy of diagnostic systems. *Science*, 1988, 240, 1285–1293.

SWETS, J.A., DAWES, R.M. & MONAHAN, J. Better decisions through science. *Scientific American*, 2000, 283, 82–87.

TENENHAUS, M. *La régression PLS théorie et pratique*. Paris, Editions Technip, 1998.

TENENHAUS, M. Statistique méthodes pour décrire, expliquer et prévoir. Dunod, Paris, 2007.

TUFFÉRY, S. *Une étude de cas en statistique décisionnelle*, Editions Technip, Paris, 2009.

TUFFÉRY, S. *Data Mining et statistique décisionnelle : l'intelligence des données*. Editions Technip, Paris, 4ème édition, 2012.

VAPNIK, V. N. *The nature of statistical learning theory*. Springer, New York, 1995.

VIENNET, E. & FOGELMAN SOULIE, F. Le traitement des refusés dans le risque crédit. *Revue des Nouvelles Technologies de l'Information (RNTI-A-1)*, 2007, 23-45.

VO THI, P. N. *Tarification du crédit : qu'apporte le nouveau ratio de solvabilité ?*. GdR Economie Monétaire et Financière, Nice, 10 et 11 juin 2004, 1-29.

WELLINK, N. Au delà de la crise: la réponse stratégique du comité de Bâle. *Banque de France, Revue de la stabilité financière*, 2009, 13, 131-141.

WOLD, H. Estimation of Principal Components and Related Models by Iterative least Squares, in Multivariate Analysis. New York, *Academic Press*, 1966, 391-420.

WOLD, S. *Three PLS algorithms according to SW, in Report from the symposium MULTDAST (multivariate data analysis in science and technology*. Sweden, Edition Umea, 1984, 26-30.

YU, G.X., OSTROUCHOV, G., GEIST, A. & SAMATOVA, N.F. An SVM-based Algorithm for Identification of Photosynthesis-specific Genome Features. *CSB*, 2003, 235-243.

ZHANG, C. & MA, Y. *Ensemble Machine Learning Methods and Applications*. Springer, 2012.

Sites Web

Bank for International Settlements. *Bâle III : dispositif international de mesure, de normalisation et de surveillance du risque de liquidité*. Décembre 2010. http://www.bis.org/publ/bcbs188_fr.pdf

Bank for International Settlements. *Nouvel accord de Bâle sur les fonds propres.* Avril 2003. http://www.bis.org/bcbs/cp3fullfr.pdf

Bank for International Settlements. *Réponse du comité de Bâle à la crise financière: Rapport au groupe des Vingt*. Octobre 2010. http://www.bis.org/publ/bcbs179_fr.pdf

Bank for International Settlements. Vue d'ensemble du nouvel accord de Bâle sur les fonds propres. Janvier 2001. http://www.bis.org/publ/bcbsca02_f.pdf

Banque de France . Participation à l'évolution du cadre réglementaire. rapport annuel 2010. http://www.acp.banque-france.fr

Basel Committee on Bank Supervision. *Sound Practices for the Management and Supervision of Operational Risk. Bank for International Settlements*, juillet 2002. http://www.bis.org/publ/bcbs91.pdf

BIS, fiche d'information, "Comité de Bâle sur le contrôle bancaire", 2010. http://www.essectransac.com/wp-content/themes/arthemia/images/2010/10/Le-Comité -de-Bâle-sur-le-contrôle-bancaire.pdf

Commission Fédérale des Banques. *Bâle II mise en application en Suisse*. Octobre 2006. http://www.finma.ch/archiv/ebk/f/regulier/rundsch/2006/Erlaeuterungen_BaselII_f.pdf

Credit Research Centre http://www.business-school.ed.ac.uk/crc

DIVVALA, S.K. *Co-training and its applications in vision*. http://homes.cs.washington.edu/~santosh/presentations/coTraining_miscRead_web.pdf

Eurogroup Consulting; « Bâle 3, quels impacts sur les métiers de la banque ? », Avril 2011. http://www.eurogroupconsulting.fr/IMG/pdf/B3_M2_20110422_VF-2-2.pdf

Financial Stability Board. *Reducing the moral hazard posed by systemically important financial institutions*. octobre 2010. http://www.financialstabilityboard.org/publications/r_101111a.pdf

Groupe de travail Bâle II. La réforme Bâle II. Club de la sécurité des systèmes d'information français, 2004, 1-28. http://www.clusif.asso.fr/fr/production/ouvrages/pdf/ReformeBale2.pdf

ISSARD, G. *L'essentiel du nouvel accord de Bâle*. Synthèse du nouvel accord de Bâle,

Banque des Règlements Internationaux, 2004.
http://www.issard.com/download/resume_baleii.pdf

KPMG. Bâle III les impacts à anticiper. Mars 2011.
http://www.kpmg.com/FR/fr/IssuesAndInsights/ArticlesPublications/Documents/Bale_III_impacts_a_anticiper_mars2011.pdf

La règlementation Bâle II.
http://denis.dupre.pagesperso-orange.fr/finance/Extrait-bale2.pdf

Reject Inference Methods
http://www.scorto.com/credit-scoring-article.htm

SAPORTA, G. *Analyse discriminante, classification supervisée, scoring*. 2009.
http://cedric.cnam.fr/~saporta/discriminante.pdf

SAPORTA, G. Classification supervisée et crédit scoring. 2006.
http://cedric.cnam.fr/~saporta/scoring.pdf

SAPORTA, G. *Introduction au Data Mining et à l'apprentissage statistique*.
http://cedric.cnam.fr/~saporta/DM.pdf

Statistical Methods in Retail Financial Services Research Group
http://www2.imperial.ac.uk/~djhand/fingroup.html

Trader-Finance.fr
http://www.trader-finance.fr/lexique-finance/definition-lettre-R/Risque-de-marche.html

ZOU, K.H. *Receiver operating characteristic (ROC) literature research*. 2002.
http://splweb.bwh. harvard.edu:8000/pages/ppl/zou/roc.html

ANNEXES

Liste des annexes

Annexe 1: Les trois piliers de Bâle II

A1.1. Le premier Pilier des accords de Bâle II

Les exigences minimales en matière de couverture des différents risques, par des fonds propres, constituent le premier pilier de Bâle II. Ce pilier introduit un nouveau ratio de solvabilité appelé "Mac Donough" où les fonds propres de la banque doivent être supérieurs à 8% des risques combinés (le risque combiné est formé de 75% du risque de crédit augmenté de 5% de risque de marché et de 20% de risque opérationnel (BIS, "Nouvel accord de Bâle sur les fonds propres", Avril 2003).

$$\frac{\text{Fonds propres effectifs}}{\text{total des actifs pondérés en fonction des risques}} = \frac{\text{Fonds propres effectifs}}{(\text{risque de crédit} + \text{risque de marché} + \text{risque opérationnel})} \geq 8\%$$

Pour chacun de ces trois risques, les établissements ont le choix entre différentes méthodes d'estimation des exigences en fonds propres comme le montre le tableau n°A1.1.

TABLEAU A1.1 : *Les méthodes d'estimation des risques*

	Types de risque		
	Risque de crédit	**Risque opérationnel**	**Risque de marché**
Méthodes	Approche standard Notations internes (NI): - Notations internes simple - Notations internes avancées	Approche de base Approche standard Approche de Mesures avancées (AMA)	Approche standard Modèles internes

(Source :Dumontier et al. 2008)

A1.1.1. Le risque de crédit

Généralement, le risque de crédit est évalué en fonction de certains critères. Bâle II prévoit trois approches pour estimer ce risque. Ces approches permettent d'estimer la solvabilité d'un demandeur de crédit et de déterminer la part des fonds propres nécessaire pour couvrir ce type de risque.

L'accord de Bâle II prévoit un choix au gré des banques entre l'approche standard et l'approche fondée sur les notations internes[21] ("Internal Rating Based Approche" ou "IRB"), cette dernière possédant deux variantes : simple " Foundation " et avancée "Advanced ".

[21] Le recours à l'approche IRB est soumis à l'agrément des autorités de contrôle sur la base de critères définis par le comité de Bâle.

Chacune de ces différentes approches possède son mode de calcul des exigences en fonds propres.

- **L'approche standard**

Selon Dumontier et al. (2008), l'approche standard est basée sur la pondération, ou "Rating", des risques des différents actifs attribuée aux banques centrales, aux assureurs de crédit et aux organismes de garantie, par les agences de notation (telle que Standard & Poor's ou Moody's). Les exigences minimales en fonds propres varient suivant la notation de la catégorie de l'emprunteur. Cette pondération a pour but de sensibiliser les banques aux risques encourus par l'octroi de crédit.

Le comité de Bâle a défini plusieurs catégories d'expositions au risque de crédit, avec pour chaque catégorie une pondération à appliquer à l'encours prêté (le montant total de la créance). Cette pondération va de 0% pour les Etats souverains donc cette catégorie est sans risque, à 150% pour les contreparties les moins bien notées.

Pour atténuer ce risque, les banques font appel à plusieurs types de couverture : prise de sûretés réelles sous la forme de liquidités ou de titres couvrant la totalité ou une partie des expositions, achat de protection sous la forme de garanties ou de dérivés de crédit ou encore accords de compensation des prêts et des dépôts sur une même contrepartie (De Servigny et al. 2010).

- **Les approches par les notations internes (IRB)**

Dans les approches internes, la pondération du risque est déterminée par l'appréciation propre de la banque basée sur une base de données relative à ses clients.

Les techniques de notation interne permettent d'obtenir un "rating" et une probabilité de défaut pour chaque actif. Le capital nécessaire dépend de quatre variables : la probabilité de défaut des emprunteurs ("Probability of Default") (PD), la perte en cas de défaut (LGD) ("Loss Given Default"), l'exposition au défaut ("Exposure At Default") (EAD) et la maturité du crédit ("Maturity") (M) (Dumontier et al., 2008).

Deux méthodes permettent de déterminer ces différents paramètres dans l'approche de notations internes, il existe:

- La première est la méthode simple qui prévoit que les banques utilisent leurs évaluations internes de la probabilité de défaut (PD) de façon à déterminer les exigences de fonds propres alors que les autres paramètres nécessaires au calcul du risque de crédit seront fournis par le comité de Bâle.

- La deuxième méthode, dite avancée, a le même principe que celui de l'approche standard, en revanche la banque calcule elle-même la quasi-totalité des paramètres qu'elle utilise pour évaluer le risque. Dans la plupart des cas, cette méthode est destinée aux plus grands établissements bancaires ayant des systèmes sophistiqués d'allocation de capital et calculant eux-mêmes l'ensemble des paramètres nécessaires.

Dans les deux cas, l'approche interne doit être contrôlée et validée par des instances de régulation qui délivrent un agrément à la banque pour qu'elle applique les pondérations issues de son système interne de notation. L'approche en notations internes avancées permet à la banque de déterminer à la fois le montant des pertes attendues PA ("Expected Losses ") et le montant des pertes inattendues PI (" Unexpected Losses") relatives au risque de crédit (Dumontier et al., 2008).

La perte attendue correspond au montant exposé au risque de défaut pondéré par la probabilité de défaut et le taux de perte en cas de défaut (voir la relation (a1.1)) :

$$\mathrm{PA} = \mathrm{EAD} \times \mathrm{PD} \times \mathrm{LGD} \qquad \text{(a1.1)}$$

Le nombre effectif des défaillances peut générer un montant supérieur aux pertes attendues notamment si les conditions économiques se détériorent, La banque se trouve alors exposée à des pertes inattendues. Pour combler les pertes inattendues, la réglementation de Bâle II considère que les banques doivent détenir un niveau de fonds propres suffisant pour couvrir ces pertes inattendues. Les pertes attendues feront l'objet de provisions.

Pour faire face à ces pertes, les régulateurs encouragent les banques à utiliser des modèles performants d'analyse du risque de crédit pour l'allocation du capital économique et la mesure de la performance telle que le "Risk Adjusted Return On Capital" (RAROC) (De Servigny et al. 2010).

A1.1.2. Le risque de marché

"*Le risque de marché est le risque de perte qui peut résulter des fluctuations des prix et des instruments financiers qui composent un portefeuille d'actifs ou éventuellement un passif. Les différents facteurs de risques liés au marché sont les taux d'intérêt, les cours de change, les cours des actions et les prix des matières premières. Les variations de ces différents éléments donnent naissance au risque de marché*." (Trader-Finance.fr)

Au niveau du choix des banques entre les différentes approches adaptées à leur besoins, Bâle II a presque reconduit la même réglementation que Bâle I. Par contre, il a mis en œuvre des instruments de réduction du risque, tels que les garanties, la compensation et les sûretés financières.

A1.1.3. Le risque opérationnel

Le risque opérationnel constitue l'une des grandes innovations de la réglementation de Bâle II, où il a été défini comme suit :

" *Le risque opérationnel se définit comme le risque de pertes résultant de carences ou de défaillances attribuables à des procédures, personnels et systèmes internes ou à des événements extérieurs. La définition inclut le risque juridique mais exclut les risques stratégiques et d'atteinte à la réputation* ". (Issard, 2004).

Ainsi, le risque opérationel peut être défini comme l'ensemble des pertes que la banque pourrait supporter à la suite d'un mauvais fonctionnement de son système de gestion interne (erreurs humaines, système de traitement de l'information, procédures de gestion, etc.), ou encore d'événements purement externes (incendie, inondation, etc.).

Le nouvel accord de Bâle II propose trois approches pour déterminer la part de fonds propres nécessaire pour couvrir le risque opérationnel : l'approche indicateur de Base, l'approche standardisée et l'approche avancée (BIS, "Nouvel accord de Bâle sur les fonds propres", Avril 2003) .

A1.2. Le second Pilier des accords de Bâle II

Le processus de surveillance prudentielle, second pilier des accords de Bâle II, est une continuité du premier pilier puisqu'il est considéré comme un élément primordial aux mesures des fonds propres réglementaires. Ce pilier repose sur quatre principes fondamentaux (BIS, "Vue d'ensemble du nouvel accord de Bâle sur les fonds propres", Janvier 2001) :

Principe 1 : les banques doivent disposer d'un système de mesure interne permettant d'évaluer le niveau de fonds propres nécessaire pour couvrir le risque et doivent pouvoir mettre en place une stratégie de maintien de cette évaluation.

Principe 2 : les autorités de contrôle doivent contrôler et évaluer le système de mesures internes des banques et les stratégies adoptées par ces dernières pour garantir le respect de la réglementation prudentielle.

Principe 3 : les autorités de contrôle exigent que les banques maintiennent un niveau de fonds propres supérieurs à la fois à ceux fixés par la réglementation prudentielle et aux exigences minimales de la banque.

Principe 4 : les autorités de contrôle doivent veiller à ce que le niveau des fonds propres des banques ne soit pas plus bas que le niveau prudentiel et en cas de défaillance, elles doivent intervenir rapidement pour ramener le niveau des fonds propres au bon niveau.

A1.3. Le troisième Pilier des accords de Bâle II

Ce troisième pilier vient renforcer les exigences minimales en fonds propres (Pilier I) et le processus de surveillance prudentielle (Pilier II).

En effet, Le nouvel accord de Bâle II "...*vise à améliorer l'information communiquée au marché par les banques et ainsi à exercer sur ces dernières une pression plus forte de nature à favoriser une meilleure gestion de leurs risques et à l'adoption de comportements plus responsables. C'est pourquoi il impose aux établissements de publier des informations quantitatives et qualitatives, de manière plus fiable et fréquente, sur le niveau de leurs fonds propres et sur leurs risques ainsi que sur les modalités d'évaluation de ces derniers. En amenant les acteurs du marché à surveiller davantage le profil de risque et le comportement des banques, le nouveau dispositif renforcera le rôle complémentaire du contrôle prudentiel, joué par la discipline de marché, qui contribuera à son tour au renforcement de l'action des autorités de tutelle."* (Noyer, 2004).

Le troisième pilier du nouveau dispositif vient donc aider en même temps les banques et les autorités de contrôle à gérer les risques et à renforcer la stabilité.

Annexe 2: La régression PLS

Cette annexe est inspirée de l'ouvrage de M. Tenenhaus, intitulé "La régression PLS théorie et pratique", (Tenenhaus, 1998).

L'algorithme de la régression PLS peut être décrit comme suit :

On commence par la recherche d'une combinaison $T_1 = \sum_{i=1}^{n} \lambda_{1i} X_i$, avec X_i la i^{eme} variable exogène, i= {1,…,p}. On cherche, par l'ensemble des variables exogènes, à maximiser à la fois la variance de T_1 et la corrélation entre T_1 et Y (Y est la variable endogène et T est l'ensemble des composantes combinaisons linéaires des variables exogènes). Ce qui revient à maximiser le carré de la covariance ($cov^2(T_1,Y)$) : c'est le critère d'optimalité de Tucker.

$$\text{Max } cov^2(T_1,Y) = cor^2(T_1, Y) * var(T_1) * var(Y) \quad \text{(a2.1)}$$

Pour ce faire, il faut déterminer la covariance $\lambda_{1i} = cov(Y, X_i)$ et normaliser le vecteur λ_{1i} pour avoir $\|\lambda_{11}, \dots, \lambda_{1p}\| = 1$, on aura donc :

$$T_1 = \sum_{i=1}^{p} cov(Y, X_i) * X_i \quad \text{(a2.2)}$$

La régression de Y sur T_1 donne un résidu Y_1 :

$$Y = C_1 T_1 + Y_1 \quad \text{(a2.3)}$$

Et la régression de X_i sur T_1 donne des résidus X_{1i} :

$$X_i = C_{1i} T_1 + X_{1i} \quad \text{(a2.4)}$$

La deuxième étape consiste à reprendre la même opération en remplaçant Y par Y_1 et les X_i par X_{1i}, on obtient alors :

$$T_2 = \sum_{i=1}^{n} \lambda_{2i} X_i \text{ ; avec } \|\lambda_{21}, \dots, \lambda_{2p}\| = 1 \quad \text{(a2.5)}$$

La régression de Y_1 sur T_2 donne un résidu Y_2 :

$$Y_1 = C_2 T_2 + Y_2 \quad \text{(a2.6)}$$

Et la régression des X_{1i} sur T_2 donne un résidu X_{2i} :

$$X_{1i} = C_{2i}\, T_2 + X_{2i} \qquad \text{(a2.7)}$$

On réitère ces opérations jusqu'à ce que le nombre de composantes T_k donne un résultat suffisant, à la fin on aura :

$$Y = C_1\, T_1 + Y_1 = C_1\, T_1 + C_2\, T_2 + Y_2 = \ldots = \sum_{j=1}^{k} C_j T_j + \text{un résidu} \qquad \text{(a2.8)}$$

On remplace l'expression (a2.8) par celle de la régression de Y en fonction des X_i à la place des T_j (a2.5):

$$Y = C_1\left(\sum_{i=1}^{p} \lambda_{1i} X_i\right) + C_2\left(\sum_{i=1}^{p} \lambda_{2i} X_i\right) + \cdots + C_k\left(\sum_{i=1}^{p} \lambda_{ki} X_i\right) + \text{résidu, soit}$$

$$Y = \sum_{i=1}^{p}\left(\sum_{j=1}^{k} C_j \lambda_{ji}\right) X_i \text{ d'où}$$

$$Y = \sum_{i=1}^{p} b_i\, X_i \qquad \text{(a2.9)}$$

Avec $b_i = C_1\lambda_{1i} + C_2\lambda_{2i} + \cdots + C_k\lambda_{ki}$.

Pour déterminer les conditions d'arrêt de l'algorithme de l'approche PLS, il faut calculer la somme des carrés résiduels (SCR) à chaque étape *(h*).

$$\text{Etape h : } SCR_h = \sum_k \left(Y_{(h-1),k} - \widehat{Y}_{(h-1),k}\right)^2 \qquad \text{(a2.10)}$$

Où $\widehat{Y}_{(h-1),k} = C_h T_{h,k}$ est la prédiction de $Y_{(h-1),k}$ calculée pour chaque observation k. Les observations sont partagées en *G* groupes et on réalise *G* fois l'étape courante de l'algorithme sur Y_{h-1} et $X_{h-1,i}$ en enlevant à chaque fois un groupe. Par la suite, on calcule la somme prédite des carrés résiduels (PSCR) à l'étape *h* :

$$PSCR_h = \sum_k \left(Y_{(h-1),k} - \widehat{Y}_{(h-1),-k}\right)^2 \qquad \text{(a2.11)}$$

On retient la $h^{éme}$ composante PLS si :

$PSCR_h \leq \gamma SCR_{h-1}$ avec $\gamma \in [0,1]$; $\gamma = 0{,}95$; le nombre d'observations <100

$\gamma = 1$; le nombre d'observations ≥ 100

Pour l'étape 0, on pose $SCR_0 = \sum(Y_i - \overline{Y})^2$ où $\overline{Y}$ est la moyenne de Y.

Annexe 3: Le modèle de Heckman

Cette annexe est inspirée de l'article de Puhani, intitulé "The Heckman correction for sample selection and its critique", (Puhani, 2000) et de l'article de Feelders, intitulé "An overview of model based reject inference for credit scoring" (Feelders, 2003).

Le problème d'inférence des refusés se définie par le biais de sélection qui est causé par le manque de données dans le modèle bivarié.

(Puhani, 2000) soient :

$$a_i = \begin{cases} 0 & \text{si le crédit est non accordé} \quad S(\mathbf{X}) < s \\ 1 & \text{si le crédit est accordé} \quad S(\mathbf{X}) \geq s \end{cases}$$

$$y_i = \begin{cases} 0 & s'ily\ a\ défaut \\ 1 & sinon \end{cases}$$

avec

a_i : mécanisme de sélection pour l'observation i.

y_i : le résultat de l'emprunt de l'observation i est observable si a_i=1.

$$\begin{cases} a_i = \mathbf{x}_i\beta + d_i \\ y_i = \mathbf{x}_i\gamma + e_i \end{cases}$$

Où "β" et "γ" sont les paramètres à estimer, et "d_i" et "e_i " sont des termes d'erreur qui suivent la loi normale $(d_i, e_i) \sim N(0, \mathbf{\Sigma})$, où $\mathbf{\Sigma}$ est la matrice de variance-covariance:

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

Le coefficient de corrélation : $\rho = \frac{cov(d_i, e_i)}{\sigma_{d_i}\sigma_{e_i}}$

La fonction de régression pour le résultat de l'emprunt, dans le cas où l'échantillon est représentatif de la population globale, est donnée par (a3.1):

$$E[y_i] = X_i\gamma \qquad \text{(a3.1)}$$

Dans le cas où l'échantillon ne contient que les dossiers acceptés, la fonction de régression s'écrit comme suit :

$$\begin{aligned} E[y_i/a_i \geq 0] &= X_i\gamma + E[e_i/a_i \geq 0] \\ &= X_i\gamma + E[e_i/d_i \geq -X_i\beta] \\ &= X_i\gamma + g(-X_i\beta) \end{aligned} \quad (a3.2)$$

avec «$g(-X_i\beta) = E[e_i/d_i \geq -X_i\beta]$».

Comme le modèle de sélection est discret, il y a alors obligation d'imposer une normalisation sur la variance «$v(d_i) = 1$», on note «$e_i = \rho\sigma_{e_i} d_i + résidu$», (en effet : $cov(d_i, e_i) = cov(\rho\sigma_{e_i} d_i, d_i) = \rho\sigma_{e_i} v(d_i) = \rho\sigma_{e_i}$.). Donc :

$$\begin{aligned} E[y_i/x; a_i \geq 0] &= X_i\gamma + E[e_i/d_i \geq -X_i\beta] \\ &= X_i\gamma + E[\rho\sigma_{e_i} d_i/d_i \geq -X_i\beta] \\ &= X_i\gamma + \rho\sigma_{e_i} E[d_i/d_i \geq -X_i\beta] \end{aligned} \quad (a3.3)$$

Le calcul de «$E[d_i/d_i \geq -X_i\beta]$» se fait comme suit:

Soit une loi normale tronquée, avec une densité :

$$f\left(d_i/d_i > s = \frac{\emptyset(d_i)}{P(d_i > s)}\right) = \frac{\emptyset(d_i)}{[1-\Phi(s)]} \quad (a3.4)$$

Avec

- «$\emptyset(.)$» est la densité de la loi normale standard
- «$\Phi(.)$» est la fonction de répartition
- «$s = -X_i\beta$» et l'espérance d'une loi normale tronquée est alors :

$$E(d_i/d_i > s) = \int_{d_i=s}^{\infty} \frac{d_i \emptyset(d_i)}{[1-\Phi(s)]} dd_i$$

$$= \frac{1}{1-\Phi(s)} \times \frac{1}{\sqrt{2\pi}} \times \int_{d_i=s}^{\infty} d_i \exp\left(-\frac{{d_i}^2}{2}\right) dd_i$$

$$= \frac{1}{1-\Phi(s)} \times \frac{1}{\sqrt{2\pi}} \times \left[-\exp\left(-\frac{{d_i}^2}{2}\right)\right]_s^{\infty}$$

$= \frac{\emptyset(s)}{[1-\Phi(s)]} = H_i$ (a3.5) qui est l'inverse du ratio de Mills

Et l'échantillon sélectionné aura alors:

$$E[y_i/a_i \geq 0] = X_i\gamma + \rho H_i \qquad \text{(a3.6)}$$

L'estimation des paramètres du modèle Heckman, soient (γ, β, ρ), est déterminée selon deux étapes. La première étape étant celle de l'estimation du paramètre «β» par le modèle Probit, qui par la suite va permettre la détermination de l'inverse du ratio de Mills, soit «H_i». Une fois «H_i» estimé, on passe alors à la deuxième étape, qui permettra l'estimation des paramètres (γ, ρ) et ceci par la méthode des moindres carrés.

A3.1. Le modèle probit

Cette étape consiste à déterminer l'estimateur du maximum de vraisemblance du paramètre «β» pour que par la suite on pourrait déterminer «H_i». Pour les données non manquantes, on a:

$$E[a_i] = \Phi\,(X_i\beta) \qquad \text{(a3.7)}$$

La fonction de vraisemblance «ll_p» est la suivante :

$$ll_p = \sum_{i=1}^{k} a_i \ln P(a_i = 1)$$

$$= \sum_{i=1}^{k} a_i \ln E[a_i]$$

x

$$= \sum_{i=1}^{k} a_i \ln \Phi (X_i \beta) \qquad \text{(a3.8)}$$

Pour obtenir l'estimation du maximum de vraisemblance de «β», il faut optimiser la fonction de vraisemblance par un vecteur «β» qui annule la dérivée de cette fonction. Une fois le vecteur «$\hat{\beta}$» déterminé, on peut alors estimer «$\widehat{H}_i$».

A3.2. L'estimation par la méthode des moindres carrés

«$\widehat{H}_i$» déterminé, le modèle peut être considéré comme un problème des moindres carrés ordinaires où les paramètres (γ, ρ), dans « $E[y_i/a_i \geq 0] = X_i\gamma + \rho\widehat{H}_i$», sont déterminés par l'estimateur des moindres carrés qui est déterminé de façon à minimiser la somme des résidus au carré (SSE) :

$$SSE = \sum_{i=1}^{k} (y_i - E[y_i/a_i \geq 0])^2 \qquad \text{(a3.9)}$$

Dans le cas où il n'y a pas de biais de sélection, les résidus «d_i» et «e_i » sont indépendants (pas de corrélation $\rho = 0$). La présence de biais de sélection peut être détecté en testant l'hypothèse nulle que «$\rho = 0$». Cette hypothèse peut être testée par le test de Wald ou le ratio de vraisemblance ou encore le test du multiplicateur Lagrangien.

Mais les recherches empiriques ont montré que les estimateurs du modèle de Heckman ne sont pas robustes car il suppose que le modèle bivarié est linéaire avec des erreurs qui suivent une distribution normale et qui sont homoscédastiques. puisque les hypothèses ne sont pas respectées, les estimateurs ne peuvent être fiables.

Annexe 4: Programmation sur SAS 9.3

A4.1. Programmation du processus de refus

Nous avons commencé par simuler un processus de refus sur les 7986 dossiers acceptés. Pour cela nous avons créé une variable uniforme U_i :

```
/*création d'une variable uniforme*/

data simul;
retain _seed_ 0;
do _i_=1 to 7986;
uniform1= 0+(1-0)* ranuni(_seed_);
output;
end;
drop _seed_ _i_;
run;
```

Nous avons comparé cette variable avec la probabilité de refus *Pr(i)* obtenue selon la discrimination acceptés refusés par une analyse discriminante:

```
/*discrimination acceptés-refusés (sur les 9892 dossiers) pour obtenir la
probabilité de refus*/

proc discrim data=___ method=normal pool=yes crossvalidate
out=___ outstat=___ list;
class ___ ;
var ___;
priors prop;
run;
```

Nous n'avons gardé que les dossiers acceptés et nous avons comparé la variable uniforme avec la probabilité de refus. Si U_i < *Pr(i)* alors l'observation i est considérée comme un dossier refusé sinon c'est comme un dossier accepté. Pour cela nous avons utilisé l'instruction " `if uniforme<Pr then delete`" pour garder les 6686 dossiers acceptés et " `if uniforme>Pr then delete`" pour garder les 1300 dossiers refusés.

A4.2. Programmation de la répartition de l'échantillon (apprentissage et test)

Nous avons départagé, d'une façon aléatoire, l'échantillon en deux, 70% pour l'échantillon d'apprentissage (soit 5590 dossiers) et 30% pour l'échantillon test (soit 2396 dossiers) et ceci par l'instruction **`"proc surveyselect".`**

A4.3. Programmation de la méthode de l'augmentation simple

Étape 1: Nous avons programmé une régression logistique par l'instruction **`"proc logistic"`** sur les dossiers acceptés uniquement. Nous avons spécifié l'option `"outmodel"` pour pouvoir récupérer les résultats de cette régression.

Étape 2: Pour pouvoir appliquer le modèle précédent sur les refusés, nous avons utilisé l'instruction **`"proc logistic"`** avec l'option **`"inmodel"`**.

Étape 3: Les dossiers refusés sont étiquetés en bon et mauvais selon le taux de défaut obtenu par l'option "pred".

Étape 4: Nous avons combiné les dossiers refusés et acceptés ensemble par l'instruction **"proc iml"** et l'option **"//"** pour construire un nouveau modèle de score.

A4.4. Programmation de la méthode de l'augmentation

Étape 1: Nous avons calculé la probabilité d'acceptation pour chaque individu (acceptés et refusés) par une régression logistique en ajoutant l'option "pred".

Étape 2: Nous avons calculé des poids par l'instruction "1/pred" et nous avons pondéré chaque dossier accepté.

Étape 3: Nous avons construit une modèle de score par l'instruction **"proc logistic"** et pour que ce modèle tient compte des pondérations nous avons ajouté l'option "weight".

A4.5. Programmation de la méthode de la reclassification itérative

Nous avons répété le processus de l'augmentation simple jusqu'à la stabilisation des score ici 10 fois (le modèle ne s'améliore plus). Nous avons calculé, à chaque itération, le nombre de bons payeurs et celui de mauvais payeurs, classés par le modèle, par une **"proc freq"** et nous nous sommes arrêtés lorsque ces nombres sont restés fixes entre l'itération 9 et 10.

A4.6. Programmation de la méthode de parceling

Étape 1: La procédure **"proc logistic"** nous a permis de déterminer la probabilité de défaut pour les dossiers acceptés. Nous avons transformé ces probabilités en score par le programme suivant:

```
/*calcul du score à partir de la probabilité prédite*/
data cscore;
smax=1;
smin=0.5;
ecart=smax-smin;
diff=pred-smin;
score=1000*diff/ecart;
run;
```

Étape 2: Nous avons réparti la population en quatre intervalles de score par l'instruction "if" et nous avons calculé le taux d'impayés pour chaque intervalle par la procédure **"proc freq"**.

Nous avons appliqué le résultat de la **"proc logistic"** précédente sur les refusés de façon à les placer chacun dans un intervalle de score sous l'hypothèse que le taux d'impayés est le même que celui trouvé pour les acceptés.

Étape 3: Nous avons affecté les refusés de chaque intervalle aléatoirement à l'une des classe. La procédure **"proc surveyselect"**, nous permet de tirer aléatoirement les bons refusés et les observations qui restent seront affectées à la classe des mauvais payeurs par la procédure **"proc iml".**

Étape 4: Nous combinons les acceptés et les refusés ainsi étiquetés par une **"proc iml"** et la procédure **"proc logistic"**, permet de construire le nouveau modèle de score.

A4.7. Programmation de la méthode de classification mixte

Nous avons commençé par utiliser la méthode des k-means sur toute la population avec la procédure **"proc fastclus"** en spécifiant l'option "drift". Nous nous sommes référés au critère de Wong pour retenir 20 classes. La procédure **"proc cluster"** nous permet d'effectuer une classification ascendante hiérarchique sur les centres finaux de la première classification (le fichier data considéré pour la CAH est l'output de la classification par les k-means). Pour fixer le nombre final de classes à retenir, nous avons utilisé plusieurs indicateurs statistiques qui s'affichent grâce à la procédure **"proc tree"** et plus exactement l'option "outtree".

Nous avons affecté les dossiers refusés, de chaque classe, à la catégorie bon ou mauvais en fonction de la catégorie dominante de la classe par la procédure **"proc freq".** Nous avons construit le modèle de score sur les acceptés et les refusés étiquetés par la procédure **"proc logistic".**

A4.8. Programmation de la méthode de Co-training

Nous avons départagé la population des acceptés, aléatoirement, en deux ensembles par la procédure **"proc surveyselect".**

Nous avons appliqué l'analyse discriminante séparément sur chaque ensemble en utilisant la procédure **"proc discrim"** avec l'option "outstat" qui nous a permis de récupérer les résultats de cette procédure pour les appliquer aux dossiers refusés. Pour cela, nous avons utilisé la procédure **"proc discrim"** avec l'option "testdata" en fixant comme fichier data celui des refusés et l'option "testout" pour récupérer les résultats de cette deuxième discrimination.

Nous avons utilisé l'output de "testout" pour calculer le taux de biens classés pour les dossiers refusés. Les dossiers refusés qui sont bien classé par l'un des classifieurs sont ajoutés aux données d'apprentissage de l'autre classifieur par la procédure **"proc iml".**

Nous avons combiné les prédictions des deux classifieurs optimaux par la procédure **"proc iml".** La procédure **"proc discrim",** qui prend comme input les sorties de la procédure **"proc iml"** précédente, nous a permis de construire le modèle de score sur l'échantillon d'apprentissage initial (par l'option "testdata").

A4.9. Programmation de l'analyse factorielle discriminante

Nous avons programmé cette méthode pour pouvoir affecter aux dossiers refusés des étiquettes et ainsi pouvoir programmer les algorithmes de Boosting. Nous avons appliqué la procédure **"proc discrim"** sur les dossiers acceptés en spécifiant l'option "outstat" pour récupérer la fonction discriminante et l'appliquer aux dossiers refusés pour une nouvelle classification. Nous avons spécifié l'option "can" car nous voulons appliquer une Analyse Factorielle Discriminante (AFD). L'option "testdata" nous a permis de spécifier la nouvelle base de données à classer qui est composée des dossiers refusés. Nous avons combinons les dossiers acceptés et les dossiers refusés par la procédure **"proc iml"**, pour obtenir la base de données qui va être utilisée pour programmer les algorithmes de Boosting.

A4.10. Programmation de l'algorithme AdaBoost

Nous avons calculé **l**es poids initiaux qui représentent les proportions réelles des dossiers acceptés et des dossiers refusés par la procédure **"proc iml"**. Nous avons combiné ces poids initiaux avec notre base de données par l'option "merge" sur lesquels nous avons appliqué la procédure **"proc discrim"** et pour prendre en considération les poids, nous avons ajouté l'option "weight". Nous avons spécifié l'option "out" pour récupérer la base de données avec les résultats de la nouvelle classification. Nous avons utilisé la procédure **"proc iml"** pour localiser les dossiers mal classés (sortie de l'option "out") par le modèle par l'option "loc". Pour les données mal classées, nous avons calculé l'erreur de classification et nous avons multiplié leurs poids par la constante α_t (déterminée dans la section 4.2.1) . Nous avons normalisé les poids par la procédure **"proc iml"** pour obtenir les nouveaux poids et nous sommes passé à l'itération suivante et nous nous sommes arrêté lorsque l'erreur de classification atteint la valeur "0.5". Une fois la condition d'arrêt est atteinte, Nous combinons toutes les valeurs prédites par la procédure **"proc discrim"** de chaque itération. Le score final pour chaque observation est obtenu par la sommation de ces prédictions multipliées par l'exponentiel de la valeur α_t par la procédure **"proc iml"**. L'affectation à l'une des catégories (bon ou mauvais payeur) est faite selon le signe du score par l'option "if".

A4.11. Programmation de l'algorithme LogitBoost

Le calcul des poids initiaux est effectué comme pour l'algorithme AdaBoost. Nous avons combiné les poids initiaux avec notre base de données par l'option "merge". Par la procédure "proc iml", nous avons déterminé les valeurs attendues z_i du classifieur (déterminées dans la section 4.2.2). Nous avons appliqué la procédure **"proc reg"** (la régression par les moindres carrés pondérés) sur les z_i et pour prendre en considération les poids, nous avons ajouté l'option "weight". L'option "out" permet de récupérer la base de données avec les nouvelles prédictions. Nous avons répété la procédure jusqu'à ce que les valeurs prédites n'ont plus varié. Le score final pour chaque observation est obtenu par la sommation de ces prédictions multipliées chacune par la valeur 0.5 en utilisant la procédure "proc iml".
L'affectation à l'une des catégories (bon ou mauvais payeur) est faite selon le signe du score par l'option "if".

A4.12. Programmation de l'algorithme Gentle AdaBoost

Le calcul des poids initiaux est effectué comme pour l'algorithme AdaBoost. Nous avons combiné les poids initiaux avec notre base de données par l'option "merge" sur lesquels nous appliquons la procédure **"proc reg"** et pour prendre en considération les poids, nous avons ajouté l'option "weight". L'option "out" permet de récupérer la base de données avec les nouvelles prédictions. La procédure **"proc iml"** permet de calculer les nouveaux poids (comme expliquer dans la section 4.2.3) et de les normaliser. L'algorithme est répété jusqu'à la stabilisation des prédictions. Le score final pour chaque observation est obtenu par la sommation de ces prédictions par la procédure **"proc iml"**. L'affectation à l'une des catégories (bon ou mauvais payeur) est faite selon le signe du score par l'option "if".

A4.13. Programmation du modèle de Heckman

Nous avons considéré l'échantillon des dossiers acceptés et refusés en même temps, sur lequel nous avons appliqué la procédure " **proc qlim ",** pour spécifier le modèle de heckman à deux étapes. La procédure est la suivante :

proc qlim data= spécifier l'échantillon (acceptés et refusés);
model : la première étape est l'estimation du modèle probit par le maximum de vraisemblance pour déterminer le mécanisme de sélection en spécifiant l'option (discrete)pour dire que c'est un modèle probit.

`model` : la deuxième étape est la régression par les moindres carrés ordinaires pour déterminer le résultat de l'emprunt, ce modèle n'est établi que pour les dossiers acceptés. L'option utilisé est "`select`" et spécifiée par la suite qu'il ne faut garder que les dossiers acceptés.

L'option " `output out`", en plus de l'option `"predicted"`, permet de récupérer la base de données avec les scores.

A4.14. Programmation du calcul de l'indice AUC et du traçage de la courbe ROC

Pour chaque méthode nous avons procédé au calcul de l'indice AUC et le traçage de la courbe ROC pour comparer les performances des modèles de score. Dans la procédure **`"proc logistic"`**, il existe une option pour déterminer la courbe ROC qui est `"plots(only)=(ROC)"` en spécifiant, dans le modèle, comme variable à prédire, la variable initiale " bon/mauvais" payeurs et une seule variable explicative qui est le score obtenu selon la méthode utilisée.

A4.15. Programmation du calcul des taux de bons classements des individus

Pour calculer le taux de bons classements des individus, nous avons utilisons la procédure **`"proc freq".`** Les variables, que nous avons utilisé, sont la variable initiale "bon/mauvais" et celle déterminée par les modèles de score de chaque méthode.

Annexe 5: Programmation sur R

Nous avons programmé la méthode de l'auto-apprentissage par les SVMs avec le logiciel R car les SVMs ne sont pas très développés sous SAS. Ils existent des programmes faciles à exécuter sous R.

Nous avons commencé par importer les données d'apprentissage qui sont les dossiers acceptés par la commande "read.table". Pour pouvoir manipuler la procédure SVM sous R, nous devons appeler le package "kernlab" par la commande " library(kernlab)".
Pour entrainer la procédure SVM sur notre base de données d'apprentissage, nous utilisons la commande "ksvm".

Pour obtenir les scores de chaque observation, nous utilisons la commande "predict" en spécifiant l'option "decision". Nous importons les dossiers refusés par la commande "read.table". Nous utilisons le modèle SVM établi pour prédire les étiquettes des dossiers refusés par la commande "predict" et nous spécifions l'option "decision" pour obtenir les scores.

Pour arriver à calculer l'indice AUC et la courbe ROC, nous devons appeler le package "ROCR" par la commande "library(kernlab)". Pour calculer l'indice AUC, nous utilisons la commande "performance" en spécifiant l'option "auc".
Pour tracer la courbe "ROC", Nous utilisons la commande "performance" avec l'option "("tpr", "fpr")".

Asma GUIZANI

TRAITEMENT DES DOSSIERS REFUSÉS DANS LE PROCESSUS D'OCTROI DE CRÉDIT AUX PARTICULIERS

Résumé

Le credit scoring est généralement considéré comme une méthode d'évaluation du niveau du risque associé à un dossier de crédit potentiel. Cette méthode implique l'utilisation de différentes techniques statistiques pour aboutir à un modèle de scoring basé sur les caractéristiques du client.
Le modèle de scoring estime le risque de crédit en prévoyant la solvabilité du demandeur de crédit. Les institutions financières utilisent ce modèle pour estimer la probabilité de défaut qui va être utilisée pour affecter chaque client à la catégorie qui lui correspond le mieux: bon payeur ou mauvais payeur. Les seules données disponibles pour construire le modèle de scoring sont les dossiers acceptés dont la variable à prédire est connue. Ce modèle ne tient pas compte des demandeurs de crédit rejetés dès le départ ce qui implique qu'on ne pourra pas estimer leurs probabilités de défaut, ce qui engendre un biais de sélection causé par la non-représentativité de l'échantillon. Nous essayons dans ce travail en utilisant l'inférence des refusés de remédier à ce biais, par la réintégration des dossiers refusés dans le processus d'octroi de crédit. Nous utilisons et comparons différentes méthodes de traitement des refusés classiques et semi supervisées, nous adaptons certaines à notre problème et montrons sur un jeu de données réel, en utilisant les courbes ROC confirmé par simulation, que les méthodes semi-supervisé donnent de bons résultats qui sont meilleurs que ceux des méthodes classiques.

Mots-clé

Réglementation prudentielle, crédit scoring, méthodes d'inférence des refusés, méthodes de l'apprentissage semi-supervisé.

Abstract

Credit scoring is generally considered as a method of evaluation of a risk associated with a potential loan applicant. This method involves the use of different statistical techniques to determine a scoring model. Like any statistical model, scoring model is based on historical data to help predict the creditworthiness of applicants. Financial institutions use this model to assign each applicant to the appropriate category : Good payer or Bad payer. The only data used to build the scoring model are related to the accepted applicants in which the predicted variable is known. The method has the drawback of not estimating the probability of default for refused applicants which means that the results are biased when the model is build on only the accepted data set. We try, in this work using the reject inference, to solve the problem of selection bias, by reintegrate reject applicants in the process of granting credit. We use and compare different methods of reject inference, classical methods and semi supervised methods, we adapt some of them to our problem and show, on a real dataset, using ROC curves, that the semi-supervised methods give good results and are better than classical methods. We confirmed our results by simulation.

Key words

prudential regulation, credit scoring, reject inference methods, semi-supervised learning methods