Avec tout respect et amour je dédie ce modeste travail

À mes chers parents
À mon petit frère Dhia et mon adorable sœur Dhekra
pour tout leur soutien moral et leur amour et affection

À tous mes amis
en Souvenir des plus beaux instants qu'on a passé ensemble

Aussi bien à tous ceux qui m'ont aidé

Merci

Waad

Remerciements

En préambule à ce mémoire, je souhaite adresser mes remerciements les plus sincères aux personnes qui m'ont apporté leur aide et qui ont contribué à l'élaboration de ce mémoire ainsi qu'à la réussite de cette année universitaire.

Ma reconnaissance s'adresse à mes deux encadreurs pour tous les efforts déployés pour me fournir les ressources documentaires dont j'ai eu besoin et aussi pour leur disponibilité et leurs encouragements permanents.

J'adresse mes vifs remerciements au Professeur Mohamed LIMAM, qui a bien accepté de m'encadrer et aussi pour son amabilité et sa bienveillance au bon déroulement de mon travail de recherche.

J'exprime aussi ma gratitude au Docteur Ghazi BEL MUFTI, qui s'est toujours montré à l'écoute et très disponible tout au long de la réalisation de ce mémoire, ainsi que pour le temps qu'il a bien voulu me consacrer et sans qui ce mémoire n'aurait jamais vu le jour.

Je tiens également à remercier Professeur Farid BENINEL, qui m'a présenté le domaine de recherche du credit scoring et qui m'a guidé dans ce travail.

Enfin, mes remerciements vont également aux membres du jury d'avoir accepter de juger mon travail.

Table des figures

Liste des tableaux

Table des matières

Introduction

Le risque de crédit est l'un des risques majeurs qu'un organisme de prêts doit gérer. Ce risque survient lorsqu'un emprunteur ne rembourse pas sa dette à l'échéance fixée ce qui conduit à des pertes supportées par le créancier. Afin d'éviter cela, un organisme bancaire n'accorde un crédit que s'il est assuré de la solvabilité de l'emprunteur et ce sur la base des informations provenant des anciens prêts.

Les gérants de la banque doivent ainsi chercher des solutions efficaces qui leurs permettent de bien distinguer les bons des mauvais demandeurs de prêt. La mise en place de ces solutions nécessite la disposition d'informations sur les emprunteurs et d'une technique objective d'évaluation. L'organisme de crédit analyse les demandes de prêts en se basant sur une grande masse d'information collectée sur la base d'anciens prêts. La masse d'information possédée par l'organisme de prêt constitue un élément de base dans la procédure d'évaluation du dossier du candidat.

Dans ce travail, nous nous intéressons à l'évaluation du risque de crédit d'une population pour qui la masse d'information disponible est considérée comme insuffisante, vu le faible effectif de celle-ci : il s'agit des emprunteurs non-clients d'une banque. Nous nous proposons d'évaluer la capacité de remboursement d'un demandeur de prêt provenant de cette population et la modélisation des risques potentiels de non remboursement en dépit du manque d'information dû au faible effectif qui caractérise cette sous-population.

En effet, les systèmes traditionnels basées sur l'expérience se trouvent

impuissants dans ce contexte face à la contrainte d'insuffisance d'observations dans l'échantillon. Il s'agit pour nous de construire un modèle statistique qui sert à prédire le comportement des potentiels emprunteurs d'une banque en matière de remboursement à partir de variables de description malgré cette contrainte.

Le comportement de chaque emprunteur est décrit par une variable binaire notée $Y : Y = 0$ lorsque l'emprunteur pose problème et $Y = 1$ sinon.La valeur prise par cette dernière fournit un élément de base dans la décision d'octroi des crédits. Nous disposons en plus de cette variable dichotomique, d'un ensemble de variables de description $(X_1, X_2, ..., X_d)$ qui renseignent sur l'emprunteur (sexe, logement, statut professionnel, ..) et sur le fonctionnement des comptes (montant en épargne, ...). Ces renseignements sont tirés à partir d'un échantillon de demandeurs de prêt provenant d'une population hétérogène formée d'emprunteurs clients de la banque en question et d'autres qui sont clients d'autres organismes bancaires.

Afin d'apporter une réponse au problème lié à la taille de l'échantillon des non-clients, l'idée est d'utiliser l'information en notre possession relative à la fois aux emprunteurs clients et non-clients. Une première approche est généralement utilisée par les banques, et qui consiste en l'usage du modèle prédictif basé sur les clients pour estimer le comportement des emprunteurs non-clients. Toutefois, l'application de cette approche ne tient pas compte de la différence entre les deux sous-populations. Une autre approche consiste à construire un modèle de prédiction pour les emprunteurs non-clients à partir d'un échantillon de cette population, ceci suppose un nombre d'emprunteurs non-clients suffisant, ce qui n'est pas le cas.

Les travaux de Biernacki et Jacques (2007), Bouveyron et Jacques (2009) ainsi que ceux de Beninel et Biernacki (2005), (2007), (2009) se sont intéressés au problème de discrimination dans le cas d'un mélange de deux sous-populations gaussiennes et ce en tenant compte des liens entre les deux populations. Ces travaux ont ainsi, apporté une solution à la contrainte liée

au faible effectif de l'une des deux sous-populations en exploitant les données de l'autre sous-population.

Dans un premier temps, en nous basant sur ces travaux, nous nous focalisons sur les modèles logistiques. Plus précisément, nous présentons sept modèles de liaison entre les paramètres des deux fonctions de score, associées respectivement aux deux populations. Dans un second temps, nous construisons d'autres modèles basés sur une méthode non-paramétrique à savoir les Séparateurs à Vaste Marge (SVM). Une fois les différents modèles implémentés sous R, nous les comparons en utilisant un jeu de données composé de 1000 observations représentant les crédits à la consommation accordés par la *DeutshBank* de Munich (LMU, 1994). Il en ressort que lorsque l'on tient compte des liaisons entre les deux sous-populations, les modèles de régression logistique sont les plus performants et s'ajustent le mieux à nos données. Dans le cas contraire, les modèles basées sur les SVMs deviennent les plus performants. Néanmoins, cette performance dépend fortement de leur paramétrage : un changement des paramètres est susceptible de dégrader considérablement les résultats.

Notre travail se subdivise en cinq chapitres. Dans un premier chapitre, nous présentons le concept du credit scoring ainsi que ces principales méthodes. Le rappel des concepts de la discrimination gaussienne, son lien avec la discrimination logistique ainsi que la modélisation des différents modèles de régression logistique font l'objet du chapitre deux. Le chapitre trois est consacré aux SVMs. Les quatrième et cinquième chapitres sont consacrés à la mise en œuvre des divers modèles et leur comparaison sur une base de données provenant d'une banque allemande.

Chapitre 1

Le credit scoring et ces méthodes

1.1 Le credit scoring

Le credit scoring a fait ses premières apparitions suite aux travaux de Durand (1941). En utilisant un test du khi-deux, Durand a pu identifier les variables qui permettent de distinguer, d'une manière significative les bons des mauvais emprunteurs. Ensuite, il a développé un indice d'efficacité qui montre l'efficience d'une variable dans la différenciation des bons et des mauvais demandeurs de prêt.

C'est en 1958 que le cabinet Fair et Isaac a développé les premiers systèmes de credit scoring pour les crédits de consommation aux Etats-Unis. Par la suite, les méthodes de credit scoring sont devenues de plus en plus importantes avec la croissance spectaculaire du crédit à la consommation, au cours de ces dernières années (Hand et Henley, 1997).

Feldman (1997) définit le credit scoring comme le processus d'assignation d'une note (ou score) à un emprunteur potentiel pour estimer la performance future de son prêt. Saporta (2006) et Thomas *et al.* (2002) définissent aussi ce terme comme l'ensemble d'outils d'aide à la décision utilisés par les organismes financiers pour évaluer le risque de non-remboursement des prêts. La Figure 1.1 décrit le processus du credit scoring.

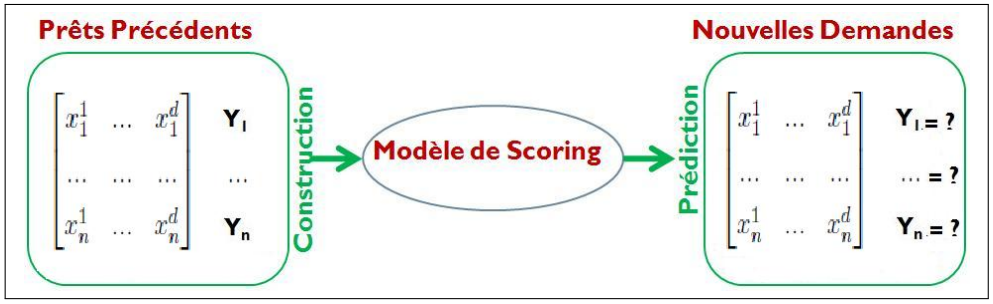

FIGURE 1.1 – *Processus de credit scoring, adapté de Yang (2001)*

1.2 Notations

Dans ce qui suit, nous présentons les principales notations qui seront utilisées dans ce travail.
Soit Ω la population des emprunteurs clients de la banque. On note χ l'espace des observations dans $\mathbb{R}^d$ défini par une variable aléatoire x.

$$\begin{aligned} x : \quad & \Omega \to \chi \in \mathbb{R}^d \\ & i \to x_i = (x_i^1, x_i^2, ..., x_i^d) \end{aligned} \tag{1.1}$$

On dispose alors de n individus décrits par d variables de description. Ces données sont résumés dans la matrice x suivante :

$$\begin{bmatrix} x_1^1 & ... & x_1^d \\ ... & ... & ... \\ x_n^1 & ... & x_n^d \end{bmatrix}$$

Les variables de description sont décrites par le vecteur X et les poids associés aux variables de description $X = (X_1, X_2, ..., X_d)$ sont définis par le vecteur β, $\beta^T = (\beta_1, \beta_2, ..., \beta_d)$.

Les observations sont réparties en g groupes connus a priori. Chaque individu appartient à un et un seul groupe. L'étiquète du groupe d'appartenance d'un individu i est représentée au travers des modalités $\{y_1, ..., y_g\}$, d'une variable catégorielle Y. Dans ce travail nous nous proposons de travailler avec $g = 2$, d'où la variable Y est une variable binaire, qui prend ses valeurs dans $\{0, 1\}$. La modalité 0 dénote un emprunteur non solvable et la modalité 1 dénote un emprunteur solvable.

1.3 Méthodes du Credit Scoring

Dans ce qui suit, nous présentons les principales méthodes de credit scoring proposées dans la littérature.

1.3.1 L'analyse discriminante

L'origine de cette méthode remonte aux travaux de Fisher (1936) et Mahalanobis (1936). L'analyse discriminante est une technique statistique qui permet d'affecter un individu i à l'un des groupes connus a priori en se basant sur une fonction de discrimination f qui est une combinaison linéaire des variables de description. La fonction s'écrit alors :

$$f(x_i) \quad = \quad \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + ... + \beta_d x_i^d \tag{1.2}$$

L'analyse discriminante est souvent utilisée par les banques pour sa simplicité (Tuffery, 2007). En effet, il suffit de consulter les coefficients des variables pour sortir des conclusions explicites sur le pouvoir discriminant de chaque variable. Cette technique est très performante une fois appliquée sur des petits échantillons. Néanmoins, son application nécessite de fortes hypothèses sur les données. L'analyse discriminante, n'est applicable qu'aux variables explicatives continues sans valeurs manquantes, tout en se basant

sur des hypothèses de multinormalité et d'homoscédasticité (i.e. matrices de variance-covariance interclasses égales).

1.3.2 La régression logistique

L'origine de cette méthode remonte aux travaux de Cox (1970), (1989). Un modèle de régression logistique fournit une fonction linéaire de descripteurs comme outil de discrimination, du même type que ceux introduis dans l'équation (1.2). Fan et Wang (1998) et Sautory *et al.* (1992) recommandent l'emploi de la régression logistique binaire, lorsque les conditions d'application de l'analyse discriminante ne sont pas réunies. Ce choix s'impose dans le cas où des variables qualitatives interviennent dans le modèle (Bardos, 2001).

Le modèle de régression logistique permet de prédire la valeur prise par la variable dépendante Y. Cette technique cherche à déterminer une probabilité a posteriori notée p_i, qui permet d'affecter un individu i à son groupe d'appartenance. Cette probabilité est définie par :

$$p_i = \frac{1}{1+\exp^{-\beta_0-\beta^T x_i}} \tag{1.3}$$

Les coefficients obtenus par cette technique ont les mêmes valeurs que ceux obtenus par l'analyse discriminante linéaire. Néanmoins, ils sont obtenus sous des hypothèses moins contrariantes (Vojtek et Kocenda, 2006). Toutefois, la mise en place de cette technique nécessite une grande masse de données d'apprentissage. Par ailleurs, elle nécessite des données complètes sans valeurs manquantes pour chaque individu comme dans le cas de l'analyse discriminante. Nous revenons plus en détails sur cette technique dans le chapitre 2, dans le quel nous étendrons le principe de cette technique au cas d'un mélange de populations.

1.3.3 Les arbres de décision

Les arbres de décision ont été introduite par Breiman *et al.* (1984). Cette technique a été utilisée à des fins de credit scoring pour la premier fois par Frydman *et al.* (1985). La technique des arbres de décision est une technique très simple et flexible. Son intérêt majeur réside dans la gestion de données hétérogènes ou manquantes (Tuffery, 2007).

Giudici (2003) définit un arbre de décision comme étant une procédure récursive dans laquelle un ensemble de n individus est progressivement divisés dans des groupes. La procédure commence par le choix d'une variable explicative $X_j \in X$ qui permet de séparer les individus dans des sous-groupes homogènes appelés nœuds. Chaque nœud contient des individus d'une seule classe, l'opération est répétée jusqu'à ce que la division en sous-populations n'est plus possible.

Cette procédure est utilisée par les banques seulement comme un outil de soutien pour les méthodes paramétriques de credit scoring, décrites précédemment, au cours de la phase de sélection de variables (Devaney, 1994). Tufféry (2007) recommande l'utilisation de cette technique que si, on dispose d'un nombre suffisant d'observations pour éviter le risque de surapprentissage (i.e. modèle avec d'excellentes performances dans la phase d'apprentissage mais trouve des difficultés de généralisation pour les nouveaux échantillons).

1.3.4 Les réseaux de neurones

La technique des réseaux de neurones date des travaux de Mcculloch et Pitts (1943). Il s'agit d'un outil d'optimisation non-linéaire composé d'un ensemble d'unités élémentaires, appelées neurones. Chaque neurone reçoit en entrée la description d'une observation i, $\{x_i^1, ..., x_i^d\}$ appelées signaux. Ensuite, il effectue une somme pondérée sur ces entrées dont le résultat est soumis à une transformation appelée fonction d'activation notée f (Devaney,

1994). La fonction d'activation d'un neurone est définie par :

$$f(x_i) \quad = \quad \sum_{j=1}^{d} \beta_j x_i^j + \beta_0 \quad = \quad \beta^T x_i \tag{1.4}$$

où $x_i = [1, x_i^1, x_i^2, , x_i^d]^T$ est le vecteur des entrées. Chaque signal d'apport est associé à un poids β_j, $j = 1, ..., d$. Le poids détermine l'importance relative du signal d'apport dans la production de l'élan final transmit par le neurone. La sortie y_i du neurone est décidée en fonction du signe de $f(x_i)$. Pour un individu i

$$y_i = \begin{cases} 0 \ si \ \beta^T x_i \leq 0 \\ 1 \ sinon \end{cases} \tag{1.5}$$

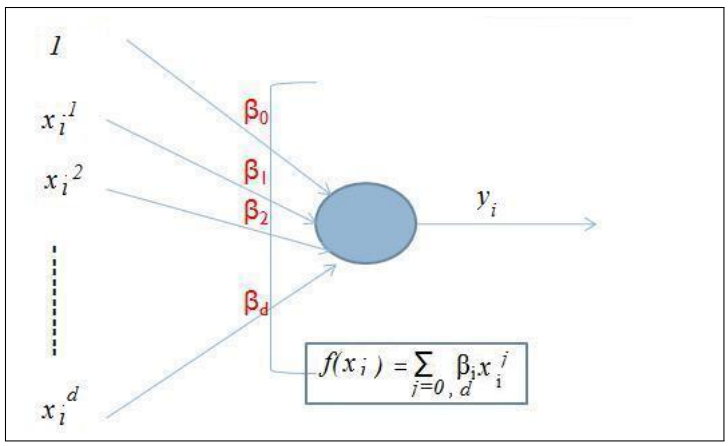

FIGURE 1.2 – *Schéma d'un perceptron à un " neurone"adapté de Cornuéjols (2002)*

Ce mécanisme de prédiction permet le traitement des données d'une manière simple sans aucune compétence statistique. Toutefois, cette méthode est une sorte de boîte noire, qui ne permet pas d'expliquer la manière dont

on arrive au résultat. Dans le cas des réseaux de neurones, il y a un risque énorme de sur-apprentissage si le nombre d'observations est faible. Cette procédure n'assure toujours pas une convergence vers la solution cherchée et ne permet pas de traiter les problèmes avec un nombre important de variables explicatives.

1.3.5 La méthode des k plus proches voisins

La méthode des k plus poches voisins est une méthode de classification non-paramétrique qui cherche pour une nouvelle observation x_j les k observations les plus proches appelées voisins à partir d'un ensemble d'apprentissage formé des couples (x_i, y_i) où les étiquettes des groupes sont connus a priori.

La détermination des k plus proches voisins de x_j est effectuée à l'aide d'une fonction de distance notée d définie par :

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{d} (x_i^r - x_j^r)^2} \tag{1.6}$$

La nouvelle observation est affectée à la classe majoritaire des k voisins. La mise en œuvre de la méthode des k plus proches voisins est simple et ne demande pas des grandes compétences particulaires. Cette technique est utilisée pour gérer les données hétérogènes ou manquantes. Toutefois, le choix du nombre k de voisins n'est pas aussi simple, ce nombre est soit fixé d'avance soit choisi par validation croisée (Merbouha et Mkhadi, 2006). Cette méthode utilise un nombre important de données d'apprentissage. Par ailleurs, elle peut être sensible à la dimension des données.

1.3.6 Les séparateurs à vastes marge

Les séparateurs à vaste marge(en anglais Support Vector Machine, SVM) ont été d'abord suggérés par Vapnik (1995) et ont été appliqués à une grande

gamme de problèmes incluant, les travaux de Pontil et Verri (1998) dans le domaine de la reconnaissance des formes, en bioinformatique par Yu *et al.* (2003) et en catégorisation de texte par Joachims (1998). Récemment les SVMs ont aussi commencé à s'intégrer dans le domaine de credit scoring grâce au travaux de Huang *et al.* (2007). Il s'agit d'une méthode très performante une fois appliquée sur des échantillons de faible effectif et qui n'exige pas d'hypothèses antérieures sur les données. Cette technique sert à classer un ensemble formé de n individus, en identifiant des points spéciaux appelés vecteurs de support à partir d'un jeu de données en entrée. Ces points permettent de décrire une frontière entre les différentes observations.

L'objectif principal des SVMs est de trouver les paramètres de l'hyperplan frontière, en maximisant la distance entre cet hyperplan et les vecteurs de supports à partir des données d'apprentissage x. Un individu i est alors affecté à l'une des classe de la variable Y selon le signe de l'équation de l'hyperplan séparateur donné par la fonction f.

$$signe(f(x_i)) \quad = \quad signe(x_i.w + b) \tag{1.7}$$

où w est la normal de l'hyperplan et b une constante à déterminer. Les SVMs sont avantageuses par rapport aux autres méthodes non-paramétriques, de point de vue que la majorité de ces méthodes retiennent tous les vecteurs d'apprentissage pour définir la frontière de décision (Louradour, 2007), à contrario des SVMs qui se contentent des vecteurs de supports. Cette méthode est très performante une fois appliquée sur des petits échantillons ou pour des problèmes avec un nombre important de variables. Néanmoins, la puissance des SVMs est très sensible à la standardisation des données. Pa ailleurs, les SVMs nécessitent des choix fondamentaux tels que, le type de noyau à utiliser et les paramètres adéquats de ce noyau. Ce genre de choix peut s'avérer parfois couteux en termes de temps à cause de la nature des

techniques utilisées à ce niveau (validation croisée).

Nous reviendrons sur ces notions dans le chapitre 3, dans le quel nous présenterons plus en détails les principes des SVMs afin, de les appliqués ultérieurement sur l'échantillon d'emprunteurs non-clients.

Chapitre 2

Modèles de régression logistique et héritage gaussien

Dans le chapitre précédent, nous avons commencé par introduire les techniques les plus utilisées en littérature pour résoudre les problèmes de credit scoring. Dans ce chapitre nous introduisons une première méthode de modélisation qui est la méthode de régression logistique. Cette technique est connue par sa robustesse en matière de prédiction de risque (Tuffery, 2007). Elle est utilisée lorsque la variable expliquée est catégorielle. Toutefois, elle s'applique uniquement à des grands échantillons. Beninel et Biernacki (2009) ont évoqué la possibilité de contourner cette contrainte de taille, en exploitant les données d'une autre sous-population et ce en supposant l'existence de liens cachées entre les vecteurs de variable de ces deux sous-populations (Beninel et Biernacki, 2005, 2007) (Biernacki et Jacques, 2007). Nous consacrons la première section à la présentation du modèle logistique classique et à la description des concepts du modèle logistique mixte. Dans la section deux nous traitons les liens entre les deux sous-populations et ce en tenant compte des travaux de Biernacki et Jacques (2007) ainsi que ceux de Beninel et Biernacki (2005), (2007), (2009) effectués dans le cas de deux sous-populations gaussiennes.

2.1 Modèle de régression logistique

2.1.1 Le modèle de régression logistique classique

La régression logistique est présentée comme étant une méthode économétrique dans la quelle la variable endogène Y correspond au codage des emprunteurs : 0 s'il s'agit d'un mauvais emprunteur, 1 sinon. x représente la matrice des variables exogènes qui servent à expliquer la variable Y. Pour un emprunteur i on a :

$$Y_i = \begin{cases} 0 \ si \ \beta_0 + \beta^T x_i + \varepsilon_i \leq 0 \ ou \ encore \ \varepsilon_i \leq -\beta_0 - \beta^T x_i \\ \\ 1 \ si \ \beta_0 + \beta^T x_i + \varepsilon_i > 0 \ ou \ encore \ \varepsilon_i > -\beta_0 - \beta^T x_i \end{cases} \tag{2.1}$$

où β_0 est une constante et β est le vecteur des coefficients des variables exogènes. Les ε_i sont les perturbations supposées indépendantes, de moyenne nulle et de variance 1. Le choix du modèle est lié au choix de la loi des ε_i. En générale ces perturbations suivent une loi logistique (Bardos, 2001), dont la fonction de répartition est la suivante :

$$F(x) = \frac{1}{1 + exp^{-x}} \tag{2.2}$$

Un modèle de régression logistique est défini en termes de valeurs interprétées comme des probabilités, il s'agit de la probabilité que l'événement arrive dans des sous-populations différentes, appellée probabilité a priori et notée, π. Dans le cadre binaire, pour un individu donné, sa probabilité a priori d'avoir la modalité 1 s'écrit comme suit :

$$\pi_i = P(Y_i = 1) \text{ , pour } i = 1, 2, .., n \tag{2.3}$$

L'objectif de la régression logistique est de connaitre la valeur moyenne de Y pour tout valeur prise par x. Cette valeur est appelée la probabilité a posteriori ou de succès, et notée p_i. La probabilité a posteriori est définie comme la probabilité qu'un individu i reçoit la modalité 1 sachant les valeurs prises par les différents descripteurs. Notre but consiste à modéliser cette probabilité pour en déduire nos règles d'affectation. La probabilité de succée est alors donnée par :

$$\begin{aligned} p_i &= P(Y_i = 1/x_i) \\ &= P(\varepsilon_i > -\beta_0 - \beta^T x_i) \\ &= 1 - F(-\beta_0 - \beta^T x_i) \\ &= \frac{1}{1 + exp^{-\beta_0 - \beta^T x_i}} \\ &= \frac{exp^{\beta_0 + \beta^T x_i}}{1 + exp^{\beta_0 + \beta^T x_i}} \end{aligned} \tag{2.4}$$

et la probabilité d'échec est donnée par :

$$\begin{aligned} 1 - p_i &= P(Y_i = 0/x_i) \\ &= P(\varepsilon_i \leq -\beta_0 - \beta^T x_i) \\ &= \frac{1}{1 + exp^{\beta_0 + \beta^T x_i}} \end{aligned} \tag{2.5}$$

Une des transformations pratique pour la régression logistique est la transformation *logit*, donnée par :

$$\begin{aligned} logit(p_i) &= \ln \frac{p_i}{1 - p_i} \\ &= \beta_0 + \beta^T x_i \end{aligned} \tag{2.6}$$

2.1.2 Le modèle de régression logistique mixte

On note Ω et Ω^*, les deux populations étudiés respectivement de clients et de non-clients aux quelles nous associerons les probabilités a posteriori suivantes p et p^*. Notre objectif consiste à prédire la solvabilité des emprunteurs non-clients en utilisons les informations de ces deux sous-populations. Nous disposons de deux échantillons d'apprentissage : $S_A = \{(x_i, Y_i) : i = 1, ..., n\}$ et $S_A^* = \{(x_i^*, Y_i^*) : i = 1, ..., n^*\}$ issues respectivement de la sous-population des candidats clients et de la sous-population des candidats non-clients, on note :

- $\tilde{x} \in \mathbb{R}^d$: le vecteur des variables observés sur l'union de Ω et Ω^*.
- $\tilde{Y}$: la variable de réponse de type binaire, $\tilde{Y} \sim B(1, \pi)$.

On a $(\tilde{x}, \tilde{Y})_{|\Omega} = (x, Y)$ et $(\tilde{x}, \tilde{Y})_{|\Omega^*} = (x^*, Y^*)$. Le modèle logistique sur Ω est donné par la fonction suivante :

$$t(x_i, \theta) \quad = \quad P(Y_i = 1|x_i, \theta) \quad = \quad \frac{exp^{\beta_0 + \beta^T x_i}}{1 + exp^{\beta_0 + \beta^T x_i}} \tag{2.7}$$

Sur Ω^*, on a

$$t^*(x_i^*, \theta^*) \quad = \quad P(Y_i^* = 1|x_i^*, \theta^*) \quad = \quad \frac{exp^{\beta_0^* + \beta^{*T} x_i^*}}{1 + exp^{\beta_0^* + \beta^{*T} x_i^*}} \tag{2.8}$$

avec

- t et t^* sont les fonctions de score respectivement sur Ω et Ω^*.
- $\theta = \{(\beta_0||\beta^T) \in \mathbb{R}^{d+1}\}$ et $\theta^* = \{(\beta_0^*||\beta^{*T}) \in \mathbb{R}^{d+1}\}$ représentent l'ensemble des paramètres à estimer respectivement sur Ω et Ω^*.
- $(\beta_0||\beta^T)$ et $(\beta_0^*||\beta^{*T})$ représentent la concaténation de la constante du modèle avec le vecteur de coefficients des variables respectivement sur Ω et Ω^*.

Le modèle logistique mixte permet l'étude de différents problèmes de discrimination, la solution de ces divers problèmes dépend principalement des données en notre possession. On défini principalement deux problèmes, le premier consiste à l'estimation des paramètres des deux sous-populations simultanément et le second correspond à l'estimation des paramètres de la deuxième sous-population, en supposant que les paramètres de la première sont connus à l'avance.

La résolution du premier problème nécessite la possession de deux échantillons d'apprentissage de taille suffisante (un échantillon pour chaque sous-population). Tandis que, la résolution du deuxième nécessite la possession d'un seul échantillon d'apprentissage de taille suffisante. Le deuxième problème présent le cadre de notre étude précisément, nous disposons des règles de discrimination dans la sous-population des clients et nous cherchons à trouver des nouvelles règles dans la sous-population des non-clients.

L'existence d'une liaison entre les vecteurs de variables des deux sous-populations implique une liaison entre les fonctions de score définis en (2.7) et (2.8). Ainsi, l'utilisation de liaisons acceptables des deux sous-populations permet d'utiliser les informations cachées dans les échantillons S_A et S_A^* pour affecter dans les groupes, les individus de Ω^*. Nous cherchons dans ce qui suit ces liens en se basant sur les résultats trouvées dans le cade des modèles gaussiens.

2.2 Héritage gaussien et modèle de liaison entre fonction de score

Afin d'estimer les paramètres de la fonction de score dans la sous-population Ω^*, nous utilisons les informations relatives aux deux sous-populations Ω et Ω^*. L'utilisation des données de ces deux sous-populations vise à modérer la faible masse de données relative aux candidats non-clients, en supposant l'existence d'un lien caché entre la distribution des variables de ces sous-

populations.

Dans ce contexte nous exploitons l'idée des travaux de Beninel et Biernacki (2007), ainsi que Bouveyron et Jacques (2009) qui ont reposé sur l'étude des liens qui peuvent exister entre deux sous-populations . L'existence de liaisons entre les distributions des variables de ces deux sous-populations mènent à une relation entre les paramètres de leurs modèles respectives de régression logistiques.

Ainsi, l'enjeu est de chercher des liaisons acceptables entre les distributions de variables respectivement dans Ω et Ω^*. Dans ce cadre, un travail préliminaire a été mené avec succès dans le cas gaussien multivarié (Beninel et Biernacki, 2005). Il s'agit ici d'étendre ce principe au modèle logistique, ce qui mène à des modèles de liaison simples entre les paramètres des deux règles de classement logistiques associées respectivement aux deux populations.

2.2.1 Modèle gaussien : liaison affine entre sous-populations

En discrimination gaussienne, les données consistent en deux échantillons : un échantillon d'apprentissage A et un échantillon de prédiction P qui proviennent respectivement des sous-populations suivantes : Ω et Ω^*. Dans notre cas, ces deux sous-populations sont différentes.

L'échantillon d'apprentissage A est composé de n couples (x_i, Y_i), $i = 1, ..., n$ où, x_i est un vecteur de $\mathbb{R}^d$ représentant les caractéristiques numériques, qui décrivent l'individu i et où, Y_i est l'étiquette de son groupe d'appartenance. Les n couples (x_i, Y_i) sont supposés être des réalisations indépendantes est identiquement distribuées (i.i.d), du couple aléatoire (x, Y) défini sur Ω de distribution jointe :

$$x_{|Y=k} \sim N_d(\mu_k, \Sigma_k)\ k = \{1, ..., K\}, \text{ et } Y \sim M_K(1, \pi_1, ..., \pi_K) \tag{2.9}$$

où, $N_d(\mu_k, \Sigma_k)$ correspond à la distribution gaussienne de dimension d, de

moyenne μ_k et de matrice de variance-covariance Σ_k . $M_K(1, \pi_1, ..., \pi_K)$ correspond à la distribution multinomial de paramètres $\pi_1, ..., \pi_K$, avec π_k repérsente la proportion du groupe k dans la sous-population et le paramètre K représente le nombre de modalités de la variable Y.

L'échantillon de prédiction P est composé de n^* individus dont nous connaissons seulement les caractéristiques numériques x_i^*, $i = 1, ..., n^*$, supposés les mêmes que pour l'échantillon L. Les n^* étiquettes $Y_1^*, ..., Y_{n^*}^*$ sont à prédire. Les n^* couples (x_i^*, Y_i^*) sont supposés être des réalisations i.i.d. du couple aléatoire (x^*, Y^*) défini sur Ω^* de distribution jointe :

$$x^*_{|Y^*=k} \sim N_d(\mu_k^*, \Sigma_k^*),\ k = 1, ..., K, \text{ et } Y^* \sim M_K(1, \pi_1^*, ..., \pi_K^*) \tag{2.10}$$

Nous cherchons alors, à estimer les n^* étiquettes inconnues en utilisant l'information provenant des deux échantillons A et P. L'enjeu est alors d'identifier une relation liant les deux sous-populations.

Afin, de sortir les liens existant entre les deux sous-populations, nous faisons recoure à l'approche proposer par Beninel et Biernacki (2009), qui suppose l'existence d'une application $\phi_k : \mathbb{R}^d \to \mathbb{R}^d$ liant en loi les vecteurs aléatoires des sous-populations Ω et Ω^*.

L'expression de cette application est donnée par :

$$x^*_{|Y^*=k} \sim \phi_k(x_{|Y=k}) = [\phi_{k1}(x_{|Y=k}), ..., \phi_{kd}(x_{|Y=k})]^T \tag{2.11}$$

Les résultats provenant de Beninel et Biernacki (2009), confirent que La fonction ϕ_k est affine ce qui conduit aux relations suivantes :

$$x^*_{|Y^*=k} \sim \Lambda_k x_{|Y=k} + \alpha_k \tag{2.12}$$

où Λ_k est une matrice diagonale définie sur $\mathbb{R}^{d\times d}$ et α_k est un vecteur de $\mathbb{R}^d$. A partir de l'expression précédente on déduit les liens entre les paramètres des deux sous-populations définis par :

$$\mu_k^* = \Lambda_k \mu_k + \alpha_k \text{ et } \Sigma_k^* = \Lambda_k \Sigma_k \Lambda_k \tag{2.13}$$

2.2.2 Héritage gaussien

Anderson (1982) a montré l'existence d'un lien entre les paramètres du modèle gaussien adapté au mélange de deux sous-populations et les paramètres du modèle logistique correspondant à ces sous-populations. Dans le cas où, les populations Ω et Ω^* sont gaussiennes homoscédatiques conditionnellement aux groupes et les matrices de variances-covariances communes sont notées de la manière suivante :

$$\Sigma = \Sigma_1 = \Sigma_2 \text{ et } \Sigma^* = \Sigma_1^* = \Sigma_2^*, \tag{2.14}$$

nous obtenons le lien suivant entre les paramètres logistiques et les paramètres gaussiens des deux sous-populations :

$$\Omega \quad : \quad \beta_0 = \tfrac{1}{2}(\mu_2^T \Sigma^{-1} \mu_2 - \mu_1^T \Sigma^{-1} \mu_1) \text{ et } \beta = \Sigma^{-1}(\mu_1 - \mu_2) \tag{2.15}$$

$$\Omega^* \quad : \quad \beta_0^* = \tfrac{1}{2}(\mu_2^{*T} \Sigma^{*-1} \mu_2^* - \mu_1^{*T} \Sigma^{*-1} \mu_1^*) \text{ et } \beta^* = \Sigma^{*-1}(\mu_1^* - \mu_2^*) \tag{2.16}$$

En remplaçant les μ_k^*, $k = 1, 2$ et les Σ^* par leur expression donnée par (2.13) et en nous limitons aux relations linéaires qui peuvent exister entre

les paramètres des deux sous-populations, on obtient les expressions suivante pour β_0^* et β^* :

$$\beta_0^* = \alpha + \beta_0 \text{ et } \beta^* = \Lambda\beta \tag{2.17}$$

En remplaçant β_0^* et β^* par leurs expressions dans l'équation (2.8), nous obtenons la fonction de score suivante pour la population des non-clients :

$$t^*(x_i^*, \theta, \varrho) = \frac{exp^{\beta_0+\alpha+(\Lambda\beta)^T x_i^*}}{1 + exp^{\beta_0+\alpha+(\Lambda\beta)^T x_i^*}} \tag{2.18}$$

avec $\varrho = \{(\alpha||\Lambda) \in \mathbb{R}^{d+1}\}$ est l'ensemble des paramètres de transition à estimer où $(\alpha||\Lambda)$ est la concaténation du scalaire α et de la diagonale de la matrice Λ.

2.2.3 Modèles de liaison entre fonctions de score

L'estimation des liens entre les deux sous-populations Ω et Ω^* se fait à travers l'écriture de six sous modèles logistiques de liaisons, inspirés par le cas gaussien précédemment évoqué. Ces modèles de liaisons sont présentés par le tableau suivant :

Tableau 2.1 – *Modèles de liaison*

Modèles	**Paramètres**	**Descriptions**
$M1$	$\alpha = 0 \quad \Lambda = I_d$	Les fonctions de score sont identiques pour les deux sous-populations.
$M2$	$\alpha = 0 \quad \Lambda = \lambda I_d$	Les fonctions de score des deux sous-populations diffèrent uniquement au travers du paramètre scalaireλ.
$M3$	$\alpha \in \mathbb{R} \quad \Lambda = I_d$	Les fonctions de score des deux sous-populations diffèrent uniquement au travers du paramètre scalaire β_0^*.
$M4$	$\alpha \in \mathbb{R} \quad \Lambda = \lambda I_d$	Les fonctions de score des deux sous-populations diffèrent au travers du couple de paramètres (β_0^*, λ).
$M5$	$\alpha = 0 \quad \Lambda \in \mathbb{R}^{d\times d}$	Les fonctions de score des deux sous-populations diffèrent uniquement au travers du paramètre vectorielle β^*.
$M6$	$\alpha \in \mathbb{R} \quad \Lambda \in \mathbb{R}^{d\times d}$	Il n'existe plus de lien stochastique entre les discriminations logistiques des deux sous-populations. Tous les paramètres sont libres.

Pour chacun des modèles précédant dans le Tableau 2.1 l'évaluation des paramètres de transition est conditionnellement faite aux paramètres associés à la sous-population Ω. Nous ajoutons un septième modèle noté $M7$, qui consiste à introduire comme observations, tous les emprunteurs (client et non-clients) et appliquer une régression logistique simple. Cela consiste dans l'estimation jointe des paramètres de Ω et ceux de transition.

Chapitre 3

Les séparateurs à vaste marge

Dans ce chapitre, nous introduisons une deuxième méthode de modélisation qui est la méthode des SVMs. Il s'agit d'une méthode non-paramétrique, apparue avec les travaux de Vapnik (1995). Cette méthode a été utilisée dans plusieurs domaines tels que, la reconnaissance des formes (Pontil et Verri, 1998), la bioinformatique (Yu et *al* ,2003), etc. Récemment les SVMs ont aussi commencé à s'intégrer dans le domaine de credit scoring grâce au travaux de Huang *et al.* (2007).

Ce chapitre présente les fondements des SVMs. Nous consacrons la première section à la description des concepts de base des SVMs, et ce pour le cas où le classifieur est linéaire. Dans la section deux nous donnons une généralisation de ces concepts.

3.1 Les SVMs linéaires

3.1.1 Cas séparable

Les SVMs sont des techniques de classification binaire, dont la tâche est de construire un modèle sur les données d'apprentissage pour prédire la classe d'un nouvel individu. L'objectif de cette technique est de trouver une solution géométrique pour séparer les deux classes.

Nous supposons toujours que l'on dispose d'un échantillon d'apprentissage formé des couples $\{x_i, Y_i\}$, $i = 1, ..., n$, où x_i est le vecteur représentant les caractéristiques de l'individu i et Y_i est l'étiquette de son groupe d'appartenance. Dans la suite la variable Y prendera la valeur 1 pour désigner un emprunteur solvable (positif) et -1 pour désigner un emprunteur non solvable (négatif).

Notre objectif, consiste à trouver une frontière de séparation linéaire, qui permet de séparer les individus positifs $(+1)$ des négatifs (-1). La frontière de séparation est donnée par l'hyperplan d'équation : $w.x + b$, où w est la normal à l'hyperplan, $\frac{|b|}{||w||}$ est la distance perpendiculaire de l'hyperplan à l'origine, b est la constante de l'hyperplan et $||w||$ est la norme euclidienne de w. La règle d'affectation des nouveaux individus est donnée par :

$$Y_i = signe(x_i.w + b) \tag{3.1}$$

qui signifie :

$$\begin{cases} x_i.w + b \geq +1 \ pour \ Y_i = +1 \\ x_i.w + b \leq -1 \ pour \ Y_i = -1 \end{cases} \tag{3.2}$$

Les contraintes précédentes peuvent être combinées en une seule définie comme suit :

$$Y_i(x_i.w + b) - 1 \geq 0, \ i = 1, ..., n \tag{3.3}$$

En supposant, qu'il existe un hyperplan permettant de séparer les individus positifs des individus négatifs, il ne s'agit pas d'en trouver un, mais de

chercher celui qui sépare le mieux nos données. Ceci revient à chercher un hyperplan avec la marge maximale définie comme étant la distance minimale notée d_+ (respectivement d_-), entre le séparateur et le plus proche individu positif (respectivement négatif).

Considérons maintenant l'ensemble des points, qui vérifient la première inégalité dans l'équation (3.2). Ces points appartiennent à l'hyperplan $H1$: $x_i.w + b = 1$ avec une normal w et une distance perpendiculaire à l'origine $\frac{|1-b|}{||w||}$. De même les points pour lesquels la deuxième inégalité dans l'équation (3.2) est vérifiée. Ces points appartiennent à l'hyperplan $H2 : x_i.w+b = -1$, avec w comme normal et $\frac{|-1-b|}{||w||}$ comme distance perpendiculaire à l'origine, donc $d_+ = d_- = \frac{1}{||w||}$, la marge est alors donnée par $\frac{2}{||w||}$. Selon Burges (1998) les deux hyperplans $H1$ et $H2$ sont parallèles et qu'aucun point de l'échantillon d'apprentissage ne figure entre eux puisque ils possèdent la même normal. Ainsi, la paire des hyperplans qui donne la marge maximale est touvée en réduisant au minimum $||w||^2$ tout en respectant la contrainte (3.3).

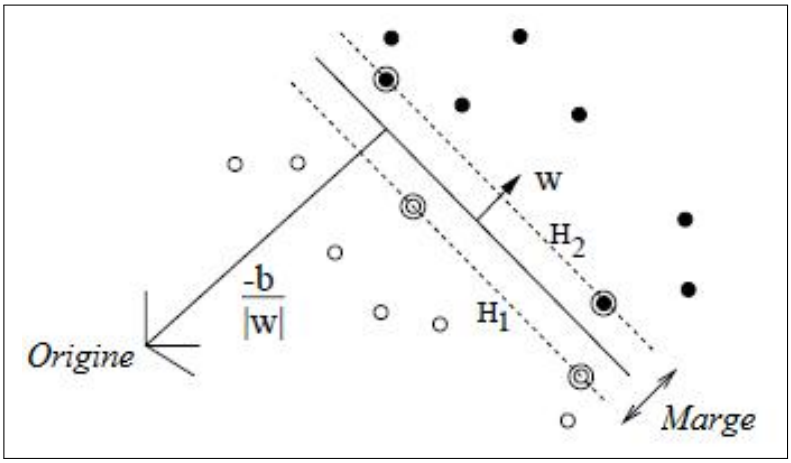

FIGURE 3.1 – *Séparateur linéaire pour le cas de données séparables adapté de Burges (1998)*

La solution est représentée par la Figure 3.1 où les points pour lesquels la

contrainte (3.3) est vérifiée et dont le déplacement change la solution trouvée, sont appelés les vecteurs de support. Ces points sont entourés par des cercles (cf Figure 3.1).

La recherche de l'hyperplan optimal revient à minimiser $||w||$, soit à résoudre le problème suivant qui porte sur les paramétres w et b :

$$\begin{cases} Min_{(w,b)} \dfrac{||w^2||}{2} \\ \text{s.c} \\ Y_i(x_i.w + b) - 1 \geq 0,\ i = 1, ..., n \end{cases} \tag{3.4}$$

Cette écriture du problème est appelée formulation primale. La résolution du problème primal revient à trouver $d + 1$ paramètres, où d est la dimension de x. Cela est possible à l'aide des méthodes de programmation quadratique lorsque la valeur de d est petite, mais devient impossible si d est élevée (Cornuejols, 2002).

Afin de trouver les paramètres w et b optimaux de l'hyperplan, Il s'agit de procédé à une formulation de Lagrange du problème. La méthode des multiplicateurs de Lagrange sert à remplacer la contrainte de l'équation (3.3) par les contraintes de Lagrange, qui sont plus faciles à traiter et qui permettent de généralisé la procédure pour les cas non linéaires (Burges, 1998). Après transformation, le problème primal L_P devient équivalent à l'expression définie par :

$$L_P \equiv \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i Y_i(x_i.w + b) + \sum_{i=1}^{n} \alpha_i \tag{3.5}$$

où α_i, $i = 1, ..., n$ représentent les multiplicateurs de Lagrange de signe positif, correspondant à chacune des contraintes d'inégalité dans (3.4). Les contraintes sous forme d'inégalité sont multipliées par des multiplicateurs

de Lagrange positifs et soustraites de la fonction objective, pour former la formulation de Lagrange. Une fois la forme de Lagrange est obtenue, le problème L_P est transformé dans sa formulation duale L_D, qui contient moins de paramètres à déterminer.

D'après la théorie de l'optimisation (Fletcher, 1987) un problème d'optimisation possède une forme duale lorsque la fonction objectif et les contraintes sont strictement convexes. Dans ce cas, la résolution du problème duale est équivalente à la résolution du problème original. Le passage au problème dual est possible dans notre car la fonction objective de (3.5) est convexe selon les conditions Karush-Kuhn-Tucker (KKT). Le problème dual est alors défini par :

$$\begin{cases} Max_{(\alpha_i)}\ L_D \equiv \sum_{i=1} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i Y_j x_i.x_j \\ \text{s.c} \\ \alpha_i \geq 0,\ i = 1, ..., n \\ \sum_i \alpha_i Y_i = 0 \end{cases} \tag{3.6}$$

L'équation de l'hyperplan optimal est alors donnée par :

$$x_i.\hat{w} + \hat{b} \tag{3.7}$$

où

$$\hat{w} = \sum_{i=1}^{n} \hat{\alpha}_i Y_i x_i \text{ et } \hat{b} = Y_j - \sum_{i=1}^{n} \hat{\alpha}_i Y_i (x_i.x_j) \tag{3.8}$$

avec x_j et Y_j étant n'importe quels points de l'échantillon s'apprentissage.

3.1.2 Cas non séparable

En réalité la séparation des classes est accompagnée d'erreurs de mauvaise classification, qui sont dues à la non-séparabilité des données. Ces erreurs sont intégrés dans la fonction objective primale par l'ajout d'un nouveau coût de mauvaise classification. Ce coût est représenté au travers d'une variable artificielle notée ξ_i, $i = 1, ..., n$. Cette variable sert à mesurer l'erreur pour un point donné. Les contraintes de (3.2) deviennent alors :

$$\begin{cases} x_i.w + b \geq +1 - \xi_i \ pour \ Y_i = +1 \\ x_i.w + b \leq -1 + \xi_i \ pour \ Y_i = -1 \end{cases} \tag{3.9}$$

L'ajout d'un terme d'erreur nécessite la satisfaction de deux objectifs, le premier consiste à maximiser la marge et le deuxième revient à minimiser l'erreur de classification sans perdre pour autant la généralité et la simplicité. Il s'agit d'un problème multiobjectifs sa résolution est effecuée en adoptant la méthode des poids. Il s'agit d'affecter un poids β_i, $i = \{1, 2\}$ à chaque objectif, la fonction objective est donnée dans ce cas par :

$$\begin{cases} Min_{(w,b)} \beta_1 \dfrac{||w^2||}{2} + \beta_2 \displaystyle\sum_{i=1}^{n} \xi_i \\ \text{s.c} \\ Y_i(x_i.w + b) \geq 1 - \xi_i, \ i = 1, ..., n \\ \xi_i \geq 0, \ \forall i = 1, ..., n \\ \beta_1 + \beta_2 = 1 \end{cases} \tag{3.10}$$

La fonction objective s'écrit aussi de la manière suivante :

$$Min_{(w,b)}\frac{||w^2||}{2} + \frac{\beta_1}{\beta_2}\sum_{i=1}^{n}\xi_i \tag{3.11}$$

La fraction $\frac{\beta_1}{\beta_2}$ est remplacée par le paramètre C, appelè paramètre régulateur : plus la valeur de C augmente plus la pénalité affectée à l'erreur augmente. En introduisant ce paramètre dans la fonction objective, l'équation de la fonction objective devient la suivante :

$$\begin{cases} Min_{(w,b)}\frac{||w^2||}{2} + C\sum_{i=1}^{n}\xi_i \\ \text{s.c} \\ Y_i(x_i.w + b) \geq 1 - \xi_i,\ i = 1, ..., n \\ \xi_i \geq 0,\ i = 1, ..., n \end{cases} \tag{3.12}$$

La formulation de Lagrange correspondante à ce problème est donnée par :

$$L_p \equiv \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\{Y_i(x_i.w + b) - 1 + \xi_i\} - \sum_{i=1}^{n}\mu_i\xi_i \tag{3.13}$$

où μ_i, $i = 1, ..., n$ représentent les multiplicateurs de Lagrange de signe positif correspondant à chacune des contraintes d'inégalité sur le signe de ξ_i. Le problème dual respectif est donné par :

$$\begin{cases} Max_{(\alpha_i)}\ L_D \equiv \sum_{i=1}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j Y_i Y_j x_i.x_j \\ \text{s.c} \\ 0 \geq \alpha_i \leq C,\ i = 1, ..., n \\ \sum_i \alpha_i Y_i = 0 \end{cases} \tag{3.14}$$

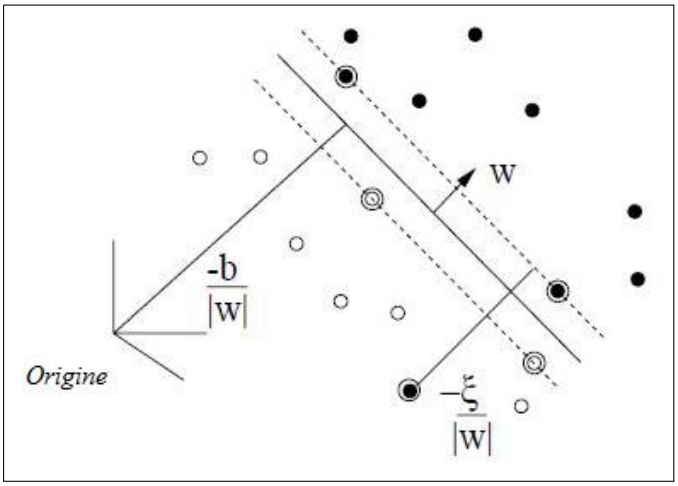

FIGURE 3.2 – *Séparateur linéaire pour le cas de données non séparables adapté de Burges (1998)*

La solution a ce problème est présentée par la Figure 3.2. la résolution du problème d'optimisation donne comme solution une estimation sur les valeurs des $\hat{\alpha}_i$ ce qui nous fourni $\hat{w}$ et $\hat{b}$, qui restent les mêmes comme définis dans l'équation (4.5). La classe d'appartenance d'un nouvel individu est alors donnée par le signe de l'expression suivante :

$$signe(\hat{w}.x + \hat{b}) \tag{3.15}$$

3.2 Les SVMs non linéaires

Dans la majorité des cas réels, il n'y a pas de vrai séparation linéaire entre les données d'apprentissage. Les traitements précédents ne peuvent pas être utilisés car ils ne sont applicables que sur des données d'apprentissage linéairement séparables.

L'idée des SVMs non linéaire est de transformer les données non linéairement séparables, en des données linéairement séparables dans un nouvel espace de données, appelé espace de redescription. La transformation des

données est effectuée en projetant les données d'entrée x dans un autre espace P de dimension D supérieure à la dimension d de l'ancien espace. Cette transformation est effectuée grâce à une fonction de transformation notée Φ, $\Phi : \mathbb{R}^d \to P$.

La transformation des x_i dans un espace de grande dimension, permet d'augmenter la probabilité de trouver une fonction de séparation linéaire dans le nouveau espace, ce qui n'a pas été possible dans l'ancien, ceci rend l'utilisation des algorithmes linéaire tels que les SVMs fort envisageable. La même procédure d'optimisation de la section précédente est alors appliquée sur les données projetées par $\Phi(x)$. Le nouveau problème d'optimisation est alors défini par :

$$\begin{cases} Max_{(\alpha_i)}\ L_D \equiv \sum_{i=1} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i Y_j \Phi(x_i).\Phi(x_j) \\ \text{s.c} \\ 0 \geq \alpha_i \leq C,\ i = 1, ..., n \\ \sum_i \alpha_i Y_i = 0 \end{cases} \tag{3.16}$$

L'équation de l'hyperplan optimal est alors donnée par :

$$\sum_{i=1}^{n} \hat{\alpha_i} Y_i \Phi(x_i).\Phi(x_j) + \hat{b} \tag{3.17}$$

L'utilisation de la fonction Φ, apporte une solution au problème de la non linéarité des donné. Néanmoins, un problème est susceptible de se produire lors du calcul du nouveau vecteur de caractéristique $\Phi(x_i)$. La difficulté réside dans le calcule des produits scalaires $\Phi(x_i).\Phi(x_j)$ qui est difficile voir impossible, si le nouveau espace est fortement dimensionnel ou de dimension infini.

Afin, de contourner ce problème, l'idée est de recourir à un type particulier de fonctions non-linéaires, appelées fonctions noyau et notées K, $K(x_i, x_j) = \Phi(x_i).\Phi(x_j)$. De telles fonctions permettent le calcul des produits scalaires dans un espace fortement dimensionnel. Il s'agit de remplacer $\Phi(x_i).\Phi(x_j)$ par $K(x_i, x_j)$, dans la fonction objective pour obtenir un hyperplan séparateur dans un espace de dimension plus forte. Ainsi, toutes les considérations des sections précédentes sont à retenir puisque il s'agit toujours d'une séparation linéaire, mais dans un espace différent. La fonction objective est alors définie par :

$$\begin{cases} Max_{(\alpha_i)}\ L_D \equiv \sum_{i=1} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i Y_j K(x_i, x_j) \\ \text{s.c} \\ 0 \geq \alpha_i \leq C,\ i = 1, ..., n \\ \sum_i \alpha_i Y_i = 0 \end{cases} \tag{3.18}$$

L'équation de l'hyperplan optimal est alors donnée par :

$$\sum_{i=1}^{n} \hat{\alpha}_i Y_i K(x_i, x_j) + \hat{b} \tag{3.19}$$

3.2.1 Fonctions de noyau

Dans ce qui précède nous avons évoqué le rôle des fonctions noyau dans la projection des données dans l'espace de redescription . La question est donc comment choisir ces fonctions et sous quels critères une fonction est jugée comme étant un noyau.

Condition de Mercer

Il existe une condition mathématique, appelée condition de Mercer (1909) qui permet de dire si une fonction est un noyau. Cette condition est définie comme suit :
Si on prend une fonction K symétrique, il existe une fonction Φ telle que :

$$K(x_i, x_j) = \Phi(x_i)\Phi(x_j) \tag{3.20}$$

Si est seulement si, pour toute fonction g telle que $\int g(x_i)^2 dx_i$, on a :

$$\int\int K(x_i, x_j)g(x_i)g(x_j)dx_i dx_j \geq 0 \tag{3.21}$$

Fonctions noyau usuelles

Si nous utilisons un noyau qui ne satisfait pas la condition de Mercer, en général la fonction objective duelle peut devenir arbitrairement grand, ce qui peut conduire à une solution infaisable. Toutefois, Vapnik (1995) a montré que dans certains cas même avec ces fonctions l'algorithme d'apprentissage converge parfaitement bien. En pratique il est très difficile de vérifier si quelques noyaux satisfont les conditions de Mercer, pour cette raison certains noyaux particuliers sont utilisés, nous citons dans ce qui suit les plus utilisés.

- Linéaire : les fonctions de type linéaire ont pour fonction noyau associée, $K(x_i, x_j) = x_i.x_j$.
- À base radiale : les fonctions à base radiale ont pour fonction noyau associée, $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$, $\gamma \in \mathbb{R}$.
- Polynômial : Les polynômes de degré d ont pour fonction noyau associée, $K(x_i, x_j) = (\gamma x_i.x_j)^d$, $d \in \mathbb{N}$.

- Sigmoïde : ce sont des réseaux de neurones qui, ont pour fonction noyau associée $K(x_i, x_j) = \tanh(\gamma x_i.x_j)$, $\gamma \in \mathbb{R}$.

Chapitre 4

Mise en place des modèles

Dans ce chapitre nous présentons les étapes nécessaires pour la mise en place des divers modèles. Nous consacrons les deux premières sections à la description des données réelles que nous allons traité et à la sélection des variables les plus pertinentes. La section trois est consacrée au choix des échantillons d'apprentissage. Dans les sections quatre et cinq nous présentons les étapes nécessaires pour calculer les paramètres des différents modèles.

4.1 Description des données

Nous allons appliquer nos modèles sur le jeu de données *DeutschBank* qui est composé de 1000 observations représentant les crédits à la consommation accordés par la *DeutchBank* de Munich. Ces données sont disponibles sur le site *http ://www.stat.uni-muenchen.de/service/datenarchiv/kredit/ kredit_e.html* (1994). Chaque candidat est représenté par la variable à expliquer appelée *kredit* qui informe sur sa crédibilité. Par ailleurs, les 1000 observations sont classées en emprunteurs solvables et non solvables selon les modalités de cette variable (cf. Figure 4.1), de ces 1000 observations : 30% sont considérés comme non solvables et 70% comme solvables.
D'autre part nous disposons de 20 variables censés influencer la capacité de

remboursement du crédit. Une variable qui nous intéresse particulièrement, c'est la variable *laufkount* qui sert à distinguer les individus qui sont clients de la banque de ceux des qui ne le sont pas.

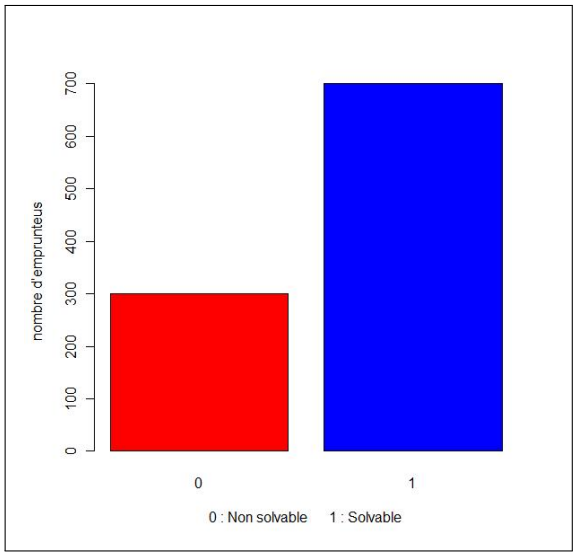

FIGURE 4.1 – *Distribution des emprunteurs selon les modalités de la variable kredit*

Les 19 variables restantes sont quant à elles des variables potentiellement explicatives. En effet, chacune de ces variables est supposée expliquer les valeurs prises par la variable *kredit*. Néanmoins, il est possible que certaines de ces variables n'aient pas d'influence significative sur le modèle. Il est donc de notre intérêt de ne pas les intégrer lors du traitement statistique. Les variables sont réparties en 3 variable quantitatives et 16 variables qualitatives, le détail de ces variables est expliqué dans l'Annexe.

4.2 Sélection des variables

Dans cette section nous disposons d'une série de variables candidates et nous cherchons les variables les plus pertinentes pour expliquer et prédire les valeurs prises par la variable cible. Ainsi, nous cherchons à éliminer les

variables explicatives, ne permettant pas d'expliquer la variable *kredit*, et ce quelles soient qualitatives ou quantitatives. L'étude du pouvoir discriminant d'une variable qualitative est effectué à l'aide du test du khi-deux et du V de Cramer. Concernant les variables quantitatives, nous allons effectuer une $ANOVA$.

L'application des tests cités précédemment sur les individus de la sous-population Ω^* des non-clients est susceptible de donner des résultats non fiables à cause du manque d'observations dans celle-ci. Par conséquence nous allons nous limiter à la sous-population Ω des emprunteurs clients dans l'étape de sélection des variables pour réaliser les tests décrits ci-dessus.

4.2.1 Pouvoir discriminant d'une variable qualitative

L'étude de la significativité de chacune des 16 variables qualitatives en notre possession est réalisée à l'aide du test du khi-deux. Ce test résume l'information contenue dans le tableau de contingence. Grâce à cette statistique nous pouvons juger la dépendance entre l'appartenance à un groupe et les modalités de la variable en question. On considère qu'une variable est significative si la $p - value$ donnée par le test est inferieure au seuil de 5%. L'hypothèse nulle d'indépendance est alors rejetée et nous pouvons alors conclure que la fiabilité du demandeur de prêt dépend de la variable testée. Les variables en rouge dans le Tableau 4.1 sont les variables pour lesquelles ce test est significatif.

D'autre part, le test du khi-deux est connu par être sensible à la taille de l'échantillon. En effet, ce test tend à favoriser les grands effectifs. Afin, d'éliminer l'effet de la taille de l'échantillon nous recourons au calcul du V de Cramer qui permet de pondérer la statistique du χ^2 et supprimer l'effet lié à la taille de l'échantillon. Les résultats de cette statistique sont aussi donnés par le Tableau 4.1 : plus la valeur du V de Cramer s'approche de 1 plus la liaison entre la variable étudiée et celle de réponse est vérifiée. En nous basons sur les travaux de Cohen (1988) nous allons considérer le seuil

de 0.1 pour évaluer la force d'association entre les variables. Les variables présentées en bleu sont les variables sélectionnées par cette procédure. Ainsi, nous allons conserver les variables explicatives pour lesquelles le test du khi-deux est significatif et dont le V de Cramer est supérieur à 0.1 et qui sont les suivantes : *sparkont*, *weitkred*, *verw*, *beruf*, *moral*, *verm*, *beszeit*, *famges* et *wohn*.

Tableau 4.1 – *Sélection des variables qualitatives*

Variables	χ^2	**p-value**	**V de Cramer**
sparkont	21.655	0.00023	0.173
beszeit	22.566	0.00015	0.176
moral	33.620	< 0.0001	0.215
rate	0.920	0.820	0.356
verm	13.820	0.003161	0.138
famges	7.972	0.04658	0.105
wohnzeit	3.094	0.3773	0.065
bishkred	2.548	0.4667	0.059
beruf	12.380	0.00618	0.131
buerge	6.016	0.04939	0.091
weitkred	15.301	0.00047	0.145
wohn	10.596	0.00500	0.121
pers	0.077	0.782	0.01
gastarb	3.270	0.07057	0.067
telef	0.072	0.7881	0.01
verw	32.518	0.00016	0.212

4.2.2 Pouvoir discriminant d'une variable quantitative

L'étude de la significativité de chacune des 3 variables quantitatives en notre possession est réalisée à l'aide de l'analyse de variance $ANOVA$. Cette procédure nous permet de tester si les valeurs de ces variables s'organisent indépendamment selon les modalités de la variable *kredit*. Le Tableau 4.2 illustre les résultats du test $ANOVA$.

Tableau 4.2 – *Sélection des variables quantitatives*

Variable	**D L**	**Somme des carrées**	**Moyenne des carrées**	**Valeur** F	$p - value$
alter	1	0.979	0.979	5.600	0.018
hoehe	1	5.047	5.047	29.837	$6.465e^{-}08$
laufzeit	1	4.68	4.68	27.587	$1.977e^{-}07$

A partir des résultats du Tableau 4.2, nous allons retenir les trois variables suivantes : *laufzeit*, *hoehe* et *alter*.

4.2.3 Sélection automatique des variables discriminantes

Afin d'affiner notre analyse nous allons procéder à une sélection pas à pas ou (Stepwise). Il d'agit d'une méthode mixte qui présente une combinaison entre la méthode de sélection ascendante et la méthode de sélection descendante. La procédure commence par un modèle vide, ensuite à chaque étape une nouvelle variable parmi les variables sélectionnées par les procédures précédentes est ajoutée. Le nouveau modèle est alors évalué à l'aide d'un critère d'optimisation. Ensuite, toutes les variables dans le modèle sont remises en question. La procédure s'arrête, lorsque toutes les variables hors modèle n'apportent aucune amélioration au modèle étudié. Pour mettre en œuvre cette méthode nous avons recouru à la procédure *Stepwise* du logiciel R. Les variables retenus à partir de cette sélection et qui seront utilisées dans la suite de ce travail sont les suivantes : *sparkont*, *moral*, *beszeit*, *weitkred* et *laufzeit*.

4.3 Construction de l'échantillon d'apprentissage et de l'échantillon test

La population dont le comportement est à prédire dans le cadre de ce travail est celle des emprunteurs non-clients de la banque. De ce fait nous

avons commencé par diviser les observations en notre disposition en deux sous-populations : une première sous-population formée de 726 emprunteurs client de la banque et une deuxième sous-population formée de 274 emprunteurs non-clients.
Par la suite, nous avons devisé la sous-population des emprunteurs non-clients en deux échantillons : un échantillon d'apprentissage S_A et un échantillon de test S_T. Le premier échantillon permet de modéliser les divers modèles et de construire les règles d'affectation d'un individu selon ces caractéristiques. L'échantillon test a pour objectif de vérifier si le modèle fondé sur l'échantillon d'apprentissage est statistiquement fiable.

4.4 Mise en place des modèles de régression logistique

La mise en place des sept modèles de régression logistique ne peut se faire directement à l'aide des logiciels de statistique et notamment à l'aide du logiciel R que nous allons utiliser. Lors de l'utilisation de ces derniers nous somme confronté aux obstacles dus à l'utilisation d'un mélange de deux sous-populations pour estimer les coefficients d'un même modèle logistique. Pour résoudre ce problème nous avons dû faire un traitement préliminaire sur les données.
Dans la suite on divise Ω^* en deux sous-ensembles Ω_1^* et Ω_2^*.

4.4.1 Mise en place du modèle logistique $M1$

Le couple de paramètres à estimer pour ce modèle est donné par le couple (β_0, β). Afin de trouver une estimation de ces paramètres nous appliquons une régression logistique simple sur les observations issus de la sous-population Ω. La régression logistique est effectuée en utilisant la fonction *glm* sous R qui admet comme paramètres : la variable à expliquer *kredit*, l'ensemble

des variables explicatives retenus au cours de la sélection des variables et l'échantillon des emprunteurs clients. Conditionnellement à la connaissance des paramètres de Ω nous évaluons les paramètres de transition entre les deux sous-populations.

4.4.2 Mise en place des modèles logistique $M2$, $M3$ et $M4$

Les couples de paramètres à estimer pour les modèles $M2$, $M3$ et $M4$ sont données par les couples (β_0^*, β^*). Ces paramètres sont estimés à partir des données d'un échantillon d'apprentissage $S_A = \Omega_1^*$, et ce en se basant sur les liens entre les deux sous-populations (cf. Chapitre 2). Rappelons que les paramètres à estimer pour ces trois modèles sont données par :

$$\beta_0^* = \alpha + \beta_0 \text{ et } \beta^* = \Lambda\beta \tag{4.1}$$

et que la fonction de score pour un individu i s'écrit :

$$\begin{aligned} t^*(x_i^*, \theta, \varrho) &= \frac{exp^{\beta_0^* + \beta^{*T} x_i^*}}{1 + exp^{\beta_0^* + \beta^{*T} x_i^*}} \\ &= \frac{exp^{\beta_0 + \alpha + (\Lambda\beta)^T x_i^*}}{1 + exp^{\beta_0 + \alpha + (\Lambda\beta)^T x_i^*}} \\ &= \frac{exp^{\beta_0 + \alpha + \Lambda^T \sum_{j=1}^{d} \beta_j \times x_i^{*j}}}{1 + exp^{\beta_0 + \alpha + \Lambda^T \sum_{j=1}^{d} \beta_j \times x_i^{*j}}} \end{aligned} \tag{4.2}$$

En notant *Somme* la variable $\sum_{j=1}^{d} \beta_j \times x_i^{*j}$, la fonnction de score t^* s'écrit alors :

$$t^*(x_i^*, \theta, \varrho) \quad = \quad \frac{exp^{\beta_0+\alpha+\Lambda^T Somme}}{1+exp^{\beta_0+\alpha+\Lambda^T Somme}} \tag{4.3}$$

où $\varrho = \{(\alpha||\Lambda) \in \mathbb{R}^{d+1}\}$ est l'ensemble des paramètres de transition à estimer avec $(\alpha||\Lambda)$ la concaténation du scalaire α et de la diagonale de la matrice Λ. $\theta = \{(\beta_0||\beta^T) \in \mathbb{R}^{d+1}\}$ représente l'ensemble des paramètres à estimer sur Ω. La variable *Somme* joue le rôle d'un opérateur linéaire permettant d'ajuster les données afin de pondérer les variables avec l'information apportée par Ω. Dans ce qui suit nous décrivons les étapes nécessaires pour mettre en place les trois modèles $M2$, $M3$ et $M4$ sous R.

- **Modèle** $M2$ **:** $\alpha = 0$ et $\Lambda = \lambda I_d$

  Il consiste à appliquer la procédure *glm* sur un modèle formé de l'unique variable *Somme* sur un l'échantillon d'apprentissage S_A. Il suffit de préciser à la fonction *glm* de ne pas évaluer la constante de ce modèle. La constante à considérer et celle du modèle $M1$. Finalement ce modèle permet d'estimer λ.

- **Modèle** $M3$ **:** $\alpha \in \mathbb{R}$ et $\Lambda = I_d$

  Les paramètres du modèle $M3$ sont déduits de ceux du modèles $M1$. Seule la constante est estimée à partir d'un nouvel échantillon d'apprentissage S_A. Nous estimons ce modèle en précisant à la fonction *glm* de considérer la variable *Somme* comme une constante, ce qui revient à n'estimer que β_0^*.

- **Modèle** $M4$ **:** $\alpha \in \mathbb{R}$ et $\Lambda = \lambda I_d$

  Le modèle $M4$ est le fruit des modifications effectuées sur $M2$ et $M3$. Il

s'agit d'appliquer la fonction *glm* sur un modèle constituer de l'unique variable *Somme* sans imposer des contraintes, et ce en utilisant les informations issus de l'échantillon d'apprentissage S_A. Ce modèles permet d'estimer le couple de paramètres (β_0^*, λ).

4.4.3 Mise en place des modèles logistique $M5$, $M6$ et $M7$

Dans ce qui suit nous décrivons les étapes nécessaires pour mettre en place les modèles restants sous R.

- **Modèle $M5$: $\alpha = 0$ et $\Lambda \in \mathbb{R}^{d\times d}$**

  Ce modèle consiste à appliquer la fonction *glm* sur l'échantillon d'apprentissage d'emprunteurs non-clients $S_A = \Omega_1^*$. Il suffit de préciser à la fonction *glm* de ne pas évaluer la constante de ce modèle. La constante à considérer et celle du modèle $M1$. Ce modèle permet d'estimer Λ.

- **Modèle $M6$: $\alpha \in \mathbb{R}$ et $\Lambda \in \mathbb{R}^{d\times d}$**

  Ce modèle consiste à appliquer une simple régression logistique sur l'échantillon d'apprentissage $S_A = \Omega_1^*$ sans aucune contrainte. Ce modèle permet d'estimer le couple de paramètres (β_0^*, Λ).

- **Modèle $M7$:**

  Le modèle $M7$ consiste à appliquer une régression logistique simple sur un échantillon formé des emprunteurs clients et des non-clients $S_A = \Omega_1^* \cup \Omega$. Ce modèle permet d'estimer le couple de paramètres (β_0^*, β^*).

Le Tableau 4.3 résume les différents paramètres à estimer pour les sept modèles avec le sous-ensemble Ω_2^* servira par la suite comme échantillon de test.

Tableau 4.3 – *Tableau récapitulatif des paramètres à estimer*

Modèles	S_A	S_T	**Paramètres de transition**	**Paramètres estimés**
$M1$	Ω	Ω_2^*		$\hat{\beta}_0^* = \beta_0$ et $\hat{\beta}^* = \beta$
$M2$	Ω_1^*		λ	$\hat{\beta}_0^* = \beta_0$ et $\hat{\beta}^* = \lambda\beta$
$M3$			α	$\hat{\beta}_0^* = \alpha + \beta_0$ et $\hat{\beta}^* = \beta$
$M4$			α et λ	$\hat{\beta}_0^* = \alpha + \beta_0$ et $\hat{\beta}^* = \lambda\beta$
$M5$			Λ	$\hat{\beta}_0^* = \beta_0$ et $\hat{\beta}^* = \Lambda\beta$
$M6$			α et Λ	$\hat{\beta}_0^* = \alpha + \beta_0$ et $\hat{\beta}^* = \Lambda\beta$
$M7$	$\Omega_1^* \cup \Omega$			$\hat{\beta}_0^*$ et $\hat{\beta}^*$

Les coefficients estimés pour les différents modèles sous R sont donnés par le Tableau 4.4.

Tableau 4.4 – *Estimation des coefficients des divers modèles logistiques sous R*

Variables	**Coefficients**						
	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$	$M7$
Constante	−0.500	−0.500	−1.609	−1.314	−0.500	−0.485	−0.651
laufzeit	−0.026	−0.0039	−0.026	−0.021	−0.039	−0.035	−0.027
moral	0.013	0.002	0.013	0.011	0.022	0.025	0.017
sparkont	0.011	0.001	0.011	0.009	0.006	0.006	0.011
beszeit	0.011	0.001	0.011	0.009	−0.001	−0.0002	0.008
weitkred	0.007	0.001	0.007	0.005	−0.004	−0.003	0.004

Après avoir estimé les coefficients des divers modèles, nous remplaçons ces coefficients par leurs valeurs dans l'équation (2.8) donnée par :

$$\frac{exp^{\beta_0^* + \beta^{*T} x_i^*}}{1 + exp^{\beta_0^* + \beta^{*T} x_i^*}}$$

On obtient ainsi une probabilité a posteriori. Cette probabilité est alors comparée au seuil de 0.5. Un individu est alors affecté à la classe solvable si sa

probabilité a posteriori est supérieure à 0.5, sinon il sera affecté à la classe non solvable.

4.5 Mise en place des modèles basés sur les SVMs

L'utilisation des SVMs pour résoudre un problème de credit scoring se défini en termes d'une recherche d'un ensemble de paramètres optimaux. Nous cherchons dans ce travail le paramètre de régularisation C qui sert à pénaliser l'erreur de mauvaise classification, une bonne famille de fonctions noyau, ainsi qu'une bonne combinaison des paramètres de ces fonction noyau.

Nous considérons quatre modèles basés sur les SVMs notés : $S1$, $S2$, $S3$ et $S4$ ces quatre modèles sont basés respectivement sur un noyau : linéaire, à base radiale, polynomial ou sigmoïde. Les paramètres à estimer pour ces différents modèles sont donnés par le Tableau 4.5.

Tableau 4.5 – *Paramètres à estimer des quatre noyaux*

Fonction noyau	Forme fonctionnelle	Paramètres	Valeur par défaut
Linéaire	$x_i.x_j$		
A base radiale	$exp(-\gamma x_i.x_j)$	$\gamma \in \mathbb{R}$	$\gamma = 1$
Polynomial	$(\gamma x_i.x_j)^d$	$\gamma \in \mathbb{R}$ et $d \in \mathbb{N}$	$\gamma = 1$ et $d = 3$
Sigmoïde	$tanh(\gamma x_i.x_j)$	$\gamma \in \mathbb{R}$	$\gamma = 1$

où x_i et x_j deux points de l'échantillon d'apprentissage.

L'approche habituelle pour déterminer la paire de paramètres optimaux proposée par Hsu *et al.* (2003) consiste à utiliser une *Grille de recherche*. On se donne un ensemble de valeurs possibles des paires (C, γ) ou (C, d). Il s'agit d'estimer pour chacune de ces paires, l'erreur de généralisation du meilleur modèle calculée par validation croisée. Le couple de paramètre choisi est celui qui permet d'obtenir la performance maximale en la mesurant empiriquement par validation croisée.

Le principe de la validation croisée consiste à diviser l'échantillon d'origine aléatoirement en $k = 10$ sous-ensembles d'apprentissage et de tests indépendants et de même taille. Des k sous-ensembles, un seul est conservé comme échantillon de test. Les $k - 1$ restants sont utilisés comme échantillons d'apprentissage (Huang et al, 2007). Chacun des $k-1$ sous-ensembles sert à estimer un modèle. Ensuite, les modèles construis sur ces échantillons d'apprentissage sont testés sur l'échantillon restant ce qui permet de calculer l'erreur de généralisation pour chaque modèle.
Selon Salzberg (1997) les avantages de la validation croisée consistent que l'impact de la dépendance de données est réduite au minimum et la fiabilité des résultats peuvent être améliorées.

4.5.1 Choix du paramètre de régularisation C

Nous nous limitons dans un premier temps aux valeurs des paramètres par défaut (cf. Tableau 4.5) associés aux quatre modèles d'SVMs et nous allons faire varier le paramètre C afin d'évaluer l'impact du type de noyau et de ce paramètres sur le pouvoir de généralisation du modèle. Nous utilisons la fonction *tune.svm* sous R. Cette fonction permet de tester plusieurs valeurs du paramètre C en estimant la performance de prédiction pour un noyau donné par validation croisée. Le Tableau 4.6 illustre les résultats donnés par cette fonction sous R pour un noyau à base radiale et pour une valeur de $C = 2^m$ avec $m \in \{2, ..., 9\}$.

A partir du Tableau 4.6, on remarque que le modèle $S2$ associé à un noyau à base radiale donne la plus petite valeur d'erreur de généralisation lorsque $\gamma = 1$ et $C = 4$.

Tableau 4.6 – *Résultat de la fonction* tune.svm *(noyau à base radiale)*

Itération	C	**Erreur**
1	4	0.284
2	18	0.293
3	16	0.301
4	32	0.319
5	64	0.354
6	128	0.403
7	256	0.456
8	512	0.557

Le résultat de l'application de la fonction *tune.svm* sur les différents noyaux est illustré par la Figure 4.2. A partir de la Figure 4.2 nous pouvons remarquer que le modèle à noyau linéaire $S1$ fournit l'erreur de généralisation la plus petite. En effet, l'erreur de généralisation pour ce noyau varie dans l'intervalle $[0.23343, 0.23344, ..., 0.23347]$ lorsque l'en varie le paramètre C. Ces valeurs d'erreur sont considérées comme faibles par rapport à l'intervalle de variation d'erreur de généralisation des autres modèles. La valeur de l'erreur varie dans l'intervalle $[0.3, 0.34, ..., 0.55]$ pour un noyau à base radiale et pour un noyau polynomial cette erreur varie dans l'intervalle $[0.4, 0.5, ..., 0.9]$.

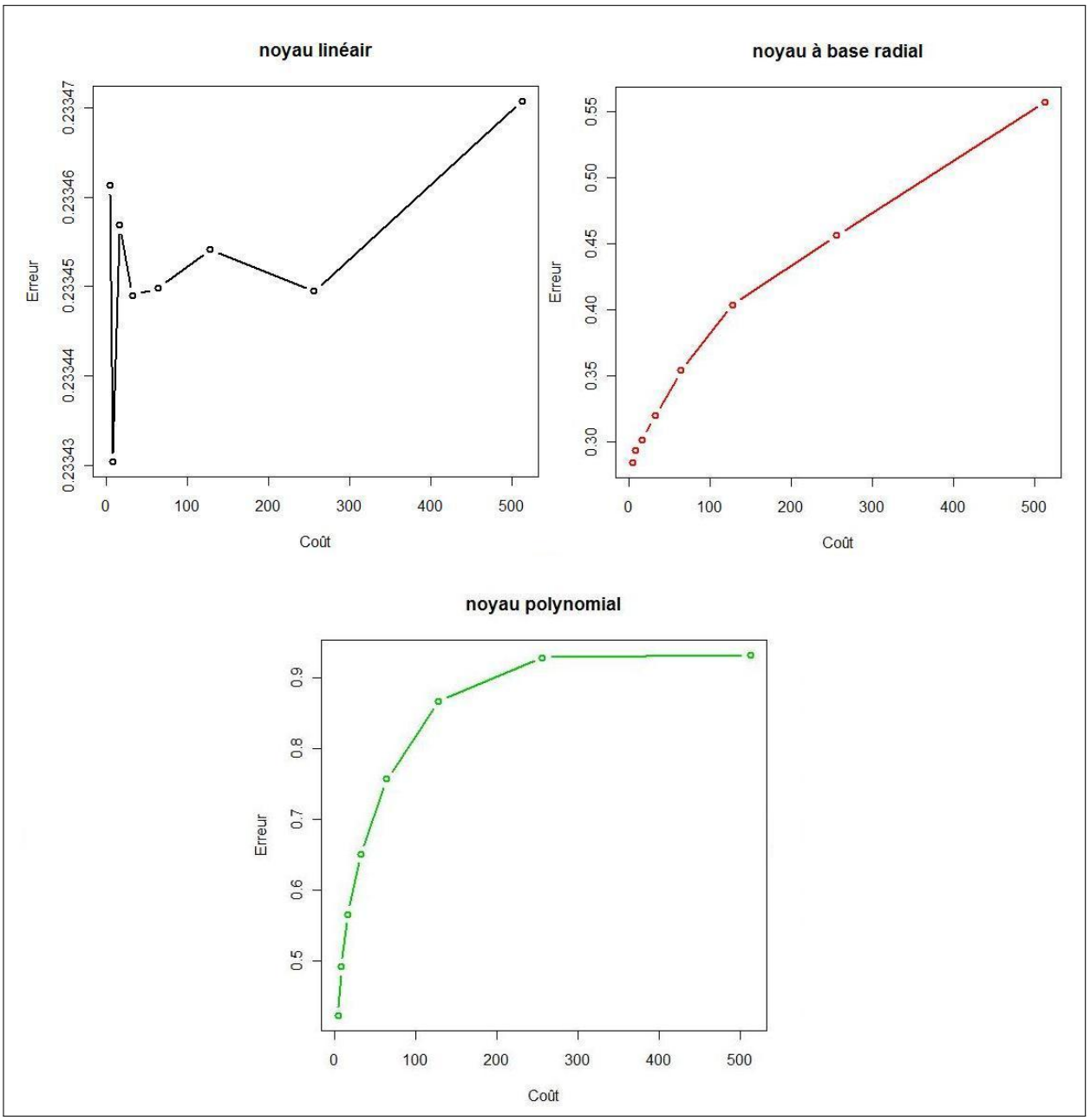

FIGURE 4.2 – *Influence de la variation de C sur la performance selon le type de noyau*

4.5.2 Choix des paramètres des divers noyaux

Le choix des paramètres appropriés des noyaux permet l'amélioration de la qualité de la classification des SVMs. Nous cherchons maintenant à varier les valeurs des paramètres des divers noyaux, tout en faisons varier la valeur du paramètre C et chercher la meilleur combinaison. Pour ce faire nous recourons à l'algorithme *Grille de recherche* évoqué précédemment (Huang et al, 2007). L'Algorithme 1 présente la *Grille de recherche* dans le cas du couple (C, γ)

Algorithme 1 *Grille de recherche()*

1: On considère un espace de recherche des couples de paramètres (C, γ) avec $\log_2(C) \in \{-5, -4, ..., 12\}$ et $\log_2(\gamma) \in \{-13, -12, ..., 7\}$.
2: Pour chaque paire d'hyperparamètres (C, γ) dans l'espace de recherche faire 10-validation croisé sur le jeu d'apprentissage.
3: choisir le couple de paramètres (C, γ) qui donnent le plus bas taux d'erreur de classification.
4: Utilisez le meilleur paramètre pour créer un modèle de prédiction.

La mise en œuvre de cet algorithme est assurée par l'application de la fonction *tune.svm* sous R. Pour des raisons techniques nous considérons les valeurs de $\log_2(C) \in \{2, ..., 9\}$ et de $\log_2(\gamma) \in \{-1, ..., 7\}$ pour les couples (C, γ) associés au modèles $S2$ et $S4$ respectivement à noyau à base radiale et sigmoïde. Pour un noyau polynomiale le degré d est considéré dans $\{1, ..., 4\}$. Le Tableau 4.7 donne les résultats de la *Gille de recherche* pour un noyau à base radiale avec variation du couple (C, γ). Les résultats sont fournis en appliquant la procédure *tune.svm* sur le même échantillon qui a servi à calculer les résultats présentés par le Tableau 4.6.
On remarque a partir du Tableau 4.7 qu'en faisant varier la valeur du paramètre γ de ce noyau, la valeur minimale de l'erreur empirique est passé de 0.284 à 0.264. Cette diminution confirme le faite que le choix des paramètres adéquats des noyaux améliore les résultats.

Tableau 4.7 – *Choix du C et γ (noyau à base radiale)*

$C = 4$					
γ	**Erreur**	γ	**Erreur**	γ	**Erreur**
0.5	0.351	1	0.3644	2	0.331
4	0.305	8	0.273	16	0.264
32	0.265	64	0.266	128	0.267
$C = 16$					
γ	**Erreur**	γ	**Erreur**	γ	**Erreur**
0.5	0.419	1	0.407	2	0.362
4	0.310	8	0.284	16	0.271
32	0.264	64	0.266	128	0.267
$C = 64$					
γ	**Erreur**	γ	**Erreur**	γ	**Erreur**
0.5	0.482	1	0.435	2	0.388
4	0.324	8	0.296	16	0.270
32	0.264	64	0.266	128	0.267
$C = 256$					
γ	**Erreur**	γ	**Erreur**	γ	**Erreur**
0.5	0.574	1	0.522	2	0.439
4	0.347	8	0.300	16	0.270
32	0.2647560	64	0.266	128	0.267
$C = 512$					
γ	**Erreur**	γ	**Erreur**	γ	**Erreur**
0.5	0.721	1	0.599	2	0.531
4	0.354	8	0.300	16	0.270
32	0.264	64	0.266	128	0.267

4.5.3 Mise en place des SVMs sous R

Le langage R offre une librairie destinée à la prédiction des paramètres des SVMs. Cette librairie à été réalisée par Chang et Lin (2001) et elle est intégrée au package e1071. Ce dernier est basé également sur la bibliothèque LIBSVM, qui est une bibliothèque qui regroupe un ensemble d'algorithmes de fouille de données.

En utilisant la procédure *svm* de la libraire e1071 nous pouvons calculer le nombre de vecteurs de supports trouvés pour un noyau donné. Ensuite en

nous basant sur ces vecteurs de support nous pouvons trouver les paramètres de l'hyperplan frontière ce qui nous permettra d'estimer l'étiquette du groupe d'appartenance d'un nouveau candidat de prêt (cf. Chapitre 3).

Chapitre 5

Evaluation des performances des différents modèles

La comparaison de la performance des différents modèles de prévision est une étape décisive pour choisir le modèle qui s'ajuste le mieux à nos données. Au cours de cette étape nous cherchons à étudier la faculté de discrimination et de généralisation de chaque modèle en nous basant sur un ensemble de critères de performance.
Nous consacrons la première section de ce chapitre à la présentation de la matrice de confusion ainsi qu'aux indicateurs de performance déductibles à partir de cette matrice. La deuxième section de se chapitre sera consacrée à l'évaluation des divers modèles en se basant sur un critère graphique.

5.1 Indicateurs déductibles à partir de la matrice de confusion

5.1.1 Matrice de confusion

La matrice de confusion est un outil de mesure du pouvoir d'un modèle à prédire avec précision les valeurs prises par la variable cible. Il s'agit d'un

tableau où les colonnes regroupent les occurrences de la classe de référence et les lignes regroupent les occurrences de la classe estimée. Le rôle de cette matrice est de confronter les valeurs estimées et les valeurs réelles. Cette matrice est présentée par le tableau suivant :

Tableau 5.1 – *Matrice de confusion*

		Estimé		
		Bon payeur $(Y=1)$	**Mauvais payeur** $(Y=0)$	**Total**
Réel	$(Y=1)$	VP	FP $(ErreurI)$	$VP+FP$
	$(Y=0)$	FN $(ErreurII)$	VN	$FN+VN$
	Total	$VP+FN$	$FP+VN$	N

avec

- ***VP (vrais positifs)*** : nombre d'emprunteurs positifs (solvables) correctement classés.
- ***FP (faux positifs)*** : nombre d'emprunteurs positifs faussement classés.
- ***VN (vrais négatifs)*** : nombre d'emprunteurs négatifs (non solvables) correctement classés.
- ***FN (faux négatifs)*** : nombre d'emprunteurs négatifs faussement classés.
- ***N*** : nombre d'emprunteurs dans l'échantillon.

On déduit à partir de cette matrice les indicateurs suivant :

- ***Taux de vrais positifs (sensibilité)*** : la proportion d'emprunteurs positifs correctement classés.

$$TVP = \frac{VP}{VP+FN} \tag{5.1}$$

- ***Taux de vrais négatifs (spécificité)*** : la proportion d'emprunteurs négatifs correctement classés.

$$TVN = \frac{VN}{VN + FP} \tag{5.2}$$

- ***Taux de faux positifs (Taux d'erreur de type I)*** : la proportion d'emprunteurs positifs faussement classés.

$$TFP = \frac{FP}{FP + VN} \tag{5.3}$$

- ***Taux de faux négatifs (Taux d'erreur de type II)*** : la proportion d'emprunteurs négatifs faussement classés.

$$TFN = \frac{FN}{VP + FN} \tag{5.4}$$

- ***Taux de biens classés (exactitude)*** : la proportion d'emprunteurs correctement classés (vrais positifs et vrais négatifs) par rapport à la totalité emprunteurs étudiés.

$$TBC = \frac{(VP + VN)}{N} \tag{5.5}$$

Les indicateurs de performance qu'on vient de citer sont plus fiables une fois appliqués sur un échantillon de test que sur un échantillon d'apprentissage, car l'évaluation d'un modèle sur un échantillon qui a servit à le construire est toujours d'un caractère optimiste (Tuffery, 2007). De ce fait nous avons construit nos divers modèles à partir du même échantillon d'apprentissage S_A tiré d'une manière aléatoire de la sous-population des emprunteurs non-clients (sauf pour les modèle $M1$ et $M7$ cf. chapitre 4). Ensuit nous les avons testé sur le même échantillon de test S_T formé des données qui n'ont pas servi à élaborer les modèles en phase d'apprentissage.

5.1.2 Taux de biens classés

Le taux d'instances biens classées (TBC) représente la proportion de vrais cas : vrais positifs et vrais négatifs dans la population. Nous utilisons ce critère afin d'évaluer le pouvoir de chacun de nos modèles à générer le plus grand nombre d'instances bien classifiées. Le Tableau 5.2 donne les taux moyens d'instances bien classées des sept modèles de régression logistique et des quatre modèles d' SVMs. Ce tableau expose l'évolution des taux d'instances biens classées sur 50 simulations en fonction de la taille de l'échantillon d'apprentissage.

Tableau 5.2 – *Pourcentages moyens d'instances biens classés en fonction de la taille de l'échantillon (50 simulations)*

Modèles	$n^* = 50$	$n^* = 100$	$n^* = 150$	$n^* = 200$
$M1$	$0,562$	$0,558$	$0,562$	$0,567$
$M2$	$0,581$	$0,567$	$0,555$	$0,537$
$M3$	$0,619$	$0,620$	$0,628$	$0,641$
$M4$	$0,618$	$0,615$	$0,623$	$0,638$
$M5$	$0,613$	$0,624$	$0,627$	$0,624$
$M6$	$0,601$	$0,610$	$0,619$	$0,635$
$M7$	$0,569$	$0,570$	$0,578$	$0,583$
$S1$	$0,590$	$0,595$	$0,595$	$0,605$
$S2$	$0,550$	$0,565$	$0,571$	$0,575$
$S3$	$0,585$	$0,594$	$0,595$	$0,604$
$S4$	$0,531$	$0,521$	$0,519$	$0,518$

La Figure 5.1 récapitule les résultats énoncés par le tableau précédent. A partir de cette figure nous constatons que les taux d'instances bien classés augmentent proportionnellement à l'augmentation de la taille de l'échantillon d'apprentissage. Cette amélioration peut être due aux paramètres des différents modèles qui deviennent plus précis avec l'augmentation de la taille de l'échantillon d'apprentissage.

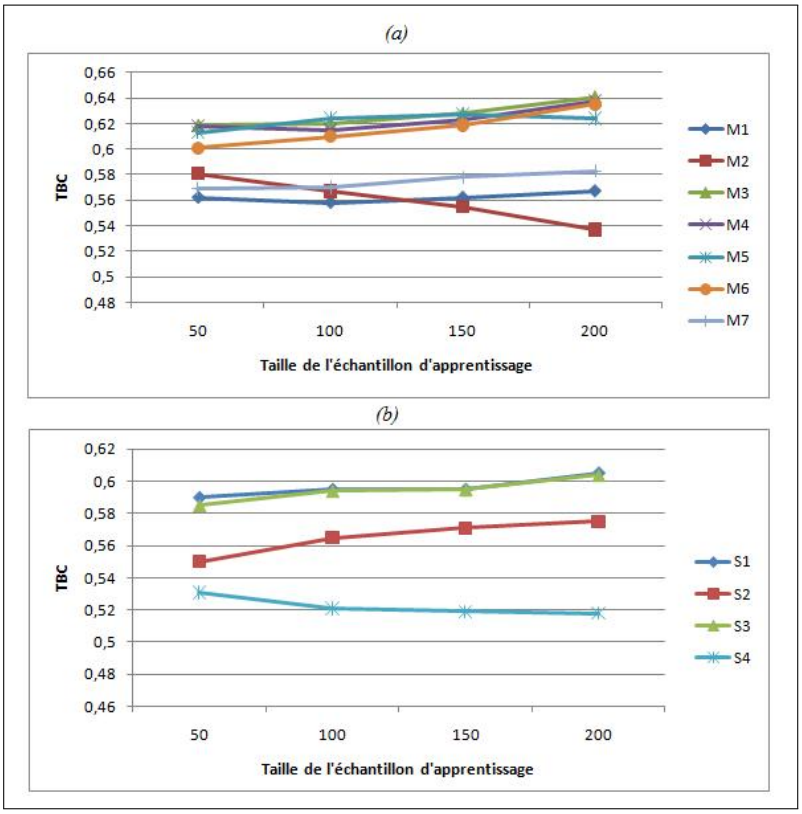

FIGURE 5.1 – *Influence de la taille de l'échantillon d'apprentissage sur l'évolution du nombre d'instances biens classés*

La Figure 5.1 (a) montre que les modèles $M3$ et $M4$ sont les modèles avec les taux d'instances biens classées les plus élevés. Ces modèles se détachent en termes de bonne prédictions par rapport aux autres modèles étudiés. Ils constituent selon ce critère, les deux meilleurs modèles de classification. Nous pouvons aussi remarquer à partir de la Figure 5.1 (a) que les modèles $M5$ et $M6$ engendrent des bons résultats, mais qui ne sont pas les meilleurs. Le reste des modèles logistiques semblent fournir des taux d'instances biens classées faibles. En particulier le modèle $M1$ fournit les taux d'instances biens classées les plus faibles. Les modèles $S1$ et $S3$ sont plus performants que les modèles de régression logistique qui ne tiennent pas compte des liens entre les deux sous-populations (cf. Figure 5.1 (b)).

5.1.3 Taux d'erreur de type I

Le taux d'erreur de type I dit aussi taux de faux positifs est la proportion de déclarer un positif (emprunteur solvable) à tort, là où il est en réalité non solvable. Ce type d'erreur survient quand l'organisme de crédit accord un prêt à un emprunteur qui fera éventuellement défaut. Les probabilités d'occurrence de ce genre d'erreur sont assez faibles (de 1 à 5%) mais le montant relatif est assez élevé. C'est pourquoi ce genre d'erreur est considéré comme l'un des risques majeurs aux quels un organisme de prêt est confronté. Il est toujours souhaité d'obtenir un taux de faux positifs le plus petit possible. Le Tableau 5.3 récapitule les taux moyens d'erreur de type I pour les différents modèles sur 50 simulations en fonction de la taille de l'échantillon d'apprentissage.

Tableau 5.3 – *Pourcentage moyen de faux positifs en fonction de la taille de l'échantillon d'apprentissage (50 simulations)*

Modèles	$n^* = 50$	$n^* = 100$	$n^* = 150$	$n^* = 200$
$M1$	$0,462$	$0,466$	$0,462$	$0,452$
$M2$	$0,266$	$0,242$	$0,191$	$0,195$
$M3$	$0,381$	$0,378$	$0,357$	$0,343$
$M4$	$0,385$	$0,384$	$0,369$	$0,346$
$M5$	$0,341$	$0,304$	$0,273$	$0,248$
$M6$	$0,400$	$0,395$	$0,379$	$0,357$
$M7$	$0,457$	$0,457$	$0,449$	$0,438$
$S1$	$0,383$	$0,389$	$0,380$	$0,374$
$S2$	$0,428$	$0,421$	$0,406$	$0,402$
$S3$	$0,391$	$0,391$	$0,379$	$0,378$
$S4$	$0,458$	$0,471$	$0,474$	$0,473$

La Figure 5.2 récapitule les résultats énoncés par le tableau précédent. A partir de cette figure nous constatons que les taux d'instances bien classés des modèles $M2$ et $M5$ (cf. Figure 5.2 (a)) diminuent avec la diminution de la taille de l'échantillon d'apprentissage. Toutefois, pour le reste des modèles la taille de l'échantillon d'apprentissage ne semble pas influencer les taux

d'erreur de type I.

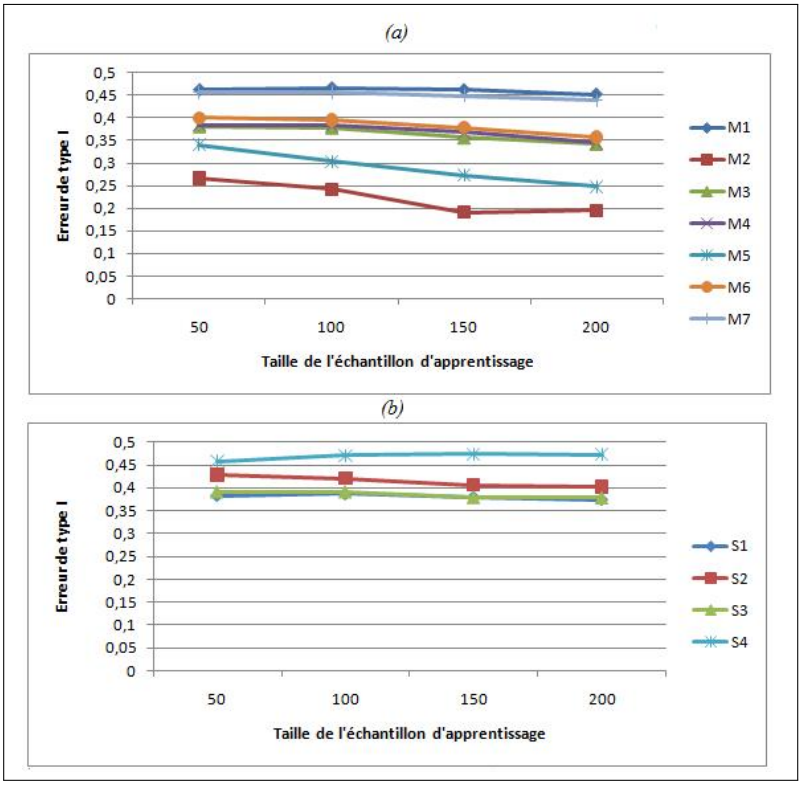

FIGURE 5.2 – *Influence de la taille de l'échantillon d'apprentissage sur l'évolution de nombre d'instances fausses positives*

A partir de la Figure 5.2 (a) nous constatons que les modèles $M2$ et $M5$ ont les taux d'erreur de type I les plus faibles. Ils constituent selon ce critère, les deux meilleurs modèles de classification puisque ils engendrent le plus grand nombre d'instance négatives bien classées. A l'opposé des modèles $M2$ et $M5$, les modèles $M1$ et $M7$ ainsi que $S2$ et $S4$ (cf. Figure 5.2 (b)) génèrent les taux d'erreur de type I les plus importants. Les modèles restants détiennent des taux de prédiction par excès acceptables.

5.1.4 Taux d'erreur de type II

Le taux d'erreur de type II dit aussi taux de faux négatifs est la proportion de déclarer un négatif (emprunteur non solvable) à tort, là où il est en réalité un positif. Ce type d'erreur n'est pas aussi grave que le premier type d'erreur, quoiqu'il soit plus fréquent que le premier. Le taux de faux négatifs se produit lorsque un organisme de prêt refuse d'accorder un crédit à un candidat qui aurait parfaitement remboursé sa dette. Il est toujours souhaité d'obtenir un taux de faux négatifs le plus petit possible afin d'éviter de rejeter des bons emprunteurs à tort. Le Tableau 5.4 représente les taux moyens d'erreur de type II calculés pour nos divers modèles sur 50 simulations.

Tableau 5.4 – *Pourcentage moyen de faux négatifs en fonction de la taille de l'échantillon d'apprentissage (50 simulations)*

Modèles	$n^* = 50$	$n^* = 100$	$n^* = 150$	$n^* = 200$
$M1$	$0,287$	$0,280$	$0,286$	$0,313$
$M2$	$0,430$	$0,441$	$0,455$	$0,482$
$M3$	$0,371$	$0,369$	$0,366$	$0,369$
$M4$	$0,366$	$0,367$	$0,366$	$0,371$
$M5$	$0,406$	$0,403$	$0,407$	$0,423$
$M6$	$0,384$	$0,367$	$0,364$	$0,367$
$M7$	$0,290$	$0,281$	$0,295$	$0,325$
$S1$	$0,406$	$0,398$	$0,407$	$0,391$
$S2$	$0,434$	$0,436$	$0,417$	$0,416$
$S3$	$0,411$	$0,398$	$0,406$	$0,391$
$S4$	$0,471$	$0,482$	$0,484$	$0,490$

La Figure 5.3 récapitule les résultats énoncés par le tableau précédent.

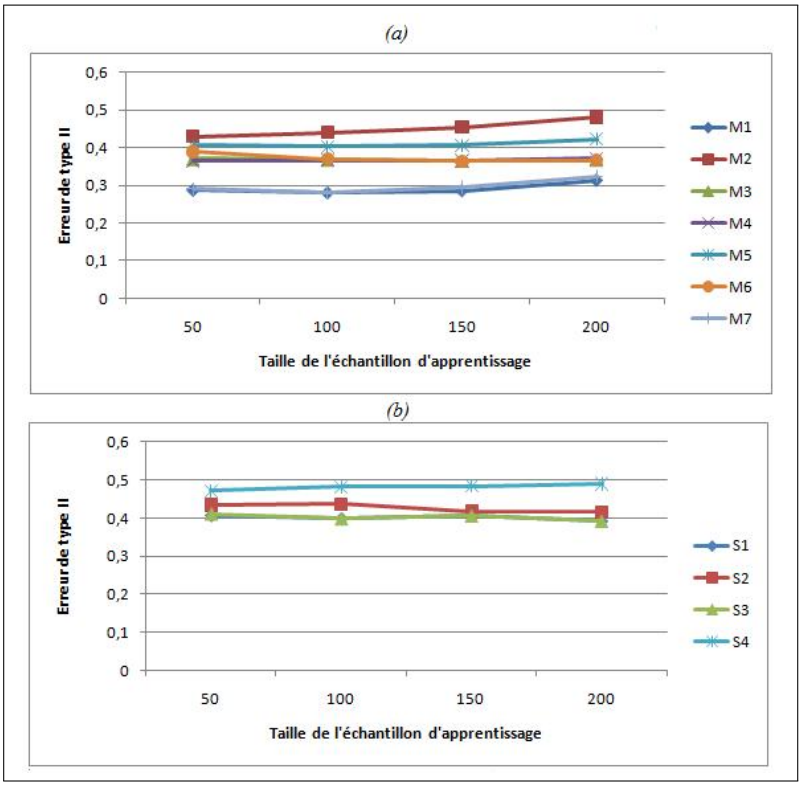

FIGURE 5.3 – *Influence de la taille de l'échantillon d'apprentissage sur l'évolution de nombre d'instances fausses négatives*

Nous constatons à partir de la Figure 5.3 (a) que les modèles $M1$ et $M7$ détiennent les taux de faux négatifs les plus petits. Ces modèles ont tendance à repérer les bons candidats de prêts facilement. Néanmoins, ces deux modèles ont aussi tendance à se tromper au sujet des emprunteurs non solvables (cf. Figure 5.2) ce qui met en doute leur pouvoir de discrimination. Les modèles $M2$ et $M5$ ainsi que $S2$ et $S4$ présentent les risques les plus élevés d'erreur de type II, ces modèles ont une tendance à refuser beaucoup de candidats. Les modèles $M3$ et $M4$ ont des taux de faux négatifs acceptables. La Figure 5.3 (b) montre que les modèles basés sur les SVMs possèdent des taux d'erreur de type II considérablement importants. Néanmoins, les modèles $S1$ et $S2$ sont plus performant que les modèles de régression logistique qui ne tiennent pas en compte des liens entre les deux sous-populations étudiées.

5.2 Critères Graphiques

Les taux d'erreur et le taux d'instances biens classées sont considérés par fois comme des mesures de performance faibles car ils ne tiennent pas en compte de la distribution des classes et des coûts de classification. Selon Hand (2001) la répartition des classes positives et négatives n'est pas toujours équilibrée. Généralement la classe la moins représentée est celle que l'on cherche à identifier, ce qui rend ces critères insuffisants sans incorporation d'un coût de mauvaise classification. Ce coût permettra de pondérer les scores en prenant en compte la distribution des classes.
La question est alors comment choisir ces coûts, le problème est que le plus souvent ces coûts sont difficiles à déterminer. Ce qui rend l'évaluation très difficile et nous conduit à chercher des critères plus compétitifs pour évaluer la performance. L'idée proposée par Hand (2001) est d'utiliser des courbes comme critères d'évaluation plutôt que des valeurs scalaires. Nous utilisons dans ce qui suit la courbe de Receiver Operating Characteristic (ROC) pour évaluer la performance de nos modèles.

5.2.1 Principe de la courbe ROC

La courbe ROC est un outil graphique qui permet l'évaluation de divers modèles sans s'intéresser aux matrices de coûts utilisées pour pondérer les distributions déséquilibrées. Pour un modèle donné, on peut être mené à favoriser la sensibilité ou la spécificité suivant la stratégie choisie : soit sélectionner les bons payeurs ou écarter les mauvais payeurs. Le choix de la meilleure combinaison sensibilité-spécificité est aidé par le tracé de la courbe ROC.

La courbe ROC permet de modéliser la force du modèle à placer les positifs devant les négatifs. Cette courbe présente d'une manière graphique, le lien ente le taux de faux positifs (1-spécificité) placé en abscisse et le taux de vrais positifs placé en ordonnée, pour toutes les valeurs-seuil envisageables.

Selon Perneger et Perrier (2004) un modèle différencierait parfaitement entre les individus solvables et non solvables s'il permet de trouver une valeur seuil ayant une sensibilité et une spécificité de 1. A l'inverse si le modèle est sans pouvoir de discrimination la proportion des emprunteurs solvables serait égale à la proportion des emprunteurs non solvables. La plupart des modèles se trouvent entre ces deux extrêmes. Un modèle est plus performant quand sa courbe ROC se rapproche du coin supérieur gauche du graphique (premier extrême).

5.2.2 Evaluation des modèles à l'aide de la courbe ROC

Nous comparons maintenant nos modèles en termes de leur courbe ROC. Le modèle qui a le meilleur pouvoir discriminant est celui qui correspond à la courbe ROC la plus haute. Nous obtenons les courbes ROC de nos différents modèles à partir de l'échantillon S_T (cf. Figure 5.4)

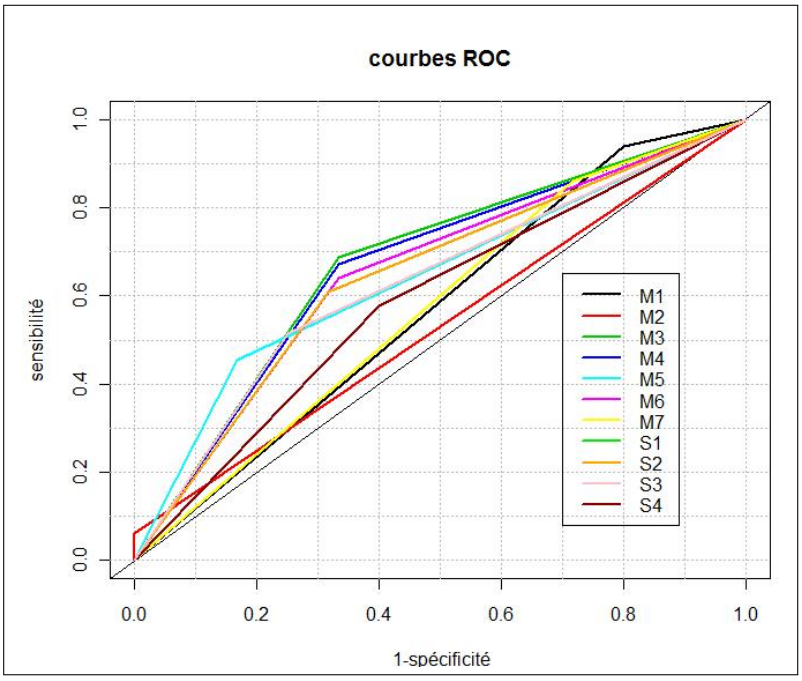

FIGURE 5.4 – *Courbes ROC des différents modèles*

Comme première constatation nous remarquons que les courbes des divers modèles sont convexes et situées au-dessus de la première bissectrice ce qui nous mène à affirmer que ces modèles sont statistiquement valides.
A partir de La Figure 5.4 on remarque que pour une sensibilité inferieur à $0,1$ le modèle $M2$ détient un aire sous la courbe ROC supérieur à ceux des autres modèles. Au-delà de ce seuil le modèle $M5$ devient le modèle avec l'aire le plus important sous la courbe ROC. Pour une sensibilité de 0.5 et une spécificité de 0.75 les courbes des modèles $M3$ et $M4$ se confondent et deviennent les courbes les plus creuses. Pour une sensibilité supérieur à 0.9 et une spécificité supérieur à 0.25 la courbe du modèle $M1$ devient la plus dominante.
Les courbes $S1$ et $S3$ se confondent (cf. Figure 5.4). De plus, l'aire sous leurs courbes est presque identique. Ces deux modèles ont alors un bon compromis entre sensibilité et spécificité et peuvent être classés parmi les modèles les plus discriminants lorsqu'on ne tient pas en compte des liens entre les deus sous-populations étudiés.

5.3 Conclusion

Les mesures de performances évoquées dans ce chapitre nous ont permit d'évaluer la validité ainsi que le caractère prédictif des différents modèles. Les résultats fournis par ces différents critères nous ont permit de dégager les conclusions suivantes : les modèles $M5$ et $M6$ ainsi que $S1$ et $S3$ sont des bons classifieurs. Néanmoins, ce sont les modèles $M3$ et $M4$ qui semblent les plus adéquats pour évaluer le comportement des emprunteurs non-clients en termes de remboursement. Ces derniers se distinguent comme étant les meilleurs modèles du point de vue exactitude. De plus, l'utilisation de la courbe ROC vient pour confirmer que ces deux modèles présentent les meilleurs compromis entre sensibilité et spécificité. Ces modèles ont été construits à partir des informations de la sous-population des emprunteurs

non-clients, et ce en tenant compte des liens avec la sous-population des clients. Les modèles logistique $M3$ et $M4$ ont apporté la preuve qu'un lien réel existe entre les deux sous-populations, venant confirmer l'intuition de départ.
Le modèle $M1$ construit à partir des données des emprunteurs clients apparaît le modèle le moins réussi une fois appliqué aux données des emprunteurs non-clients. Ce modèle en plus des modèles $S2$ et $S4$ se distingues par une faculté de discrimination faible par apport aux autres modèles étudiés.
Les modèles $M2$ et $M5$ possèdent les taux d'erreur de type I les plus faibles et les taux d'erreur de type II les plus élevés, ces modèles sont considérés comme prudents. Toutefois, leur utilisation peut mener à la perte d'emprunteurs fiables.

Conclusion générale

La qualité d'information possédée par une banque sur les clients est cruciale dans l'évaluation des demandes de prêts. En effets les modèles de prédiction utilisés par la majorité des banques ne permettent pas de prédire le comportement d'un emprunteur (solvable/non solvable) provenant d'un échantillon de taille restreinte.

Dans le cadre de ce mémoire nous nous somme intéressé à l'évaluation du risque de crédit en présence d'une population à faible effectifs, représentée par les emprunteurs non-clients d'une banque. L'objectif était de prouver l'incapacité des systèmes adoptés par les banques en termes d'évaluation des dossiers de crédit pour ce genre de population et d'apporter une amélioration en développant des nouvelles règles d'inférence. Pour cela, nous avons considéré deux techniques de modélisation.

La première technique est la technique de la régression logistique qui est l'une des techniques les plus utilisées en prédiction de risque lorsque la variable de réponse est dichotomique. L'utilisation de cette première technique nécessite la possession d'un échantillon d'apprentissage de taille suffisante, ce qui n'est pas le cas de la population des non-clients. Afin d'apporter une réponse à cette contrainte de manque d'observations nous avons dû utiliser les informations relatives à la population des emprunteurs clients de la banque, et ce en supposant l'existence de liens cachés entre ces deux populations. Pour sortir ces liens nous nous somme basé sur les travaux de Beninel et Biernacki (2005), (2007), (2009) effectués dans le cas d'un mélange de deux popula-

tions gaussiennes. Ceci nous a permis de dégager sept modèles de régression logistique fondés sur les informations des deux populations.

La deuxième technique considérée dans ce travail est la technique des séparateurs à vaste marge (SVM). En se basant sur cette technique nous avons élaboré quatre autres modèles basés respectivement sur un noyau linéaire, à base radiale, polynomial et sigmoïde. Les onze modèles ont été implémentés dans le langage R sur un jeu de données composé de 1000 observations représentant les crédits à la consommation accordés par la *DeutshBank* de Munich (LMU, 1994).

Les résultats trouvés dans cette étude viennent pour confirmer l'intuition de départ sur l'incapacité des systèmes traditionnels dans la prédiction du comportement des non-clients en termes de remboursement. En effet, le modèle $M1$ qui ne tient pas en compte des liens entre les deux sous-populations et qui est le modèle le plus utilisé par les banque se caractérise par une faculté de discrimination relativement faible une fois appliqué sur les données des emprunteurs non-clients. Les modèles $M3$ et $M4$ sont les modèles qui s'ajustent le mieux à nos données et qui se détachent en terme de bonne prédictions par rapport aux autres modèles. Ces derniers modèles ont été construits à partir des données des deux populations et ce en tenant compte des liens qui les relient. Ceci vient pour confirmer l'existence d'un lien entre la population des clients et des non-clients.

Lorsque l'on ne tient pas en compte des liens entre les emprunteurs clients et les non clients, les modèles $S1$ et $S3$ basés sur les SVMs se sont avérer les plus performants. Néanmoins, cette performance reste dépendante du choix du paramètre de régularisation C, ainsi que du choix des paramètres adéquats des noyaux. Dans certains temps les modèles basés sur les SVMs ont pu nous donner des meilleurs résultats en présence d'une autre procédure dans la phase de la sélection des paramètres que la procédure de la validation croisée qui est couteuse en termes de temps.

Bibliographie

J.A. Anderson : Logistic discrimination. *In Handbook of Statistics*, 2:169–191, 1982.

M. Bardos : *Analyse discriminante : Application au risque et scoring financier*. Dunod, 01 2001. ISBN 2 10 004777 9.

F. Beninel et C. Biernacki : Analyse discriminante généralisée : Extension au modèle logistique. *In Cloque Data Mining et Apprentissage Statistique Applications en Assurance*, Niort, France, 2005.

F. Beninel et C. Biernacki : Modèles d'extension de la régression logistique. *In Revue des Nouvelles Technologies de l'Information, Data Mining et apprentissage statistique : application en assurance, banque et marketing*, pages 207–218, France, 2007.

F. Beninel et C. Biernacki : Updating a logistic discriminant rule : Comparing some logistic submodels in credit-scoring. *In International Conference on Agents and Artificial Intelligence*, pages 267–274, France, 2009.

C. Biernacki et J. Jacques : Analyse discriminante sur données binaires lorsque les populations d'apprentissage et de test sont différentes. *In Revue des Nouvelles Technologies de l'Information, Data Mining et apprentissage statistique : application en assurance, banque et marketing*, pages 109–125, France, 2007.

C. Bouveyron et J. Jacques : Modèles adaptatifs pour les mélanges de régressions. *In 41èmes Journées de Statistique, SFdS*, Bordeaux, France, 05 2009. inria-00386638, version 1-22.

L. Breiman, J. Friedman, R. Olshen et C. Stone : *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

J.C. Burges : A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, 1998. ISSN 1384-5810.

C.C. Chang et C.J. Lin : *LIBSVM : a library for support vector machines*, 2001. URL `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

J. Cohen : *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Routledge Academic, 2 édition, January 1988. ISBN 0805802835.

A. Cornuéjols : Une nouvelle méthode d'apprentissage : Les svm. séparateurs à vaste marge. Rapport technique 51, Association Française d'Intelligence Artificielle, France, 06 2002.

D.R. Cox : *The analysis of binary data [by] D. R. Cox.* Methuen London,, 1970. ISBN 0416104002.

D.R. Cox et E.J. Snell : *Analysis of binary data.*

S. Devaney : The usefulness of financial ratios as predictors of household insolvency : Two perspectives. *Financial Counseling and Planing*, 5:15–24, 1994.

D. Durand : *Risk elements in consumer instalment financing. (Technical edition) By David Durand.* National bureau of economic research [New York], 1941.

X. Fan et L. Wang : Comparing linear discriminant function with logistic regression for the two-group classification problem. *In Annual Meeting of American Educational Research association*, pages 265–286, 1998.

R. Feldman : Small business loans, small banks and big change in technology called credit scoring. *The Region*, (Sep):19–25, 1997.

R.A. Fisher : The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

R. Fletcher : *Practical methods of optimization ; (2nd ed.).* Wiley-Interscience, New York, NY, USA, 1987. ISBN 0-471-91547-5.

H. Frydman, E. I. Altman et D-L. Kao : Introducing recursive partitioning for financial classification : The case of financial distress. *Journal of Finance*, 40(1):269–91, March 1985.

P. Giudici : *Applied Data Mining : Statistical Methods for Business and Industry.* John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, 2003.

D. J. Hand et W. E. Henley : Statistical classification methods in consumer credit scoring : a review. *Journal Of The Royal Statistical Society Series A*, 160(3):523–541, 1997.

D.J. Hand : Measuring diagnostic accuracy of statistical prediction rules. *Statistica Neerlandica*, 55(1):3–16, 2001.

C.W. Hsu, C.C. Chang, C.J. Lin et Department Of Computer : A practical guide to support vector classification. Rapport technique, 2003.

C-L. Huang, M-C. Chen et C-J. Wang : Credit scoring with a data mining approach based on support vector machines. *Expert Syst. Appl.*, 33(4):847–856, 2007. ISSN 0957-4174.

T. Joachims : Text categorization with support vector machines : learning with many relevant features. *In* Claire Nédellec et Céline Rouveirol, éditeurs : *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142, Heidelberg et al., 1998. Springer.

LMU : Datasets at the department of statistics, university of munich, and the $sfb386$, determining the solidness of borrowers via creditscoring. url http ://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit_e.html. 1994.

J. Louradour : *Noyaux de séquences pour la vérification du locuteur par Machines à Vecteurs de Support.* Thèse de doctorat, Université Toulouse III, 2007.

P.C. Mahalanobis : On the generalized distance in statistics. *Natl. Inst. Science*, 12:49–55, 1936.

W. Mcculloch et W. Pitts : A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, 5(4):115–133, December 1943.

A. Merbouha et A. Mkhadri : Méthodes de scoring non-paramètriques. *Revue de Statistique Appliquée*, 56(1):5–26, 2006.

J. Mercer : Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London*, 209:415–446, 1909.

T. Perneger et A. Perrier : Analyse d'un test diagnostique : courbe roc, ou " receiver operating characteristic ". *Revue des Maladies Respiratoires*, 21(2):398–401, 4 2004.

M. Pontil et A. Verri : Support vector machines for 3d object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:637–646, 1998. ISSN 0162-8828.

S.L. Salzberg : On comparing classifiers : Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1:317–327, 1997.

G. Saporta : Credit scoring, statistique et apprentissage. *In EGC*, pages 3–4, 2006.

O. Sautory, W. Chang et V. Sébastien : Une étude comparative des méthodes de discrimination et de régression logistique. *In Journées de Métodologie Statistique 1992*, 1992. INSEE Méthodes N 46-47-48.

L.C. Thomas, J. Crook et D. Edelman : *Credit Scoring and Its Applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002. ISBN 0898714834.

S. Tufféry : *Data mining et statistique décisionnelle : l'intelligence des données*. Editions Ophrys, 2007. ISBN 2710808889, 9782710808886.

V.N. Vapnik : *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0387945598.

M. Vojtek et E. Kocenda : Credit scoring methods. *Finance a uver - Czech Journal of Economics and Finance*, 56:152–167, 03 2006.

L. Yang : New issues in credit scoring application. Working Papers 16/2001, Institut Fur Wirtschaftsinformatik, 2001.

G-X. Yu, G. Ostrouchov, A. Geist et N.F. Samatova : An svm-based algorithm for identification of photosynthesis-specific genome features. *In CSB '03 : Proceedings of the IEEE Computer Society Conference on Bioinformatics*, page 235, Washington, DC, USA, 2003. IEEE Computer Society. ISBN 0-7695-2000-6.

Annexe : Description des variables

Nom des variables	Description des variables	Type des variables	Description des modalités
Alter	Age	Quantitative	pas de modalités
Beruf	Profession	Catégorielle	1 : Sans qualification et sans domicile fixe 2 : Sans qualification avec domicile fixe 3 : Salarié qualifié / Fonctionnaire 4 : Cadre / Patron
Beszeit	Stabilité dans l'emploi	Catégorielle	1 : Chômeur 2 : ≤ 1 an 3 : $1 \leq .. < 4$ an 4 : $4 \leq .. < 7$ an 5 : ≥ 7 ans
Beurge	Autres garants pour le rembourssement du crédit	Catégorielle	1 : Aucuns 2 : Co-responsable 3 : Garant
Bishkred	Nombre de précédents crédit à la banque	Catégorielle	1 : Zéro 2 : Un ou deux 3 : Troix ou quatre 4 : Cinque ou plus
Famges	Statut marital / Sexe	Catégorielle	1 : Homme divorcé 2 : Femme divorcée / mariée 3 : Homme célibataire / marié / fiancé 4 : Femme célibataire
Gastarb	Travailleur étranger	Catégorielle	1 : Oui 2 : Non

Hoehe	Montant du crédit	Quantitative	pas de modalités
Kredit	Solvabilité du client	Quantitative	0 : Non solvable 1 : Solvable
Laufkount	Etat de compte	Catégorielle	1 : Ne posede pas de compte courant (Non-client) 2 : Absence de débit 3 : $0 \leq .. < 200$ DM 4 : 200 DM / compte courant plus d'un an
Laufzeit	Durée du crédit (mois)	Quantitative	pas de modalités
Moral	Comportement passé pour rembourser d'autres crédits	Catégorielle	0 : Payements difficiles des crédits précédents 1 : Comptabilité problématique ou autre crédits extérieurs à la banque 2 : Pas de crédits antérieurs ou extérieurs remboursés 3 : Aucuns problèmes de crédits courants 4 : Crédits passés à cette banque remboursés
Pers	Nombre de personnes gérant le crédit	Catégorielle	1 : de 0 à 2 personnes 2 : plus de 2 personnes
Rate	Pas de mensualités dans le revenu disponible en %	Catégorielle	1 : ≥ 35 2 : $25 \leq ... < 35$ 3 : $20 \leq ... < 25$ 4 : < 20
Telef	Possession d'une ligne téléphonique	Catégorielle	1 : Oui 2 : Non
Sparkont	Valeur des ressources financières du client	Catégorielle	1 :Pas d'économies 2 : < 100 DM 3 : $100 \leq ... < 500$ DM 4 : $500 \leq ... < 1000$ DM 5 : ≥ 1000 DM
Verm	Actif en dur	Catégorielle	1 : Pas disponible / aucuns 2 : Voiture / autre 3 : Contrat d'épargne avec une société immobilière / assurance vie 4 : Propriétaire dune maison / terrain

Verw	Objectif du credit	Catégorielle	1 : Voiture neuve 2 : Voiture d'occasion 3 : Divers 4 : Radio / télévision 5 : Electroménager 6 : Réparation 7 : Education 8 : Vacances 9 : Recyclage 10 : Travail
Weitkred	Autres crédits en cours	Catégorielle	1 : à d'autres banques 2 : Crédit dans un grand magasin ou vente par correspondance 3 : Aucuns
Wohn	Type de logement	Catégorielle	1 : appartement en location 2 : Propriétaire d'un appartement 3 : Aucuns
Wohnzeit	Temps passé par le client dans son logement actuel	Catégorielle	1 :$<$ 1 an 2 : $1 \leq ... < 4$ ans 3 : $4 \leq ... < 7$ ans 4 : ≥ 7 ans