

# **EVALUATION STATISTIQUE DU RISQUE DE CREDIT PAR LA TECHNIQUE DU SCORING : Cas de Afriland First Bank**

**Présenté par :**

**TENE Georges Colince**

*Maître ès-Sciences en Mathématiques Pures*

**Sous la direction de**

**Eugène-Patrice NDONG NGUEMA**

*Chargé de cours à l'ENSP de Yaoundé*

**Sous l'encadrement professionnel de**

**Célestin GUELA SIMO**

*Directeur des Etudes et du Corporate Banking, AFRILAND FIRST BANK*

---

---

# DEDICACES

---

A la mémoire de mon père. Papa : Que ton âme repose en paix.

A ma mère, Mme **TCHOUNDA Madeleine** ; Maman, c'est sûr que tu ne comprendras pas grand-chose au sujet que j'ai traité dans ce document, mais saches que chaque mot, chaque phrase, chaque ponctuation et chaque lettre que j'y ai inscrit ont une seule et même signification : « tu es la meilleure des mères ».

A notre chef de famille, Mr **FOYO Jean-Paul**, pour tes conseils, ton soutien inconditionnel que tu m'as toujours apporté comme à tous tes enfants.

A mes sœurs, Mme **WOUAGOU Juliette** et Mme **MASSO Flore**, pour votre amour, votre soutien moral et financier, votre patience et votre dévouement sans faille qui m'ont permis d'arriver jusqu'à ce point. Puisse ce diplôme nous réserver à tous des lendemains meilleurs.

---

# REMERCIEMENTS

---

*« Louange à Dieu, le très clément et le très miséricordieux ».*

## **Au Pr. Henri GWÉT**

De prime à bord, nous voudrions lui exprimer notre grande considération à travers les grands efforts fournis pour nous procurer le savoir et le savoir être dans des conditions universitaires favorables.

## **Au Dr. Eugène-Patrice NDONG NGUEMA**

Une mention toute particulière d'admiration et d'un grand respect à son endroit, dont les nombreux conseils méthodologiques et la constante disponibilité ont été plus que déterminant durant notre formation et pour la réalisation de ce mémoire.

## **Au Pr. Philippe BESSE**

**Laboratoire de Statistique et Probabilités, UNIV Paul Sabatier de Toulouse III.**

Pour sa disponibilité, son aide et les précieux conseils qu'il m'a donné via le Net.

A tout le personnel enseignant du Master 2 de Statistique Appliquée de l'ENSP.

Nous voulons ici exprimer nos sincères gratitude à toutes les personnes qui nous ont été d'un apport positif pendant notre stage à la First Bank.

Nous remercions d'abord **M. GUELA SIMO** Célestin pour avoir bien voulu nous parrainer pendant notre séjour au sein de **Afriland First Bank**.

Merci à **M. MOUAHA YEKEL, SIME Brice** pour leur encadrement, et surtout pour avoir bien voulu lire notre travail, contribuer par leurs nombreuses critiques positives à son amélioration.

Merci aussi à tout le reste du staff de la Direction des Etudes et du Corporate Banking, et plus particulièrement à **MM El Hadj OUSMANE MAHAMAT** et **TACHOULA TSOGNO Saturnin** pour nous avoir bien accepté parmi eux, et nous avoir guidé dans les tâches que nous avons eu à réaliser pendant notre stage.

Sincères remerciements à Mr **Raymond TACHAGO** pour l'encadrement et le soutien inconditionnel dont j'ai toujours bénéficié à ses côtés.

A Mr **Guillaume EYOUM** pour le soutien dont il a toujours faire montre à mon égard.

A mes amis et connaissances : **Clotilde DJOTUE, Léopold NGUETGNIA, Nicanor NYAND-JOU, Narcisse ZEBAZE, Samuel MBE** et **Raphaël FONGANG**.

A tous mes camarades du Master 2 de Statistique Appliquée de l'ENSP.

Enfin, à tous ceux qui nous ont oeuvré dans quelque circonstance que ce soit pour la conception et la réalisation de ce document, qu'ils trouvent ici l'expression de notre profonde gratitude.

---

# AVANT - PROPOS

---

Le stage académique de fin de formation fait partie du système d'évaluation de l'étudiant en Master 2 de Statistique Appliquée de l'Ecole Nationale Supérieure Polytechnique de l'Université de Yaoundé I. Il présente pour celui-ci un double intérêt : ce stage permet à l'apprenant d'une part de se trouver dans un cadre mieux indiqué pour pouvoir confronter la théorie reçue pendant la formation à la pratique sur le terrain de son futur métier. D'autre part, c'est aussi l'occasion de se familiariser avec son milieu de demain, le monde du travail. C'est la raison pour laquelle le stage académique est une étape indispensable pour les futurs diplômés en Statistique Appliquée que nous sommes. A l'issue de ce stage, l'étudiant présentera un mémoire de fin de formation qui sera sanctionné par le diplôme de master 2 de Statistique Appliquée.

Durant notre séjour qui a duré trois mois (25 juin au 25 septembre 2007) à la First Bank, notre objectif était de fournir un outil statistique pouvant permettre de réduire le taux d'impayés élevé par rapport à la moyenne nationale, subie par cette banque en 2006, en mettant objectivement sur pied un outil qui permettrait une détection automatique des clients à risque qui sont la principale cause de ces impayés. Il s'agissait pour nous de construire un modèle statistique de décèlement précoce du statut «bon» ou «mauvais» client d'un nouvel emprunteur de la First Bank. L'orientation de notre travail était portée sur la conception d'un modèle statistique d'octroi de crédit par la technique du scoring : C'est le credit scoring. Ce terme désigne un ensemble d'outils d'aide à la décision utilisés par les organismes financiers pour évaluer le risque de nonremboursement des prêts. Un scoring est une note de risque, ou une probabilité de défaut. Le modèle construit devrait nous permettre d'évaluer le risque de crédit des emprunteurs de la First Bank.

Nous n'avons pas la prétention d'avoir cerné les contours du sujet, bien au contraire nous pensons que plusieurs études doivent encore être faites pour l'amélioration de ce travail. Pour cette raison, nous restons assujettis à vos remarques et critiques.

*“ il arrive que les grandes décisions ne se prennent pas, mais se forment d'elles mêmes ”*

Henri Bosco (1888-1976)

---

# RESUME

---

Comment les banques sont-elles censées évaluer, prévoir et gérer efficacement le risque crédit, face à l'incroyable diversité des dangers et menaces qui pèsent désormais sur leur activité ? Comment peuvent-elles répondre avec succès aux nouvelles contraintes qui émanent de la clientèle tout en préservant leur rentabilité future ? Ces deux questions sont au coeur des enjeux liés à la mesure du risque de crédit, et ne sont pas sans effet sur la capacité future des banques à gérer ce type de risque. Encore aujourd'hui, seules les banques et institutions financières de premier plan sont capables d'évaluer leur risque de crédit avec un certain degré de confiance ou disposent d'une base de données fiable pour le scoring ou la segmentation comportementale des emprunteurs. Spécifier des modèles de risque plus robustes que les méthodes traditionnelles, en intégrant davantage de facteurs de risque de crédit et en améliorant la précision de la mesure de ce risque, tel est le défi que doivent aujourd'hui relever les banques.

Dans le cadre de ce mémoire, notre travail consiste à la mesure du risque de crédit par une notation statistique des emprunteurs à Afriland First Bank. On y développe deux méthodes paramétriques de construction d'un scoring, puis une comparaison finale des qualités de prévision sur la base du taux de mal classés y est faite pour l'optimisation des modèles.

*Mots clés : Banque, Risque de crédit, scoring, emprunteur, modèle.*

---

---

# ABSTRACT

---

How banks are supposed to evaluate, forecast and manage efficiently credit risk, given the multiple dangers and threats they have to face now ? How can they answer successfully to the new constraints arising from supervisors while preserving their future profitability ? These two questions are the most challenging issues related to credit risk, and they can impact on the future banks' ability to manage this type of risk. Even now, only first-ranked financial institutions are really able to evaluate their credit risk with an acceptable level of confidence or have a reliable database for the scoring or the behavioural segmentation of the borrowers. Building more robust credit risk models than traditional methods, by including more risk factors and improving the accuracy of operational risk measures and indicators, such are the challenges banks have to deal with in the near short term.

Within the framework of this memory, our work consists with the measurement of risk credit, by a statistical notation of the borrowers at Afriland First Bank. We develop two parametric methods of construction of a scoring, then a final comparison of qualities of forecast on the basis of rate of evil classified is made for the optimization of the models.

*Keywords : Banks, credit Risk, scoring, model, borrower*

---

# SIGLES ET ABBREVIATIONS

---

**BTP** : Bâtiment Travaux Publics.

**CA** : Chiffre d’Affaire.

**DECB** : Direction des Etudes et du Corporate Banking.

**ENSP** : Ecole Nationale Supérieure Polytechnique .

**ESDC** : Étude Statistique des Dossiers de Crédit .

**Ets** : Etablissement (entreprise personnelle) .

**FCFA** : Franc de la Communauté Financière de l’Afrique Centrale .

**GIC** : Groupement d’Intérêt Commun .

**IC** : Intervalle de Confiance.

**M** : millions .

**RN** : Résultat Net.

**SA** : Société Anonyme .

**SARL** : Société à Responsabilité Limitée .

**SN** : Situation Net.

**VA** : Valeur Ajoutée.

**i.e** :c’est-à-dire

---

# LEXIQUE DES TERMES TECHNIQUES[15]

---

- **Actifs** : c'est l'ensemble des biens ou droits constituant le patrimoine de l'entreprise, i.e tout ce qu'elle possède.
- **Autonomie financière d'une entreprise** : c'est la capacité de s'autogérer elle-même. Elle est évaluée par les ratios : *capital / dettes* ; *SN/dettes* ; *SN/total passif*.
- **Cash flow** : Le cash flow d'une entreprise permet de mesurer sa capacité à autofinancer ses investissements. C'est le moyen le plus « sain » puisqu'il correspond aux liquidités dégagées par l'entreprise.
- **Charges directes** : une charge est dite directe par rapport à un produit lorsqu'elle participe sans ambiguïté à la fabrication de ce produit. Parmi les charges directes, on a entre autres les matières premières et fournitures qui entrent en fabrication des produits et la main d'œuvre directe composée des frais de personnel résultant des travaux effectués sur un seul produit.
- **Charges indirectes** : ce sont les charges qui concernent plusieurs produits (et parfois même tous les produits) de l'entreprise et qui sont réparties (ou "imputées") entre ces produits à l'aide de clés de répartition.

Ce sont par exemple : certains frais d'usine (bâtiment, entretien, assurances, etc.), les coûts des services généraux de l'entreprise (direction générale, direction de la recherche, direction commerciale, etc.), les campagnes publicitaires portant sur plusieurs produits de l'entreprise ou sur l'entreprise elle-même (publicité corporate).

- **Chiffre d'affaire** : Le chiffre d'affaire désigne le total des ventes de biens et de services facturés par une entreprise sur un exercice comptable.
- **Credit scoring** : c'est un ensemble d'outils d'aide à la décision utilisés par les organismes financiers pour évaluer le risque de nonremboursement des prêts.
- **Fonds propres ou capitaux propres** : ils correspondent aux ressources stables de l'entreprise i.e dans une optique fonctionnelle, les capitaux propres participent, concurremment avec les éléments du passif externe, au financement de l'entreprise
- **Montant des investissements** : Pour une entreprise, C'est le montant placé essentiellement dans une opération économique pour acquérir des biens durables utilisés à court ou à moyen terme.
- **Passifs** : ce sont les éléments du patrimoine ayant une valeur économique négative pour

l'entreprise, ie les obligations de l'entreprise à l'égard d'un tiers dont il est probable ou certain qu'elle provoquera une sortie de ressources au bénéfice de ce tiers, sans contrepartie au moins équivalente attendue de celui-ci. Les passifs comprennent les provisions et les dettes.

- **Taux d'intérêt hors taxes en %** : Le taux d'intérêt d'un prêt ou d'un emprunt est le pourcentage, calculé selon des conventions prédéfinies, qui mesure de façon synthétique, sur une période donnée, la rentabilité pour le prêteur ou le coût pour l'emprunteur de l'échéancier de flux financiers du prêt ou de l'emprunt.
- **Valeur ajoutée** : C'est la contribution additionnelle d'une ressource, d'une activité ou d'un processus dans la réalisation d'un produit ou d'un service. En comptabilité elle est donnée par : Valeur Ajoutée = Chiffre d'affaire - Valeur des consommations intermédiaires.
- **Rentabilité d'une entreprise** : c'est l'aptitude à donner des résultats (positif ou négatif). La rentabilité permet d'évaluer l'efficacité, ou plutôt l'utilisation rationnelle de ressources limitées. Elle est évaluée via les ratios suivants :  $VA/CA$  ;  $RN/CA$  ou *taux de marge nette* ;  $RN/capitaux\ propres$ .
- **Résultat net** : Le résultat net d'une entreprise sur une période donnée (par exemple : une année) est égal à :

la somme des produits réalisés par celle-ci sur la période, (chiffre d'affaires) de laquelle on a déduit l'ensemble des charges (directes et indirectes) engagées sur la même période, ainsi que l'impôt sur les sociétés.

Le résultat net peut donc prendre la forme d'une perte (résultat net négatif) ou d'un bénéfice (résultat net positif).

- **Scoring** : c'est une note de risque, ou une probabilité de défaut.
- **Siège social ou lieu d'exploitation** : Le siège social d'une entreprise est un lieu, précisé dans les statuts d'une société, qui constitue son domicile et détermine son domicile juridique.
- **Solvabilité d'une entreprise** : c'est sa capacité à payer ses dettes ou ses créanciers. Elle est évaluée par le ratio *actif total/dettes*.

---

# RESUME EXECUTIF

---

Cette note propose une application aux techniques de « credit scoring » à partir d'une étude de cas sur les difficultés financières des emprunteurs de la First Bank servant de support à des formations initiales et continuées en analyse des données. On présente tout d'abord la problématique de l'évaluation du risque de crédit, les contraintes qu'impose la collecte de données comptables dans un tel contexte, et la batterie des critères micro-économiques retenus pour mesurer le degré d'insolvabilité des microcréditeurs. L'information fournie par cette batterie de variables financiers est ensuite analysée aux moyens de techniques statistiques telle que la régression logistique et la discrimination linéaire au sens de Fisher. Les résultats fournis par ces techniques d'analyse discriminante, et de classement permettent de montrer l'intérêt méthodologique de ces outils pour ce type d'étude micro-économique. Les résultats obtenus sont interprétés directement à partir des sorties du logiciel **R**.

## **Objectif de l'étude :**

Proposer une base méthodologique de mesure du risque de crédit applicable aux emprunteurs à l'intention de la First Bank.

### **0.1 Problème**

Dans un contexte de transition issu des résultats comptables peu satisfaisants du rapport annuel 2006, il est vraisemblable que la problématique de l'évaluation du risque de crédit bancaire connaisse un regain d'intérêt compte tenu des multiples sollicitations de crédit auxquelles fait face la First Bank. Les créances douteuses ont ainsi pesé assez lourdes sur le résultat net qu'a connu la First Bank. Il s'avère dès lors indispensable de mettre en place des moyens efficaces qui puissent permettre autant qu'il est possible de réduire les risques liés aux crédits accordés par la First Bank, faute de pouvoir les éviter complètement.

### **0.2 Données**

Les données sont collectées à la DECB-division des projets et investissements, l'unité statistique étant un dossier de crédit. Malgré les difficultés de collectes auxquelles nous avons été confrontés pendant la période de stage, nous avons pu collecter 130 dossiers de crédit pour un total de 25 variables par dossier de crédit. Ceci nous a permis de confectionner notre base de données sous forme d'un tableau individus-variables pour en faire une analyse.

### 0.3 Méthodologie

Il s'agit d'une classification supervisée à deux groupes : « bons clients » et « mauvais clients ». Afin de discriminer au mieux les deux groupes d'emprunteurs répertoriés du point de vue des critères financiers et comptables, nous avons utilisé l'analyse discriminante sur la base des variables financières les plus pertinentes, pour prédire l'appartenance de chaque emprunteur ou client au groupe défini par la valeur de la variable qualitative Y « statut du client » à deux modalités :

- si ( $Y=0$ ), l'entreprise(emprunteur) est considérée comme financièrement saine ;
- sinon ( $Y=1$ ), l'entreprise est considéré comme défaillante.

À partir de combinaisons des caractéristiques financières utilisées comme variables explicatives (exogènes) dans l'analyse, l'analyse discriminante construit des fonctions discriminantes ou credit scoring permettant d'affecter l'emprunteur à l'un des groupes prédéfinis sur la base d'une règle probabiliste bayésienne. Les méthodes utilisées sont la régression logistique à deux classes et la discrimination au sens de Fisher.

Afin de valider les résultats obtenus, nous utiliserons une procédure de validation croisée qui consiste pour chaque individu de l'échantillon à réaliser son classement sur la base de la fonction linéaire discriminante obtenue avec les autres individus de l'échantillon. Cela revient à effectuer autant d'estimations qu'il y a d'individus dans l'échantillon. Selon cette procédure, chaque individu classé sert d'échantillon-test pour le calcul du pourcentage de bien-classés et le classement s'effectue sur la base d'un échantillon d'apprentissage constitué par les n-1 individus restants.

### 0.4 Résultats

Le scoring obtenu par la régression logistique binaire semble être la mieux appropriée pour la notation statistique des emprunteurs à la First Bank. Ainsi, l'expression mathématique du scoring est :

$\hat{S}(X) = 0.9209897 \text{CREDIT} - 0.4030249 \text{R1} + 0.7310701 \text{R3} - 0.0876921 \text{R6} + 0.061002 \text{R7}$   
qui est une probabilité de défaut.

La méthode théorique nous fournit un seuil  $s=0$  et un pourcentage de plus de 80% de bons classements, conséquemment la règle de décision suivante :

- $\hat{S}(X) \leq 0$  alors  $\hat{Y} = 0$ , ie que l'emprunteur est non risqué, il est donc considéré comme bon.
- $\hat{S}(X) > 0$  alors  $\hat{Y} = 1$ , ie que l'emprunteur est risqué, il est mauvais client.

La construction de ce scoring a généré un seuil  $s= 80.3$ . En supposant que *la politique économique de la First Bank* est de ne pas prendre de risque ie on est au seuil de 80.3, alors on a la règle de décision suivante :

- Si  $\hat{S}(X) \leq 80.3$ , alors le client est considéré comme non risqué ie bon.
- Si  $\hat{S}(X) > 80.3$ , alors le client est peut-être risqué car ici on rencontre les bons et les mauvais clients.

Il est possible de faire encore varier ce seuil, cela signifie qu'on accroît le risque et l'erreur.

L'examen statistique de la situation économique et financière des entreprises (emprunteurs), en vue de la détection précoce des difficultés de la clientèle, est extrêmement fructueux. Par

l'analyse multicritères, il permet la construction d'un scoring qui fournit une image synthétique du profil de l'entreprise empreunteuse. Celui-ci est, dans la très grande majorité des cas, révélateur de la santé de l'entreprise. Si un tel outil ne peut se substituer au jugement de l'expert, il peut contribuer à l'informer rapidement sur le niveau de risque de l'entreprise et concourir au diagnostic, grâce aux aides à l'interprétation qui l'accompagnent. L'analyste pourra alors se concentrer sur des aspects plus délicats et moins quantifiables de l'évaluation, en particulier les aspects qualitatifs. Ainsi, expertise et utilisation d'un scoring ne sont pas contradictoires ; au contraire, elles se complètent et permettent d'affiner l'analyse du *risque de crédit*. De même, lorsque plusieurs outils d'évaluation du risque sont disponibles, généralement fondés sur des systèmes d'information différents, il est très fructueux de les examiner tous. En effet, les renseignements qu'ils apportent relativisent les points de vue, accroissent la fiabilité de la prévision et renforcent le diagnostic.

---

# INTRODUCTION

---

## *Contexte et problématique*

Le risque de crédit[6] est le risque (vu comme une probabilité) que l'emprunteur ne rembourse pas sa dette en partie ou en totalité, à l'échéance fixée. De nos jours, sa maîtrise est l'une des principales préoccupations pour la plupart des organismes bancaires, notamment via les créances qu'elles accordent à leurs clients, qui sont pour la plupart des formes de prêt à court terme, et pour cette raison, de nombreuses banques sont aujourd'hui amenées à l'intégrer dans leur gestion afin de le minimiser. Ce risque est en effet lourd de conséquences pour la banque, car toute dette non remboursée est économiquement une perte sèche que supporte le créancier. Comptablement parlant, les créances et emprunts accordés à des tiers constituent ainsi un poste spécifique dans le bilan de l'entreprise et toute évolution négative obère d'autant la survie de l'entreprise à moyen ou long terme. Très tôt, les établissements bancaires ont donc cherché à s'immuniser contre ce risque de crédit. En amont, ce risque peut faire l'objet d'une évaluation grâce à différents critères et des techniques mêlant calcul et intuition. Suite à cette évaluation, les banques disposent ensuite de différents moyens de protection pour minimiser, voire annuler ce risque économique.

Dans le cadre de leur fonction d'intermédiation financière, les banques s'exposent au risque de ne pas recouvrer la totalité des fonds engagés dans les délais impartis. La First Bank, 4<sup>me</sup> banque en total du bilan au Cameroun en 2006, a dû constituer FCFA 4 milliards de provisions pour faire face aux mauvaises créances au titre du même exercice, pour un résultat net(RN) d'exploitation de FCFA 1 milliard[3]. Ce qui représente un taux moyen de créances en souffrance ou taux d'impayés d'environ 17% (taux supérieur à la moyenne nationale qui est de 12%). Les créances douteuses ont ainsi pesé assez lourd sur ce résultat net qu'a connu la First Bank. Il s'avère dès lors indispensable de mettre en place des moyens efficaces qui puissent permettre autant qu'il est possible de réduire les risques liés aux crédits accordés par la First Bank, faute de pouvoir les éviter complètement. C'est la raison pour laquelle l'un des défis économiques majeurs pour la First Bank en 2007 est réduire de manière considérable ces impayés.

Le marché du crédit bancaire mettant en relation le banquier et le client emprunteur est caractérisé par une imperfection d'information, source de rationnement du crédit aux yeux de Christophe Godlewski[11]. Le banquier se doit ainsi de chercher les moyens efficaces qui lui permettent de bien faire la sélection de ses clients. La pratique de cette sélection nécessite que le banquier dispose d'au moins deux choses : l'information sur les clients, et une technique

objective de sélection elle-même. Pour détenir cette information, il y a une source officielle représentée par les documents comptables et sociaux, et une source privée nécessitant que le banquier soit effectivement en relation avec l'emprunteur. Pour analyser l'information qu'un banquier détient sur ses clients, on dispose à la First Bank d'une méthode « subjective » dont les exigences majeures sont le jugement et le bon sens, ce qui ne permet pas à la First Bank de déceler judicieusement les clients susceptibles de ne pas honorer à leurs engagements avec la banque. Ainsi, Pour analyser l'information que le banquier détient sur son client, il faut trouver une autre approche complémentaire pour l'étude des dossiers de crédit, amélioratrice du taux d'impayés. Ceci passe objectivement par la mise sur pied d'un modèle statistique d'évaluation du risque de non remboursement (risque de crédit) des emprunteurs de la First Bank.

L'utilisation de la statistique pour étudier les dossiers de demande de crédit passe par un travail de synthèse d'une grande masse d'informations collectée dans le passé. En effet, les techniques statistiques permettent de retracer le profil des bons clients et des mauvais clients à travers leur passé à partir duquel il est possible de pronostiquer le risque de défaut d'un nouveau client. Si un modèle d'évaluation est utilisé, les variables discriminantes contenues dans ce modèle doivent être statistiquement représentatives. La fiabilité du modèle et ses paramètres doivent être contrôlés *à priori* (mesure de la performance prédictive) et *à posteriori* (back-testing).

### **Enjeu :**

La mesure du risque de crédit sur les emprunteurs est un enjeu important, surtout lorsqu'il s'agit des besoins traditionnels tel que le crédit bancaire. La nécessité pour les banques de disposer d'outils fiables est encore plus forte dans la période actuelle de montée du risque de crédit et de doutes sur les comptes de la clientèle. La réalisation d'un modèle de notation statistique d'octroi de crédit par le scoring (*credit scoring*) est d'une grande importance en ce sens que sa capacité de pronostiquer facilite l'évaluation des risques des candidats aux microcrédits. Le credit scoring est objectif, cohérent et explicite, il permet de quantifier le risque comme probabilité et suppose qu'une bonne partie des risques est liée aux caractéristiques quantifiées dans la base de données.

### **Plan de travail :**

Notre travail est divisé en cinq principaux chapitres. Le squelette se présente comme suit : dans un premier temps, nous présentons la banque Afriland First Bank, sa Direction des Etudes et du Corporate Banking(DECBC) et les différents risques auxquelles font face la plupart des banques en mettant un accent particulier sur le risque de crédit. La description de nos données fait l'objet du chapitre deux. Le chapitre trois est consacré à un exposé sur quelques applications statistiques version paramétrique du credit scoring à savoir la régression logistique et la discrimination linéaire-quadratique au sens de Fisher, ensuite un quatrième chapitre est consacré à la technique pratique de construction et représentation d'un scoring, on fait varier le seuil  $s$  de discrimination et on propose un algorithme pour estimer les mal classés lors de la prédiction, il s'agit des erreurs de première et deuxième espèce. Le chapitre cinq enfin, est essentiellement porté sur les applications informatiques via le logiciel **R** des différentes méthodes annoncées aux chapitres trois et quatre, en essayant d'interpréter les sorties obtenues. Un paragraphe pour les recommandations y est inséré à la fin pour conclure ce travail.

# PRESENTATION DE LA STRUCTURE D'ACCUEIL ET CONCEPT DE RISQUE BANCAIRE

---

Ce premier chapitre de notre travail est d'une part consacré à une présentation sommaire de la structure dans lequel nous avons effectué notre stage académique. D'autre part, on y présente dans sa généralité le concept de *risque bancaire* en y mettant un accent particulier sur le risque de crédit, la raison d'être de notre travail.

## 1.1 Présentation de la structure d'accueil

### 1.1.1 Afriland First Bank

Afriland First Bank est un établissement bancaire de 6 500 000 000 FCFA de capital social. C'est une Société Anonyme (SA) dont l'histoire remonte au 4 octobre 1987, date de création de la Caisse Commune d'Épargne et d'Investissement (CCEI) qui allait être rebaptisée Afriland First Bank en abrégé First Bank 15 années plus tard. Son siège social est à Yaoundé. Le tableau 1.1 présente les principales caractéristiques de la First Bank.

Notre stage s'est déroulé au siège social de la First Bank à Yaoundé, précisément au sein de la Direction des Études et du Corporate Banking (DECB) dont les missions et l'organisation sont sommairement présentées dans les lignes qui suivent :

### 1.1.2 La Direction des Etudes et du Corporate Banking(DECB)

#### a- Les missions de la DECB

Plusieurs missions sont assignées à la DECB, notamment :

- l'étude de faisabilité des projets ;
- l'évaluation des entreprises ;
- l'étude de la restructuration des entreprises ;
- l'élaboration d'une banque de données économiques et statistiques ;

**TAB. 1.1 – Fiche d'identification de Afriland First Bank**

<b>Raison sociale : Afriland First Bank</b>
<b>Forme juridique : S.A</b>
<b>Siège social : Yaoundé, Hippodrome, Place de l'indépendance, B.P : 11834 Tel. : 22 23 30 68 / 22 22 37 34/22 23 63 27 Fax : 22 22 17 85 Telex : 8907 KN Web : <a href="http://www.afrilandfirstbank.com">www.afrilandfirstbank.com</a></b>
<b>Capital social : 6 500 000 000 F CFA</b>
<b>Vocation : La volonté d'être et de rester une banque africaine</b>
<b>Ambitions :</b> <ul style="list-style-type: none"><li>- rester le partenaire de l'entreprise gagnante ;</li><li>- entretenir la flamme de l'innovation ;</li><li>- rester la banque de proximité ;</li><li>- nourrir la croissance par une bonne liquidité.</li></ul>

**Source :** [www.afrilandfirstbank.com](http://www.afrilandfirstbank.com)

- l'analyse des filières économiques ;
- l'alimentation permanente de la banque des projets ;
- l'organisation / le conseil / le suivi des entreprises ;
- la gestion des lignes de financement ;
- la promotion des entreprises ;
- la promotion des fonds de garanties mutuelles ;
- la recherche des solutions aux problèmes spécifiques de financement des entreprises ;
- la recherche des subventions pour le financement du suivi/conseil des entrepreneurs ;
- la recherche des lignes de financement moyen et long terme ;
- la recherche des partenaires étrangers ainsi que l'assistance technique pour les projets ;
- le développement des diverses relations avec les bailleurs de fonds ;
- la promotion et le suivi des microstructures ;
- la gestion du portefeuille des participations locales.

#### **b- L'organisation de la DECB**

La DECB est dirigée par un directeur qui en assure le suivi et la gestion. Elle comprend trois Départements :

##### **Le Département Micro banque organisé en cinq divisions :**

- la Division Micro banque Ouest et Nord-ouest ;
- la Division Micro banque Grand Nord ;
- la Division Micro banque Littoral, Est et Sud-Ouest ;
- la Division Micro banque Sud-Centre ;
- la Division Audit

##### **Le Département des Études, des Projets et des Investissements avec trois divisions :**

- la Division des Études ;
- la Division des Projets et Investissements ;
- la Division Documentation et Archivage.

##### **Le Département du Corporate Banking et des Marchés Financiers qui comprend trois divisions :**

- la Division des Marchés Financiers ;
- la Division Asset Management / Gestion Actif ;
- la Division du Corporate Banking

### **1.1.3 Contexte de l'étude**

#### **Le dispositif actuel d'étude des dossiers de crédit à la First Bank**

Le réemploi des ressources collectées au titre des crédits accordés aux agents économiques à besoin de financement est la raison d'être de la First Bank. En effet, plusieurs types de clients, personnes physiques ou morales peuvent, au besoin, solliciter le concours de la First Bank pour le financement de leurs projets ou diverses activités économiques.

Cependant, la First Bank ne répond pas favorablement à toutes les demandes exprimées

par ses clients potentiels. Seuls les clients jugés aptes à retourner les fonds reçus aux conditions convenues peuvent être financés. Cette aptitude à respecter ses engagements vis-à-vis de la banque s'évalue à travers l'étude des dossiers de demande de crédit introduits par les clients auprès de la banque. Cette étude est conduite respectivement par les analystes, les contre analystes des dossiers de crédit et les comités de crédit.

#### **a. Les analystes et les contre analystes des dossiers de crédit**

Les analystes des dossiers de crédit débutent l'analyse de tout dossier de crédit introduit auprès de la First Bank par les clients. Mais, parallèlement à cette analyse des dossiers de demande de crédit, les analystes assurent le conseil et l'orientation du client afin de lui permettre de bien circonscrire l'objet de sa demande.

Ce début d'analyse consiste notamment à :

- faire une description des caractéristiques du client et de son besoin exprimé ;
- faire l'état de la situation des engagements en cours du client vis-à-vis du système bancaire en général, et en particulier ses engagements vis-à-vis de la First Bank ;
- décrire le projet objet de la demande, et en analyser les risques, la rentabilité et la solvabilité ;
- recenser les types de garanties que le client propose pour la couverture d'éventuels engagements de la banque ;
- résumer les points forts et les points faibles susceptibles d'orienter une appréciation du dossier en traitement ;
- faire une proposition de décision vis-à-vis du financement sollicité par le client, ainsi que les conditions de banque que sont l'échéance, les garanties, le mode d'amortissement du crédit et le taux d'intérêt.

Ce travail des analystes est par la suite présenté à un contre analyste pour des critiques en vue de son amélioration. Le dossier étudié par l'analyste et le contre analyste est alors prêt à être présenté aux comités de crédit pour son appréciation.

#### **b. Les comités de crédit**

Ce sont les seules instances à même de valider définitivement un dossier de crédit devant bénéficié du concours de la banque dans les limites de leurs compétences. Il y a à cet effet 5 comités de crédit chacun habilité à valider les dossiers de crédit portant des montants compris dans un intervalle donné. Un autre critère distinctif de ces comités est la qualité des membres.

Le comité 1 commence l'analyse de tout dossier étudié par l'analyste et le contre analyste. Les membres débattent du dossier de crédit sur la base d'une fiche d'analyse rédigée par l'analyste. Il valide la demande de financement en précisant les conditions de banque (échéance, garanties et taux), ou la rejette, si le montant se trouve dans les limites de ses compétences. Si non, il donne son avis, favorable ou non, pour le financement du besoin du client, puis transmet le dossier au comité 2 qui suit le même processus. Cette démarche se poursuit jusqu'au comité 5 pour les montants des crédits pour lesquels les 4 premiers comités ne peuvent se prononcer définitivement. Enfin, notons que la décision de chaque comité est motivée, et accompagnée d'un procès verbal.

## 1.2 Concept de risque bancaire :

Dans cette partie, nous abordons sommairement le concept de *risque bancaire* et nous nous articulons essentiellement sur la zoologie du risque financier.

La principale mission des banques est d'assurer la fonction d'intermédiaire financier. Lorsqu'une banque combine des ressources d'origines diverses pour financer plusieurs emplois distincts, cette fonction est qualifiée d'allocation. Cette fonction d'intermédiation dans un environnement instable fait ainsi supporter à l'établissement financier quatre types de risques [7] :

1. *Les risques commerciaux* : ce sont les risques résultant de l'insolvabilité d'un acheteur privé dans le cadre d'une vente de marchandises ou d'une prestation de service, ou d'un fournisseur privé dans le cadre d'une opération de préfinancement. La couverture de ce risque peut être limitée à l'insolvabilité juridiquement constatée ou élargie à l'insolvabilité de fait (présomée) ou à la carence pure et simple (défaut).

2. *Les risques de positionnement concurrentiel* : C'est un type de risque principalement caractérisé par la situation d'un produit ou une entreprise à produit unique par rapport à la concurrence et de pouvoir tirer les enseignements qui s'imposent quant à la position concurrentielle de la firme et à l'attrait du marché.

3. *Les risques opérationnels* : Ce sont les risques que l'organisation, ses acteurs et l'environnement externe font courir à la banque. Ils se décomposent en 4 sous-ensembles :

- Le risque lié au système d'information : défaillance matérielle, bogue logiciel, obsolescence des technologies (matériel, langages de programmation, SGBD,...).
- Le risque lié aux processus (saisies erronées, non respect des procédures,...) ;
- Le risque lié aux personnes (absentéisme, fraude, mouvements sociaux,... mais aussi capacité de l'entreprise à assurer la relève sur les postes clés) ;
- Le risque lié aux événements extérieurs (terrorisme, catastrophe naturelle) .

4. *Les risques financiers* : Ce sont les plus importants, ces risques, pouvant entraîner des pertes sérieuses pour la banque, doivent être pris en compte dans sa gestion interne. On en distingue six principaux types :

**Le risque de liquidité**, pour une banque, représente l'éventualité de ne pas pouvoir faire face, à un moment donné, à ses engagements ou à ses échéances.

**Le risque de taux** d'un établissement financier est celui de voir sa rentabilité ou la valeur de ses fonds propres affectées par l'évolution des taux d'intérêt,

**Le risque de marché** est le risque de pertes sur les positions du bilan et du hors bilan à la suite de variations des prix de marché.

**Le risque de change** traduit le fait qu'une baisse des cours de change peut entraîner une perte de valeur libellée en de vises étrangères.

**Le risque de solvabilité** est l'éventualité de ne pas disposer de fonds propres suffisants pour absorber les pertes éventuelles.

**Le risque de crédit ou de contrepartie** : c'est le risque pour un créancier de perdre définitivement sa créance dans la mesure où le débiteur ne peut pas, même en liquidant l'ensemble de ses avoirs, rembourser la totalité de ses engagements.

Ce dernier est la principale typologie de risque à laquelle sont confrontés les établissements de crédit (banque) de nos jours. Le crédit comporte toujours un certain degré de risques. La banque ne peut les éliminer totalement, mais plutôt en les mesurant correctement, elle doit

pouvoir les gérer, par exemple les couvrir s'ils sont importants. Lorsque la banque octroie un crédit, elle pose un acte de confiance vis-à-vis du débiteur. Faire crédit, c'est essentiellement faire confiance : la banque croit au remboursement ultérieur de la somme empruntée. Mais il n'y a jamais de certitude absolue que le débiteur remboursera dans les délais convenus. C'est la raison pour laquelle la gestion du risque de crédit requiert une attention de grande envergure car le contraire pourrait entraîner la faillite de la banque.

### 1.2.1 Le risque de crédit : veiller aux défauts de paiement[13]

L'événement risqué est le non-respect par un client ou par une contrepartie de ses obligations financières ou, de manière plus générale, la détérioration de la qualité crédit de cette contrepartie. Tout produit bancaire pour lequel un défaut de paiement du client entraînerait une perte pour la banque doit donc faire l'objet d'un calcul de risque crédit. L'horizon de temps pertinent pour le risque de crédit s'étale donc jusqu'à l'expiration des contrats, mais il est souvent ramené à un an, période de recapitalisation de la banque.

### 1.2.2 Le risque de crédit : niveaux de gestion[7]

Le risque de crédit est géré à plusieurs niveaux :

1- **Les systèmes de gestion des limites** permettent de diversifier le risque et d'éviter la concentration des encours crédit sur un pays (risque géofigure), un secteur économique, un groupe international, etc.

2- **Les systèmes de scoring** évaluent la probabilité de défaut de paiement pour un client ex-ante (avant même de lui octroyer le crédit) ; ces systèmes sont souvent basés sur des statistiques de défaillances et permettent de segmenter les clients suivant le risque.

3- **Un système de gestion de portefeuille**, au-delà des systèmes utilisés pour l'octroi du crédit, permet d'optimiser les transactions. La notion de pertes moyennes intervient à nouveau, mais couplée à la notion de capital économique, pour dériver un « Risk Adjusted Return On Capital » (RAROC). Ce RAROC sera disponible à plusieurs niveaux : par transaction, par client et par entité de la banque. Ce système permet également, grâce aux investisseurs et aux marchés financiers, de redistribuer le portefeuille des crédits pour un rapport *rendement / risque* optimal.

# DESCRIPTION STATISTIQUE DE LA BASE DE DONNEES

---

## Introduction :

Les outils de la Statistique descriptive fournissent des résumés synthétiques de séries de valeurs adaptées à leur type (qualitatives ou quantitatives), et observées sur une population ou un échantillon. Ce chapitre se propose de présenter quelques moyens permettant de résumer les caractéristiques (tendance centrale, dispersion, boîte à moustaches, histogramme, tests statistiques) d'une variable statistique ou les relations entre variables de même type quantitatif (coefficient de corrélation, nuage de points) ou qualitatif. De types différents (rapport de corrélation, diagrammes en boîtes parallèles). Les notions présentées sont illustrées sur un jeu de données typique d'un credit scoring en marketing bancaire. C'est ensuite la recherche de prétraitements des données afin de les rendre conformes aux techniques de modélisation ou d'apprentissage qu'il sera nécessaire de mettre en oeuvre afin d'atteindre les objectifs fixés :

- Codage en classe ou recodage de classes,
- Imputations ou non des données manquantes,
- Classification supervisée et premier choix de variables.

Dans le cas d'une seule variable, Les notions les plus classiques sont celles de médiane, quantile, moyenne, fréquence, variance, écart-type définies parallèlement à des représentations figures : diagramme en bâton, histogramme, diagramme-boîte, figures cumulatifs, diagrammes en colonnes, en barre ou en secteurs. Dans le cas de deux variables, on s'intéressera à la corrélation, au rapport de corrélation ou encore à la statistique d'un test du khi deux associé à une table de contingence. Ces notions sont associées à différentes figures comme le nuage de points (scatterplot), les diagrammes-boîtes parallèles.

Dans ce qui suit, nous présentons la méthodologie de collecte des données et nous nous proposons simplement de produire via certains outils moins classiques mais efficaces et présents dans la plupart des logiciels statistiques comme le logiciel **R**. Cela nous permettra également d'illustrer les premières étapes exploratoires à réaliser sur notre jeu de données.

## 2.1 Méthodologie de collecte des données

La constitution d'un échantillon pour notre étude s'avère très coûteuse en temps du point de vue de la collecte des données, pour des raisons techniques liées à l'harmonisation des pratiques comptables de la First Bank ; mais également en raison des règles strictes de confidentia-

lité imposées par la nature des informations comptables et financières. Les données financières et comptables sont issues de la DECB - division des projets et investissements de la First Bank. La sélection s'est effectuée sur la base de données physiques comptables, disponibles et fiables pour des exercices datant jusqu'en fin 2006 <sup>1</sup>.

Pour notre étude, nous avons considéré comme unité statistique un dossier de crédit. A la First Bank, un dossier de crédit est un fichier physique dans lequel on retrouve toutes les informations comptables et financières sur un client. Les dossiers de crédit qu'on a pu consulter étaient essentiellement les dossiers sur les projets d'investissements pour lesquels la First Bank s'était engagée en mettant à la disposition du promoteur le crédit (en totalité ou en partialité suivant les recommandations des différents comités de crédit) dont il avait besoin pour le financement de son affaire. Dans la conduite d'une analyse statistique des dossiers de crédit, le premier travail a consisté à constituer un fichier qui contient des informations complètes sur des dossiers de prêts. La constitution de la base de données sous forme d'un tableau à deux entrées individus-variables nécessaire à cette analyse a commencé avec la définition des variables de natures diverses à collecter à partir des dossiers de crédit déjà octroyés par la First Bank et qui étaient arrivés à échéance. L'étape de la collecte sera suivie par celle de la saisie afin de disposer d'une base de données sous forme de fichier électronique pour les besoins d'analyse. Malgré la difficulté de collecte de données à laquelle nous avons été confrontés pour des raisons de « secret bancaire » émis par la banque, nous avons tout de même recueilli pendant une durée de trois semaines un total de 130 dossiers de crédit, bien évidemment peu satisfaisant en nombre, mais aussi, suffisant pour mener à terme notre travail. A l'issue de cette collecte, des 130 dossiers de crédit obtenus, on en dénombre 19 dont le crédit octroyé par la First Bank n'était pas remboursé totalement, en partie ou même pas remboursé après la date de l'échéance convenue avec la banque.

## 2.2 Description des variables d'analyse

Le choix des variables d'analyse se doit d'obéir à la seule logique de couverture maximale, autant que faire se peut, de l'information susceptible d'aider à distinguer les bons dossiers de crédit des mauvais dossiers, ou les mauvais clients des bons clients. Les variables à retenir doivent donc contenir l'essentiel de l'information sur le client. La batterie de critères économiques et financiers comporte 25 variables explicatives et une variable qualitative Y à expliquer dont les sélections sont faites selon les thèmes décrits dans le tableau 2.1 .

### Remarque

Dans le tableau 2.1, les variables FORJU, NACTI, SISO et Y sont qualitatives et les 22 autres sont quantitatives. Il est à noter la variable Y = statut d'un client est la variable qualitative binaire à prédire dont les modalités sont 0 = « bon client ou client non risqué » et 1 = « mauvais client ou client risqué ». Nous déclarons un client bon lorsqu'il a remboursé la totalité de son emprunt à l'échéance fixée avec la banque, sinon il est considéré comme mauvais client.

<sup>1</sup> Etant donné qu'un crédit mis sur pied en 2007 ne pouvait pas encore être à son terme au moment de notre étude, en effet la plupart des dossiers de crédit avait une durée de remboursement supérieure à 12 mois.

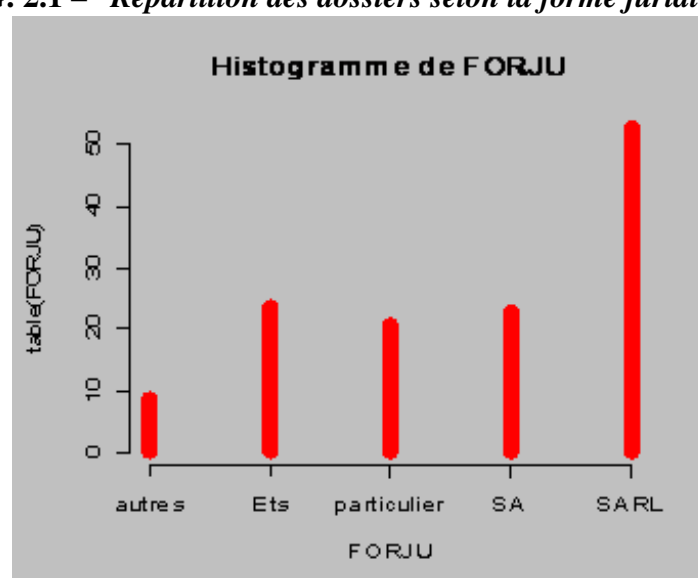
TAB. 2.1 – codage des variables d'étude

NOM	VARIABLE	CODAGE
FORJU	forme juridique	1= SARL ; 2= Ets ; 3=SA ; 4=particulier ; 5= autres(association et GIC)
CAP	montant du capital	en FCFA
NACTI	nature de l'activité	1= commerce général ; 2=BTP ; 3= industrie ; 4=santé publique ; 5=agriculture ; 6=commerce spécialisé ; 7=autres
SISO	siège social	1=Yaoundé ; 2=Douala ; 3=bafoussan ; 4=batouri ; 5=autres
CREDIT	montant du crédit octroyé	en FCFA
EFF	effectif de l'entreprise	en nombre de personnes
DUREMB	durée du remboursement	en mois
GAR	montant des garanties	en FCFA
AGE	âge du promoteur	en années
EXP	expérience du promoteur	en années
THT	taux d'intérêt hors taxes	en %
INVEST	montant des investissements	en FCFA
CHDI	charges directes et indirectes	en FCFA
MASA	masse salariale ou frais du personnel	en FCFA
CAF	cash flow	en FCFA
VA	valeur ajoutée	en FCFA
CA	chiffre d'affaire	en FCFA
RN	résultat net	en FCFA
R1	ratio 1 de rentabilité=CA/VA	numérique
R2	ratio 2 de rentabilité=RN/CA =taux de marge nette	numérique
R3	ratio 3 de rentabilité =RN/capitaux propres	numérique
R4	ratio 1 d'autonomie financière =capital/dettes	numérique
R5	ratio 2 d'autonomie financière =SN/dettes	numérique
R6	ratio 3 d'autonomie financière =SN/total passif	numérique
R7	ratio de solvabilité =actif total/dettes	numérique
Y	statut d'un client	0=<< bon client >> ; 1=<< mauvais client >>

TAB. 2.2 – Répartition des dossiers de crédit suivant la forme juridique des entreprises.

Forme juridique(FORJU)	effectif	(%)
Société à responsabilité limité(SARL)	53	40.77
Etablissement(Ets)	24	18,46
Société Anonyme(SA)	23	17,69
particulier	21	16.15
autres	9	6,92
<b>Total</b>	<b>130</b>	<b>100,00</b>

FIG. 2.1 – Répartition des dossiers selon la forme juridique.



Dans ce qui suit, nous décrivons d'abord les variables endogènes qualitatives, ensuite la description est portée sur certaines variables quantitatives endogènes en privilégiant les figures et en recherchant les éventuelles liaisons entre elles.

#### – La forme juridique(FORJU)

Le tableau 2.2 donne la répartition des dossiers de crédit enregistrés dans notre étude suivant la forme juridique des entreprises ayant initié ces dossiers. Le plus gros lot de dossiers (40,77 %) est issu des SARL. Les Ets suivent avec 18,46% de ces dossiers, les SA occupent 17,69%, les particuliers avec un peu plus de 16,15% des dossiers dans chaque cas. Environ 6,92 % des dossiers proviennent des autres i.e des groupements et associations.

Cette répartition des dossiers de crédit suivant la forme juridique peut être expliquée par des facteurs tels que le nombre de demandes exprimées, la qualité des projets présentés, ou le passé des entreprises auprès de la banque. Nous associons au tableau 2.2 un histogramme de la variable FORJU.(Cf. figure 2.1)

#### – Nature de l'activité (NACTI)

Une riche gamme d'activités est couverte par les entreprises ayant sollicité avec succès le

TAB. 2.3 – Répartition des dossiers par activités principales des entreprises.

Activité principale des entreprises	Effectif
commerce général	28
BTP	26
industrie	14
santé publique	9
agriculture	8
enseignement	7
commerce spécialisé	7
import-export	6
hôtellerie	4
médecine et chirurgie	2
restauration	4
prestations de services	4
services de transport	2
télécommunications	2
gestion immobilière	1
distribution des hydrocarbures	1
social	1
services financiers	1
pharmacie	1
imprimerie	1
communication audiovisuelle	1
<b>Total</b>	<b>130</b>

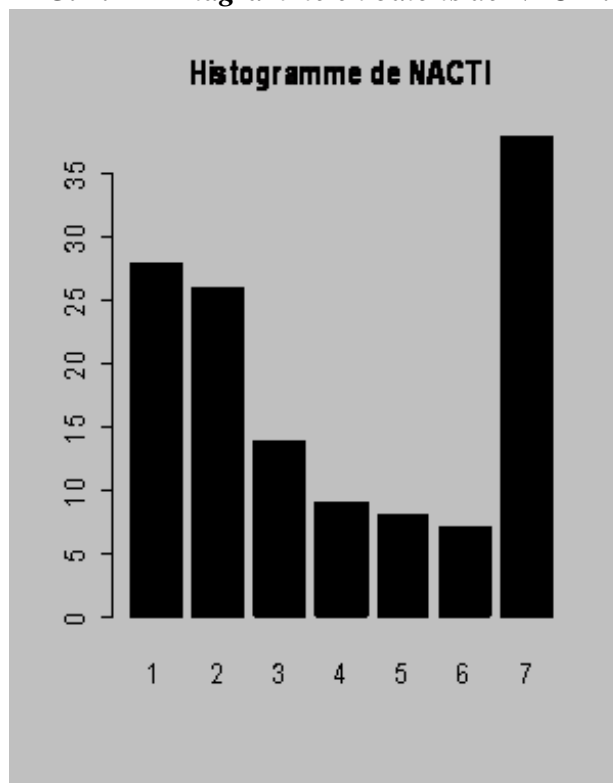
concours de la First Bank. La liste des domaines d'activités principales des entreprises dont les dossiers de crédit sont validés est donnée dans le tableau 2.3.

Les commerçants, les entreprises de bâtiment et travaux publics (54 dossiers) ont enregistré près de la moitié des dossiers validés. Les autres activités sont très peu représentées dans notre base de données. Nous illustrons ces propos par une représentation de la variable NACTI.(Cf figure 2.2)

#### – Le siège social ou le lieu d'implantation des entreprises financées (SISO)

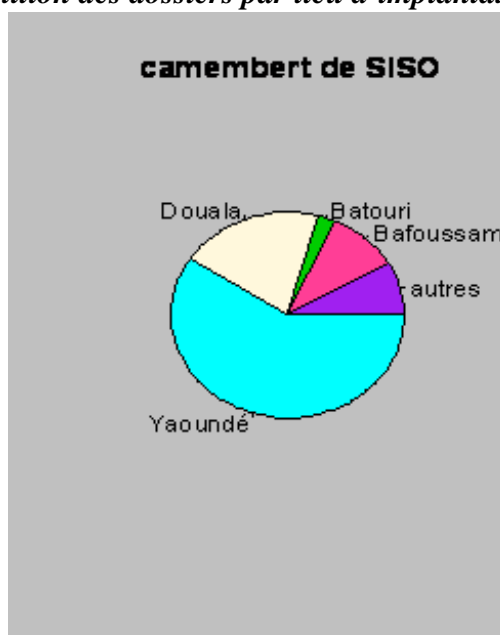
Nous avons utilisé 5 modalités pour cette variable lors de la collecte. Ces modalités sont notamment : Yaoundé, Douala, Bafoussam, Batouri et les autres villes. Sur le plan national, cette répartition suit la logique de concentration des entreprises et d'intensité de l'activité économique, comme le montre la figure 2.3 .

FIG. 2.2 – Diagramme en bâtons de NACTI.



1=commerce général | 2=BTP | 3=industrie | 4=santé publique | 5=agricultures | 6=commerce spécialisé | 7=autres.

FIG. 2.3 – Répartition des dossiers par lieu d'implantation des clients (%)



**TAB. 2.4 – Répartition des dossiers de crédit par les montants des besoins exprimés.**

Classe de besoin	Nombre de clients
10 M et moins	8
]10M ;25M]	30
]25M ;50M]	20
]50M ;500M]	62
]500M ;900M]	6
supérieur à 900M	4
<b>Total</b>	<b>130</b>

D'après la figure 2.3, On observe que la plus grande part de dossiers validés proviennent des entreprises basées à Yaoundé (59,23 %). Les dossiers provenant des entreprises implantées à Douala, où la concentration des entreprises est la plus grande à l'échelle nationale n'est que d'environ 20 %. Les autres villes et les zones rurales camerounaises sont représentées à hauteur de 8,46 % environ des avis de financement.

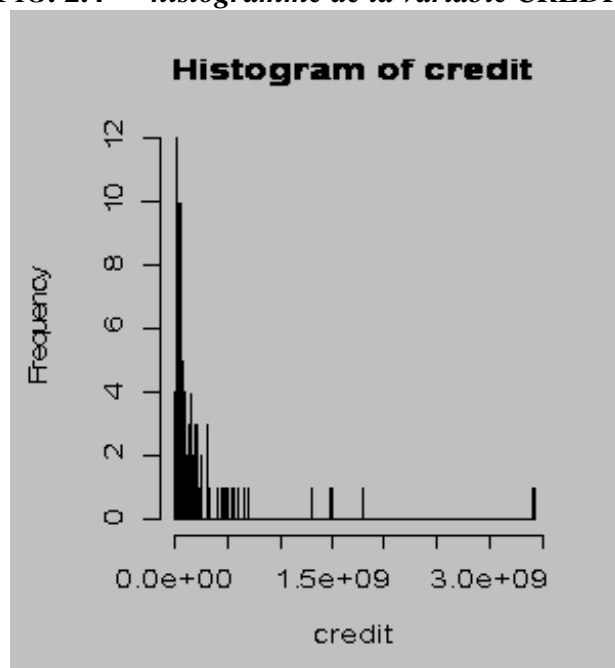
#### – Crédit octroyé (CREDIT)

La répartition des dossiers de crédit par montant de besoin exprimé du Tableau 2.4 montre que près de la moitié des emprunteurs de la First Bank sont ceux qui se sont vus accordés un crédit dont le besoin est compris entre 50 millions et 500 millions. Par contre, peu de clients ont eu un avis favorable à leur demande de crédit lorsque le besoin du financement est élevé (supérieur à 900 millions). La réticence de la First Bank à ce type de crédit provient peut-être du fait qu'elle ne veut pas financer les projets à coût trop élevé à cause du grand risque encouru pouvant engendrer des pertes énormes pour la banque. Par ailleurs, il est à remarquer aussi que la First Bank n'est pas intéressée par les clients dont la demande de crédit en besoin est faible (inférieur à 15 millions), une raison pouvant expliquer ce fait est que la banque estime le rendement de ce type de projet négligeable pour sa prospérité au vu des efforts investis et du temps consacré par les analystes pour l'étude d'un dossier d'un crédit. Le tableau 2.4 et la figure 2.4 illustrent ces propos.

#### – Variable garantie (GAR)

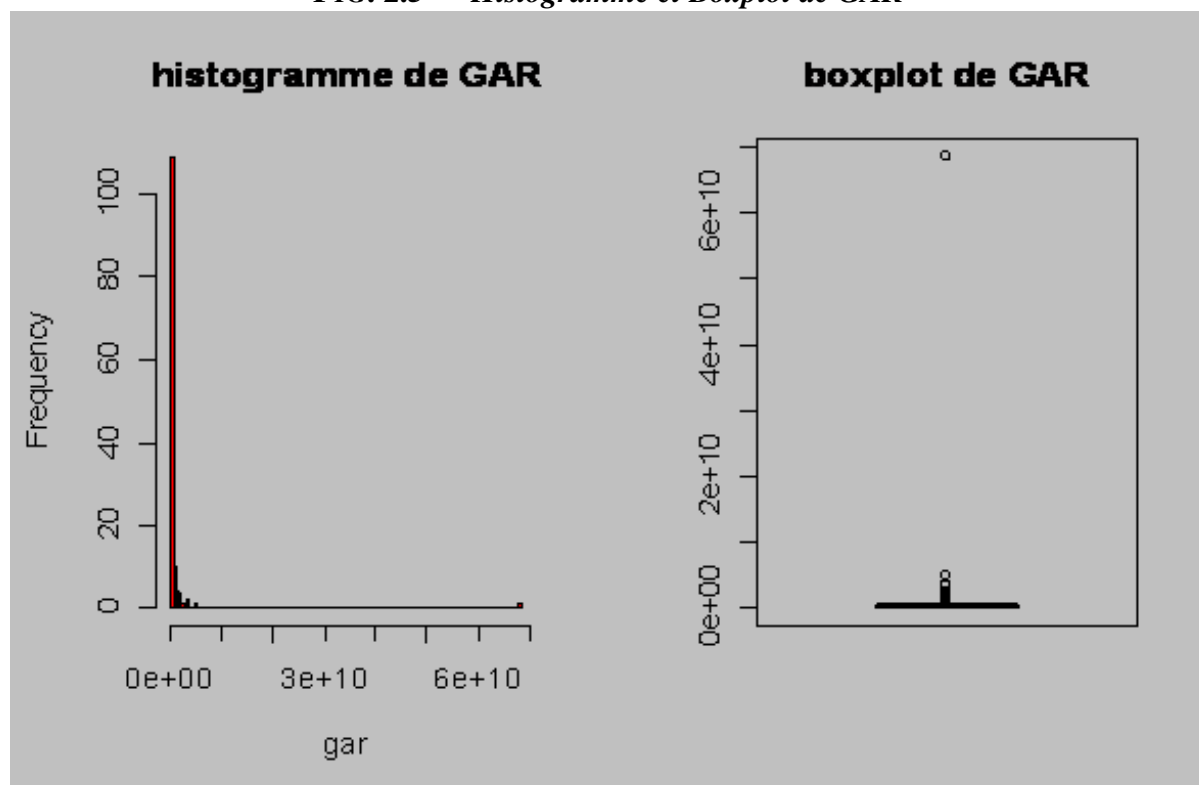
La répartition des garanties du Tableau 2.5 montre que plus de la moitié, soit 56% des dossiers de crédit porte une garantie comprise entre 100 millions et 500 millions, avec une moyenne des garanties=889.200.000 et un maximum=68.480.000.000 qui se présente comme une valeur aberrante.

Le diagramme-boîte (boxplot) et l'histogramme de la variable GAR illustre la distribution de la variable cumulant les garanties des emprunteurs. On constate une forte concentration de la variable GAR à la base de la figure de droite et une valeur atypique à l'extrémité supérieure, ce que confirme l'histogramme de GAR. Très peu de concours de crédit ayant eu une faible garantie ont été acceptés. On conclut donc que l'octroi d'un crédit à la First Bank est aussi déterminé par une masse matérielle assez imposante de garanties.

FIG. 2.4 – *histogramme de la variable CREDIT*TAB. 2.5 – *Répartition des dossiers de crédit par les montants des garanties.*

Classe de la garantie	Nombre de clients
15M et moins	3
]15M ;50M]	16
]50M ;100M]	17
]100M ;500M]	73
]500M ;1000M]	10
supérieur à 1000	11
<b>Total</b>	<b>130</b>

FIG. 2.5 – Histogramme et Boxplot de GAR



– Différents ratios de notre base de données

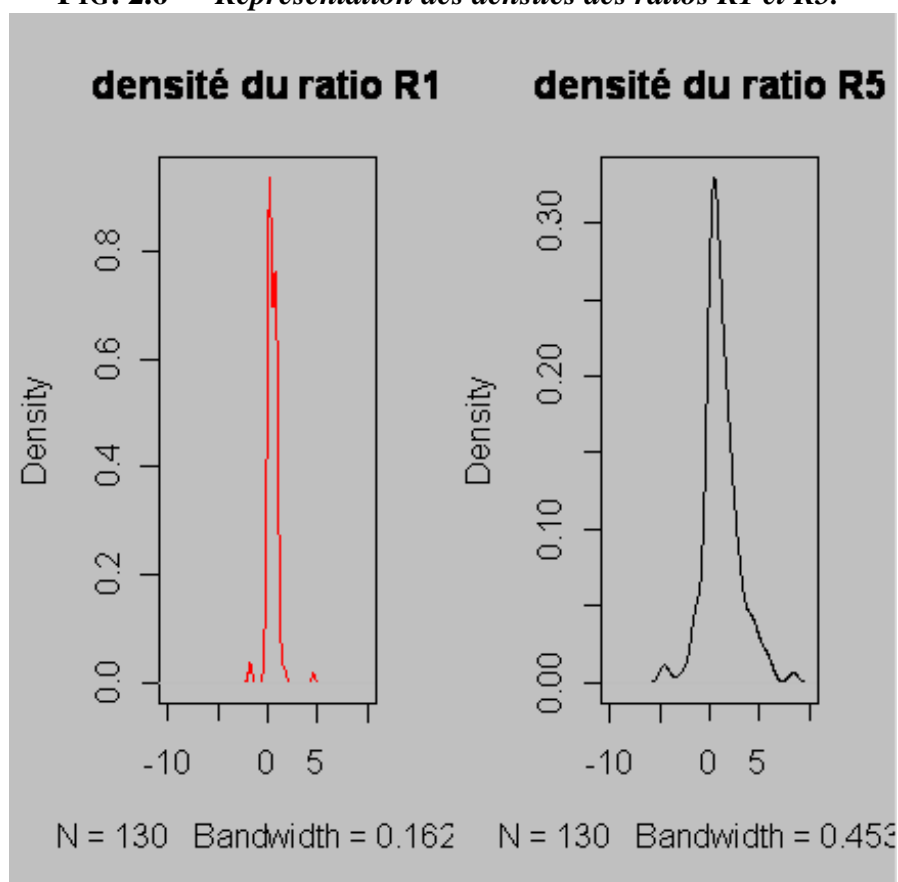
Nous résumerons la description des ratios financiers R1, R2, R3, R4, R5, R6, R7 dans le tableau 2.6 .

On constate que la plupart des ratios ont des valeurs qui fluctuent autour de zéro et ils ont une p-value résultant du test de Shapiro-Wilk inférieure au seuil 5% ; on rejette l'hypothèse nulle ( $H_0$ ) : le ratio suit une loi normale, donc on conclut qu'au seuil 5%, les ratios financiers de notre base de données ne sont pas des variables gaussiennes. Les représentations figures des fonctions de densité des ratios R1 et R5 contenues dans le figure 2.6 confirment les résultats du

TAB. 2.6 – Résumé et test de normalité des ratios

	Min	Median	Mean	Max	p-value du test de Shapiro au seuil 5%
R1	-1.8000	0.4350	0.4793	4.6000	$2,257.10^{-13}$
R2	-3.0300	0.0800	0.1351	2.9100	$pvalue < 2,2.10^{-16}$
R3	-0.7800	0.3500	0.5040	3.7100	$1,696.10^{-12}$
R4	-9.6200	0.2300	0.5249	8.6400	$2,894.10^{-12}$
R5	-4.810	0.845	1.146	8.360	$1,240.10^{-05}$
R6	1.3600	0.6800	0.6352	3.2500	$2,779.10^{-07}$
R7	-5.8600	1.5500	1.8980	9.3800	0.01168

FIG. 2.6 – Représentation des densités des ratios R1 et R5.



test précédent :

– **Le taux d'intérêt hors taxes (THT)**

Les taux d'intérêt hors taxes appliqués aux crédits à la First Bank sont très diversifiés en nombres, allant d'un minimum de 3 % (appliqué à un seul dossier à long terme) à 16.8% (appliqué deux dossiers court et moyen terme). L'évolution du nombre de dossiers validés en fonction du taux d'intérêt est représentée dans le tableau 2.7. Le taux de 13,75 % apparaît plus fréquemment aussi bien dans les contrats de courte période que dans les contrats de moyen terme. Pour l'ensemble des dossiers portant les taux d'intérêt (130 dossiers au total), on dénombre 49 dossiers portant ce taux de 13,75 %.

D'après le tableau 2.7, on constate que l'intervalle de THT qui est ]13,50 ;14,00] regorge une forte concentration des dossiers de crédit.

– **Liaison entre les variables de nos données**

Afin de vérifier s'il y a un éventuel lien entre les différentes variables prises en compte dans notre étude, nous avons calculé le coefficient de corrélation. Le choix de couple<sup>2</sup> de variables

<sup>2</sup>La matrice de variance-covariance de notre tableau nous permettait de voir les variables qui étaient liées.

**TAB. 2.7 – Répartition des dossiers suivant les taux de crédit et les échéances de remboursement.**

Taux de crédit HT en %	Echéances de remboursement			Total
	Court terme ([0 ;2 ans[)	Moyen terme ([2 ;10 ans[)	Long terme (≥ 10 ans)	
3,000	0	0	1	1
8,000	1	3	0	4
]8,000 ;8,500]	5	0	0	5
]8,500 ;9,000]	7	3	0	10
]9,000 ;9,500]	6	1	0	7
]9,500 ;10,00]	2	0	0	2
]10,00 ;10,50]	0	0	0	0
]10,50 ;11,00]	1	0	0	1
]11,00 ;11,50]	0	0	0	0
]11,50 ;12,00]	1	0	0	0
]12,00 ;12,50]	2	1	0	3
]12,50 ;13,00]	12	11	0	23
]13,00 ;13,50]	4	2	0	6
]13,50 ;14,00]	44	7	0	51
]14,00 ;14,50]	3	1	0	4
]14,50 ;15,00]	3	0	0	3
]15,00 ;15,50]	4	1	0	5
]15,50 ;16,00]	2	0	0	2
]16,00 ;17,00]	1	1	0	2
<b>Total</b>	<b>98</b>	<b>31</b>	<b>1</b>	<b>130</b>

a été guidé par des présomptions de relation entre celles-ci qui nous sont apparues logiques. Ainsi, Le coefficient de corrélation linéaire entre le taux d'intérêt hors taxes et la durée de remboursement du crédit vaut 0.13 avec un  $IC_{95\%} = [-0.0441833 ; 0.2946267]$ . On peut donc dire que la caractéristique « durée de remboursement du crédit » est faiblement corrélée au taux d'intérêt, en sorte que les échéances de plus en plus courtes correspondent aux taux d'intérêt de plus en plus élevés. Cette relation peut nous paraître plutôt surprenante, étant entendu que le taux d'intérêt est traditionnellement une fonction croissante du temps : « les taux d'intérêt à long terme sont généralement, mais pas toujours, supérieurs aux taux d'intérêt à court terme ». [13] (Gregory N. Mankiw, 2003, P 70).

Un résultat similaire s'établit aussi en utilisant le montant du crédit accordé et le taux d'intérêt hors taxes (coefficient de corrélation  $r = -0,011$ ,  $IC_{95\%} = [-0.1827414 ; 0.1615911]$ ). Les taux diminuent avec les montants de financement élevés. Mais à la différence du résultat précédent, ce second résultat paraît plus vraisemblable. Les plus gros clients représentent parfois une bonne opportunité de réemploi des ressources détenues par la banque, et ceux d'entre eux jugés « bons » méritent dès lors un assouplissement des contraintes de crédit, notamment en termes de coût de financement. A l'inverse, les clients sollicitant des concours de crédit assez petits supportent des taux d'intérêt de plus en plus élevés. On est ici face à un « système de rationnement du crédit » destiné à éviter le mécanisme d'« antisélection », et de permettre à la banque de garder ses « bons » clients. [15] (Patrick Villieu, 2000, P 56).

En considérant la corrélation positive entre le montant du crédit sollicité et l'échéance de remboursement (coefficient de corrélation  $r = 0,27$ ,  $IC_{95\%} = [0.1009225 ; 0.4211627]$ ) d'une part, et étant donné les résultats précédents, l'effet du montant des fonds prêtés sur le taux d'intérêt hors taxes semble plus probable. Une très forte corrélation entre la variable EFF et les variables CHDI et MASA, avec des coefficients de corrélation respectifs  $\text{cor}(\text{EFF}, \text{CHDI}) = 0.99$  avec  $IC_{95\%} = [0.9819636 ; 0.9909637]$  et  $\text{cor}(\text{EFF}, \text{MASA}) = 0.98$  avec  $IC_{95\%} = [0.9750032 ; 0.9874546]$ . Résultat prévisible à ce niveau car les charges d'une entreprise dont la masse salariale fait partie sont fonction de l'effectif du personnel de cette entreprise [13].

### CONCLUSION :

Cette étude importante permet de mettre en exergue le fait qu'il n'existe pas une méthode unique permettant de traiter des données d'expression ; la question "Quelle méthode dois je utiliser pour traiter mes données d'expression ?" n'a pas de sens. Il apparaît ainsi que face à des données d'expression, un statisticien seul, un analyste de crédit seul ou un comptable seul n'est pas en mesure de proposer des méthodes pertinentes ; la solution réside dans la collaboration des trois spécialités.

---

# DEUX METHODES DE DISCRIMINATION POUR LE CREDIT SCORING

---

## Introduction :

Grosso modo, le scoring consiste à affecter une note globale à un individu à partir de notes partielles, calculées sur des variables isolées ou en interaction. Cette note est utilisée essentiellement pour classer les individus par ordre ascendant ou descendant afin d'en sélectionner une partie pour une action marketing, par exemple le credit scoring. La construction d'un scoring fait appel à la modélisation prédictive, et l'on ne parle d'un scoring que lorsque la variable à prédire n'a que deux modalités[12]. Côté technique, le scoring est basé sur des méthodes classiques et qui n'évoluent que très peu d'un point de vue mathématique. En revanche, les possibilités d'évolution se trouvent dans leur application pour réaliser des analyses complexes. Et là toutes les possibilités ne sont pas encore explorées.

Dans ce chapitre, nous exposons deux approches mathématiques « classiques » du credit scoring pour la modélisation du risque de crédit à partir de l'étude du concept central de Data Mining pour les modèles paramétriques. Cependant, il existe plusieurs méthodes statistiques de construction d'un scoring dont les plus reconnues sont : l'analyse discriminante (linéaire, quadratique de Fisher), la régression logistique discriminante, les arbres de classification, méthode k-nn, les réseaux de neurones, les Séparateurs à Vaste Marge (SVM), etc. . . . .

Dans le cadre de notre étude, notre modèle sera construit à base deux modèles paramétriques à savoir la régression logistique discriminante et l'analyse discriminante (linéaire et/ou quadratique) de Fisher à cause de leur grande robustesse et leur facile interprétabilité. En effet, il sera question pour nous dont le but est d'identifier les clignotants du risque de crédit permettant de prévoir les défaillances, de construire pour chacune de ces méthodes un modèle et finalement mettre en compétition les deux modèles pour en retenir celui qui s'ajustera le mieux du point de vu prédictif à nos données.

## 3.1 Le modèle probabiliste de prédiction

Nous sommes en présence de  $n$  observations  $\{X_{i1}, \dots, X_{ip}, Y_i\}_{i=1}^p$  d'un couple  $(Y, X)$  dans une population  $\Omega$ . Pour la  $i^e$  observation notée  $(Y_i; X_i)$ ,  $Y_i$  est un label qui dénote l'appartenance à un groupe  $\in \{0; 1\}$ .

Une nouvelle observation  $x_0$  arrive, nous mesurons les variables explicatives, cette mesure est noté  $x_0 \in \mathbb{R}^p$  et nous souhaitons prédire son groupe  $Y = y_0$  à partir de l'observation de ses attributs  $\{X_i\}_{i=1}^p = \{x_{0i}\}_{i=1}^p$  avec une probabilité de se tromper dans cette prédiction aussi faible que possible. Ceci revient à mettre en évidence une fonction :

$$g : \mathbb{R}^p \longrightarrow \{0; 1\}$$

telle que l'erreur  $\varepsilon(g) = \mathbb{P}(g(X) \neq Y)$  soit aussi petite que possible.

Dans l'idéal, il faudrait chercher une fonction

$$g^* : \mathbb{R}^p \longrightarrow \{0; 1\} \text{ vérifiant } \varepsilon(g^*) = \min_{g: \mathbb{R}^p \rightarrow \{0;1\}} \varepsilon(g).$$

Si une telle fonction  $g^*$  existe, le prédicteur  $g^*(X)$  serait le meilleur pour prédire  $Y$  à partir de l'observation de  $X$ . [1]

## Prédicteur de Bayes- Erreur de Bayes

Pour  $g : \mathbb{R}^p \longrightarrow \{0; 1\}$ , on a :

$$\varepsilon(g) = \mathbb{P}(g(X) \neq Y) = \mathbb{P}(g(X) = 0; Y = 1) + \mathbb{P}(g(X) = 1; Y = 0) = \mathbb{E}(1_{g(X) \neq Y}) \quad (3.1)$$

Posons alors  $\varepsilon(g|X = x) = P(g(X) \neq Y|X = x)$  = probabilité de se tromper dans la prédiction de la valeur de  $Y$  pour un individu connaissant déjà les valeurs de ses attributs  $X = x$ .

On a donc d'après (3.1)

$$\begin{aligned} \varepsilon(g) &= \int_{\Omega} \mathbb{P}(g(X) \neq Y | X = x) dP_X(x) \\ &= \int_{\Omega} \varepsilon(g|X = x_0) dP_X(x) \end{aligned}$$

Ainsi  $g^*$  rend minimum  $\varepsilon(g)$  parmi les fonctions  $g : \mathbb{R}^p \longrightarrow \{0; 1\}$  si et seulement si  $g^*$  rend minimum  $\varepsilon(g|X = x_0)$ ,  $\forall x_0 \in \mathbb{R}^p$ , parmi les fonctions  $g : \mathbb{R}^p \longrightarrow \{0; 1\}$

### **définition**

1. un prédicteur  $g^*(X)$  de  $Y$  qui vérifie

$$\varepsilon(g^*|X = x_0) = \min \varepsilon(g|X = x_0); \forall g : \mathbb{R}^p \longrightarrow \{0; 1\}$$

est appelé *prédicteur de Bayes* pour prédire  $Y|X = x_0$ .

2. C'est le meilleur prédicteur de  $Y|X = x_0$  car

$$\varepsilon(g^*) = \min \varepsilon(g); \forall g : \mathbb{R}^p \rightarrow \{0; 1\}$$

3.  $\varepsilon^* = \varepsilon(g^*)$  est appelé *erreur de Bayes*.

### Vocabulaire

Pour  $j = 0; 1$  on a :

- $\mathbb{P}_j = \mathbb{P}(Y = j)$  = probabilité à priori de la classe ( $Y=j$ ) dans l'échantillon.
- $\mathbb{P}(Y = j|X = x_0)$  = probabilité à postériori de la classe ( $Y=j$ ) pour un individu dans la population.
- $\mathcal{V}_j(x_0) = \mathbb{P}(X = x_0|Y = j)$  = probabilité que  $X=x_0$  dans la classe ( $Y=j$ ). C'est aussi la vraisemblance de la valeur  $Y=j$  lorsqu'on a observé  $X = x_0$ .

Le théorème de Bayes nous permet d'écrire :

$$\text{pour } j=0;1 \quad \mathbb{P}(Y = j|X = x_0) = \frac{\mathbb{P}_j \times \mathbb{P}(X = x_0|Y = j)}{P_1 \times \mathbb{P}(X = x_0|Y = 1) + P_0 \times \mathbb{P}(X = x_0|Y = 0)}.$$

Le prédicteur de Bayes  $g^*(X)$  peut aussi se définir par :

$$\forall j = 0, 1 \quad g^*(X) = j \iff \mathbb{P}_j \times \mathbb{P}(X = x_0|Y = j) = \max_{k=0;1} \mathbb{P}_k \times \mathbb{P}(X = x_0|Y = k)$$

*i.e*  $j = \arg \max_{k=0;1} \mathbb{P}_k \times \mathbb{P}(X = x_0|Y = k);$

On demontre que la règle de décision finale s'écrit comme suit :

1. si  $\mathbb{P}(Y = 1|X = x_0) < \frac{1}{2}$ , alors  $g^*(x_0) = 0$
2. si  $\mathbb{P}(Y = 1|X = x_0) > \frac{1}{2}$ , alors  $g^*(x_0) = 1$
3. si  $\mathbb{P}(Y = 1|X = x_0) = \frac{1}{2}$ , alors  $g^*(x_0) = 1$  ou  $g^*(x_0) = 0$ , peu importe.

## 3.2 Analyse discriminante linéaire et quadratique

Les probabilités à priori des groupes  $j$ , notées  $\mathbb{P}(Y = j)$ ,  $j = 0; 1$  sont connues. Quand on n'a pas d'à priori, on peut, soit choisir que les groupes sont équivalents  $\mathbb{P}(Y = j) = \frac{1}{2}$ , soit l'estimer à partir des fréquences de chaque groupe dans les observations  $\{Y_i\}_{i=1}^n$ . Afin de spécifier le modèle de discrimination linéaire et quadratique, nous allons supposer l'hypothèse de normalité ci-dessous.

- **Discrimination quadratique** : La densité des variables explicatives dans chaque groupe  $j$  suit une loi multinormale  $f(x|y = j) \sim \mathcal{N}_p(\mu_j; \Sigma_j); \Sigma_j \in \mathcal{M}_p(\mathbb{R}), j = 0; 1$ .

Ensuite, nous pouvons ajouter une hypothèse supplémentaire pour obtenir le modèle de discrimination linéaire.

- **Discrimination linéaire** : La densité des variables explicatives dans chaque groupe  $j$  suit une loi multinormale de même matrice de variance-covariance  $\Sigma$  dans chacun des groupes :

$$f(x|y = j) \sim \mathcal{N}_p(\mu_j; \Sigma); \Sigma \in \mathcal{M}_p(\mathbb{R}), j = 0; 1$$

Une fois estimés tous les paramètres des lois normales, il suffit alors d'utiliser la règle de prédiction de Bayes pour connaître les probabilités d'affectation de la nouvelle observation aux différents groupes. Evidemment la prévision par la méthode sera donnée par le groupe le plus probable i.e

$$j = \operatorname{argmax} \mathbb{P}(Y = k|X = x_0) = \operatorname{argmax} f(x_0|y = k) \mathbb{P}(Y = k); \forall k \in \{0; 1\}.$$

### 3.2.1 Estimation des paramètres

Nous devons dans les 2 groupes, estimer  $(\mu_j; \Sigma_j); j = 0; 1$  où  $\mu_j \in \mathbb{R}^p$  et  $\Sigma_j \in \mathcal{M}_p(\mathbb{R})$ . Il y a donc 2 moyennes à estimer et 1 ou 2 matrices de variance-covariance à estimer. Il existe de nombreuses procédures d'estimations plus ou moins classiques.

Citons par exemple :

- La méthode des moments.
- La méthode de vraisemblance.

#### a) La méthode des moments.

Les moyennes par groupes  $\mu_j$  sont estimés par le centre de gravité de chacun des groupes

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i \in J} x_{0i};$$

où  $J$  est l'ensemble des numéros d'observations qui sont dans le groupe  $j$  et  $n_j$  le nombre d'observations dans le groupe  $j$  (ce qui est le cardinal de  $J$ ).

Pour les matrices de variance-covariance (méthode discriminante quadratique), elles sont estimées par :

$$\hat{\Sigma}_j = \frac{1}{n_j - 1} \sum_{i \in J} (x_{0i} - \hat{\mu}_j) (x_{0i} - \hat{\mu}_j)^T$$

Pour la discrimination linéaire, la matrice de variance-covariance est estimée par

$$\hat{\Sigma}_j = \frac{1}{n - 2} \sum_{j=1}^g \sum_{i \in J} (x_{0i} - \hat{\mu}_j) (x_{0i} - \hat{\mu}_j)^T$$

#### b) La méthode du maximum de vraisemblance

Les moyennes par groupes  $\mu_j$  sont encore estimés par le centre de gravité de chacun des groupes

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i \in J} X_i$$

où  $J$  est l'ensemble des numéros d'observations qui sont dans le groupe  $j$  et  $n_j$  le nombre d'observations dans le groupe  $j$  (ce qui est le cardinal de  $J$ ). Par contre les variances sont estimées par :

- Discrimination quadratique

$$\hat{\Sigma}_j = \frac{1}{n_j} \sum_{i \in J} (x_{0i} - \hat{\mu}_j) (x_{0i} - \hat{\mu}_j)^T$$

- Discrimination linéaire

$$\hat{\Sigma} = \frac{1}{n} \sum_{j=1}^g \sum_{i \in J} (x_{0i} - \hat{\mu}_j) (x_{0i} - \hat{\mu}_j)^T$$

Dans le cadre de notre travail, Nous avons posé  $Y = 1$  ou  $Y = 0$  selon que  $X$  suit une loi multinomiale  $\mathcal{N}(\mu_1; \Sigma_1)$  (de densité  $f_{X|Y=1}$ ) ou  $\mathcal{N}(\mu_0; \Sigma_0)$  (de densité  $f_{X|Y=0}$ ). Supposons

de plus  $\Sigma_0 = \Sigma_1$  ie que la discrimination devra être linéaire. Comme, nous souhaitons avoir une mesure quantitative entre 0 et 1, donnant la propension à être 1, nous nous intéressons à la probabilité à posteriori de  $Y=1$ ,  $\mathbb{P}(Y = 1|X)$ .

### 3.2.2 Calcul du seuil théorique $s$

Si nous souhaitons savoir si un individu est franchement estimé à 1, alors  $\mathbb{P}(Y = 1|X)$  sera élevé par rapport à  $\mathbb{P}(Y = 0|X)$ . On a alors

$$\mathbb{P}(Y = 1|X = x) \succ \mathbb{P}(Y = 0|X = x)$$

$$f_{X|Y=1}\mathbb{P}(Y = 1) \succ f_{X|Y=0}\mathbb{P}(Y = 0)$$

or nous connaissons les 2 densités qui sont celles de 2 lois normales  $\mathcal{N}(\mu_1; \Sigma)$  et  $\mathcal{N}(\mu_0; \Sigma)$ . Nous avons donc :

$$\mathbb{P}(Y = 1) \frac{1}{(2\pi)^{\frac{p}{2}} \sqrt{|\Sigma|}} \exp \left\{ -\frac{1}{2} (x - \mu_1)^T \Sigma^{-1} (x - \mu_1) \right\} \succ$$

$$\mathbb{P}(Y = 0) \frac{1}{(2\pi)^{\frac{p}{2}} \sqrt{|\Sigma|}} \exp \left\{ -\frac{1}{2} (x - \mu_0)^T \Sigma^{-1} (x - \mu_0) \right\}.$$

En passant au log, nous avons alors :

$$x^T \Sigma^{-1} (\mu_1 - \mu_0) + \log(\mathbb{P}(Y = 0)) - \log(\mathbb{P}(Y = 1)) - \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0 \succ 0$$

Ce qui s'écrit comme  $S(x) \succ s$  ;

avec  $S(x) = x^T \Sigma^{-1} (\mu_1 - \mu_0)$

et  $s = \log(\mathbb{P}(Y = 1)) - \log(\mathbb{P}(Y = 0)) + \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 - \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0$ .

$S(x)$  est appelée *fonction discriminante de Bayes*. C'est la *fonction scoring* de l'analyse discriminante linéaire à 2 classes et  $s$  est le *seuil*. Ce seuil dépend des probabilités à priori de ( $Y = 1$ ) et celle de ( $Y = 0$ ).

En général, ces probabilités sont inconnues à priori. Si des études ont été menées sur d'autres données, il est alors possible de connaître ces 2 probabilités. Mais, en l'absence de connaissance, elles sont posées égales à  $\frac{1}{2}$  chacune. La détermination du seuil séparant le choix ( $Y = 1$ ) du choix ( $Y = 0$ ) est donc délicat. Par ailleurs, le fait de ne plus considérer la probabilité à postérieure, mais un scoring permet d'éviter le calcul de  $f_X(x)$  qui est une densité difficilement calculable.

L'analyse discriminante linéaire est très répandue dans les logiciels de Statistique et d'apprentissage et est très utilisée. Les raisons de son succès sont les suivantes :

- elle offre souvent un très bon compromis *pertinence/complexité* ; autrement dit, elle permet souvent de bien résoudre le dilemme biais-variance. Elle est ainsi souvent supérieure à l'analyse discriminante quadratique qui dépend d'un nombre notamment plus important de paramètres.

- dans le cadre de l'analyse discriminante linéaire, la sélection de variables peut être réalisée de manière quasi optimale en utilisant une statistique F de Fisher[4]. En fait, les critères classiques de sélection de variables supposent de manière sous-jacente les hypothèses gaussiennes de l'analyse discriminante gaussienne. Ainsi, dans une prédiction à deux classes, on peut montrer que, la probabilité optimale d'erreurs s'écrit  $\Phi(-\Delta/2)$ ,  $\Phi$  étant la fonction de répartition d'une loi normale centrée réduite et  $\Delta$  représentant la distance de Mahalanobis entre deux groupes :

$$\Delta^2 = \|\mu_1 - \mu_0\|_{\Sigma^{-1}} = (\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0);$$

-l'analyse discriminante linéaire fournit des résultats *stables* (peu sujets aux fluctuations d'échantillonnage) et *robustes* (i.e supportant bien des écarts assez importants à ces hypothèses de normalité des groupes et d'égalité des matrices variances).

### 3.3 Analyse discriminante logistique

#### 3.3.1 Définition

##### a) introduction

L'analyse discriminante logistique est une méthodologie statistique qui a pour objectif, à partir d'observations, de produire un modèle permettant de prédire les valeurs prises par une variable catégorielle, à partir d'une série de variables explicatives continues et/ou binaires. Il s'agit ici pour nous de prévoir à l'aide de  $p$  variables explicatives l'appartenance à un groupe. Comme il existe une incertitude, nous la modélisons comme une probabilité et nous cherchons  $\mathbb{P}(Y = j|X = x_0)$  i.e la probabilité que l'observation soit dans le groupe  $j$  sachant nous avons en main l'observation  $x_0$  des variables explicatives. On pourra poser  $x_0 = (1, x_{01}, \dots, x_{0p})$ .

Le premier problème est que nous modélisons des probabilités discrètes, nous avons donc une contrainte :

$$\sum_{j=1}^g \mathbb{P}(Y = j|X = x_0) = 1$$

Une fois déterminées  $(g - 1)$  probabilités, la dernière est donc connue. Pour tenir compte de cette contrainte, nous allons donc considérer un groupe témoin, par exemple le  $g^e$  groupe, ensuite, nous allons modéliser non pas  $\mathbb{P}(Y = j|X = x_0)$ , mais le rapport de cette probabilité à la probabilité témoin  $\frac{\mathbb{P}(Y=j|X=x_0)}{\mathbb{P}(Y=g|X=x_0)}$ .

Ce rapport est toujours positif et il est compris entre 0 et  $+\infty$ . En passant au log, nous obtenons une mesure qui sera dans  $\mathbb{R}$  et que nous pouvons relier aux variables explicatives

$X_1, \dots, X_p$  via une fonction  $f$ . Cette fonction est choisie dans la classe la plus simple, à savoir les fonctions linéaires. Et on écrit donc :

$$\ln \frac{\mathbb{P}(Y = j|X = x_0)}{\mathbb{P}(Y = g|X = x_0)} = f(x_0) = x_0^T \beta_j \quad (3.2)$$

Ce type de modélisation est appelé *analyse discriminante logistique multiclasse* ou *régression logistique multiclasse*.

Cependant le cas le plus classique est le cas où il existe  $g = 2$  classes. Dans ce cas la notation standard veut que  $Y = 0$  ou  $Y = 1$  et que l'on prenne comme référence le groupe  $Y = 1$ . Nous ne traiterons par la suite que le cas binaire, cas qui est utilisé dans l'élaboration d'un scoring.

### b) Régression logistique (binaire)

#### définition(2.1.1) : (Régression logistique)

Nous sommes en présence d'une variable à expliquer binaire  $Y$  et de variables explicatives  $(X_1, \dots, X_p) = X \in \mathbb{R}^p$ .

Le modèle de la régression logistique s'écrit :

$$\ln \frac{\mathbb{P}(Y = 1|X = x_0)}{1 - \mathbb{P}(Y = 1|X = x_0)} = x_0^T \beta \quad (3.3)$$

ou

$$\text{logit}(\mathbb{P}(Y = 1|X = x_0)) = x_0^T \beta$$

Son nom provient du fait que la fonction  $p \mapsto \ln \frac{p}{1-p}$  est appelée fonction logit, qui est une fonction dérivable bijective de  $]0; 1[$  dans  $\mathbb{R}$ .

Remarquons que nous pouvons réécrire (3.3)  $\mathbb{P}(Y = 1|X = x_0) = \frac{\exp(x_0^T \beta)}{1 + \exp(x_0^T \beta)}$ .

#### Remarque

Nous pouvons aussi écrire  $\mathbb{P}(Y = j|X = x_0)$  dans le modèle multiclasse comme suit :

$$\mathbb{P}(Y = j|X = x_0) = \frac{\exp(x_0^T \beta_j)}{1 + \sum_{k=1}^g \exp(x_0^T \beta_k)}$$

### 3.3.2 Lien avec les GLM :

Nous modélisons deux probabilités  $\mathbb{P}(Y = 1|X = x_0)$  et  $\mathbb{P}(Y = 0|X = x_0)$  ie que la loi de  $(Y|X = x_0)$  est simplement une Bernouilli de paramètre  $\mathbb{P}(Y = 1|X = x_0)$  qui dépend de la valeur  $x_0$  de  $X$ .

L'espérance d'une Bernouilli est simplement son paramètre,  $\mathbb{E}(Y|X = x_0) = \mathbb{P}(Y = 1|X = x_0)$ .

Dans un modèle de régression logistique, nous effectuons donc 2 choix :

1. le choix d'une loi pour  $Y|X=x_0$ , ici la loi de Bernoulli,
2. le choix de la modélisation de  $\mathbb{E}(Y|X = x_0)$  par *logit* ( $\mathbb{E}(Y|X = x_0) = x_0^T \beta$ ). La fonction *logit*(.) est *bijective*, *dérivable* et est appelée *fonction de lien*. C'est une fonction de lien spéciale, appelée canonique (pour la loi de Bernoulli). La variance d'une loi de Bernoulli est  $\mathbb{V}(Y|X = x_0) = \mathbb{P}(Y = 1|X = x_0) [1 - \mathbb{P}(Y = 1|X = x_0)]$

La variance des observations décrites par le modèle n'est donc pas constante et varie selon la valeur de X. La fonction de  $x_0$  qui a pour valeur  $\mathbb{P}(Y = 1|X = x_0) [1 - \mathbb{P}(Y = 1|X = x_0)]$  est appelée *fonction de variance*.

### Remarque

Il est possible de choisir d'autres fonctions de lien bijectives. Les choix classiques sont la fonction *probit*(.) (Qui est l'inverse de la fonction de répartition d'une loi normale  $\mathcal{N}(0; 1)$ ).

Une généralisation de la méthode de régression logistique (ou régression *probit*) est appelée GLM (generalized linear model). Cette méthode revient à choisir une loi parmi un ensemble restreint de loi (les lois exponentielles GLM), puis une fonction de lien  $\varphi(\cdot)$  parmi un ensemble réduit de fonctions bijectives dérivables. Ensuite nous avons  $\varphi(\mathbb{E}(Y|X = x_0)) = x_0^T \beta$ .

### 3.3.3 Estimation des paramètres

L'estimation des paramètres se fait ici par maximum de vraisemblance. Dans le cas général (multiclasse), cette maximisation fait appel à des procédures itératives de minimisations classiques comme la méthode de Newton. Dans le cas de la régression logistique (binaire), il existe une procédure spécifique dite IRLS (Iterative Reweighted Least Squares).

Nous sommes en présence de n observations des variables notées  $\{X_{i1}, \dots, X_{ip}, Y_i\}_{i=1}^n$ , dont la  $i^e$  est notée  $(x_i, y_i)$ ,  $y_i \in \{0; 1\}$ . La vraisemblance conditionnelle de  $Y|X = x_i$  associée à l'observation  $i$  s'écrit :

$$\mathcal{V}(y_i, \beta) = \mathbb{P}(Y = 1|X = x_i)^{y_i} \mathbb{P}(Y = 0|X = x_i)^{1-y_i}$$

Et donc la vraisemblance conditionnelle de l'échantillon  $y = (y_1, \dots, y_n)$  de taille  $n$  s'écrit sous la forme :

$$\mathcal{V}(y, \beta) = \prod_{i=1}^n \mathbb{P}(Y = 1|X = x_i)^{y_i} \mathbb{P}(Y = 0|X = x_i)^{1-y_i}$$

Dans la pratique, il est plus aisé de se servir de la Log-vraisemblance notée  $\mathcal{L}(y, \beta)$ .

En passant au log, nous avons alors

$$\mathcal{L}(y, \beta) = \sum_{i=1}^n \left\{ y_i \ln \frac{\mathbb{P}(Y = 1|X = x_i)}{\mathbb{P}(Y = 0|X = x_i)} + \ln (\mathbb{P}(Y = 0|X = x_i)) \right\}$$

Grâce à la définition du modèle logistique (3.3), nous avons alors :

$$\mathcal{L}(y, \beta) = \sum_{i=1}^n \{y_i x_i^T \beta - \ln(1 + \exp(x_i^T \beta))\}$$

Ainsi, si les estimations des probabilités  $p_i = \mathbb{P}(Y = 1|X = x_{i0})$  sont en accord avec les observations, la vraisemblance sera maximisée. Il revient donc de chercher la valeur de  $\beta$  qui maximise  $\mathcal{L}(y, \beta)$ .

La fonction logarithme étant *continue* et *strictement croissante*, la Log-vraisemblance se maximise avec la valeur de  $\beta$  en même temps que la vraisemblance. Pour avoir le maximum  $\hat{\beta}$ , il ne reste plus qu'à annuler le gradient de la fonction  $\mathcal{L}(y, \beta)$  : Mais du point de vue pratique, à cause de la macroforme de  $\mathcal{L}(y, \beta)$  et de la grandeur de la taille  $n$  ; on utilise des méthodes numériques d'optimisation pour obtenir la valeurs estimée de  $\beta$ .

**Prédicteur de Bayes estimé  $\widehat{g^*(X)}$  :** On a :  $\mathbb{P}(Y = 1|X = x_0) = \frac{\exp(x_0^T \beta)}{1 + \exp(x_0^T \beta)}$  par (3.3), on en déduit que  $\mathbb{P}(Y = 0|X = x_0) = \frac{1}{1 + \exp(x_0^T \beta)}$ .

On en déduit  $\widehat{g^*(X)}$  en remplaçant  $\beta$  par sa valeur estimée  $\hat{\beta}$  dans les expressions de  $\mathbb{P}(Y = 1|X = x_0)$  et  $\mathbb{P}(Y = 0|X = x_0)$  ci-dessus :

- $\widehat{g^*(x_0)} = 0 \iff \mathbb{P}(Y = 0|X = x_0) > \mathbb{P}(Y = 1|X = x_0) \iff x_0^T \hat{\beta} < 0$ .
- $\widehat{g^*(x_0)} = 1 \iff \mathbb{P}(Y = 1|X = x_0) > \mathbb{P}(Y = 0|X = x_0) \iff x_0^T \hat{\beta} > 0$ .

Le scalaire  $x_0^T \hat{\beta}$  est appelé *fonction discriminante logistique binaire*, aussi appelé *scoring*. On constate dans ce cas que le seuil  $s$  apparaît clairement et il vaut  $s=0$ .

Dans la pratique, il serait maladroit de se figer sur ce seuil théorique  $s=0$ , il est vivement conseillé de faire varier le seuil et de conserver celui qui rend la prédiction meilleure.

### 3.3.4 Précision des estimations :

Puisque nous utilisons le maximum de vraisemblance, il est alors possible de bâtir des intervalles de confiance pour  $\beta_j$  au seuil  $\alpha$  selon la formule suivante :

$$IC_\alpha(\beta_j) = \left[ \hat{\beta}_j - U_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{n}} \sigma_{\hat{\beta}_j}; \hat{\beta}_j + U_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{n}} \sigma_{\hat{\beta}_j} \right]$$

Où  $U_{1-\frac{\alpha}{2}}$  représente le quantile de niveau  $(1 - \frac{\alpha}{2})$  de la loi normale  $\mathcal{N}(0; 1)$ ,  $\sigma_{\hat{\beta}_j}^2$  est égal à  $\left[-I(\hat{\beta})\right]_{jj}^{-1}$  et  $\left[I(\hat{\beta})\right]_{jj}^{-1}$  est l'élément  $(j, j)$  de l'inverse de la matrice de Fisher  $\mathbb{E}\left(\frac{\partial^2 \mathcal{L}}{\partial \beta^2}\right)$ .

La validité de ces intervalles est toute relative puisqu'il s'agit d'une approximation valable asymptotiquement et dont la variance dans le cas de l'approximation normale, doit être évaluée à la vraie valeur du paramètre inconnu.

Il est toujours possible de compléter cette étude par bootstrap afin d'obtenir d'autres intervalles de confiance dans le cas où ceux-ci sont particulièrement importants. Cela dit, en pratique, on se contente de l'intervalle de confiance bâti grâce à la matrice d'information de Fisher.

### 3.3.5 La qualité du modèle

#### 3.3.5.1 Un outil spécifique : la déviance

Comme la vraisemblance n'est jamais à la même échelle (cela dépend des données), il n'est pas facile d'avoir une idée de la qualité d'ajustement. Pour cela, un outil spécifique est introduit : la déviance. Elle compare la vraisemblance obtenue à celle que l'on obtiendrait dans un modèle parfait : le modèle saturé. Dans le modèle saturé, la prévision est parfaite, il n'existe donc aucune incertitude et la probabilité estimée par le modèle au point  $X = x_i$  est donc 1 pour le groupe observé et 0 sinon. Dans le cas où plusieurs observations seraient disponibles au point  $X = x_i$ , alors, si le modèle était parfait,  $\hat{y}_i$  serait la moyenne des  $y_i$  au point  $X = x_i$ . Ce modèle est appelé *modèle saturé* par définition.

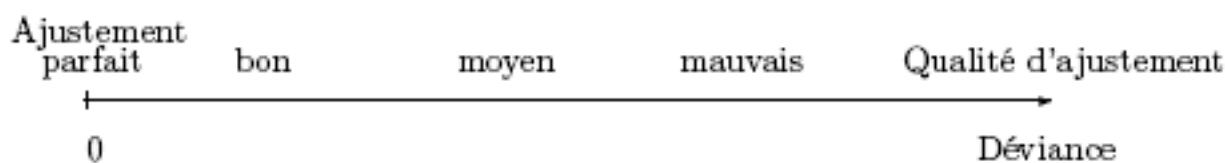
Pour le modèle logistique binaire, la vraisemblance pour l'observation  $i$  pour ce modèle saturé est égale par définition à :

$$\mathcal{L}_{\text{satur}} = \sum_{i=1}^n Y_i \log Y_i + (1 - Y_i) \log (1 - Y_i).$$

La déviance d'un modèle notée  $D$ , est définie par rapport au modèle saturé correspondant comme

$$D = 2 \left[ \sum_{i=1}^n (\mathcal{L}_{\text{satur}} - \mathcal{L}(\beta)) \right] \geq 0$$

La déviance est égale à 2 fois une différence de vraisemblance. Elle constitue un écart en terme de log-vraisemblance entre le modèle saturé d'ajustement maximum et le modèle considéré :



La déviance dans le cas binaire est donnée par :

$$D = 2 \sum_{i=1}^n Y_i \log \frac{Y_i}{\hat{P}_i} + (1 - Y_i) \log \frac{1 - Y_i}{1 - \hat{P}_i}$$

#### Test d'adéquation par la déviance

Puisque nous élaborons un test, définissons hypothèses nulle et alternative :

- $H_0$  le modèle considéré à  $p$  paramètres est adéquat.
- $H_1$  le modèle considéré à  $p$  paramètres n'est pas adéquat.

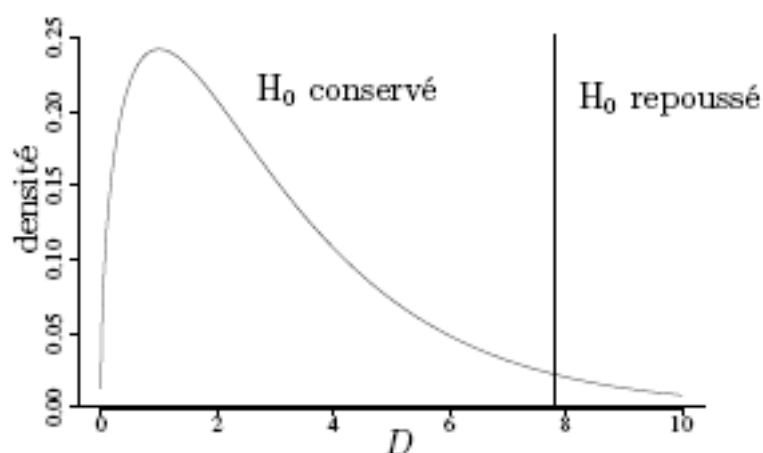


FIG. 3.1 – Test de déviance, la droite verticale représente le seuil de rejet  $D_c = q_{1-\alpha}(n - p)$ .

Ici, nous allons comparer le modèle saturé au modèle considéré au moyen de la déviance. Nous savons que si la déviance est grande, alors le modèle considéré est loin du modèle saturé et donc il n'est pas très adéquat. Par contre si la déviance est proche de 0, le modèle considéré sera adéquat. Pour quantifier cette notion de "proche de 0" et de "grande déviance", la loi de la déviance sous  $H_0$  (le modèle considéré est le vrai modèle) va nous être utile. En effet  $H_0$  si est vraie, le modèle considéré est vrai par définition. La déviance sera répartie sur  $\mathbb{R}^+$ , mais avec plus de chance d'être proche de 0. Par contre si  $H_0$  n'est pas vraie la déviance sera répartie sur  $\mathbb{R}^+$  mais avec plus de chance d'être éloignée de 0. Nous nous accordons  $\alpha$  % de chance de se tromper sous  $H_0$  donc si, l'on connaît la loi de D sous  $H_0$  alors en prenant le quantile de niveau  $1 - \alpha$  nous excluons les  $\alpha$  % d'erreur tout en excluant les déviations les plus grandes, ie les cas qui se présenteront vraisemblablement si  $H_0$  n'est pas vraie.

La déviance est en fait le test de rapport de vraisemblance et sous des hypothèses techniques ([8]Schervish, 1995, p. 459), D suit donc une loi du  $\chi^2(n - p)$  degrés de liberté, où p est le nombre de paramètres du modèle et n le nombre d'observations. Le test se déroule alors de la manière classique :

1. Les hypothèses sont fixées
  - $H_0$  le modèle considéré à p paramètre est adéquat
  - $H_1$  le modèle considéré à p paramètres n'est pas adéquat
2.  $\alpha$  est choisi (en général 5%)
3. L'observation de D est calculée, notons la  $D_{obs}$
4. Calcul du quantile de niveau  $(1 - \alpha)$  de la loi du  $\chi^2(n - p)$ , noté  $q_{1-\alpha}(n - p)$ .
  - Si  $D > q_{1-\alpha}(n - p)$  alors  $H_0$  est repoussé au profit de  $H_1$ , le modèle considéré n'est pas adéquat.
  - Si  $D_{obs} \leq q_{1-\alpha}(n - p)$  alors  $H_0$  est conservé, le modèle considéré est adéquat.

### Remarques

La validité de la loi et donc du test n'est qu'asymptotique, il est donc nécessaire d'avoir un peu de recul quant aux conclusions.

Lorsque les données sont binaires et qu'aucune répétition n'est présente au point  $X_i = x_i, \forall i$ ,

alors  $D$  ne suit pas une loi du  $\chi^2$ . Pour les données binaires le test d'adéquation d'Hosmer Lemershow est à conseiller.

### Test d'Hosmer Lemershow

Ce test permet de vérifier l'adéquation d'un modèle quand la variable à expliquer est une variable binaire uniquement. Il permet donc de vérifier l'adéquation dans les cas où le test d'adéquation par la déviance est particulièrement déconseillé.

Pour cela, les  $\hat{P}_i = \mathbb{P}(Y = 1 | X = x_i)$  sont ordonnés par ordre croissant. Ensuite  $K$  groupes de tailles égales sont créés, en général  $K = 10$  et le dernier groupe, celui des  $\hat{P}_i$  les plus grands, possède un effectif inégal aux autres. Notons  $m_k^*$  l'effectif du groupe  $k$ . Ensuite une statistique du type  $\chi^2$  est calculée sur ces groupes. L'effectif observé des cas ( $Y = 1$ ) dans le groupe  $k$  est décompté, ce qui donne de manière mathématique  $o_k = \sum_{j \in gpe\ k} y_j$ . La fréquence théorique est simplement la moyenne des probabilités estimées par le modèle, pour toutes les observations du groupe :  $\bar{\mu} = \sum_{j \in gpe\ k} \hat{P}_j$ . La statistique de test est alors

$$C^2 = \sum_{k=1}^K \frac{(o_k - m_k^* \bar{\mu}_k)^2}{m_k^* \bar{\mu}_k (1 - \bar{\mu}_k)},$$

où  $m_k^*$  est l'effectif du groupe  $k$ .

Le test se conduit de manière identique au test de déviance, la statistique  $C^2$  suivant approximativement un  $\chi^2$  à  $K - 1$  degrés de liberté. Cette approximation ayant été validée uniquement par simulation[4] (Collett, 2003, p. 88), il semble donc important de ne pas appliquer trop strictement la procédure de test, mais plutôt de la considérer comme une indication.

### Critère de choix de modèles

L'objet de ces critères de choix est de comparer des modèles entre eux et qui ne sont pas forcément emboîtés les uns dans les autres.

Par définition l'AIC (Akaike Informative Criterion) pour un modèle à  $p$  paramètres est

$$AIC = -2\mathcal{L} + 2p.$$

La philosophie est simple : plus la vraisemblance est grande, plus grande est donc la log-vraisemblance  $\mathcal{L}$  et meilleur est le modèle. Cependant si l'on met le nombre maximum de paramètres (ce qui est le modèle saturé) alors  $\mathcal{L}$  sera maximum. Il suffit donc de rajouter des paramètres pour la faire augmenter. Pour obtenir un modèle de taille raisonnable il sera donc bon de la pénaliser par une fonction du nombre de paramètre, ici  $2p$ . Un autre critère de choix de modèle le BIC (Bayesian Informative Criterion) pour un modèle à  $p$  paramètres estimé sur  $n$  observations est défini par :

$$BIC = -2\mathcal{L} + p \log(n).$$

L'utilisation de ces critères est simple. Pour chaque modèle concurrent le critère de choix de modèle est calculé et le modèle qui présente le plus faible est sélectionné.

Remarquons que certains logiciels utilisent  $-AIC$  et  $-BIC$ , il est donc prudent de bien vérifier dans quel sens doivent être optimisés ces critères (maximisation ou minimisation). Ceci peut être fait aisément en comparant un modèle très mauvais (sans variable explicative) à un bon modèle (à une variable) et de vérifier dans quel sens varie les critères de choix.

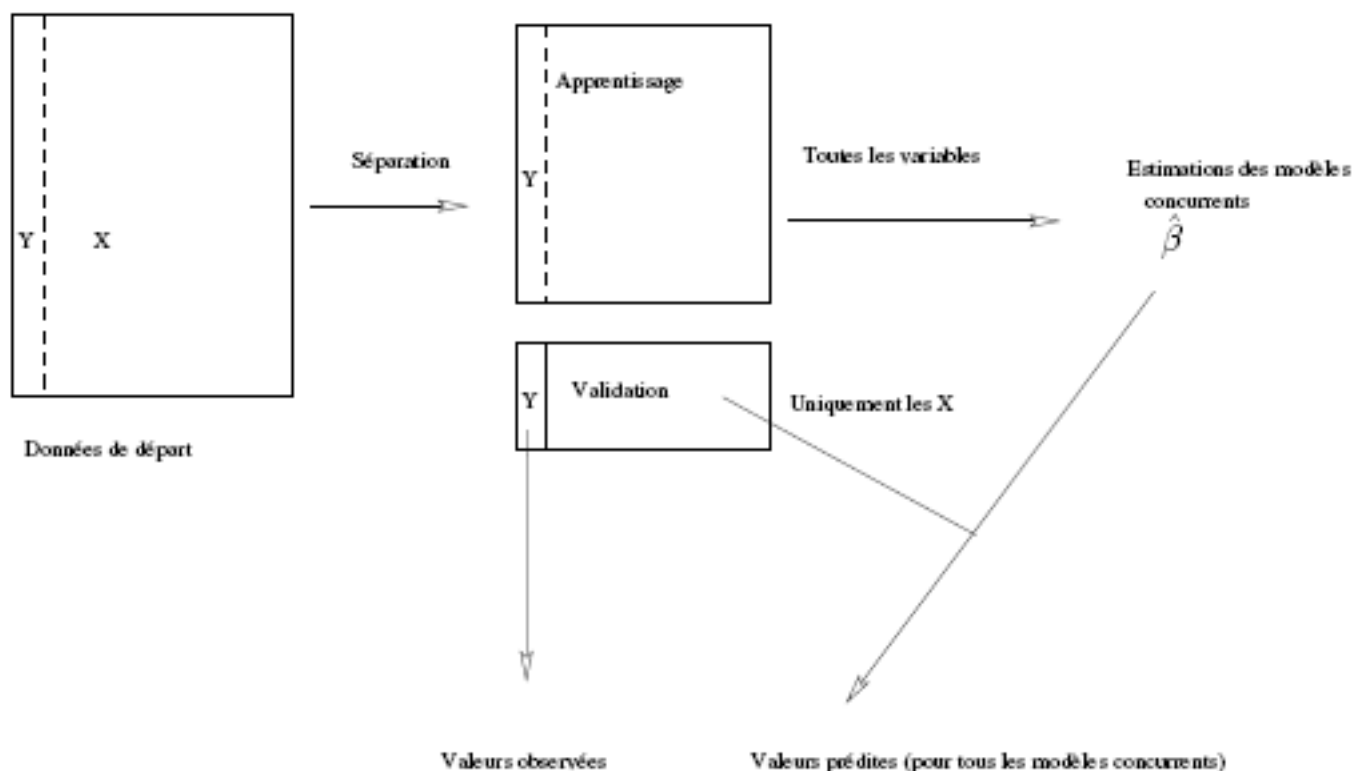


FIG. 3.2 – Procédure d’apprentissage/validation

### 3.3.5.2 Evaluation des performances

L’analyse discriminante décisionnelle vise à proposer une règle de décision destinée à être appliquée pour le classement dans le futur d’observations de provenance inconnue. Il est donc très important d’être capable de mesurer le taux d’erreur que l’on risque lors de l’application d’une règle de décision construite sur la base d’un échantillon d’apprentissage. Dans ce qui suit, nous passons en revue quelques moyens d’estimer le taux d’erreur réel d’une règle de décision.

#### Apprentissage/validation

La procédure de validation consiste à séparer de manière aléatoire les données en deux parties distinctes  $(y_a, X_a)$  et  $(y_v, X_v)$ . Tous les modèles concurrents sont construits avec le jeu d’apprentissage  $(y_a, X_a)$  (figure 3.2). Ensuite en utilisant tous ces modèles et les variables explicatives  $X_v$ , les valeurs de la variables à expliquer sont prédites  $\hat{y}_v(j)$  pour tous les modèles  $j$  concurrents. Comme le modèle de régression logistique binaire donne des estimations des probabilités  $\hat{\mathbb{P}}(Y = 1|X = X_v)$ , pour avoir une prévision binaire, on prend la modalité dont la probabilité estimée est la plus élevée. Nous noterons cette prévision ainsi obtenue  $\hat{y}_v$ .

La qualité du modèle est ensuite obtenue en mesurant la distance entre les observations prévues et les vraies observations par un critère. Le plus connu est le PRESS

$$PRESS(j) = \|\hat{y}_v(j) - \hat{y}_v\|^2,$$

mais dans le cas d'observations binaires il n'est pas utilisé. En général, on utilise le nombre de mal classé, ie

$$MC = \|\hat{y}_v(j) - y_v\|_1,$$

où  $\|x\|_1 = \sum_i |x_i|$ . Comme les valeurs de Y sont 0 ou 1, cette méthode est bien le nombre de mal classés.

Le modèle optimal  $k$  choisi est celui qui conduit au MC minimum (ou au critère choisi minimum). Cette procédure semble la plus indiquée mais elle nécessite beaucoup de données puisqu'il en faut suffisamment pour estimer le modèle et pas trop pénaliser les modèles avec beaucoup de variables dont les coefficients seront moins bien estimés, mais il faut aussi beaucoup d'observations dans le jeu de validation  $(y_v, X_v)$  pour bien évaluer la capacité de prévision dans de nombreux cas de figure. De plus, comment diviser le nombre d'observations dans le jeu d'apprentissage par rapport au jeu de validation ? Là encore aucune règle n'existe mais l'on mentionne souvent la règle 3/4 dans l'apprentissage et 1/4 dans la validation. De plus, il faut pouvoir calculer le MC sur chacun des modèles concurrents ce qui dans certains cas est impossible, lorsque le nombre de variables possibles  $p$  est grand.

### Validation croisée

Lorsque l'on n'a pas assez de données pour l'apprentissage/validation, la validation croisée est utilisée pour évaluer le taux d'erreur. La validation croisée, dans sa version la plus classique, connue sous le nom de *leave-one-out*, procède comme décrit ci-dessous.

Pour  $i = 1, \dots, n$  on construit la règle de décision sur la base de l'échantillon d'apprentissage privé de son  $i^e$  élément et on affecte ce dernier à l'un des groupes suivant cette règle. Le taux d'erreur estimé est alors la fréquence de points de mal classés de la sorte. L'estimation du taux d'erreur ainsi obtenue est pratiquement sans biais. Mais la variance de l'estimation est d'autant plus importante que  $n$  est grand puisque, dans ce cas, les différentes règles de décision construites à partir de  $n-2$  observations communes auront tendance à se ressembler. De plus, cette procédure est également assez coûteuse même si, du fait qu'à chaque étape l'échantillon soit amputé d'une seule observation, il est en général possible de recalculer la règle de décision.

Aussi, on peut lui préférer la procédure suivante. On divise l'échantillon aléatoirement en  $L$  parties (approximativement) égales. Pour  $l=1, \dots, L$ , on construit la règle de décision sur la base de cet échantillon privé de sa  $l^{eme}$  partie, ensuite pour cette  $l^{eme}$  partie donnée, on utilise la procédure d'apprentissage/validation, la  $l^{eme}$  partie étant le jeu de validation et les autres observations formant le jeu d'apprentissage. Si  $L = n$ , on retombe sur la procédure standard de *leave-one-out*. On évalue la qualité du modèle par un critère, le nombre de mal classés MC par exemple, donnant ainsi  $MC(j)_l$  et ensuite on itère le procédé sur toutes les parties  $l$  variant de 1 à  $L$ . Le critère final à minimiser est alors

$$MC_{CV}(j) = \sum_{l=1}^L MC(j)_l,$$

On en déduit une estimation de l'erreur de prédiction pour le modèle  $j$  :

$$\hat{\varepsilon}(j) = \frac{1}{L} MC_{CV}(j)$$

et le modèle  $k$  retenu est celui qui conduit au minimum sur  $\{MC_{CV}(j)\}$ , donc sur  $\hat{\varepsilon}(j)$ . Bien entendu le choix du nombre  $L$  parties n'est pas anodin. Plus le nombre  $L$  est faible, plus la capacité de prévision sera évaluée dans de nombreux cas puisque le nombre d'observations dans la validation sera élevé, mais moins l'estimation sera précise. Au contraire, un  $L$  élevé conduit à peu d'observations dans la validation et donc à une plus grande variance dans les nombres de mal classés.

### 3.3.6 Sélection automatique

La sélection de modèle peut être vue comme rechercher le modèle optimum au sens d'un critère choisi parmi toutes les possibilités. Cela peut donc être vu comme une optimisation d'une fonction objectif (le critère). Pour cela et à l'image des possibilités en optimisation, on peut soit faire une recherche exhaustive car le nombre de modèles possibles est fini, soit partir d'un point de départ et utiliser une méthode d'optimisation de la fonction objectif (recherche pas à pas).

Remarquons qu'en général trouver le minimum global de la fonction objectif n'est pas garanti dans les recherches pas à pas et que seul un optimum local sera trouvé dépendant du point de départ choisi. en général, on utilise l'une des méthodes suivantes [10] :

- Recherche pas à pas, méthode descendante (backward selection)
- Recherche pas à pas, méthode progressive (stepwise selection)
- Recherche pas à pas, méthode ascendante (forward selection) dont l'algorithme figure de procédure se présente dans la figure 3.3.

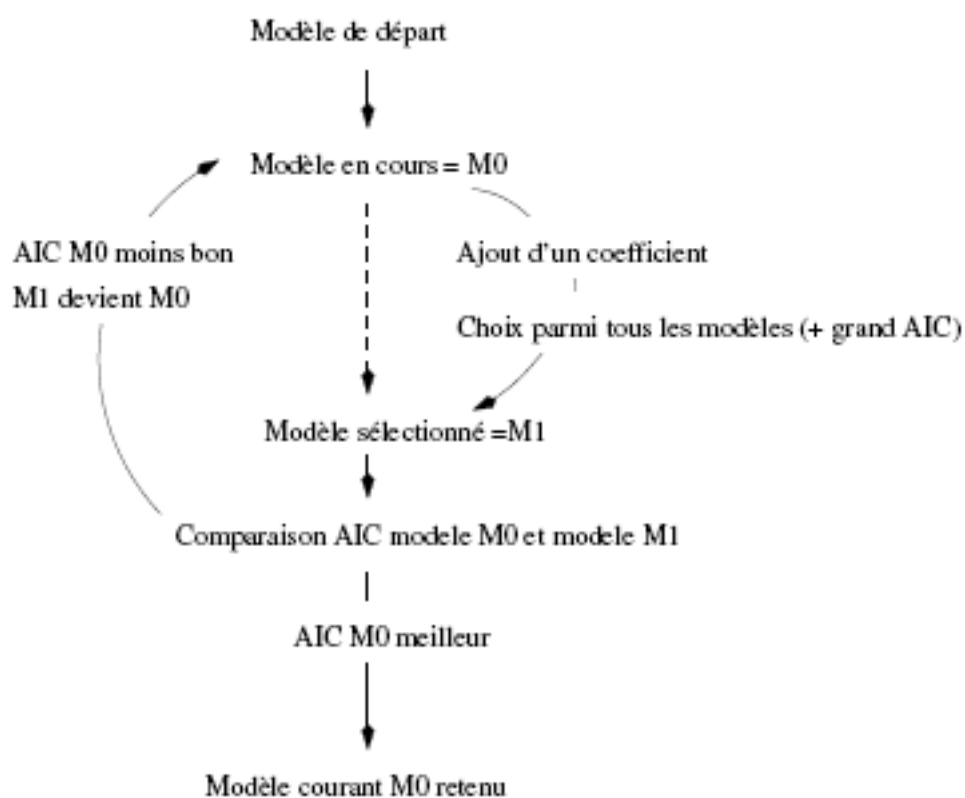


FIG. 3.3 – Technique ascendante utilisant l’AIC

# REPRESENTATION D'UN SCORING

Si l'analyse discriminante permet de prévoir correctement (ou non) la variable binaire  $Y$ , il est rare qu'il n'y ait pas d'erreur. Ainsi, dans une banque, chaque client est particulier et son comportement dépend du temps présent de son environnement, des circonstances extérieures qu'il rencontre etc.. Le modèle discriminant fournit, sur les données d'apprentissage, des erreurs que l'on peut résumer par un tableau de contingence.

Ainsi nous avons le nombre  $a$  de bien classés pour  $Y = 1$  et  $d$  le nombre de bien classés pour  $Y = 0$ . Les erreurs sont elles résumées par  $c$  et  $b$ . (Cf. figure 4.1)

Dans la représentation du scoring, le seuil  $n$ 'est pas fixé à priori comme il l'est dans l'analyse discriminante classique ou théorique.

## 4.1 Représentation théoriques

### 4.1.1 Présentation sous forme de densité

Toutes les variables explicatives  $X_1, \dots, X_p$  sont aléatoires et donc le scoring  $S(X)$  est une variable aléatoire à valeur dans  $\mathbb{R}$ . En théorie, nous pouvons tracer sa densité sachant que  $Y = 0$  et sa densité sachant que  $Y = 1$ . Le tracé des densités dans les 2 cas renseigne sur le pouvoir discriminant du scoring. Rappelons que pour un seuil donné, nous choisissons la valeur prévue par le modèle. Nous pouvons donc avoir des renseignements sur le pouvoir discriminant d'un scoring, pour un seuil donné, grâce aux erreurs de première et seconde espèce :

$\alpha = \mathbb{P}(S(X) > s | Y = 0)$ , prévoir 1 alors qu'en réalité  $Y=0$ .

$\beta = \mathbb{P}(S(X) \leq s | Y = 1)$ , prévoir 0 alors qu'en réalité  $Y=1$ .

		$\hat{Y}$		
		0	1	
	0	a	b	
	1	c	d	
$Y$				$n$

FIG. 4.1 – Tableau de contingence résumant la capacité d'ajustement de l'analyse discriminante

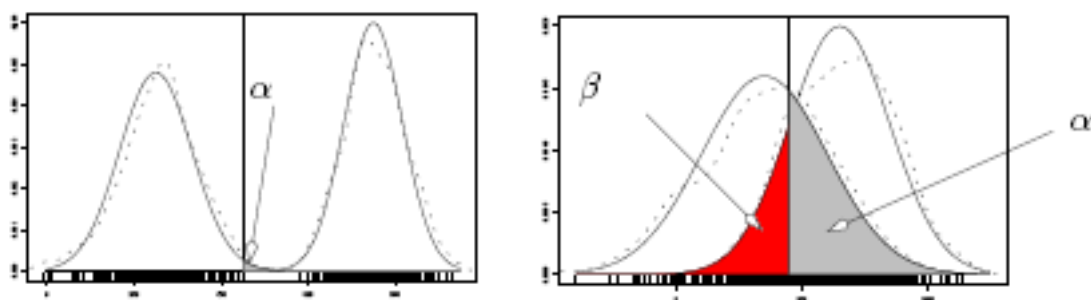


FIG. 4.2 – Densité de  $S(X)$  sachant  $Y=0$  et  $Y=1$  (traits pleins) et leurs estimations (traits pointillés). Le premier dessin figure un cas où des erreurs risquent d'apparaître. Les aires colorées correspondent au choix d'un seuil de  $s=450$  et aux erreurs  $\alpha$  et  $\beta$ .

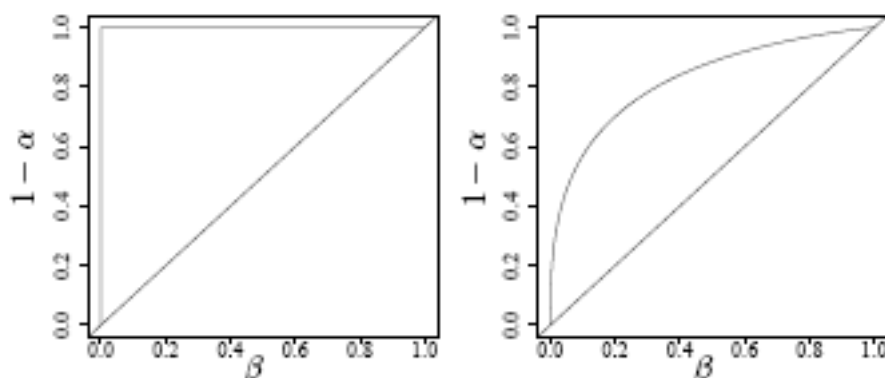


FIG. 4.3 – Courbe ROC d'évolution des erreurs en fonction du seuil, à gauche scoring parfait et à droite scoring avec des erreurs de classement.

Plus ces erreurs sont faibles, meilleur est le scoring (Cf figure 4.2)

Dans le cas de notre étude, l'erreur  $\alpha$  est ici ne pas prendre un nouveau client (prévoir 1) alors qu'il ne serait jamais à découvert ( $Y=0$ ) et l'erreur  $\beta$  sera de prendre un nouveau client alors qu'il sera à découvert.

La présentation sous forme de densité permet de montrer les dispersions potentielles des scorings sous les 2 hypothèses. Cependant il est difficile de voir l'influence d'un changement de seuil.

#### 4.1.2 Receiver Operating Curve (ROC)

La courbe ROC est une courbe paramétrée ayant en abscisse  $\beta(s)$  et en ordonnée  $(1 - \alpha(s))$ . Elle permet de synthétiser de manière plus simple l'évolution des erreurs en fonction de  $s$ . Pour  $s$  donné, plus  $\beta(s)$  est faible et  $(1 - \alpha(s))$  est fort, meilleur est le scoring. (figure 4.3)

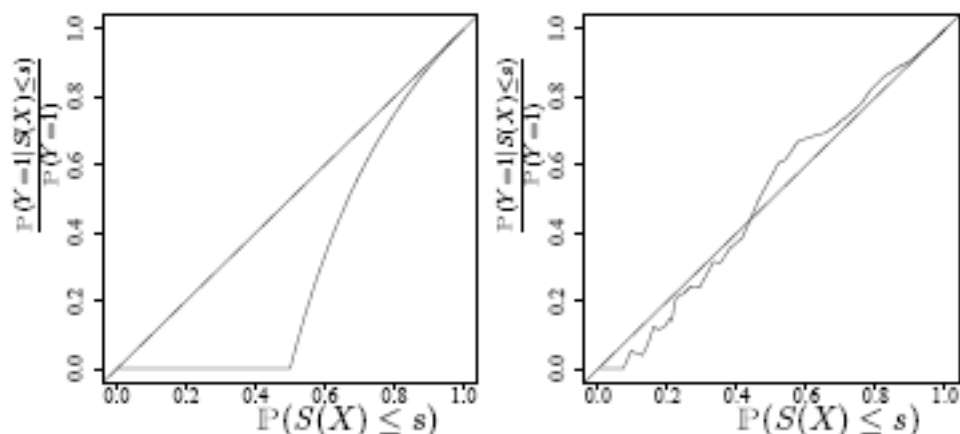


FIG. 4.4 – Courbe de performance, à gauche un scoring parfait et à droite un scoring avec des erreurs.

Ce type de courbe est très facile à lire mais ne dépend pas de la probabilité des  $Y = 0$  et des  $Y = 1$ . Pour les faire intervenir, les praticiens utilisent les courbes de performance ou les courbes de sélection.

### 4.1.3 Courbe de performance

Nous allons cette fois ci nous intéresser à la proportion des individus dont le scoring est inférieur au seuil  $s$ , ie  $\mathbb{P}(S(X) \leq s)$ . Cette grandeur sera l'abscisse de notre courbe paramétrée :  $x(s) = \mathbb{P}(S(X) \leq s)$ . Cette abscisse peut s'interpréter comme le pourcentage (ici la probabilité) de la population (les observations dont  $(Y = 0)$  et celles dont  $(Y = 1)$  dont le scoring est inférieur à  $s$ ).

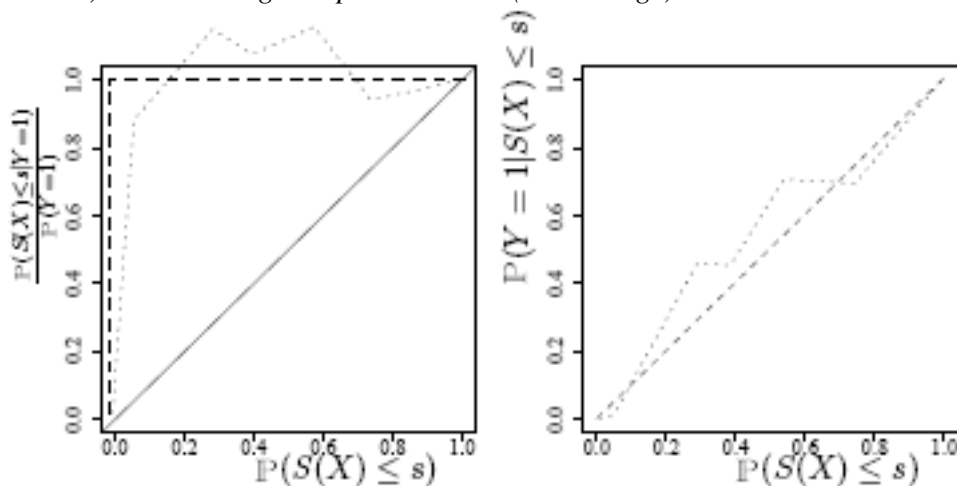
En ordonnée nous allons regarder la probabilité de  $(Y = 1)$  dans cette population des individus dont le scoring est inférieur à  $s$ . Afin d'avoir une abscisse qui soit égale à 1 quand  $s$  est maximum, nous choisissons  $y(s) = \mathbb{P}(Y = 1|S(X) \leq s)/\mathbb{P}(Y = 1)$ . (figure 4.4)

Remarquons qu'un score parfait "touche" l'axe des abscisses au seuil  $s_0$  qui sépare parfaitement les deux populations. Cependant un score qui touche l'axe au point  $(p_0, 0)$  (*avec*  $p_0 > 0$ ) n'est pas forcément un score parfait. Un score possède une courbe de performance qui touche l'axe des abscisses au seuil  $s_0$  en deçà duquel il n'existe plus de  $Y = 1$ . Au dessus de  $s_0$  peuvent cohabiter des 0 et des 1 (si il n'y a que des 1, le score est alors parfait).

#### Remarques

- Si le score est fantaisiste, alors la courbe peut remonter au dessus de la droite  $y = 1$ . En effet, si dans tous les scores élevés, au lieu de regrouper les individus  $\{Y = 1\}$ , on ne regroupe que des  $\{Y = 0\}$ , ce qui constitue un score tout à fait fantaisiste, alors la probabilité

FIG. 4.5 – Courbe de performance (à gauche) et de sélection (à droite) pour un scoring fantaisiste (pointillés) et un scoring indépendant de  $Y$  (tirets longs)



$P(Y = 1 | S(X) \leq s)$  va diminuer avec  $s$ . Ce cas est bien sûr un cas “limite” que l’on ne souhaite pas voir.

- Si le score est simplement indépendant de  $Y$ , ie que l’on cherche à discriminer  $Y$  mais aucune variable n’explique  $Y$ , alors nous avons  $\mathbb{P}(Y = 1 | S(X) \leq s) = P(Y = 1)$  et nous avons alors que le scoring est la droite  $y = 1$ . (Cf. figure 4.4)

#### 4.1.4 Courbe de sélection

Afin de contenir la courbe dans un demi carré, une autre courbe voisine existe. Elle est peut être plus naturelle à lire, dans le sens où elle ressemble à une courbe de concentration. L’abscisse reste toujours identique  $x(s) = \mathbb{P}(S(X) \leq s)$ , mais l’ordonnée est la probabilité d’avoir un score inférieur à  $s$  sachant que  $(Y = 1)$ , cette ordonnée peut se réécrire comme

$$\mathbb{P}(S(X) \leq s | Y = 1) = \frac{\mathbb{P}(S(X) \leq s \cap (Y = 1))}{\mathbb{P}(Y = 1)} = \frac{\mathbb{P}(Y = 1 | S(X) \leq s)}{\mathbb{P}(Y = 1)} \mathbb{P}(S(X) \leq s)$$

ie l’ordonnée de la courbe de performance multipliée par une probabilité (comprise entre 0 et 1).

- Si le score est un tant soit peu réaliste, il est en dessous de la première bissectrice.
- Si le score est parfait, il “touche” l’axe des abscisses à la valeur du paramètre  $s_0$  (figure 4.5).
- Un score possède une courbe de sélection qui touche l’axe des abscisses au seuil  $s_0$  en deçà duquel il n’existe plus de  $Y = 1$ . Ce n’est pas forcément un score parfait (au dessus de  $s_0$  peuvent cohabiter des 0 et des 1).
- Si le score est indépendant de  $Y$  (ie pas de pouvoir explicatif des variables  $X_1, \dots, X_p$  sur  $Y$ ), alors la courbe est la première bissectrice (figure 4.6).

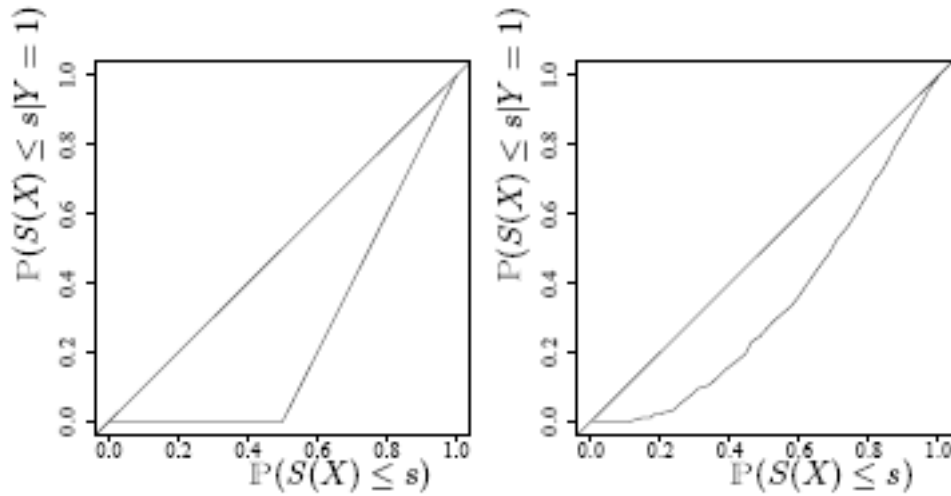


FIG. 4.6 – Courbe de sélection. A gauche scoring parfait et à droite scoring avec des erreurs de classements.

## 4.2 Estimations

Nous n'avons jamais les scores vrais mais des estimations que nous noterons  $\hat{S}(\cdot)$ . De même nous ne pouvons avoir la densité de  $\hat{S}(\cdot)$  sachant  $Y = 0$  ou de  $\hat{S}(\cdot)$  sachant  $Y = 1$ , mais nous pouvons l'estimer par un estimateur à noyau ou un histogramme. Enfin les probabilités  $\mathbb{P}(S(X) \leq s)$  sont simplement estimées par des pourcentages.

### 4.2.1 Qualité d'ajustement

Nous possédons  $n$  mesures des variables notées  $\{X_{i1}, \dots, X_{ip}, Y\}_{i=1}^n$ . A partir de ces mesures, nous estimons un scoring, par exemple par régression logistique avec choix de variables. Nous avons donc un scoring estimé  $\hat{S}(\cdot)$ .

1. La première étape est d'ordonner les observations selon les valeurs du scoring :

$$X_{(1)1}, \dots, X_{(1)p}, Y_{(1)}; \dots; X_{(n)1}, \dots, X_{(n)p}, Y_{(n)}.$$

2. Il faut choisir une grille  $s_1, s_2, \dots, s_K$  de valeurs de scoring. En général, on choisit les valeurs extrêmes du scoring,  $s_1 \approx \hat{S}(X_{(1)})$  et  $s_K \approx \hat{S}(X_{(n)})$ .

3. Pour chaque intervalle, on dénombre le nombre d'observations  $n_1, \dots, n_k$  qui sont dans l'intervalle  $]s_k; s_{k+1}]$  pour  $k \in \{1, \dots, K\}$ . De même, on dénombre le nombre d'observations qui possèdent une valeur de  $Y$  égale à 1 que nous noterons  $n_1^1, \dots, n_k^1$ .

4. Les totaux sont le nombre total d'observation  $n = \sum_{k=1}^K n_k$  et le nombre total d'observations avec  $Y_i = 1$ , noté  $n^1 = \sum_{k=1}^K n_k^1$ . L'estimation de  $\mathbb{P}(Y = 1)$  est alors  $\frac{n^1}{n}$ .

5. Les dénombrements sont ensuite cumulés donnant  $N_k = \sum_{l=1}^k n_l$  et  $N_k^1 = \sum_{l=1}^k n_l^1$ .

Les probabilités sont estimées par

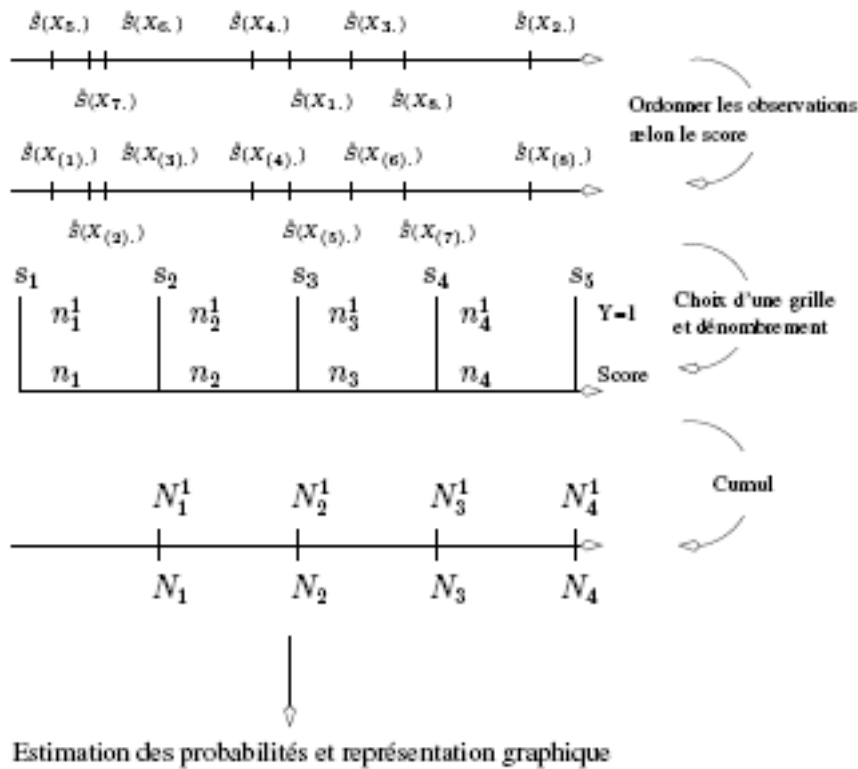


FIG. 4.7 – Etapes de construction d’une représentation d’un scoring.

$$\mathbb{P}(S(X) \leq s_{k+1}) \approx \frac{N_k}{n}$$

$$\mathbb{P}(Y = 1 | S(X) \leq s_{k+1}) \approx \frac{N_k^1}{N_k}$$

$$\mathbb{P}(S(X) \leq s | Y = 1) \approx \frac{N_k^1}{N_k} \frac{n}{n^1} \frac{N_k}{n}$$

6. Enfin un type de courbe est choisi et on représente alors  $K$  points  $(x(s_k), y(s_k))$ ,  $k \in \{1, \dots, K\}$  et on ajoute le point  $(1; 1)$  (Cf. figure 4.7)

**Remarques**

- Plus nous avons de points dans la grille de découpage (ie plus  $K$  est grand) plus nous aurons de points dans la représentation finale et plus elle aura l’aspect d’une courbe.
- Plus nous avons de points dans un intervalle  $]s_k, s_{k+1}]$  plus les estimations des probabilités seront précises.

- Comme nous utilisons les  $Y_i$  pour construire l'estimation du score  $\hat{S}(\cdot)$  et pour estimer les probabilités, nous n'avons que des courbes "optimistes", c'est *le problème de l'ajustement qui est toujours meilleur que la prévision...*

## 4.2.2 Qualité de prévision

Nous possédons  $n$  mesures des variables notées  $\{X_{i1}, \dots, X_{ip}, Y_i\}_{i=1}^n$ . A partir de ces mesures nous estimons un scoring, par exemple par une régression logistique avec choix de variables. Nous avons donc un score estimé  $\hat{S}(\cdot)$ .

Ensuite nous avons un jeu de données de validation qui n'a pas encore été utilisé. Nous avons les observations suivantes  $\{X_{i1}^*, \dots, X_{ip}^*, Y_i^*\}_{i=1}^n$ .

1. La première étape est d'ordonner les observations selon les valeurs du scoring.

$$X_{(1)1}^*, \dots, X_{(1)p}^*, Y_{(1)}^*; \dots; X_{(n)1}^*, \dots, X_{(n)p}^*, Y_{(n)}^*.$$

2. Il faut choisir une grille  $s_1, s_2, \dots, s_{K+1}$  de valeurs du scoring. En général, on choisit les valeurs extrêmes du scoring,  $s_1 \approx \hat{S}(X_{(1)}^*)$  et  $s_K \approx \hat{S}(X_{(n)}^*)$ .

3. Pour chaque intervalle on dénombre le nombre d'observations du jeu de validation  $n_1^*, \dots, n_k^*$  qui sont dans les intervalles  $[s_k; s_{k+1}]$  pour  $k \in \{1, \dots, K\}$ . De même, on dénombre le nombre le nombre d'observations du jeu de validation qui possèdent une valeur de  $Y^* = 1$  que nous noterons  $n_1^{1*}, \dots, n_k^{1*}$ .

4. Pour chacun des intervalles, on dénombre le nombre d'observations  $n^* = \sum_{k=1}^K n_k^*$  et le nombre total d'observations avec  $Y_i^* = 1$ , noté  $n^{1*} = \sum_{k=1}^K n_k^{1*}$ . L'estimation de  $\mathbb{P}(Y^* = 1) = \mathbb{P}(Y = 1)$  est alors  $\frac{n^{1*}}{n}$ .

5. Les dénombrements sont ensuite cumulés donnant  $N_k = \sum_{l=1}^k n_l^*$ .  $N_k^{1*} = \sum_{l=1}^k n_l^{1*}$ . Les probabilités sont alors estimées par

$$\mathbb{P}(S(X) \leq s_{k+1}) = \frac{N_k^*}{n^*}$$

$$\mathbb{P}(Y = 1 | S(X) \leq s_{k+1}) = \frac{N_k^{1*}}{N_k^*}$$

$$\mathbb{P}(S(X) \leq s | Y = 1) = \frac{N_k^{1*}}{N_k^*} \frac{n^*}{n} \frac{N_k^*}{n}$$

6. Enfin un type de courbe est choisi et on représente alors  $K$  points  $(x(s_{k+1}), y(s_{k+1}))$ ,  $k \in \{1, \dots, K\}$  et on ajoute le point  $(1; 1)$ .

# MODELISATION

## Introduction :

Dans ce chapitre, on s'intéresse au thème de la *modélisation du credit scoring* par un traitement de nos données en appliquant les méthodes paramétriques exposées au chapitre 3. Compte tenu de la variété des outils pouvant être mis en jeu, nous avons fait le choix d'insister sur la pratique des méthodes considérées ainsi que sur la compréhension des sorties proposées par le logiciel **R**<sup>1</sup>. **R** comme la plupart des logiciels en Statistique supposent implicitement les hypothèse de normalité, les distributions des estimateurs et donc les statistiques de test comme valides[5](*Data Mining 1, p 67*). Plus rigoureusement, ces résultats sont justifiés par les propriétés des distributions asymptotiques des estimateurs, propriétés qui ne sont pas développées dans ce mémoire. Nous allons estimer une fonction de scoring par régression logistique puis par discrimination linéaire. Nous appliquerons la validation croisée pour estimer l'erreur de prédiction.

*Warning* : Recodage des variables qualitatives.

Le cas où les variables explicatives sont qualitatives ont nécessité un traitement particulier. En effet comment faire une combinaison linéaire de variables qualitatives et quantitatives ? Cela n'a pas évidemment pas de sens. La solution retenue est basée sur ce qu'on appelle la *forme disjonctive d'une variable X à m modalités*[12]. On définit les m variables indicatrices des modalités ( $1_1, 1_2, \dots, 1_m$ ) telles que  $1_j$  vaut 1 si on appartient à la modalité j, 0 sinon. Seule une des indicatrices vaut 1, celle qui correspond à la modalité prise. Les m indicatrices sont donc équivalentes à la variable qualitative. Au cas où l'une de ces variables figurerait dans le scoring, celui-ci serait alors une combinaison des indicatrices. Les variables explicatives qualitatives qui interviennent dans le scoring sont donc les indicatrices de variables qualitatives. Cependant, une difficulté intervient : la matrice  $\Sigma$  n'est pas de plein rang et n'est donc pas inversible car la somme des indicatrices des modalités de chaque variable vaut 1. Cela signifie qu'il existe une infinité de solutions équivalentes pour estimer les coefficients : une des solutions couramment utilisée consiste alors à ne prendre que m-1 indicatrices pour chaque variable qualitative puisque la dernière est redondante.

## 5.1 Régression logistique

Nous allons diviser aléatoirement notre ensemble d'apprentissage  $\mathcal{A}$  en  $L = 5$  parties  $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4, \mathcal{A}_5$  distinctes, soit 26 individus pour chaque bloc. Pour chacune de ces parties, on va construire le

<sup>1</sup>dont la version 2.5.1 est téléchargeable sur le site [www.cran.at.r-project.org](http://www.cran.at.r-project.org)

TAB. 5.1 – Coefficients du modèle logistique.

Variabes retenues	Coef. Estimate	Std. Error	z value	p-value	IC <sub>95%</sub>
CREDIT	0.9209897	0.0479373	19.21	$< 10^{-3}$	0.8270343   1.014945
R1	-0.4030249	0.0263697	-15.28	$< 10^{-13}$	-0.4547085   -0.3513413
R3	0.7310701	0.0404733	18.06	$< 10^{-8}$	0.6517439   0.8103964
R6	-0.0876921	0.0289502	-3.03	0.002	-0.1444334   -0.0309509
R7	0.061002	0.0066751	9.14	$< 10^{-5}$	0.047919   0.0740851

prédicteur de Y et on va calculer l'erreur de prédiction par validation croisée en utilisant la fonction *cv.glm* qui se trouve dans la bibliothèque **boot** du logiciel R.

## Construction du modèle et estimation de l'erreur

La construction du modèle logistique se fait en utilisant la fonction *glm* du logiciel R, ensuite une sélection automatique des variables par minimisation de l'AIC est faite par la fonction *step* pour la sélection des variables pertinentes du modèle.

Le modèle construit a retenu les variables CREDIT, R1, R3, R6, R7 comme significatives.

L'estimation de l'erreur par une validation croisée en 5 blocs est 0.44957681.

Nous jugeons l'erreur un peu élevée. Etant donné la taille réduite de notre échantillon, on peut se permettre d'augmenter le nombre L de blocs, voire appliquer une estimation de l'erreur par leave-one-out. On obtient alors une estimation de l'erreur égale à 0.1886792, soit près 82% des observations originales classées correctement ce qui est acceptable car cette erreur est plus petite que celle obtenue par une segmentation des données en 5 blocs. La qualité d'ajustement peut être donnée par le taux de mal classés. En effet, le scoring estimé  $\hat{S}(X)$  étant déterminé, il est possible pour chaque individu de notre échantillon d'estimer son scoring<sup>2</sup>. Au delà du seuil  $s=0$ , l'estimation par le modèle est 1. Le scoring estimé par la régression logistique avec une erreur de 0.1886792 est :

$$\hat{S}(X) = 0.9209897 \text{CREDIT} - 0.4030249 \text{R1} + 0.7310701 \text{R3} - 0.0876921 \text{R6} + 0.061002 \text{R7}.$$

## Construction d'un scoring à partir de la régression logistique

Dans le modèle de scoring ci-dessus, l'erreur et la règle de décision finale sont obtenues avec un seuil théorique  $s=0$ . Maintenant, le seuil  $s$  est variable. On va estimer les scoring de tous les individus de l'échantillon. Il est d'usage de ramener ces scoring estimés entre 0 et 100 (en %), ce qui peut se faire en utilisant la transformation :

$$\text{scoring} = (\text{scoring} - \min(\text{scoring})) * 100 / (\max(\text{scoring}) - \min(\text{scoring})).$$

Ainsi pour les 130 individus de notre échantillon, on obtient une suite de 130 nombres réels inférieurs entre 0 et 100. Ensuite, on les ordonne par valeurs croissantes du scoring. Puis un

<sup>2</sup>Dans R, cela se fait en utilisant la fonction *predict*

TAB. 5.2 – Découpage du scoring en classes en classes d'effectifs approximativement égaux.

decoupage en classes	effectif par classe
[0 ; 2.59]	12
]2.59 ;80.3]	11
]80.3 ;84.6]	12
]84.6 ;86.3]	12
]86.3 ;87.6]	12
]87.6 ;90.1]	12
]90.1 ;91]	11
]91 ;92]	12
]92 ;93.2]	12
]93.2 ;95.2]	12
]95.2 ;100]	12

découpage est fait en 11 classes (par exemple) d'effectifs égaux comme expliqué dans l'algorithme de représentation d'un scoring du paragraphe 4.2.1 du chapitre précédent<sup>3</sup>.

Au niveau numérique, cela donne les résultats suivants :

Pour connaître le nombre d'observations pour lesquelles la variable à expliquer vaut 0 dans chaque classe, ou connaître le nombre d'observations pour lesquelles la variable à expliquer vaut 1 dans chaque classe, il suffit de faire le tableau de contingence entre les observations de Y.

D'après le tableau 5.3, nous jugeons donc la qualité d'un scoring basé sur la régression logistique. Nous voyons qu'en deçà d'une probabilité  $\mathbb{P}(S < s_0)$  d'environ 0.2, aucun client risqué ou mauvais client ( $Y=1$ ) n'est trouvé. Cette probabilité correspond à la fin de la seconde classe ie à un seuil  $s_0$  de 80.3.

A partir de ces dénombrements par classe, nous devons estimer les probabilités  $\mathbb{P}(S \leq s)$  et  $\mathbb{P}(S > s)$  grâce aux cumulés comme présenté au paragraphe 4.2.1. Puis nous pouvons estimer les probabilités suivantes  $\mathbb{P}(Y = 1|S \leq s)$ ,  $\mathbb{P}(Y = 0|S > s)$ . Enfin nous estimons  $\mathbb{P}(Y = 0)$  et  $\mathbb{P}(Y = 1)$ . Nous pouvons maintenant tracer la courbe ROC, la courbe de performance et la courbe sélection contenues dans la figure 5.1.

La figure 5.1 montre que les deux derniers graphiques touchent l'axe des abscisses au point(0.2 ;0), ce qui concorde bien avec les résultats obtenus dans le tableau 5.3 ie en deçà de la probabilité 0.2 aucun mauvais client n'est trouvé, ce qui correspond à un seuil de 80.3.

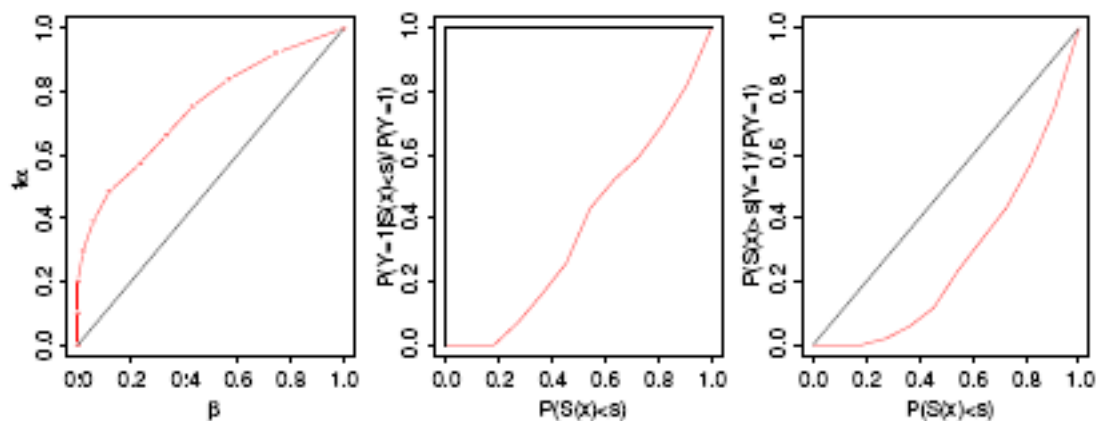
Si la politique de la banque consiste à ne pas prendre de risque, ce seuil est celui à sélectionner. Le seuil "pas de mauvais client"est 80.3 car la fin de la seconde classe correspond à une probabilité  $\mathbb{P}(S < s_0) \approx 0.2$ .

<sup>3</sup>Voir en ANNEXE pour les commandes du logiciel R qui fournissent les résultats de cet algorithme.

TAB. 5.3 – Découpage du scoring suivant les modalités de Y

Découpage	Y	effectif par classe
[0 ;2.59]	0	12
	1	0
]2.59 ;80.3]	0	11
	1	0
]80.3 ;84.6]	0	11
	1	1
]84.6 ;86.3]	0	11
	1	1
]86.3 ;87.6]	0	10
	1	2
]87.6 ;90.1]	0	9
	1	3
]90.1 ;91]	0	10
	1	1
]91 ;92]	0	10
	1	2
]92 ;93.2]	0	10
	1	2
]93.2 ;95.2]	0	8
	1	4
]95.2 ;100]	0	9
	1	3

FIG. 5.1 – Courbes ROC, de performance et de sélection liée à la regression logistique



TAB. 5.4 – Coefficients estimés du modèle linéaire

Variables	Coefficients	IC <sub>95%</sub>
Intercept	$1,367.10^{-2}$	$1,221465.10^{-2}   4.750795.10^{-2}$
CAP	0,1951406	-0,1697578   0,2243187
CREDIT	4,205561	-2,404566   6,006555
EFF	2,175276	1,320829   3,029722
DUREMB	-1,641947	-1,777509   -0,506386
GAR	-0,1855474	-0,2673828   -0,103712
AGE	-0,2934967	-0,3563134   -0,23068
EXP	-0,8431434	-0,9661575   -0,7201294
THT	$1.029068.10^{-2}$	-0,0455745   0,568789
INVEST	$8.923639.10^{-2}$	$-2.36664.10^{-5}   9.215487.10^{-10}$
CHDI	$-1,041.10^{-2}$	$-2.063303.10^{-2}   3.093465.10^{-2}$
MASA	0,7230202	-0,5667727   0,9223421
CAF	0,6557301	-0,5473226   0,7856098
VA	0,7451563	-0,6188638   0,8972216
CA	0,9554785	-0,0111789   1,125447
RN	0,4991191	-0,2016577   1,23536
R1	1,805436	-0,7369857   4,422879
R2	-1,075136	-1,9113738   -0,268324
R3	0,1880326	-0,1593717   0,218477
R4	2,175276	-1,320829   3,029722
R5	-1,641947	-1,777509   0,506386
R6	-0,1855474	-0,2673828   0,103712
R7	-0,0165511	-0,0964645   0,0633623

## 5.2 Analyse discriminante

### – Le cas quadratique :

Le logiciel **R** nous signale un message d'avertissement que la commande *qda* de la bibliothèque *MASS* ne peut être appliquée ceci étant dû à la taille réduite de nos données.

### – Le cas linéaire :

Les coefficients de l'analyse discriminante linéaire sont obtenues par la commande *lda* de la bibliothèque *MASS*, dans laquelle figure l'option *CV* de la validation croisée.

Le tableau 3.5 montre qu'en observant les intervalles de confiance<sup>4</sup>, les variables qu'on peut retenir sont : EFF, DUREMB, AGE, EXP, CHDI, R2 et l'intercept. Le scoring pour la discrimination linéaire est :

$$\hat{S}(X) = 2,175276\text{EFF} - 1,641947\text{DUREMB} - 0,2934967\text{AGE} - 0,8431434\text{EXP} + (-1,041.10^{-9})\text{CHDI} - 1,075136\text{R2} + 1,367.10^{-2}.$$

<sup>4</sup>Les intervalles de confiance qui contiennent 0 sont non significatifs pour les coefficients correspondants.

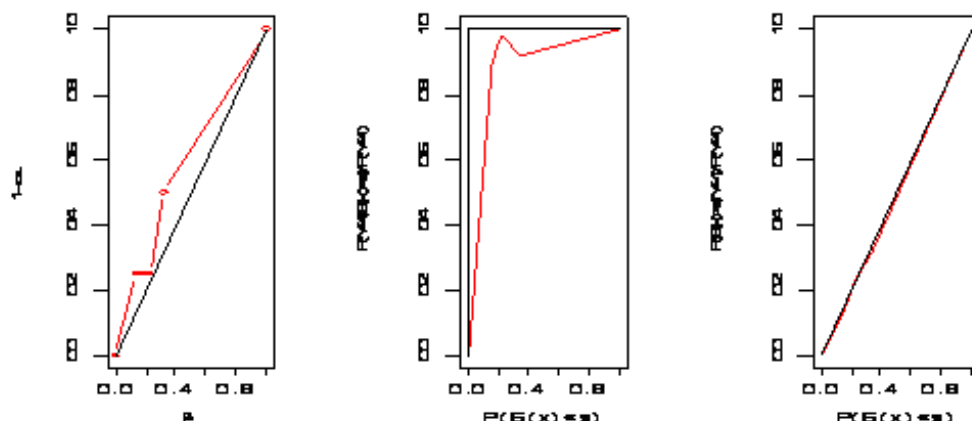


FIG. 5.2 – Courbes ROC, de performance et de sélection liée à la discrimination linéaire.

La transformation

$$scoring = (scoring - \min(scoring)) * 100 / (max(scoring) - \min(scoring))$$

nous permet d’avoir le scoring comme une probabilité(entre 0 et 100 %).

L’option *CV* de la commande *lda* nous permet d’avoir l’erreur de prédiction par leave-one-out. L’erreur de prédiction ainsi obtenu est de 0.62547, soit 0.37453 des observations originales classées correctement, ce qui n’est pas intéressant.

Le calcul du seuil par la formule

$$s = \log(\mathbb{P}(Y = 1)) - \log(\mathbb{P}(Y = 0)) + \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 - \frac{1}{2}\mu_0^T \Sigma^{-1} \mu_0$$

obtenu ci-dessus nous donne  $s = 0,4125833$ .

### Construction du scoring à partir de la discrimination linéaire

La procédure est identique que celle élaborée à la régression logistique. On décide de faire varier le seuil en espérant pouvant améliorer l’erreur de prédiction.

On constate d’après la figure 5.2 que la construction d’un scoring par discrimination linéaire est dégradée, l’allure des trois courbes montre que la déduction d’un seuil par représentation du scoring n’est pas possible. Il est pratiquement difficile d’en tirer des conclusions.

---

# CONCLUSIONS ET RECOMMANDATIONS

---

Le but de notre travail était de construire via des outils statistiques, une méthode de notation des emprunteurs à la First Bank.

Au terme de ce travail, il en ressort que des deux méthodes annoncées au chapitre 3 (à savoir la régression logistique et la discrimination linéaire), seule la régression logistique binaire nous donne un résultat satisfaisant. Le problème avec la discrimination linéaire, est qu'il est un peu difficile de se prononcer eu égard de la figure 5.2 qui montre 3 courbes dont les allures sont loin de fournir un indice sur le calcul graphique du seuil. On pourrait se contenter du seuil théorique  $s=0,4125833$ , mais l'erreur estimée ( $=0.62547$ ) par validation croisée (leave-one-out) est élevée. Cependant, l'étude faite avec la régression logistique nous fournit des résultats concrets au seuil  $s=0$ . On a obtenu une erreur de  $0.1886792$ , soit plus de 80% de bons classements. En faisant varier le seuil par construction du scoring, on se rend compte qu'avec un seuil  $=80.3$ , la discrimination est parfaite ie qu'on est sûr de l'affectation d'un emprunteur à l'une des deux classes (erreur nulle). Toutefois, il est à remarquer que ce seuil nous paraît un peu trop rigoureux, elle n'est applicable que si la banque ne veut courir aucun risque, ce qui n'est pas toujours une bonne politique bancaire car cette stratégie peut avoir un impact négatif sur la rentabilité de la banque<sup>5</sup>. Cette méthode de seuil variable permet de prendre des risques pour la sélection des clients dans la banque et ce selon les objectifs du moment.

Pour la First Bank, il convient d'adopter le scoring évalué par la régression logistique :

$$\hat{S}(X) = 0.9209897 \text{CREDIT} - 0.4030249R1 + 0.7310701R3 - 0.0876921 R6 + 0.061002 R7.$$

$\hat{S}(X)$  est une fonction des ratios de rentabilité  $R1=CA/VA$ ,  $R3=RN/\text{capitaux propres}$ , du ratio d'autonomie financière  $R6=SN/\text{total passif}$ , du ratio de solvabilité  $R7= \text{actif total}/\text{dettes}$  et du financement (CREDIT) que l'emprunteur a bénéficié auprès de la First Bank.

## Recommendations

Pour un nouvel emprunteur ou un client de crédit qui sollicite un soutien financier dans le cadre d'un projet d'investissement :

1. La First Bank recueille auprès du client les ratios  $R1$ ,  $R3$ ,  $R6$ ,  $R7$  et le crédit sollicité (CREDIT).
2. La banque calcule son scoring  $\hat{S}(X) = 0.9209897 \text{CREDIT} - 0.4030249R1 + 0.7310701R3 - 0.0876921 R6 + 0.061002 R7$  qui est une probabilité de défaut.

---

<sup>5</sup>Etant donné que l'octroi de crédits fait partie de l'activité principale des banques, chercher à annuler le risque de crédit influence sur les bénéfices de la banque.

En considérant le seuil théorique  $s=0$ , avec une erreur de 0.1886792 de mauvais classement, on a le choix d'affectation d'un nouvel emprunteur régis de manière suivante :

- Si  $\hat{S}(X) < 0$  alors  $\hat{Y} = 0$ , ie que l'emprunteur est non risqué, il est donc considéré comme bon.
- Si  $\hat{S}(X) > 0$  alors  $\hat{Y} = 1$ , ie que l'emprunteur est risqué, il est mauvais client.
- Si  $\hat{S}(X) = 0$  alors  $\hat{Y} = 0$  ou  $\hat{Y} = 1$ , peu importe.

Cependant, en supposant que *la politique économique de la First Bank* est de ne pas prendre de risque ie on est au seuil de  $s=80.3$ , alors on a la règle de décision suivante :

- Si  $\hat{S}(X) \leq 80.3$ , alors le client est considéré comme non risqué ie bon.
- Si  $\hat{S}(X) > 80.3$ , alors le client est peut-être risqué.

L'examen statistique de la situation économique et financière des entreprises (emprunteurs), en vue de la détection précoce des difficultés de la clientèle, est extrêmement fructueux. Par l'analyse multicritères, il permet la construction d'un scoring qui fournit une image synthétique du profil de l'entreprise emprunteuse. Celui-ci est, dans la très grande majorité des cas, révélateur de la santé de l'entreprise. Si un tel outil ne peut se substituer au jugement de l'expert, il peut contribuer à l'informer rapidement sur le niveau de risque de l'entreprise et concourir au diagnostic, grâce aux aides à l'interprétation qui l'accompagnent. L'analyste pourra alors se concentrer sur des aspects plus délicats et moins quantifiables de l'évaluation, en particulier les aspects qualitatifs. Ainsi, expertise et utilisation d'un scoring ne sont pas contradictoires ; au contraire, elles se complètent et permettent d'affiner l'analyse du *risque de crédit*. De même, lorsque plusieurs outils d'évaluation du risque sont disponibles, généralement fondés sur des systèmes d'information différents, il est très fructueux de les examiner tous. En effet, les renseignements qu'ils apportent relativisent les points de vue, accroissent la fiabilité de la prévision et renforcent le diagnostic.

---

# ANNEXES

---

## Codes R des fonctions utilisées

### Chapitre 5

```
##### importation du fichier base
tab=read.table("donnee.txt",header=TRUE,sep="\t",dec=",")
#####on rend visible le fichier des donnees
attach(tab)
##### résumé des données
summary(tab)
##### modele simple
modelsimple =glm(Y~1,data=donnee,family=binomial)
summary(modelsimple)
##### modele complet
modelcomplet=glm(Y~.,data=donnee,family=binomial)
summary(modelcomplet)
##### Sélection ascendante des variables du modèle final
Modelfinal=step(modelcomplet,scope=list(upper=formula("Y~(FORJU+CAP +NACTI+SISO+CREDIT
+INVEST+CHDI+MASA+CAF+VA+CA+RN+R1+R2+R3+R4+R5+R6+R7)", direction="forward"),
lower=formula("Y~1")))
##### ajustement
table(modelfinal$fitted.values>0,tabapp$Y= =0)
##### calcul du scoring de chaque individu
score=predict(modelfinal)
##### on ramène le scoring entre 0 et 100.
score= (score - min(score))*100/(max(score)-min(score))
##### on ordonne les valeurs du scoring par valeurs croissantes
ordre = order(score)
y.ordonne = donnee[ordre,"Y"]
score.ordonne = score[ordre]
decoupage=quantile(score.ordonne,seq(0,1,length=12))
score.decoupage =cut(score.ordonne,breaks=decoupage)
table.score=table(score.decoupage)
table.y = table(y.ordonne,score.decoupage)
```

```

## Proba (S()<s )
px.v= cumsum(table.score)/sum(table.score)
## Proba( S()>s )
px2.v = rev( cumsum(rev(table.score))/sum(table.score) )
## Proba (Y=1 | S()<s )
py.v =cumsum(table.y.[2,])/cumsum(apply(table.y,2,sum))
## Proba (Y=0 | S()>s )
py2.v =rev(cumsum(rev(table.y[1,])/cumsum(rev(apply(table.y,2,sum)))) )
## Proba (Y=0 )
p0.v = table(y.ordonne)[1]/sum(table(y.ordonne))
## Proba (Y=1)
p1.v = table(y.ordonne)[2]/sum(table(y.ordonne))
### construction des courbes ROC, de sélection et de performance
par(mfrow=c(1,3))
### courbe ROC
plot(c(0,py.v/p1.v*px.v),1-c(py2.v*px2.v/p0.v,0),type="b",xlab=expression(beta), ylab=expression(1-
alpha),col="red")
segments(0,0,1,1)
##### courbe de sélection
plot(c(0,px.v),c(0,py.v/p1.v),type="l",xlab="P(S(x)<s)", ylab="P(Y=1|S(x)<s)/P(Y=1)", xlim=c(0,1),ylim=
segments(c(0,0),c(0,1),c(0,1),c(1,1))
##### courbe de performance
plot(c(0,px.v),c(0,py.v/p1.v*px.v),type="l",xlab="P(S(x)<s)", ylab="P(S(x)>s|Y=1)/P(Y=1)",
xlim=c(0,1),ylim=c(0,1),col="red")
segments(0,0,1,1)

```

---

---

# REFERENCES

---

- [1] NDONG NGUEMA, *Cours de Data Mining*(2007), Ecole Nationale Supérieure Polytechnique de Yaoundé.
- [2] NDONG NGUEMA, *Cours de Statistique Mathématique*(2007), Ecole Nationale Supérieure Polytechnique de Yaoundé.
- [3] *RAPPORT ANNUEL 2006*, Afriland First Bank.
- [4] Collett D. (2003). *Modelling binary data*. Chapman & Hall/CRC, 2 ed.
- [5] Philippe BESSE *Data Mining 1*, Laboratoire de Statistique et Probabilités, Université Paul Sabatier, Toulouse III.
- [6] *Glossaire 2007* de la banque centrale tunisienne.
- [7] Duffie (D.), Singleton (K. J.). *Credit Risk*. Princeton University Press (2003)
- [8] Schervish M.J. (1995). *Theory of statistics*. Springer-Verlag, New-York.
- [9] Mallows C.L. (1986). *Augmented partial residuals*. *Technometrics*, 28, 313–319.
- [10] Schwarz G. (1978). *Estimating the dimension of a model*. *Annals of statistics*, 6, 461–464.
- [11] Christophe J. GODLEWSKI, « *Rôle de la Nature de l'Information dans l'Intermédiation Bancaire* », LaRGE, Avril 2004
- [12] Gilbert SAPORTA, « *La Notation Statistique des Emprunteurs (ou scoring)* », CREM, mars 2003
- [13] Gregory N. MANKIW, « *Macroéconomie* », Nouveaux Horizons, De Boeck, 3ème Édition, Bruxelles, 2003.
- [14] Mark SCHREINER, « *Les Vertus et Faiblesses de l'Évaluation Statistique en Micro*

*finance*», Septembre 2003.

[15]Cohen, E. (1999), *Dictionnaire de Gestion*, Al Manar, Dictionnaires Repères.

[16]Patrick VILLIEU, « *Macroéconomie : l'Investissement* », (Repères, 276), La Découverte, Paris, 2000 .

[17] [www.afrilandfirstbank.com](http://www.afrilandfirstbank.com)

[18] [www.microfinance.com](http://www.microfinance.com)

---

# Table des matières

---

<b>1</b>	<b>PRESENTATION DE LA STRUCTURE D'ACCUEIL ET CONCEPT DE RISQUE BANCAIRE</b>	<b>15</b>
1.1	Présentation de la structure d'accueil . . . . .	15
1.1.1	Afriland First Bank . . . . .	15
1.1.2	La Direction des Etudes et du Corporate Banking(DECB) . . . . .	15
1.1.3	Contexte de l'étude . . . . .	17
1.2	Concept de risque bancaire : . . . . .	19
1.2.1	Le risque de crédit : veiller aux défauts de paiement[13] . . . . .	20
1.2.2	Le risque de crédit : niveaux de gestion[7] . . . . .	20
<b>2</b>	<b>DESCRIPTION STATISTIQUE DE LA BASE DE DONNEES</b>	<b>21</b>
2.1	Méthodologie de collecte des données . . . . .	21
2.2	Description des variables d'analyse . . . . .	22
<b>3</b>	<b>DEUX METHODES DE DISCRIMINATION POUR LE CREDIT SCORING</b>	<b>33</b>
3.1	Le modèle probabiliste de prédiction . . . . .	33
3.2	Analyse discriminante linéaire et quadratique . . . . .	35
3.2.1	Estimation des paramètres . . . . .	36
3.2.2	Calcul du seuil théorique $s$ . . . . .	37
3.3	Analyse discriminante logistique . . . . .	38
3.3.1	Définition . . . . .	38
3.3.2	Lien avec les GLM : . . . . .	39
3.3.3	Estimation des paramètres . . . . .	40
3.3.4	Précision des estimations : . . . . .	41
3.3.5	La qualité du modèle . . . . .	42
3.3.5.1	Un outil spécifique : la déviance . . . . .	42
3.3.5.2	Evaluation des performances . . . . .	45
3.3.6	Sélection automatique . . . . .	47
<b>4</b>	<b>REPRESENTATION D'UN SCORING</b>	<b>49</b>
4.1	Représentation théoriques . . . . .	49
4.1.1	Présentation sous forme de densité . . . . .	49
4.1.2	Receiver Operating Curve (ROC) . . . . .	50

---

4.1.3	Courbe de performance . . . . .	51
4.1.4	Courbe de sélection . . . . .	52
4.2	Estimations . . . . .	53
4.2.1	Qualité d'ajustement . . . . .	53
4.2.2	Qualité de prévision . . . . .	55
<b>5</b>	<b>MODELISATION</b>	<b>56</b>
5.1	Régression logistique . . . . .	56
5.2	Analyse discriminante . . . . .	60

---

# Table des figures

---

2.1	<i>Répartition des dossiers selon la forme juridique.</i> . . . . .	24
2.2	<i>Diagramme en bâtons de NACTI.</i> . . . . .	26
2.3	<i>Répartition des dossiers par lieu d'implantation des clients (%)</i> . . . . .	26
2.4	<i>histogramme de la variable CREDIT</i> . . . . .	28
2.5	<i>Histogramme et Boxplot de GAR</i> . . . . .	29
2.6	<i>Réprésentation des densités des ratios R1 et R5.</i> . . . . .	30
3.1	<i>Test de déviance, la droite verticale représente le seuil de rejet <math>D_c = q_{1-\alpha}(n - p)</math>.</i> . . . . .	43
3.2	<i>Procédure d'apprentissage/validation</i> . . . . .	45
3.3	<i>Technique ascendante utilisant l'AIC</i> . . . . .	48
4.1	<i>Tableau de contingence résumant la capacité d'ajustement de l'analyse discriminante</i> . . . . .	49
4.2	<i>Densité de <math>S(X)</math> sachant <math>Y=0</math> et <math>Y=1</math> (traits pleins) et leurs estimations (traits pointillés). Le premier dessin figure un cas où des erreurs risquent d'apparaître. Les aires colorées correspondent au choix d'un seuil de <math>s=450</math> et aux erreurs <math>\alpha</math> et <math>\beta</math>.</i> . . . . .	50
4.3	<i>Courbe ROC d'évolution des erreurs en fonction du seuil, à gauche scoring parfait et à droite scoring avec des erreurs de classement.</i> . . . . .	50
4.4	<i>Courbe de performance, à gauche un scoring parfait et à droite un scoring avec des erreurs.</i> . . . . .	51
4.5	<i>Courbe de performance (à gauche) et de sélection (à droite) pour un scoring fantaisiste (pointillés) et un scoring indépendant de <math>Y</math> (tirets longs)</i> . . . . .	52
4.6	<i>Courbe de sélection. A gauche scoring parfait et à droite scoring avec des erreurs de classements.</i> . . . . .	53
4.7	<i>Etapes de construction d'une représentation d'un scoring.</i> . . . . .	54
5.1	<i>Courbes ROC, de performance et de sélection liée à la regression logistique</i> . .	59
5.2	<i>Courbes ROC, de performance et de sélection liée à la discrimination linéaire.</i>	61

---

# Liste des tableaux

---

1.1	<b>Fiche d'identification de Afriland First Bank</b>	16
2.1	<b>codage des variables d'étude</b>	23
2.2	<i>Répartition des dossiers de crédit suivant la forme juridique des entreprises.</i>	24
2.3	<i>Répartition des dossiers par activités principales des entreprises.</i>	25
2.4	<b>Répartition des dossiers de crédit par les montants des besoins exprimés.</b>	27
2.5	<b>Répartition des dossiers de crédit par les montants des garanties.</b>	28
2.6	<i>Résumé et test de normalité des ratios</i>	29
2.7	<i>Répartition des dossiers suivant les taux de crédit et les échéances de remboursement.</i>	31
5.1	<i>Coefficients du modèle logistique.</i>	57
5.2	<i>Découpage du scoring en classes en classes d'effectifs approximativement égaux.</i>	58
5.3	<i>Découpage du scoring suivant les modalités de Y</i>	59
5.4	<i>Coefficients estimés du modèle linéaire</i>	60