République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Abderrahmane Mira de Béjaia
Faculté des Sciences Exactes
Département de Recherche Opérationnelle

Mémoire Présenté

Pour l'obtention du Diplôme de Master

En Recherche Opérationnelle

Option :Mathématique Financière

Par : HAMMOU Hammou

Évaluation de risque de non-remboursement de crédit bancaire.

Soutenue le : 16/07/2022 devant le jury :

D^{r}Y. Ziane	M.C. classe/ A	Président	à l'UAMB
D^{r}L. Asli	M.C. classe/ A	Encadreur	à l'UAMB
D^{r} M. Soufit	M.C. classe/ B	Examinateur	à l'UAMB
D^{r}S. Ziani	M.C. classe/ B	Examinateur	à l'UAMB

Année Universitaire 2021 – 2022

Remerciements

Nous remercions Dieu de nous avoir donné le courage et la patience afin de terminer ce travail.

Je tiens à exprimer toute ma gratitude envers mon encadreur Dr : **Asli Larbi** pour son soutien et son aide. En m'écoutant patiemment, et en discutant en maintes fois de la nature et l'avancement de mon travail, il ma permis de synthétiser, comprendre et expliquer un grand nombre de questions. Ses conseils et sa gentillesse ma apporté un précieux soutien, qu'il soit chaleureusement remercie ici.

Mes remerciements sont aussi adressés aux membres du jury qui m'ont fait l'honneur d'accepter de juger mon travail.

Ma gratitude va également à Dr : **Y.Ziane** pour avoir accepté de présider le jury de la soutenance.

Mes remerciements s'adressent également à Dr :**M.Soufit** et Dr :**S.Ziani** pour l'honneur qu'ils ma fait en acceptant d'examiner ce mémoire.

Je remercié aussi tous les enseignants du département de Recherche Opérationnelle qui nous ont permis d'améliorer notre formation.

Merci à mes parents et à mes frères et soeurs pour m'avoir inculqué le goût d'apprendre, de m'avoir enseigné à penser et de m'avoir encouragé sans cesse pour aller plus loin.

Enfin, je n'oubliers pas de remercier ceux qui m'ont aidé d'une manière où d'une autre d'élaborer ce travail.

H. Hammou

Dédicaces

Je dédie ce modeste travail

Á ma très chère honorable et aimable mère qui représentes pour moi le symbole de la bonté par excellence.

Á mon très cher père. Rien au monde ne vaut les efforts fournis jour et nuit pour mon éducation.

Á mes frères Farid , Yasser , Yousef et Djamel.

Á mes sœurs Lyakout , Salima , Tounes et Rahima.

Á toute ma famille, à tous mes amis et aux étudiants de notre promotion.

H. Hammou

Table des matières

Liste des tableaux

Table des figures

Introduction générale

La banque est une entreprise dans le commerce de l'argent, d'une part elle reçoit les dépôts de public, collecte des épargnes, d'une autre part, elle joue le rôle d'une l'intermédiaire financier entre le déposant de l'argents et le demandeur de prêt financier.

La crise financière qui a secoué le monde, s'est exprimée par la faillite de grandes banques, elle a provoqué une remise en cause des modèles de gestion des risques bancaires, notamment le risque crédit. Ce risque doit être géré dorénavant par des méthodes fiables en mesure de réduire son impact négatif sur la rentabilité des banques.

Le système bancaire algériens utilise des méthodes classiques pour faire face aux risques crédits. Parmi ces méthodes, le diagnostic financier et la prise de garantie occupent sans doute une place centrale. Cette situation engendre des effets néfastes sur le gonflement des impayés ce qui peut mettre en cause la survie même de la banque. Or, il existe actuellement des méthodes sophistiquées destinées à la gestion du risque de crédits dont l'analyse discriminante et la régression logistique. Ces méthodes correspondent toute à une méthode d'analyse financière qui tente à synthétiser un ensemble de ratios pour parvenir à un indicateur unique permettant de distinguer d'avance les bons clients des clients défaillants.

Le credit-scoring a vu le jour suite aux travaux pionniers de BEAVER en 1966 et ALTMAN en 1968, et sur la base de ces recherches, que le crédit scoring s'est développé partout dans le monde, et a évolué au cours de ces 20 dernières années. On notera surtout l'évolution de la fonction Z de ALTMAN en 1968 qui devient la fonction ZETA après les améliorations de ALTMAN, HALDEMAN et NARAYANAN en 1977[10].

En France, dans la lignée des travaux d'ALTMAN en 1968, la Banque de France a développé plusieurs fonctions score. L'ancienne fonction était connue sous le nom de fonction Z mais elle a été réactualisée en plusieurs fonctions différenciées par secteur d'activité et disponibles à partir du module 38 de FIBEN (BARDOS, 2001).La conception d'un modèle de scoring suit une procédure relativement standard[10].

Elle se fonde sur l'observation ex-post du devenir des entreprises (à partir de données historiques généralement comptables et financières), dont on sait avec certitude si elles ont été défaillantes ou non. Le but est de sélectionner les variables les plus discriminantes individuellement, puis de construire un modèle statistique établissant une relation dicho-

tomique entre ces variables et le fait d'avoir connu la faillite ou non.

Ce mémoire est constitué d'une introduction générale et de trois chapitres

Dans le premier chapitre nous allons donner les définitions des principaux notions de la finance : banque, crédit et risque de crédit.

Dans le deuxième chapitre les déférente méthode de résolution des problèmes associés.

Le dernier chapitre sera consacré pour l'application et l'implémentation des résultats du travail.

Enfin, on termine ce travail par une conclusion générale.

1

Notions et préliminaires sur les banques, crédits et risques banquaire

Introduction

Ce premier chapitre sera consacré à la présentation d'un certain nombre d'éléments et notions fondamentales, la définition de la notion de banque, son rôle et son activité ainsi que les crédits. Dans le cadre de son activité, la banque est sujette à différents risques, ceux-ci seront brièvement définis et finalement, on présentera les approches pour l'identification et le contrôle des risques.

1.1 Notion générale sur les banques

1.1.1 Définition d'une banque

Une banque est une entreprise particulière qui s'occupe des dépôts d'argent et des moyens de paiement. Au sens juridique, c'est une institution financière qui dépend du Code monétaire et financier. Une banque a pour fonction de proposer des services financiers, recevoir des dépôts d'argent, collecter l'épargne, gérer les moyens de paiement, accorder des prêts. Une banque fonctionne généralement sous forme de plusieurs agences constituant le réseau de la banque [6].

1.1.2 Typologie des banques

Il existe différents types de banques, selon leur statut juridique [11] ;

1. **les banques coopératives**.
   Les banques coopératives sont des sociétés qui appartiennent à leurs clients ou sociétaires, c'est-à-dire qu'ils en sont les actionnaires. Ceux-ci peuvent être des personnes morales ou physiques. Une banque coopérative est une entreprise dont la propriété est collective et dans laquelle le pouvoir est démocratique. Les clients sociétaires ont la caractéristique d'être à la fois associés et usagers, à la fois propriétaires et clients de leur banque.
2. **les banques commerciales**
   Les banques commerciales sont des sociétés constituées d'un capital détenu par des actionnaires extérieurs à leur clientèle, par opposition aux banques coopératives. Les banques commerciales ont pour but de réaliser des bénéfices commerciaux et peuvent être cotées en bourse. Une banque commerciale est une entreprise privée qui collecte de l'argent par les dépôts et sur le marché monétaire, et les redistribuais sous forme de liquidité ou de crédit. Elle propose ainsi différents produits financiers et non financiers :
   Crédits (crédit personnel ; prêt immobilier), placement, épargne et assurances (assurance-vie ; assurances automobile et habitation).
3. **les banques publiques**
   Une banque publique est une société bancaire dont l'État ou des acteurs publics sont propriétaires. Elle se distingue d'une banque commerciale par son type d'actionnariat, mais aussi souvent par certaines missions qui lui sont confiées par la puissance publique

1.1.3 Servies fournies par les banques

La banque met à la disposition de ses clients divers outils pour qu'ils puissent gérer leur argent, c'est ce qu'on appelle les services bancaires. Ils comprennent notamment les moyens de paiement.

La banque est tenue de fournir a ses clients des services de base, et parmi eux [9] ;

- l'ouverture, la tenue et la clôture d'un compte ;
- la délivrance de relevés d'identité bancaire ;
- le changement d'adresse ;
- l'envoi du relevé de compte ;
- le paiement par prélèvement ;
- une carte (de paiement ou de retrait) ;
- l'encaissement des chèques et des virements ;
- la domiciliation de virements bancaires ou postaux ;
- les dépôts et les retraits d'espèces au guichet et au distributeur automatique.

1.1.4 Ressources de la banque

Il existe deux grand catégories de ressources : les ressources clientèles et les ressources hors clientèle[11].

1. **Ressources de la clientèle**
   Ces ressources sont principalement formées par :
   - Les dépôts (à vue et à terme) sont des liquidités placées en banque par les clients. Les dépôts à vue peuvent être restitués à la demande les dépôts à terme ne peuvent être restitués avant délais ;
   - Les bons de caisse (nominatifs ou anonymes) sont des titres émis par la banque contre un placement de fond à rembourser à une échéance définie avec payement d'un intérêt ;
   - Les bons d'épargne sont des titres émis par la banque pour la collecte de ressources, ils sont payés en plus des intérêts produits à leur épargne.
2. **Ressources hors clientèle**
   Ces ressources sont formées principalement par le marché interbancaire, les avances de la banque centrale ex.. .

1.1.5 Rôle des banque

La banque joue un rôle d'intermédiaire entre les détenteurs et les demandeurs de capitaux. Son activité principale consiste à collecter les capitaux disponibles pour son propre compte et les utiliser sous sa responsabilité à des opérations de crédit. Elle peut également effectuer d'autres opérations de banque : les services bancaires de paiement, les opérations de change [11].

1.1.6 Classification des banque

En général, on distingue trois catégories essentielles de banque : Les banques de dépôts, les banques d'investissement et les banques d'affaires [11].

1. **Les banques de dépôts** : L'activité principale de ce type de banque consiste à effectuer des opérations de crédits et à recueillir les dépôts de fonds à vue et à terme. Au quotidien, elles gèrent les comptes des particuliers et des entreprises. Elles sont garantes de la sécurité des transactions financières.

2. **Les banques d'investissement** : Les banques d'investissement sont des banques dont l'activité consiste à accorder des crédits dont la durée est supérieure à deux ans.

3. **Les banques d'affaires** : En plus de l'octroi des crédits, Les banques d'affaires participent à la prise et la gestion de participations dans des affaires existantes ou en

formation. Les opérations de financement engagées par ce type de banques immobilisent des capitaux pour une longue période.

1.2 Notion générale sur les crédit bancaire

1.2.1 Définition de crédit bancaire

Le crédit est une mise à disposition d'une ressource (une somme d'argent ou un bien) sous forme de prêt (opération par laquelle des fonds sont remis par un créancier à un bénéficiaire, moyennant en générale le paiement par ce dernier d'un intérêt versé au créancier, et assorti de l'engagement de remboursement de la somme prêtée), consentie par un créancier (une personne ou une organisation) à un débiteur. Lorsque la ressource est un bien on parle de crédit fournisseur ; lorsque c'est une somme d'argent accordée par une banque on parle de crédit bancaire. Le crédit est fortement lié à la notion de confiance il repose sur la confiance que le créancier accorde au débiteur[18].

1.2.2 Objectif de crédit bancaire

Le domaine du crédit est extrêmement vaste. Il s'étale dans le temps et l'espace, s'étend à toutes sortes d'activités et répond à de multiples besoins économiques, Il peut donc avoir pour objectif aussi bien le financement des investissements des entreprises et des particuliers que les besoins temporaires de trésorerie. Il permet de faire face à tous les décalages, entre recettes et les dépenses quelle que soit l'origine des unes et des autres[6].

1.2.3 Classification des crédit

Différents critères peuvent être pris en compte pour classifier les crédits, les principaux étant la durée (critère le plus utilisé), le bénéficiaire et la destination [2] :

- **la durée** : elle va dépendre du type d'opération pour laquelle le crédit est utilisé. On relève :
  1. le crédit à très court terme (au jour le jour) qui est utilisé par les banques pour ajuster quotidiennement leur trésorerie.
  2. le crédit à court terme, de 3 mois à deux ans, utilisé par les ménages et les entreprises.
  3. le crédit à moyen terme, entre deux et sept ans.
  4. le crédit à long terme, plus de sept ans, concernant les ménages, les entreprises et les collectivités locales (communes, département...).
- **les bénéficiaires** : ce sont essentiellement les ménages, les entreprises et les administrations publiques.

- **la destination** : il s'agit de l'utilisation ou le domaine de l'application .

1.2.4 Types d'amortissement d'un crédit

L'amortissement ou le remboursement d'un crédit bancaire peut prendre plusieurs formes. En général, les échéances de paiement sont mensuelles. On recense plusieurs types de remboursements [2] :

- Le remboursement à mensualités constantes, ou remboursement progressif du capital (les mensualités sont toujours les mêmes, mais au début elles comportent une part majoritaire d'intérêts, et à la fin une part majoritaire de capital) ;

- Le remboursement à mensualités dégressives, ou remboursement constant du capital (tous les mois, le même montant de capital est remboursé, ce qui fait que le montant mensuel des intérêts associés décroît dans le temps) ;

- Le remboursement infime (on ne paye tous les mois que les intérêts, et on rembourse la totalité du capital au terme du crédit).

1.2.5 Rôle des crédit

Le crédit est un moteur de l'économie, c'est un facteur important du développement des entreprises. Il permet de faire face à tout les décalages entre les recettes et les dépenses quelques soit leurs origine. Le crédit joue un rôle considérables dans les économies modernes car ils [2] :

- Permet d'accroitre la qualité du production ;
- Met à la disposition d'une personne un pouvoir d'achat immédiat, ce qui facilite les échanges entre les entreprises et les particuliers ;
- Permet d'assurer la continuité dans un processus de production et de commercialisation ;
- Est un moyen de création monétaire.

1.2.6 Type de crédit bancaire

Les types de crédit sont nombreux, ce qui offre à l'emprunteur plusieurs possibilités de choisir la forme qui lui convient, on distingue plusieurs formes de crédit à s'avoir [2] :

1. Le crédit d'exploitation : Ce type de crédit est destiné à rééquilibrer l'équation de trésorerie, c'est-à-dire qu'un déficit de trésorerie s'il existe, peut être comble par des crédits.

2. Les crédits d'investissements : Les crédits d'investissement sont destinés à financer la partie haute du bilan, entre autres les immobilisations, outil de travail de l'entreprise. Le remboursement de ces crédits ne peut être assuré que par l'enjeu des bénéfices.

Les crédits d'investissement se décomposent en crédits à moyen et à long terme. Il existe une autre forme de crédits permettant à l'entreprise d ?acquérir des investissements, c'est le crédit-bail appelé aussi leasing.

3. Les crédits aux particuliers : Il s'agit de différents types de crédit que les particuliers utilisent pour financer des besoins très variés, on repère notamment plusieurs pratiques, les plus important sont : le crédit à la consommation et le crédit immobilier.
   – Le crédit à la consommation : C'est la catégorie de crédit accordée à des particuliers par des établissements bancaires pour financer les achats de biens et services, comme les grosses dépenses en biens d'équipements (automobile, équipement de maison). Il se caractérise par des montants de prêt plus faible, une durée de remboursement relativement courte [17].
   – le crédit immobilier : désigne d'une manière générale un emprunt destiné à financer tout ou une partie de l'acquisition d ?un bien immobilier, de l'opération de construction, ou des travaux sur le bien. Ce genre de crédit est destiné au particulier pour l'achat, la rénovation, ou pour faire des travaux de construction.

1.3 Notion générale sur les risques de crédit bancaire

1.3.1 Notion de risque

Le risque est une notion complexe, de définitions multiples car ce dernier est à usage multidisciplinaire. Étymologiquement le mot risque vient du latin "rescare" qui signifie "couper ". Ainsi nous pouvons définir le risque dans le sens commun comme étant un événement, un inconvénient qui vient " couper " ou perturber une activité habituelle. La notion de risque est également liée à la gravité des conséquences de l'aléa (chance bonne ou mauvaise) dont la survenue est probable. Prédire ou prévoir les conséquences des aléas fait partie de l'analyse et la gestion des risques.

Le risque est également vu comme la probabilité d'un événement négatif combinée avec l'impact chiffré qu'il peut y avoir. C'est cette dernière définition qui nous ait intéressante car elle fait appelle à la " probabilité " qui nous laisse sous l'effet d'une analyse quantitative avant de prendre des décisions concernant d'important risque ou de menace à l'organisation[[16],[8]].

1.3.2 Définition du risque de crédit

Le risque de crédit ou de contrepartie est le risque le plus important et le plus dangereux auquel est exposé une banque. Il peut donc se définir ainsi comme étant le risque que l'emprunteur (particulier ou entreprise) ne rembourse pas sa dette à l'échéance fixé.

Ce risque est en effet lourd de conséquence pour toute banque ou établissement bancaire : toute dette non rembourser est économiquement une perte sèche que supporte la banque. Les créances et emprunt accordé aux entreprises ou aux particuliers constituent

ainsi un poste spécifique dans le bilan de la banque et toute évolution négative obère d'autant la survie de la banque à court ou à long terme.

En effet le risque de crédit représente la perte consécutive à l'incapacité par un débiteur de respecter ses engagements. Cet engagement peut aussi être de livrer des fonds ou des titres dans le cadre d'une opération à terme ou d'une caution ou garantie donnée. Ce risque est alors enregistré dans le hors-bilan.

Ainsi le risque de crédit est caractérisé par l'incertitude temporelle d'un événement ayant une certaine probabilité de survenir et de mettre en difficulté la banque.

D'après Godlewiski C.J " le risque de crédit peut être défini comme une non performance de la contrepartie engendrant une perte probable au niveau de la banque ".

Dans les affaires de crédit la banque à l'obligation de respecter la règle d'or des banques cette règle dite principe de l'adossement stipule que : " Les finances les prêts à court terme avec des fonds à court terme et les prêts à long terme avec des passifs à long terme ". Du moment où la banques ne met pas cette règle en vigueur lors de ses transactions, elle est donc exposée aux risques de crédit sur différentes formes.

Il faut donc noter que le risque de crédit est une perturbation de la banque qui peut faire succomber cette dernière ou de la mettre en difficulté en fonction du type de risque[15].

1.3.3 Types de risque des crédits

Le risque de crédit comprend trois types de risque qui sont les suivants[16] :

- **Le risque défaut clients**
  Ce type de risque est caractérisé par le l'incapacité qu'un débiteur à assurer le payement de ses échéances. Un débiteur est en défaut lorsque l'un ou plusieurs des événements suivants est aperçus
  - L'emprunteur ne remboursera pas en totalité ses dettes ; La constatation d'une perte portant sur l'une de ses facilités ;
  - Le débiteur est en défaut de payement depuis quatre-vingt-dix (90) jours sur l'un de ses crédits ;

- **Le risque de dégradation de la qualité du crédit.**
  Il se traduit par la dégradation de la situation financière d'un emprunteur, ce qui accroit la probabilité de défaut, même si le défaut proprement dit ne survient pas nécessairement. Le risque de dégradation de la qualité du crédit est le risque de voir se dégrader la qualité de la contrepartie (dégradation de sa note) et donc l'accroissement de sa probabilité de défaut. Cela conduit à une hausse de sa prime de risque, d'où la baisse de la marge sur intérêts. Ce risque peut être mesuré d'une façon séparée pour chaque contrepartie ou globalement sur tout le portefeuille de crédit.
  Il correspond à la détérioration de la qualité du crédit qui se traduit par la prime de risque liée à l'emprunteur sur le marché des capitaux. En outre, si celui-ci bénéficie d'un rating auprès d'une agence de notation, sa note est susceptible de se détériorer. D'ailleurs ces signaux sont très corrélés avec le risque de défaut et sont utilisés par

les marchés comme indicateur d'un risque éminent.

– **Le risque de taux de recouvrement.**
Le taux de recouvrement permet de déterminer le pourcentage de la créance qui sera récupéré en entreprenant des procédures judiciaires, suite a la faillite de contrepartie. Le recouvrement portera sur le principal et les intérêts après déduction du montant des garanties préalablement recueillies.
Le taux de recouvrement constitue une source d'incertitude pour la banque dans la mesure où il est déterminé à travers l'analyse de plusieurs facteurs :
– La durée des procédures judiciaires qui varient d'un pays à un autres ;
– La valeur réelle des garanties ;
– Le rang de la banque dans la liste des créanciers
Il correspond à l'incertitude liée aux taux de recouvrement postérieur à un défaut constaté. Le taux de recouvrement permet de déterminer la proportion des créance qui sera récupérée par des procédures judiciaires, la valeur réelle des garanties et la priorité donnée au règlement de certaines créances.

1.3.4 Composantes de risque de crédit

Généralement les composantes du risque de crédit sont les suivantes[16] :

– **le défaut** : événement par lequel l'emprunteur n'honore pas une échéance fixée .

– **l'exposition à la date du défaut** : c'est le montant pour lequel la banque est en risque et qui inclut le capital restant.

– **la perte en cas de défaut** : elle correspond à la fraction de l'exposition qui ne pourra être récupérée, elle dépend fortement du taux de recouvrement (ou de récupération) en cas de défaut, lui-même lié à la situation de l'entreprise, à la législation et à la présence d'éventuelles garanties en faveur du créancier financier.

1.3.5 Analyse et l'identification du risque de crédit

Le risque lié à l'activité de crédit peut dépendre de l'emprunteur ou du prêteur. Si le risque provient du débiteur, il s'agit d'un cas d'insolvabilité. Dans ce cas de risque externe, la banque n'est pas responsable de la dégradation de la situation du client. Si le risque provient du créancier, le problème repose sur la politique de distribution des crédits de la banque. Dans ce contexte de risque interne, la banque est responsable de la diffusion des crédits sur le marché. Cette étape révèle une menace plurielle, la banque comme le client peut avoir sa responsabilité engagée [11].

Avant de pouvoir gérer les risques il est nécessaire de les identifier. Elle permet de rechercher les sources ou facteurs de risques liés à l'activité de crédit. Cette analyse permet de vérifier la réalisation, les objectifs poursuivis et de mettre en place des mesures correctrices

si nécessaire. Pour mener ces recherches la banque va s'intéresser sur toutes les données relatives au client ainsi que sur le crédit demandé.

Conclusion

A l'issue de ce chapitre, nous pouvons dire que la banque joue un rôle important d'intermédiaire financière, c'est un interlocuteur de choix pour les entreprises et les particuliers qui constituent une demande sur plusieurs types de services bancaire, tel que le crédit sous ses différentes formes, qui est l'activité de base de chaque banque.

Nous avant décrit les principaux concepts sur le crédit ainsi que la couverture du risque. En accordant des crédits bancaires, le banquier convient de connaitre comment faire de ce crédit un générateur de profit et de gain et non celui de risque et de perte. Pour cela, il met en place des moyens pour pouvoir gérer les risques de ces crédits.

dans le chapitre suivant nous allons exposés les principaux moyennes et méthodes de L'analyse et l'identification du risque de crédit.

2

Présentation des méthodes de résolution

Introduction

Suite à l'étape de l'identification des éventuels risques de contrepartie sur un portefeuille, les établissements bancaires cherchent à se prémunir au maximum avant de devoir passer à une possible gestion curative. La gestion préventive est majeure pour les banques car elle permet de réduire le plus possible la situation de non remboursement d'un client.

Dans ce chapitre nous présentons le crédit scoring d'une manière générale, en suite, on passe à la présentation de deux méthodes d'évaluation de risque qui nous servirons pour élaborer deux fonctions scores en utilisant les outils mathématiques, nous présenterons l'analyse discriminante de Fisher puis on parlera de la régression logistique.

2.1 Méthode Crédit scoring

Le crédit scoring se trouve parmi les modèles de prévisions des risques les plus utilisés dans la micro finance notamment dans les pays en développement. Cet outil est manifesté dans les travaux d'ALTMAN E I, les deux véritables pionnières de l'application des techniques de crédit Scoring à l'activité d'octroi de crédit aux entreprises[1].

Le crédit scoring est un outil d'aide à la décision, basé sur l'utilisation des techniques statistiques pour prédire la probabilité de défaillance d'un demandeur de prêt. Elle vise à associer à chaque demande de crédit une note proportionnelle à la probabilité de l'emprunteur de faire défaut[[10],[20]].

2.1.1 Objectif de la méthode

Le principe du scoring est de déterminer les variables clés qui discriminent le plus les deux groupes d'entreprise (emprunteurs) (entreprises saines et entreprises défaillantes), Ensuite un indicateur appelé **score** est calculé nous permet de juger rapidement la situation d'une entreprise. Cet indicateur est élaboré sur la base de deux échantillons d'entreprises, jugées à priori saines ou défaillantes.

La fonction score peut se présente sous la forme suivante :

$$S(score) = \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_p X_p.$$

Avec

- X : sont les paramètres à choisir (ratio de comptable).
- α : sont les coefficients à estimer

Ils existe plusieurs fonction de score ,elles présentent plusieurs avantage pour le secteur bancaire et celle à travers[[7],[20]] :

1. La simplicité : L'utilisation du score est obtenu à partir d'un certain nombre d'informations synthétisées et offre une rapidité dans la prise de décision, ce qui constitue un double avantage : une charge de travail réduite et une réponse rapide pour le client.

2. L'homogénéité : Le crédit scoring donne la même décision quelque soit l'agence ou le temps de la prise de décision.

2.1.2 Etapes de la méthode de scoring

l'objectif principal de l'application de cette méthode est la discrimination des nouveaux clients demandeur de crédits en deux classes(bon payeurs/mauvais payeurs) on basent sur les quatre (04) étapes suivent[4] :

1. Étape 1 : Dans Cette étape, trois (03) éléments nécessaire à déterminer sont :
   - Un échantillon des clients demandeurs de crédit qui sont de la classe "bon payeurs".
   - Un échantillon des clients demandeurs de crédit qui sont de la classe "mauvais payeurs"
   - Un ensemble de ratios.
2. Étape 2 : Cette étape consiste à déterminer :
   - La sélection des ratios les plus significatifs par la méthode discriminante.
   - La détermination des coefficients de pondération cohérents à chaque ratio.
3. Étape 3 : Déterminer une fonction score noter (S) qui exprime une combinaison entre les ratios et les coefficients de pondération déterminer dans l'étape président avec la formule suivent :

   $$S = \alpha_1 R_1 + \alpha_2 R_2 + ... + \alpha_n R_n.$$

   avec :

- R_i :Les ratios .
- α_i :sont les coefficients de pondérations

4. Étape 4 : Dans cette étape, on définit un score limite servant de seuil de classement et de validation de la fonction.

2.1.3 Techniques scoring utiliser à l'application du crédits

Il existe plusieurs techniques pour la construction des modèles de score :

- **A : Les techniques fondées sur les méthodes paramétriques de classification**
  Les méthodes paramétriques de classification établissent une relation fonctionnelle entre les variables explicatives dont la loi de distribution est supposée connue et la variable expliquée, relation dont la forme est donnée a priori. Dans cette catégorie, on peut trouver trois grandes familles de méthodes : la méthodologie unidimensionnelle, l'analyse discriminante (linéaire et non linéaire) et la régression sur variables qualitatives.

  1. **La méthodologie unidimensionnelle**
     La mise en oeuvre d'une approche unidimensionnelle illustrée par l'étude de W.BREAVER en 1966, est considérée comme un premier effort sur l'application de méthode statistique.Cette méthode de classification est fondée sur un ratio unique. L'objectif est de classer les entreprises parmi l'un des deux groupes : défaillantes ou non défaillantes sur la base du ratio le plus discriminant. BREAVER a procédé de la manière suivante : il a classé les entreprises en fonction des valeurs prises par chaque ratio. Ensuite, il a choisi un seuil critique de telle sorte que toute entreprise présentant un ratio inférieur à ce seuil est considérée comme défaillante et toute celle ayant un ratio supérieur est considérée comme saine. Le seuil critique est déterminé de manière à maximiser le taux de bon classement. C'est ce taux qui va déterminer le ratio le plus discriminant[5].

  2. **L'analyse discriminante (Modèle d'Altman(1968))**[13]
     Contrairement à la méthode unidimensionnelle qui utilise un seul ratio, l'analyse discriminante est une technique qui permet de définir à partir d'un ensemble d'entreprises réparties en deux groupes (les saines et les défaillantes) et caractérisées par un nombre d'indicateurs financiers.
     Altman (1968)[1] a développé un modèle de score établi sur la base d'un échantillon de 66 entreprises, dont 33 sont considérées comme défaillantes et 33 comme saines. Nous remarquons que la taille de l'échantillon utilisée dans le modèle d'Altman est faible ce qui peut altérer la qualité des résultats.
     La technique statistique adoptée dans ce modèle est celle de l'analyse discriminante multivariée. Elle repose sur une fonction de score combinaison linéaire des cinq ratios financiers jugés les plus pertinents pour départager au mieux les deux groupes d'entreprises (saines ou défaillantes).

la fonction d'Altman peut se représenter sou la forme suivent :

$$Z = 1.2R_1 + 1.4R_2 + 3.3R_3 + 0.6R_4 + 0.99R_5.$$

Avec
– les X_i sont des variables explicatives.
– Z représente le résulta de la discrétisation.
Le risque encouru par la banque varie dans le sens contraire de Z,la figure suivent résume cette variance .

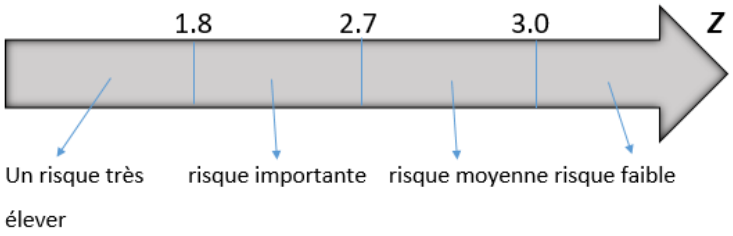

FIGURE 2.1 – Règles de décision pour le modèle d'Altman [12]

Analyse

A partir de la figure présidente on remarque que :
– Si la valeur de Z inférieur à 1.8 alors la probabilité de faire défaut d'un problème financier est très élevée.
– Pour une valeur du score compris entre 1.8 et 2.7, la probabilité de faire défaut est importante donc l'entreprise est jugée à haut risque.
– Pour une valeur du score compris entre 2.7 et 3.0, la probabilité de faire défaut est peu donc l'entreprise est jugée à un risque moyenne .
– Si la valeur de Z est supérieur à 3, l'entreprise à peu de risque de faire défaut.[1]

3. **Les modèles de régression**
   Les modèles de régression sont utilisés dans le cas où la variable à expliquer est une variable qualitative, qui prend la valeur zéro ou un (0 ou 1), selon que l'entreprise est défaillante ou non. Le modèle explique cette variable en fonction d'un vecteur de variables exogènes qui est composé de K ratios économiques et financiers retenus pour leur qualité discriminante et leur faible corrélation entre elles.

– **B : Les techniques d'intelligence artificielle (réseaux de neurones)**
Les réseaux de neurones sont des algorithmes d'intelligence artificielle qui permettent à partir de l'expérience de déterminer la relation entre les caractéristiques d'un des emprunteurs et leur probabilité de défaut. Cette technique prend en compte l'effet de non-linéarité entre la variable à expliquer et les variables explicatives, mais sa modélisation, son utilisation et l'interprétation des résultats peuvent être complexes, comme on lui reproche souvent le manque de stabilité de ses résultats.
Le principe des réseaux de neurones consiste en l'élaboration d'un algorithme dit d'apprentissage qui imite le traitement de l'information par le système neurologique humain. Trois sortes de neurones existent : les neurones d'entrée, les neurones de sortie et les neurones cachés. Les neurones d'entrée ont pour input les K ratios comptables présélectionnés , les neurones de sortie ont pour output la variable dichotomique défaillante / non défaillante. Les neurones cachés sont des neurones qui traitent l'information entre les neurones d'entrée et de sortie.

2.1.4 Processus d'application de crédit scoring

Pour évaluer et prédire si un emprunteur sera bon ou mauvais payeur la plupart des institutions financières sont dotées d'un système de notation interne qui lui facélite cette tache.Ce processus comporte trois (03) étape sont les suivent [13] :

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

2. **Quantification et la détermination des variables du crédit**
   Apres avoir terminer la collecte des information concernent l'emprunteur, Le banquier étudie le dossier du client et prend sa décision d'accorder ou, non ce crédit en se basant sur le score calculé. Avant la phase de modélisation, les données collectées sont analysées pour déterminer les variables les plus pertinentes et discriminantes des clients. la plupart des variables sont de type qualitatif et ne peuvent être intégrés

directement dans un modèle statistique.

3. **l'étape de modélisation**
   lorsque la phase de préparation des données terminé, on commence la phase de modélisation qui permet le calcul des scores. On utilise les différentes techniques de scoring, on peut l'applique pour la construction de la fonction de score.
   Dans le cas pratique, une fois l'étape de collecte de données terminé, des scores partiels sont attribués à chaque variable du dossier et le score final est obtenu par leur sommation.

4. **L'étape de décision**
   Dans la dernière étape du processus consiste à comparer ce score au seuil fixé par la banque. Si le score dépasse ce seuil alors la banque accorde le crédit sinon, la demande sera rejetée.

Condition d'utilisation de la méthode

Pour utiliser la méthode scoring, certaine condition doive être vérifier on citent par exempl[6] :

1. Les données historiques doivent couvrir une période assez longue pour couvrir un cycle économique (autour de 7 ans) ;
2. L'échantillon de construction doit contenir un grand nombre d'individus pour qu'il soit représentatif du portefeuille de crédit ;
3. Le modèle doit prévoir le défaut : le taux de bon classement doit être le plus élevé possible ;
4. Les coefficients doivent être significatifs et conformes à la logique comptable et économique ;
5. L'utilisation des scores en dynamique : il est nécessaire d'examiner un peu plus en détail la situation financière du client afin de lutter contre la dérive temporelle.

2.1.5 Avantages et limites du crédit scoring

– **A : Avantages et limites du crédit scoring**
  L'utilisation du crédit scoring au sein des établissements de crédit en tant qu'outil d'aide à la décision offre plusieurs avantages [13] :

  1. En proposant une appréciation synthétique de la situation d'une entreprise, la méthode des scores permet, d'anticiper le risque de défaillance de l'entreprise et de diminuer par conséquent les impayés, aussi " parce qu'il fondé sur une appréciation objective des critères de risque, l'utilisation des scores permet à l'établissement de crédit de disposer en fonction de sa sensibilité aux risques le niveau d'impayés qu'il tolère ".

2. Les modèles de score par rapport aux autres méthodes traditionnelles permettent, grâce à la rapidité de décision qu'ils présentent, un traitement de masse de populations nombreuses d'emprunteurs et leur usage réduit de manière significative la durée du traitement des dossiers de crédit (de 15 jours à quelques heures, pour la plupart des crédits standard).
   Ce gain de temps permet à l'analyste financier de concentrer son attention sur d'autres aspects comme l'étude de demandes de crédit plus délicates et plus complexes.
3. Le scoring contribue à résoudre les difficultés induites par la multiplicité des indicateurs d'équilibre financier, en orientant vers une sélection qui échappe aux pièges de la subjectivité.
4. Les outils de scoring sont peu coûteux.

– **B : Les limites des modèles de score**
  Au-delà du problème de biais de sélection ou du problème de la réintégration des refusés, nous pouvons indiquer les limites suivantes des modèles de score :

1. Le système de crédit scoring apparaît figé dans le temps, car le secteur pour lequel il a été construit ainsi que la situation économique peuvent évoluer, de ce fait au-delà d'une certaine durée d'utilisation, il peut perdre son pouvoir discriminant .
2. Les modèles de score capturent mal les changements de toute nature qui modifient l'attitude des emprunteurs par rapport au défaut (en augmentant par exemple le hasard moral) .
3. Les modèles omettent des éléments qualitatifs liés à la qualité des dirigeants ou aux caractéristiques particulières des marchés sur lesquels opèrent les emprunteurs.
4. Les modèles de score sont des outils statistiques. Ils comportent deux types d'erreurs, l'erreur (de type II) qui consiste à classer en défaut des emprunteurs sains et l'erreur (de type I) qui consiste à classer comme sain un emprunteur dont la probabilité de défaut est en réalité élevée.
5. Ces erreurs ont naturellement un coût pour le prêteur utilisant un modèle de score. C'est pourquoi, généralement, les résultats du score peuvent être corrigés ex post en traitant des informations complémentaires, à la manière des systèmes experts.
6. La méthode des scores peut aussi, accélérer la défaillance d'une entreprise qui aurait un mauvais score. Il est très probable que le comportement des partenaires de celle-ci se modifie, ce qui accélérera le processus de dégradation.

Dans la suite de travaille en présentons juste deux outille mathématique pour l'élaboration d'une fonction score. Nous présenterons l'analyse discriminante de Fisher puis la régression logistique.

2.2 Analyse discriminante de Fisher

Cette méthode est la plus ancienne des méthodes statistiques de classement. Remontant aux travaux de Fisher en 1936, elle permet de classer les individus d'une population entre différents groupes définis à priori au vu de données relatives à des variables quantitatives. C'est donc une méthode où les probabilités conditionnelles à estimer sont supposées relever de lois de probabilités données mais dépendant néanmoins de paramètres inconnus à estimer à partir des données mises à disposition (estimation paramétrique). Pour ce qui est de notre travail nous allons utiliser un système binaire pour déclarer les bonnes entreprises et les mauvaise, à cet effet nous utiliserons 0 pour dire l'entreprise est défaillantes et 1 pour dire que l'entreprise est en bonne état.

2.2.1 Modélisation

Considérons : $E_1 \cup E_2 = E$ et $E_1 \cap E_2 = \varnothing$ où cardinal de E est égale à n, cardinal de E_1 est égale à n_1 et cardinal de E_2 est égal à n_2 ($n_1 + n_2 = n$).

Tel que E_1 représente le groupe des clients en bonnes états et E_2 représente le groupe des clients défaillants. Mesurons simultanément p variables explicatives (X_i tel que $1 \leq i \leq p$) (c'est-à-dire, chaque client de chaque sous-échantillon est caractérisé par p variables explicatives) et une variable à expliquer Y à deux modalités (bon client, mauvais client). Nous obtenons

$$S(X) = b + \sum_{i=1}^{p} a_i X_i \tag{2.1}$$

avec b est une constante et les a_i sont le poids ou les coefficients associes aux variables X_i, La fonction score $S(X)$ ainsi définie permet de calculer le score de chaque individu et de l'affecter à un groupe. Dans la suite, nous ferons l'hypothèse suivante :

H_1 :Les variables explicatives sont continues, indépendantes et normalement distribuée.

2.2.2 Classifieur Bayésien

1. **Élément de la théorie de décision**[14].
   On considère une population E de n individus répartis entre groupe E_1, E_2 définis comme suite :

   $$\cap_{k=1}^{2} E_k = E \quad et \quad E_1 \cap E_2 = \varnothing$$

   Soit un client e de E, dont on ne connait pas le groupe d'appartenance et qu'on cherche à classer dans l'un des 2 groupes. Ce client peut être considéré comme le résultat d'une expérience aléatoire de tirage au hasard d'un élément de E.

   Vu de cette manière, le problème de classement peut être placé dans le cadre de la théorie probabiliste. L'ensemble E se présente ainsi comme un ensemble de résultats

possibles d'une expérience aléatoire auquel on peut adjoindre une tribu A et une probabilité P pour former un espace probabilisé.

2. **État de la nature**
   Pour un client e dont on ne connait pas le groupe d'appartenance, on définit m états de nature c'est-à-dire des éventualités, concernant son groupe d'appartenance. Ces états de la nature sont notés : θ_k "l'individu $e \in E_k$" et on désigne : $\theta = \{\theta_1, \theta_2\}$ l'ensemble des états de la nature. Soit T une application de E dans θ qui à chaque individu l'associe a son état de la nature.
   On peut considérer T comme une valeur quantitative, prenant les modalités θ_1, θ_2 et les probabilités à priori d'appartenance au groupe K sont : $P_k = P(T = \theta_k)$ comme sa loi de probabilité.

   Remarque 3.2.1 : Il faut noter que T est non observable.

3. **Espaces des observations**
   Soit $x = (x_1, x_2, ..., x_j, ..., x_p)$ un vecteur de p observations relevées auprès du client e. On peut considérer x comme une réalisation du vecteur aléatoire $x = (x_1, x_2, ..., x_j, ..., x_p)$. On note $\Gamma = \{x \in \mathbb{R}^p;\ \ x\ \ ralisation\ \ de\ \ X\}$. C'est l'espace des observations. La variable X est une application de E dans Γ.

4. **Espaces des décisions**
   On à affecter un client e dans l'un des 2 groupes. C'est une décision, on note a_k la décision d'affecter l'individu e dans le groupe k. On note A l'ensemble de décisions :$A = \{a_1, a_2\}$.

5. **Règle de décision**
   C'est une méthode de classement. Formellement, c'est une application de Γ dans A. On note δ cette application, techniquement c'est un procédé permettant de prendre une décision.

   $$\ll a \gg \text{ au vu de la réalisation x de X : } a = \delta(x).$$

   Il faut noter que comme x résulte du hasard, $a = \delta(x)$ aussi résulte du hasard, on définit aussi la variable aléatoire : $Y = \delta(X)$ qui prend les valeurs a_1, a_2 avec des probabilités définies par :

   $$P(Y = a) = P(x \in \Gamma; a = \delta(x)) = P(\delta^{-1}(a)).$$

6. **Fonction de pèrte**
   A chaque règle de décision on associe une fonction de perte définie par une application L de $(\mathrm{A}\,;\,\theta)$ dans $\mathbb{R}^+$ telle que $L(a_k, \theta_1) \geq 0$. On l'interprète comme la perte ou le

coût supporté en affectant le client e au groupe k alors qu'en réalité il appartient au groupe l. On note $L(a_k, \theta_k) = 0$ pour tout k = 1 à 2. D'autre part, comme a_k et θ_1 résulte du hasard, la perte encourue $z = L(a_k, \theta_l)$ résulte aussi du hasard c'est une réalisation d'une variable aléatoire $Z = L(Y, T)$.
Dans la suite on est amené à calculer la perte moyenne d'une règle de décision :

$$\begin{aligned} E(Z) &= \sum_A \sum_B P(Y = y, T = \theta) L(y, \theta) \\ &= \sum_\Gamma \sum_\theta P(X = x, T = \theta) L(\delta(x), \theta) \end{aligned}$$

Définition 3.2.1 : Étant donné un espace d'état de la nature θ, un espace d'observation Γ, un espace de décision A et une fonction de perte L, le clasifieur de Bayes est la règle de décision minimisant la perte moyenne parmi toute les règles de décisions possibles.
Soit Δ l'ensemble de toute les règles de décisions possibles de Γ dans A. Le classifieur de bayes noté δ^* est donc :

$$E(L(\delta^*(X), T)) \leq E(L(\delta(X), T)) \qquad \forall \delta \in \Delta$$

2.2.3 Caractérisation

Proposition 3.2.1 : Soit $\delta_0 \in \Delta$ minimisant $E(L(\delta_0(X), T)/X = x) \ \ \forall x \in \Gamma$ alors δ_0 minimise $E(L(\delta(X), T))$.

Preuve :

$$\begin{array}{ll} E(L(\delta_0(X), T)/X = x) & \leq E(L(\delta(X), T)/X = x) \quad \forall x \in \Gamma \ \ et \ \ \forall \delta \\ E_x(E(L(\delta_0(X), T)/X = x)) & \leq E_x(E(L(\delta(X), T)/X = x)) \quad \forall \delta \in \Delta \\ E(\delta_0(X), T) & \leq E(\delta(X), T) \quad \forall \delta \in \Delta \end{array}$$

La règle de Bayes est donc telle que :

$$E(L(\delta^*(X), T)/X = x) \leq E(L(\delta(X), T)/X = x) \quad \forall x \in \Gamma \ \ et \ \ \forall \delta \in \Delta$$

Soit donc,

$$\sum_{k=1}^{2} L(\delta^*(x), \theta_k) P(T = \theta_k / X = x) \leq \sum_{k=1}^{2} L(\delta(x), \theta_k) P(T = \theta_k / X = x) \ \ \forall x \in \Gamma \ \ et \ \ \forall \delta \in \Delta$$

Or d'après la formule de bayes

$$P(T = \theta_k / X = x) = \frac{p_k f_k(x)}{\sum_{k=1}^{2} p_k f_k(x)}$$

En remplaçant

$$\sum_{k=1}^{2} L(\delta^*(x), \theta_k) p_k f_k(x) \leq \sum_{k=1}^{2} L(\delta(x), \theta_k) p_k f_k(x).$$

Ainsi, étant donné un client e et x_e la réalisation de X chez cette individu. Notons a_{l^*} la décision prise au vue de x_e si on applique la règle de Bayes ($a_{l^*} = \delta^*(x_e)$) et a_l la décision prise au vue de x_e si on applique une autre règle de décision ($a_l = \delta(x_e)$) . Comme $L(a_k, \theta_k) = 0$ pour tout $k = 1$ à 2, on a ainsi :

$$\sum_{k=1,k\neq l^*}^{2} L(a_{l^*}, \theta_k) p_k f_k(x) \leq \sum_{k=1,k\neq l}^{2} L(a_l, \theta_k) p_k f_k(x) \quad \forall l^* \neq l.$$

D'où la règle pratique suivante, étant donné un individu e de caractéristique x :
- On commence par calculer $\sum_{k=1,k\neq l^*}^{2} L(a_{l^*}, \theta_k) p_k f_k(x)$ pour chaque $l = 1$ à 2.
- Le groupe à retenir pour l'affectation de l'individu e est celui pour lequel cette quantité est plus faible.

Remarque 3.2.2 : Les coûts varient d'une application a une autre. Si l'on suppose que les coûts sont égaux, la règle de Bayes à une formulation assez simple. En effet en développant la formule précédant on trouve :

$$L(a_{l^*}, \theta_l) p_l f_l(x) + \sum_{k=1,k\neq l^*,k\neq l}^{2} L(a_{l^*}, \theta_k) p_k f_k(x) \quad \leq L(a_{l^*}, \theta_{l^*}) p_{l^*} f_{l^*}(x) +$$

$$\sum_{k=1,k\neq l^*,k\neq l}^{2} L(a_{l^*}, \theta_k) p_k f_k(x) \forall l^* \neq l$$

Comme les coûts sont égaux alors,

$$L(a_{l^*}, \theta_l) = L(a_{l^*}, \theta_{l^*}) et L(a_{l^*}, \theta_k) = L(a_l, \theta_k) \ \ \forall l^* \neq l \ \ \forall k = 1, 2$$

On a donc :

$$\begin{aligned} \sum_{k=1,k\neq l^*,k\neq l}^{2} L(a_{l^*}, \theta_k) p_k f_k(x) \ &= \sum_{k=1,k\neq l^*,k\neq l}^{2} L(a_{l^*}, \theta_k) p_k f_k(x) \ \ \forall l^* \neq l \\ &et \ \ L(a_{l^*}, \theta_l) = L(a_{l^*}, \theta_{l^*}) \end{aligned} \tag{2.2}$$

En simplifiant les expressions on a, $p_l f_l(x) \leq p_{l^*} f_{l^*}(x)$

en divisant (3.2) par $\sum_{k=1}^{2} p_l f_l(x)$ membre par membre et on trouve :

$$P(T = \theta_l / X = x) \neq P(T = \theta_{l^*} / X = x)$$

Ce qui signifie que le classifier de Bayes affecte le client e au groupe pour lequel la probabilité d'appartenance à postériori est la plus élevé. Bien que l'hypothèse d'égalité des coûts ne soit pas possible, c'est cette règle qui est la plus retenue en pratique. Ainsi, étant donné un client e et x c'est caractéristiques ;
- On commence par calculer la quantité $C_k(x) = p_k f_k(x)$ pour k = 1 à 2.
- Le groupe d'affectation de e est celui pour lequel la quantité $C_k(x)$ est la plus élevée ou maximal

Soit l'hypothèse :

H_2 : La loi conditionnelle de X sachant Y suit la loi multi-normale de moyenne μ_i et de matrice de variance-covariance $\sum$

$$X/(Y = K) \rightsquigarrow N(\mu_i, \sum)$$

Théorème 3.2.1 (Marie Chavent, 2015)[3] :

La règle de décision Bayésienne est donnée par : $g_k = argmax P_k f_k(x)$ Sous l'hypothèse H_2,la densité conditionnelle de X|Y s'écrit :

$$f_k(x) = \frac{1}{(2\Pi)^{\frac{p}{2}}(\det \sum)^{\frac{1}{2}}} exp(-\frac{1}{2})(x-\mu_k)' \sum^{-1}(x-\mu_k)$$

Théorème 3.2.2 : la règle de décision bayésienne devient alors :

$$g_k = argmax c_k(x) \quad avec \quad c_k(x) = \ln(P_k) - \frac{1}{2}(x-\mu_k)' \sum^{-1}(x-\mu_k)$$

Preuve : on sait que $g_k = argmax P_k f_k(x)$ toute transformation monotone de $P_k f_k(x)$ ne change rien sur la maximisation de cette expression en particulier la transformation logarithmique on a :

$$\begin{aligned} \ln(P_k f_k(x)) &= \ln(P_k) + \ln(f_k(x)) \\ &= \ln(P_k) - \tfrac{1}{2}(x-\mu_k)' \sum^{-1}(x-\mu_k) - \ln\left((2\Pi)^{\frac{p}{2}}(\det \sum)^{\frac{1}{2}}\right) \end{aligned}$$

Puisque $\ln\left((2\Pi)^{\frac{p}{2}}(\det \sum)^{\frac{1}{2}}\right)$ ne dépend de k, maximiser $\ln(P_k f_k(x))$ revient à maximiser $c_k(x) = \ln(P_k) - (x-\mu_k)' \sum^{-1}(x-\mu_k)$. C'est-à-dire maximiser $P_k f_k(x)$ revient à maximiser $c_k(x)$.

On dira donc qu'un client est bon si : $g_1 > g_2$

$$\begin{aligned} g_1 > g_2 \quad &\Rightarrow \quad \ln(P_1) - \tfrac{1}{2}(x-\mu_1)' \sum^{-1}(x-\mu_1) > \ln(P_2) - \tfrac{1}{2}(x-\mu_2)' \\ &\quad \sum^{-1}(x-\mu_2) \\ &\Rightarrow \quad \ln(P_1) - \tfrac{1}{2}x' \sum x - \tfrac{1}{2}\mu_1' \sum \mu_1 + x' \sum \mu_1 > \ln(P_2) \\ &\quad -\tfrac{1}{2}x' \sum x - \tfrac{1}{2}\mu_2' \sum \mu_2 + x' \sum \mu_2 \\ &\Rightarrow \quad \ln(P_1) - \tfrac{1}{2}\mu_1' \sum \mu_1 + x' \sum \mu_1 > \ln(P_2) - \tfrac{1}{2}\mu_2' \sum \mu_2 \\ &\quad +x' \sum \mu_2 \\ &\Rightarrow \quad x' \sum(\mu_1 - \mu_2) > \tfrac{1}{2}(\mu_1 - \mu_2)' \sum(\mu_1 - \mu_2) + \ln\left(\frac{p_1}{p_2}\right) \end{aligned}$$

En posant

$$S(x) = x' \sum(\mu_1 - \mu_2) \text{ et } s = \tfrac{1}{2}(\mu_1 - \mu_2)' \sum(\mu_1 - \mu_2) + \ln\left(\tfrac{p_1}{p_2}\right)$$

Un client e au vu de ses observations x est donc dit bon si $S(x) > s$ et il sera dit défaillant ou mauvais si $S(x) < s$, s est appelé seuil.

Pour la mise en application du classifieur de Bayes, il faut disposer des probabilités à priori p_k, des moyennes μ_k et la matrice de variance-covariance $\sum$. En pratique, ces grandeurs sont en général inconnues. Il convient en conséquence de les estimer à partir de l'échantillon. On peut procéder à une estimation directe de ces probabilités dans le cas où $X = (X_1, X_2, ..., X_j, ..., X_p)$ est discret et p petit (estimation non paramétrique).

2.2.4 Estimation

Soit X la variable aléatoire que nous souhaitons étudier. Nous effectuons pour cela n réalisations de X. Les résultats de ces réalisations sont des variables aléatoires notées $X_1, X_2, ..., X_n$, qui ont même loi que X. Nous considérerons dans la suite que les mesures sont effectuées de manières indépendantes. On dit alors que les variables $X_1, X_2, ..., X_n$ sont i.i.d. : indépendantes et identiquement distribuées. Le but est ici d'estimer la valeur du paramètre θ, c'est-à-dire de déterminer la valeur de θ la plus vraisemblable pour la loi de $X_1, X_2, ..., X_n$. Et une fois cette grandeur estimée, une question serra posée : quelle est la précision de l'approximation réalisée ?. Soit X une variable aléatoire dont la densité de probabilité $f(x, \theta)$ dépend d'un paramètre θ appartenant à $I \subset \mathbb{R}$. A l'aide d'un échantillon issu de X,il s'agit de déterminer au mieux la vraie valeur θ_0 de θ. On pourra utiliser deux méthodes :

1. Estimation ponctuelle : on calcule une valeur vraisemblable $\hat{\theta}$ de θ_0.
2. estimation par intervalle : on cherche un intervalle dans lequel θ_0 se trouve avec une probabilité élevée.

1. **Estimation ponctuelle**

Définition 2.3.1 : Un n-échantillon de X est un n-uplet $(X_1, X_2, ..., X_n)$ tel que les X_k ont la même loi que X et sont indépendantes. Une réalisation de l'échantillon est alors un n-uplet $(x_1, x_2, ..., x_n)$ de valeurs prises par l'échantillon.

Définition 2.3.2 : Une statistique de l'échantillon est une variable aléatoire $\phi(x_1, x_2, ..., x_n)$ où ϕ est une application de $\mathbb{R}^n$ dans $\mathbb{R}$.

Un estimateur T (ou bien $\widehat{\theta}$) de θ est une statistique à valeurs dans I. Une estimation est la valeur de l'estimateur correspondant à une réalisation de l'échantillon.

Définition 2.3.3 : Le biais de l'estimateur T de θ est $E[T] - \theta_0$. S'il est nul, on dit que T est un estimateur sans biais.

Définition 2.3.4 : On dit que L'estimateur T_n est asymptotiquement sans biais si

$$\lim E[Tn] = \theta_0.$$

On note souvent le biais $b_\theta(T)$.

Définition 2.3.5 : L'estimateur est dit convergent si la suite (T_n) converge en probabilité vers θ_0 :

$$\forall \epsilon > 0, P(|T_n - \theta_0| > \epsilon) \underset{n \longrightarrow +\infty}{\Longrightarrow} 0.$$

On parle d'estimateur fortement convergent lorsqu'on a convergence presque sûre.

- **A :Estimation de la moyenne empirique**
  On considère un n-échantillon $(X_1, X_2, ..., X_n)$ issu d'une loi de moyenne μ et de variance σ^2, toutes deux inconnues.
  - D'après la loi de grand nombre, la moyenne empirique X_n est un estimateur convergent de μ.
  - L'estimateur X_n est sans biais.
  - par indépendance : $Var(X_n) = \frac{\sigma^2}{2}$.
  - si $X \rightsquigarrow N(\mu, \sigma^2)$, alors $X_n \rightsquigarrow N(\mu, \frac{\sigma^2}{2})$.

- **B :Estimation de la variance empirique**
  La variance empirique associée à un n-échantillon $(X_1, X_2, ..., X_n)$ est définie par

  $$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X - \overline{X}_i\right)^2.$$

  (a) S_n^2 est un estimateur convergent de la variance σ^2.

  (b) S_n^2 est sans biais.

  (c) loi de S_n^2 n'a pas de résultat général. Cependant,si $X \rightsquigarrow N(\mu, \sigma^2)$, alors la v.a $X_n \rightsquigarrow N(\mu, \frac{\sigma^2}{2})$ suit une loi du chi-deux à $n-1$ degrés de libertés et on note

  $$\chi^2(n-1)$$

  .

- **C :Estimation par le maximum de vraisemblance**
  La vraisemblance d'un modèle d'un échantillon correspondent à la probabilité d'avoir obtenu cet échantillon lorsqu'on a ce modèle. Ainsi, la vraisemblance des observations $x_1, x_2, ..., x_n$ s'écrit sous la forme :

$$L(\theta, x_1, x_2, ..., x_n) = \begin{cases} P_\theta(x_1), ..., P_\theta(x_n) si \ \ on \ \ a \ \ une \ \ loi \ \ dicrte; \\ \\ f_\theta(x_1), ..., f_\theta(x_n) si \ \ on \ \ a \ \ une \ \ loi \ \ continue; \end{cases}$$

La forme de produit est justifiée par l'hypothèse que les observations sont indépendantes. Le principe de l'estimation par maximum de vraisemblance est de se dire que plus la probabilité d'avoir obtenu les observations est forte, plus le modèle est proche de la réalité. Ainsi, on retient le modèle pour lequel la vraisemblance de notre échantillon est la plus élevée :

$$\hat{\theta}_n = argmax L(\theta, \{x_1, x_2, ..., x_n\}).$$

En pratique, le problème ci-dessus est compliqué à résoudre directement en raison de la présence du produit mais il suffit de prendre le logarithme :

$$\hat{\theta}_n = argmax \ln(L(\theta, \{x_1, x_2, ..., x_n\})).$$

Pour trouver le maximum, on résout l'équation du premier ordre :

$$\frac{\partial \ln(L(\theta, \{x_1, x_2, ..., x_n\}))}{\partial \theta}|_{\theta=\hat{\theta}_n} = 0$$

La théorie nous dit que la solution de cette équation nous donne toujours un maximum. On obtient $\hat{\theta}_n$ sous forme $\hat{\theta}_n = g(x_1, x_2, ..., x_n)$.
L'estimateur du maximum de vraisemblance est alors $\hat{\theta}_n = g(X_1, X_2, ..., X_n)$ et l'estimation du maximum de vraisemblance est obtenue en remplaçant $x_1, x_2, ..., x_n$ par leurs valeurs numériques dans $\hat{\theta}_n = g(x_1, x_2, ..., x_n)$.

2. **Estimation par intervalle**
   Pour l'estimation par intervalle, on considère :
   (a) un paramètre inconnu θ,
   (b) un ensemble de valeurs observées $(x_1, ..., x_n)$, réalisations d'un n-échantillon aléatoire $(X_1, ..., X_n)$, et son estimation ponctuelle noté $\overline{x}_n = \frac{1}{n}\sum_{i=0}^{n} x_i$ Les estimations ponctuelles n'apportent pas d'information sur la précision des résultats, c'est-à-dire qu'elles ne tiennent pas compte des erreurs dues aux fluctuations d'échantillonnage. Pour évaluer la confiance que l'on peut avoir en une valeur, il est nécessaire de déterminer un intervalle contenant, avec une certaine probabilité fixée au préalable, la vraie valeur du paramètre, c'est l'estimation par intervalle de confiance.

Définition 3.3.5 : Soit $(X_1, ..., X_n)$ un n-échantillon aléatoire et θ un paramètre inconnu de la loi des X_i.
Soit $\alpha \in]0,1[$. S'il existe des v.a $\theta_{min(X_1,...,X_n)}$ et $\theta_{max(X_1,...,X_n)}$ telles que

$$P\left(\theta \in \left[\theta_{min(X_1,...,X_n)}, \theta_{max(X_1,...,X_n)}\right]\right) = 1 - \alpha,$$

On dit alors que $\left[\theta_{min(X_1,...,X_n)}, \theta_{max(X_1,...,X_n)}\right]$ est un intervalle de confiance pour θ, avec coefficient de sécurité $1 - \alpha$. On le note $IC_{1-\alpha}(\theta)$.

2.2.5 Estimation des paramètres

On cherche à estimer le paramètre θ défini par $\theta = (P_k; \mu_k; \Sigma)$
L'estimation de θ par la méthode du maximum de vraisemblance donne :

$$\begin{cases} \widehat{p}_k = \frac{n_k}{n} \; ou \; n_k = card(E_k) \\ \widehat{\mu}_k = g_k, k = 1, 2 \\ \widehat{\Sigma} = \frac{n}{n-k} W \end{cases}$$

où les g_k et W sont définies à partir d'observations sur les p variables dans un échantillon de n clients répartis en deux groupes.

2.2.6 Validation du modèle

– **A :Test de lambda wilk's :**
Il permet de tester si les variables explicatives $X = (X_1, X_2, ..., X_p)$ sont liées à une variable à expliquer Y. Mais aussi, il permet de valider le modèle obtenu par analyse discriminante de Fischer.

Définition 3.3.6 : La statistique du Lambda de Wilk's notée Λ est le rapport du déterminant de la matrice de variances covariances intra-groupe au déterminant de la somme des matrices de variances covariances inter et intra-groupe.
Donc :

$$\Lambda = \frac{\det W}{\det(W + B)}$$

Ce rapport est compris entre les nombres 0 et 1 ; où 0 signifie que les variables explicatives expliquent parfaitement la variable Y , et 1 signifie que les variables explicatives n'expliquent pas le modèle. Ainsi, plus le lambda de Wilk's est proche de 0, le modèle est meilleur . Cette statistique de test suit une loi de paramètre $(p; n-p-1)$

– **B :Test du M de Box**
C'est une approche paramétrique qui permet de tester si les matrices de variances covariances associées à $X|Y = k$ sont égales.

Définition 3.3.7 (Statistique du M de Box) : La statistique du M de Box est la statistique définie par :

$$M = (n-2)\ln(\det(H)) - \sum_{k=1}^{2}(n_k - 1)\ln(det(V_k)).$$

avec

$$H = \frac{1}{n-2}\sum_{k=1}^{2}(n_k - 1)V_k$$

Afin de pouvoir rapporter la statistique du M de Box à la table de la loi du χ^2, il convient de lui appliquer la transformation suivante :

$$\chi^2 = M(1-\delta)\ avec\ \delta = \frac{2(p+1)(p+2)}{\theta_p + 1}\left[\sum_{k=1}^{2}\frac{1}{n_k - 1} - \frac{1}{n_k - 2}\right]$$

La statistique de χ^2 test du M de Box, suit alors une loi du khi-deux à $\frac{p(p+1)}{2}$ degrés de liberté. Le M de Box doit être le plus élevé possible.

2.2.7 Analyse d'efficacité du modèle

Une fois un modèle ou plusieurs modèles de scoring sont estimés, il convient d'analyser leurs performances avant de les valider pour être utilisés comme outil d'aide à la décision.

L'analyse de performances, à l'issue de laquelle une méthode de scoring est validée, permet notamment d'améliorer un modèle en comparant plusieurs de ses variantes (ajout ou retrait de variables explicatives, etc.)

L'analyse des performances d'un modèle gagnerait à être conduite sur un jeu de données différent de celui qui a été utilisé pour l'estimation. On doit en effet, lorsque cela est possible, distinguer entre l'échantillon d'apprentissage et l'échantillon de test ou de validation. Ce dernier doit nécessairement contenir les valeurs réelles de la variable cible (appartenance aux groupes). D'une manière générale, il s'agit de comparer entre les valeurs réelles de la variable cible avec celles prédites par le modèle.

- **Concepts de base**
  Comme vue plus haut nous disposons d'un échantillon de n client partitionnée en deux sous-groupes E_1 et E_2. On appelle (par convention) les bons clients, les clients de E_1 et les clients défaillants ou mauvais clients, les clients de E_2. On dispose par ailleurs d'une fonction de score (issue du modèle) notée S et d'un seuil s définies tels que :

  1. On affecte le client présentant l'observation x au groupe E_1 si $S(x) > s$. Autrement dit, on considère ce client comme bon.
  2. Sinon, on l'affecte au groupe E_2, on le considère donc comme défaillant On appelle :
     (a) Faux bon client, un client défaillant considéré par la méthode de score comme bon client.

(b) Faux défaillant, un bon client considéré par la méthode de score comme client défaillant.

On appelle coefficient de spécificité et on note $1-\alpha$ la probabilité suivante :

$$\begin{aligned} 1-\alpha &= Pr(S(x) < s/x \in E_2) \\ &= Pr(S(x) < s/Y = 0) \end{aligned}$$

C'est donc la probabilité de bien détecter un négatif ou encore c'est la proportion des négatifs dans la population pouvant être détecté par la méthode.
La quantité $\alpha = Pr(S(x) \geq s/x \in E_2)$ désigne donc la probabilité de considérer un client comme bon alors qu'il est défaillant (faux bon client). C'est un premier type de risque d'erreur d'affectation.

– **Sensibilité**
On appelle coefficient de sensibilité et on note $1-\beta$
la probabilité suivante :

$$\begin{aligned} 1-\alpha &= Pr(S(x) > s/x \in G_1) \\ &= Pr(S(x) > s/Y = 1) \end{aligned}$$

C'est donc la probabilité de bien détecter un bon client ou encore c'est la proportion des bons clients dans la population pouvant être détecté par la méthode.
La quantité $\beta = Pr(S(x) \leq s/x \in E_1)$ désigne par conséquent la probabilité de considérer un client comme défaillant alors qu'il est bon (faux mauvais client). Il s'agit donc d'un deuxième type de risque d'erreur d'affectation.

Remarque 3.3.1 :

1. Le meilleur modèle (et donc la meilleure fonction de score) est celui qui minimise les deux types de risque d'affectation (les quantités α et β).
2. Les coefficients α et β changent lorsque le seuil s change. On les expriment comme des fonctions de s :$\alpha(s)$ et $\beta(s)$. Le seuil s est déterminé à l'extérieur du modèle notamment par des considérations d'ordre économique

2.2.8 Principaux outils de mesures de performances

Plusieurs outils de mesure de performance sont proposés par la littérature statistique. On présente dans ce qui suit deux de ces outils qui sont les plus connus : la matrice de confusion, la courbe ROC.

– **a :Matrice de confusion**
On l'appelle aussi tableau de classement. Elle prend la forme suivante :
Où n_{11} est le nombre de mauvais client, n_{22} est le nombre de bon client, n_{12} est le nombre de mauvais client considéré comme bon client, n_{21} est le nombre de bon

	Mauvais clients (Y = 0)	Bons clients (Y = 1)	Total
Considérés mauvais ($\widehat{Y} = 0$)	n_{11}	n_{12}	$n_{1.}$
Considérés bons ($\widehat{Y} = 1$)	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

TABLE 2.1 – Tableau de classment

client client considéré comme mauvais client, $n_{2.}$ est le nombre de client considérés (prédit) comme mauvais et $n_{1.}$ est le nombre de client considérés (prédit) comme bon. L'avantage d'une telle matrice est qu'elle montre rapidement si le système parvient à classifier correctement les individus dans leurs groupes.
A partir de ce tableau, on calcule :

1. Le taux d'erreur de classement donné par : $m_c = \frac{n_{12}+n_{21}}{n}$.
2. Le taux de biens classés donné par : $b_c = \frac{n_{11}+n_{22}}{n}$.

– **Courbe de ROC**
L'appellation ROC vient des abréviations du nom anglais donné à cette courbe : (Receiver operating caracteristics) [19].

Définition 3.3.8 : La courbe ROC est défini par la représentation graphique de la proportion $(1-\beta(s))$ des positifs détectés par la méthode en fonction de la proportion des faux positifs $\alpha(s)$ lorsque s varie. On peut noter d'après le graphique ci-dessous, que :

- Lorsque $\alpha(s) = 0{,}1 - \beta(s) = 0$
- Lorsque $\alpha(s) = 1{,}1 - \beta(s) = 1$
- $1 - \beta(s)$ et $\alpha(s)$ évoluent dans le même sens

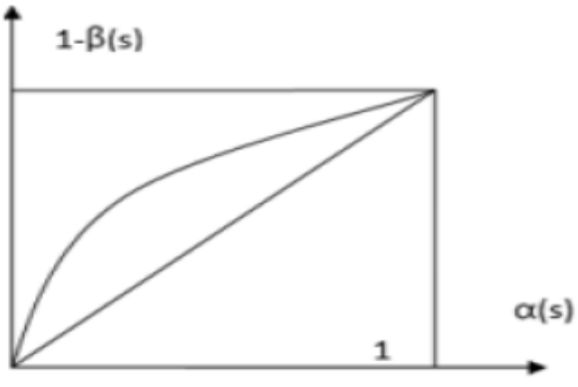

FIGURE 2.2 – Courbe ROC

On peut noter que :

- Lorsque les deux distributions de $S(x)$(dans les deux groupes)sont bien distinctes, la courbe ROC est horizontale. En effet, lorsque $\alpha(s)$ passe de 0 à 1,1 $-\beta(s)$ prend toujours la valeur 1(Modèle le plus performant).
- Lorsque les deux distributions de $S(x)$ sont confondues, la courbe ROC coincide avec la première bissectrice.(Modèle le moins performant). Ces remarques conduisent à considérer la surface sous la courbe ROC(AUC) comme indicateur synthétique de la performance d'un modèle :
  - $AUC \simeq 1$ modèle très performant.
  - $AUC \simeq \frac{1}{2}$modèle non performant

Cet indicateur permet ainsi de choisir entre modèles : On retient le modèle ayant le AUC le plus élevé.

La surface AUC peut être calculée en utilisant la méthode de trapèzes une fois que la courbe ROC est tracée. Mais en pratique, on utilise la méthode des paires concordantes. On démontre en effet que :$AUC = P(S_1 > S_2)$ où $S1$ et $S2$ sont respectivement les scores de deux clients tirés d'une manière indépendante dans le groupe des bons clients puis dans le groupe des clients défaillants.

Dans les applications, cette probabilité est estimée par la proportion des paires concordantes. Le nombre de paires s'élève à n_1, n_2. Parmi ces paires, celles où le score du positif dépasse celle du négatif sont appelées paires concordantes.

2.3 Régression logistique

La fonction logistique a été inventée au 19^{me} siècle pour la description de la population et l'évolution des réaction chimiques autocatalytique par Malthus en 1789 la notion de fonction logistique sera revue par l'astronaute statisticien belge Alphonse Quételet (1795-1874) par le fait que l'extrapolation de la croissance de la population devrait conduire à l'impossible de valeur. Il demande également a son élève Pierre François Verhulst de réfléchir au problème ce dernier publiera trois articles entre 1834 et 1847 donc le premier sera éditer par son prof Alphonse Quételet en 1838.

2.3.1 Modélisation

Considérons toujours :$E_1 \cup E_2 = E$ et $E_1 \cap E_2 = \varnothing$ où cardinal de E est égale à n , cardinal de E_1 est égale à n_1 et cardinal de E_2 est égal à $n_2(n_1 + n_2 = n)$;E_1représente le groupe des clients en bonnes états et E_2 représente le groupe des clients défaillants . Nous voulons prédire le groupe d'appartenance d'un client à partir des p variables explicatives. La probabilité apriori est $P(Y = k)$ avec $k \in \{0; 1\}$. Soit x une observation de la variable explicative X. Nous cherchons la probabilité a postériori $P(Y = k/X = x)$ avec $k \in \{0; 1\}$.

Définition 3.4.1 : (Rouvière, 2015) : Soit Y une variable à expliquer binaire et $X = (X_1, X_2, ..., X_p)$ p variables explicatives. Le modèle de régression logistique est défini comme suit :

$$ln\left(\frac{P(Y=1/X=x)}{1-P(Y=1/X=x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p = x'\beta$$

Simplement Logit $(P(Y=1/X=x) = x'\beta)$.

- Où $\beta = (\beta_0, \beta_1, ..., \beta_p)$ est une matrice uni-colonne de dimension $(p+1)\times 1$ de terme général $(\beta_i)_{0\leq i\leq p}$ et à valeurs dans $\mathbb{R}$.
- *Logit* $:p \to \ln(\frac{p}{1-p})$ est une bijective de]0,1[dans $\mathbb{R}$. C'est la fonction de lien qui permet d'exprimer le $\ln(\frac{P(Y=1/X=x)}{1-P(Y=1/X=x)})$ en fonction des variables explicatives.

La Représentation graphique de la fonction logistique est comme suite :

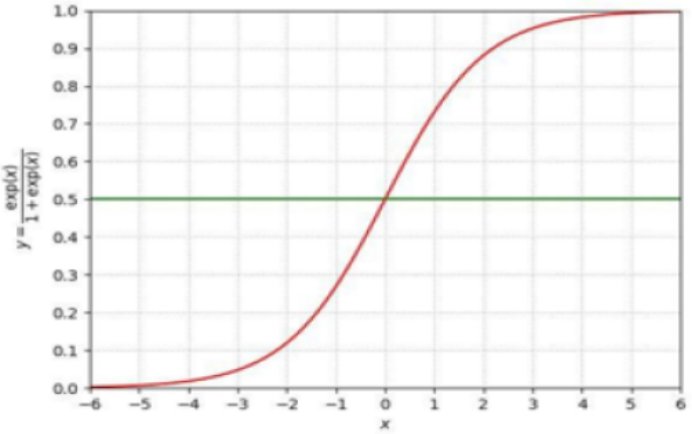

FIGURE 2.3 – Graphe de la fonction logistiqueest

Proposition 3.4.1 :

$$P(Y=1/X=x) = \frac{exp(x'\beta)}{1+exp(x'\beta)}$$

Preuve : On a

$$P(Y=1/X=x) = \frac{p_1 f_1(x))}{p_1 f_1(x)+p_2 f_2(x))}$$

grâce à la formule de Bayes en divisant les membres de l'égalité de droite par $p_2 f_2(x)$. On obtient :

$$P(Y=1/X=x) = \frac{\frac{p_1 f_1(x)}{p_2 f_2(x)}}{1+\frac{p_1 f_1(x)}{p_2 f_2(x)}} \tag{2.3}$$

Où f_1 etf_2 sont des densités du vecteur X respectivement dans les groupes E_1 et E_2.

Or :

$$\frac{1}{P(Y=1/X=x)} = 1 + \frac{p_2 f_2(x)}{p_1 f_1(x)}$$

$$\begin{aligned} \frac{1}{P(Y=1/X=x)} = 1 + \frac{p_2 f_2(x)}{p_1 f_1(x)} &\iff \frac{1}{P(Y=1/X=x)} - 1 = \frac{p_2 f_2(x)}{p_1 f_1(x)} \\ &\iff \frac{1-P(Y=1/X=x)}{P(Y=1/X=x)} = \frac{p_2 f_2(x)}{p_1 f_1(x)} \\ &\iff \frac{P(Y=1/X=x)}{1-P(Y=1/X=x)} = \frac{p_1 f_1(x)}{p_2 f_2(x)} \end{aligned} \tag{2.4}$$

En passant au Logarithme de part et d'autre de l'équation (3.4) on obtient :

$$\ln\left(\frac{P(Y=1/X=x)}{1-P(Y=1/X=x)}\right) = \ln\left(\frac{p_1 f_1(x)}{p_2 f_2(x)}\right)$$

Or par définition nous avons :

$$\ln\left(\frac{P(Y=1/X=x)}{1-P(Y=1/X=x)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p = x'\beta$$

ce qui nous permet d'écrire

$$\ln\left(\frac{p_1 f_1(x)}{p_2 f_2(x)}\right) = x'\beta$$

c'est-à-dire

$$\frac{p_1 f_1(x)}{p_2 f_2(x)} = exp(x'\beta)$$

.

en remplaçant donc

$$\frac{p_1 f_1(x)}{p_2 f_2(x)} par exp(x'\beta) \tag{2.5}$$

dans (3.3) on obtient :

$$P(Y=1/X=x) = \frac{exp(x'\beta)}{1+exp(x'\beta)}$$

La linéarité d'un score constitue un avantage important en pratique du fait notamment de la facilité de la mise en oeuvre informatique. Néanmoins les hypothèses de normalité et d'homoscédasticité peuvent s'avérer dans certain cas peu réalistes et inadaptés. Cependant en restant dans le cadre Bayésien, il peut être noté que la linéarité du score peut être obtenue sous autre hypothèse concernant les lois conditionnelles. L'analyse discriminante logistique, n'impose aucune loi particulière à suivre par les descripteurs. Elle se donne comme hypothèse de base la linéarité du logarithme du rapport de vraisemblance :

$$\ln\left(\frac{f_1(x)}{f_2(x)}\right) = \beta' x \tag{2.6}$$

Elle est ainsi plus générale que l'analyse discriminante bayésienne avec normalité et homoscédasticité des descripteurs.

2.3.2 Expression de la probabilité à postériori

Lorsque le score est linéaire, les probabilités à postériori prennent une forme particulière qui est celle de la loi logistique. En effet, notons $p(x)$ (respectivement $q(x)$) la probabilité à postériori d'appartenance au groupe E_1(respectivement E_2) :

$$p(x) = P(e \in E_1/X = x) = \frac{f_1(x)p_1}{f_1(x)p_1 + f_2(x)p_2}$$

Soit en divisant le numérateur et le dénominateur par $f_2(x)p_2$ et compte tenu de l'hypothèse de linéarité de logarithme du rapport de vraisemblance :

$$P(x) = \frac{e^{x'\beta}}{1 + e^{x'\beta}}.$$

on a :

$$q(x) = P(e \in E_2/X = x) = \frac{f_2(x)p_2}{f_1(x)p_1 + f_2(x)p_2}$$

En divisant toujours le numérateur et le dénominateur par $f_2(x)p_2$ on obtient :

$$q(x) = \frac{1}{1 + e^{x'\beta}}$$

.

Ce qui montre que les probabilités à postériori prennent la forme de la fonction de répartition d'une loi logistique. Pour les estimations ici on utilise la méthode du maximum de vraisemblance pour ses bonne propriété asymptotique.

1. **Estimation des paramètres**
   On utilise la méthode du maximum de vraisemblance pour ses bonnes propriétés asymptotiques. Soit un échantillon indépendant de n observations :

$$(y_{1,};x_{1,1};x_{1,2};...;x_{1,j};...;x_{1,p}),\quad (y_{2,};x_{2,1};x_{2,2};..;x_{2,j};...;x_{2,p}),...,$$
$$(y_{i,};x_{i,1};x_{i,2};...;x_{i,j};...;x_{i,p}),\quad (y_{n,};x_{n,1};x_{n,2};...;x_{n,j};...;x_{n,p}).$$

La vraisemblance de cette échantillon, est par définition, sa probabilité de réalisation :

$$L(y,x,\beta)=P((Y_1=y_1,X_1=x_1),...,(Y_i=y_i,X_i=x_i),...,(Y_n=y_n,X_n=x_n))$$

Soit compte tenu de l'indépendance de l'échantillon,

$$\begin{aligned}L(y,x,\beta) &= \textstyle\prod_{i=1}^{n} P(Y_i=y_i,X_i=x_i)\\ &= \textstyle\prod_{i=1}^{n} P(Y_i=y_i|X_i=x_i)P(X_i=x_i)\ car\\ &P(Y_i=y_i|X_i=x_i)=\frac{P(Y_i=y_i|X_i=x_i)}{P(X_i=x_i)}\end{aligned}$$

en passant aux logarithmes :

$$ln(L(y,x,\beta))=\textstyle\sum_{i=1}^{n}\ln(P(Y_i=y_i|X_i=x_i))+\sum_{i=1}^{n}\ln(P(X_i=x_i))$$

D'où en remplaçant

$$\begin{aligned}\ln(L(y,x,\beta)) &= \ln P(x_i)^{y_i}(1-P(x_i))^{1-y_i}+\textstyle\sum_{i=1}^{n}\ln(P(X_i=x_i))\\ &= \textstyle\sum_{i=1}^{n} y_i(\ln(P(x_i))-\ln(1-P(x_i)))\\ &+\textstyle\sum_{i=1}^{n}\ln(P(X_i=x_i))\\ &= \textstyle\sum_{i=1}^{n} y_i(x_i'\beta)\ln(e^{x_i'\beta})+\sum_{i=1}^{n}\ln(P(X_i=x_i))\end{aligned}$$

Notons $\ln L(y,x,\beta)$ par LL par la suite.
La maximisation de L passe par l'annulation de ses dérivées premières.
La solution de cette équation ne peut pas être déterminée explicitement. On utilise à cet effet un algorithme de résolution numérique, le plus connu étant l'algorithme de Newton-Raphson.β est ainsi estimé par $\widehat{\beta}$.

2. **Estimation des probabilités apostérioris**
D'après la proposition On a :

$$P(Y=1|X=x)=\frac{\exp(x'\beta)}{1+\exp(x'\beta)}\ \ et\ P(Y=0|X=x)=\frac{1}{1+exp(x'\beta)} \tag{2.7}$$

$$\widehat{p}(Y=1|X=x)=\frac{\exp(x'\widehat{\beta})}{1+\exp(x'\widehat{\beta})}\ \ et\ \widehat{p}(Y=0|X=x)=\frac{1}{1+exp(x'\widehat{\beta})} \tag{2.8}$$

La règle Bayésienne permet d'affecter une nouvelle observation à une classe d'appartenance . Ainsi, pour déterminer la classe d'appartenance d'une nouvelle observation, on compare les probabilités, $\widehat{p}(Y=1|X=x)$ avec $\widehat{p}(Y=0|X=x)$
en appliquant le logarithme on obtient :

$$\ln\left\{\frac{\widehat{p}(Y=1|X=x)}{\widehat{p}(Y=0|X=x)}\right\} = ln\left(\widehat{p}(Y=1|X=x)\right)$$

En effet le rapport de ses probabilités est comparé à 1 en passant la fonction ln on a donc la règle de décision suivante :

$$ln\left(\widehat{p}(Y=1|X=x)\right) > 0 \Leftrightarrow x'\beta > 0$$

donc la nouvelle observation est affectée au groupe des mauvais clients ;

$$ln\left(\widehat{p}(Y=1|X=x)\right) < 0 \Leftrightarrow x'\beta < 0$$

donc la nouvelle observation est affectée au groupe des bons clients ;
dans le cas ou $x=0$ on ne peut rien dire.
La fonction score est donc $S(x) = x'\beta$.

2.3.3 Test de paramètres

La théorie du maximum de vraisemblance nous donnant la loi (asymptotique) des estimateurs, il est possible de tester la significativité des variables explicatives. Pour cela, deux tests sont généralement utilisés :

1. Le test de Wald ;
2. Le test du rapport de vraisemblance ou de la déviance.

Les hypothèses sécrivent : $\begin{cases} H_0 : \beta H_0 : \beta_0 = \beta_1 = ... = \beta_{q-1} = 0 \\ H_1 : \exists k \in 0, ..., q-1 | \beta_k \neq 0 \end{cases}$

(a) Test de Wald

Il est basé sur la précision des estimateurs et test.
On note
$\beta_0, .., \beta_{q-1}$ le vecteur composé des q premières composantes de β et $\widehat{\Sigma}_0^{-1}, ..$
$., \widehat{\Sigma}_{q-1}^{-1}$, la matrice et un bloc composée des q premières lignes et $q-1$ de colonnes. Il est facile de voir que sous H_0,

$$\beta_0', ..., \beta_{q-1}'\widehat{\Sigma}_0^{-1}, .., \widehat{\Sigma}_{q-1}^{-1}\beta_0, ..., \beta_{q-1} \rightsquigarrow \chi_q^2$$

(b) Le test du rapport de vraisemblance ou de la déviance

La statistique de test est basée sur la différence des rapports de vraisemblance entre le modèle complet et le modèle sous H_0. On note β_{H_0} l'estimateur du maximum de vraisemblance contraint par H_0 (il s'obtient en supprimant les q premières variables du modèle). On a alors sous H_0,

$$\begin{aligned} LR &= -2\ln\left(\frac{\textit{modèle reduit}}{\textit{modèle complèt}}\right) \\ &= [-2LL(\textit{modèle reduit})] - [-2LL(\textit{modèle complet})] \\ &= LL(\widehat{\beta}) - LL(\widehat{\beta}_{H_0}) \end{aligned}$$

danc

$$LR \rightsquigarrow \chi^2_q$$

2.3.4 Sélection des variables explicatives significatives

Dans les études réelles, beaucoup de variables disponibles, plus ou moins pertinentes, concurrente.Trop de variables tue l'interprétation, il y a le danger du sur-apprentissage anssi,

Le critère de choix du modèle a pour objectif de comparer les modèles qui ne sont pas forcément emboîtés les uns dans les autres. Il permet donc de choisir le modèle le plus parcimonieux (c'est-à-dire un modèle dans lequel les variables explicatives expliquent parfaitement la variable à expliquer). Pour cela, on pénalise la vraisemblance par une fonction de nombre de paramètres. En effet, le meilleur modèle est celui qui a la plus grande log-vraisemblance. Cependant, en présence d'un modèle saturé (Modèle ayant un nombre maximum de paramètres), la log-vraisemblance sera maximum, car la log vraisemblance augmente avec la complexité du modèle, et choisir le modèle qui maximise la log-vraisemblance revient à choisir le modèle saturé. Ce modèle est clairement surparamétrés. C'est pourquoi, dans la perspective d'obtenir un modèle de taille raisonnable, il sera donc bon de la pénaliser par une fonction du nombre de paramètres. Deux critères répondent à ces spécifications.

- Par définition, Le critère de choix Akaike Informative Critérion (AIC) pour un modèle à p variables est défini par :

  $$AIC = 2L + 2p$$

  Où

  L est la log-vraisemblance du modèle de régression logistique et $2p$ est la fonction de nombre de paramètres ;
- Un autre critère de choix du modèle est le Bayesiam Informative Critérion (BIC), pour un modèle à p paramètres, il est défini comme suit :

$$BIC = 2L + p\ln(n)$$

Pour chaque modèle concurrent, le critère de choix est calculé et le modèle qui possède le plus petit AIC ou BIC est celui que l'on retient.

- **Sélection ascendante (option Forward)**
  Elle consiste à ne présenter aucune variable a l'étape initiale . A chaque étape on fait entrer la variable qui contribue le plus au pouvoir discriminant du modèle, la sélection s'arrête lorsque toutes les variables sont intégrées dans le modèle ou lorsque le critère de choix ne diminue plus .

- **Sélection descendante (option Backward)**
  On démarre avec le modèle complet (construit avec toutes les variables) a chaque étape, la variable contribuant le moins au pouvoir discriminant du modèle est éliminée. La sélection sarrête On s'arrête lorsque toutes les variables sont retirées du modèle ou lorsque le critère de choix ne diminue plus.

- **Sélection mixte (option Stepwise)**
  Elle consiste à démarrer comme dans la procédure ascendante ; dès qu'une variable entre dans le modèle, on vérifie compte tenu de cette entrée si l'une des variables déjà présentes est susceptible d'être éliminée. La sélection s'arrête quand on ne plus ajouter ou éliminer de variables.

2.3.5 Validation du modèle

1. **Test d'Hosmer-Lemeshow**
   Il permet de déterminer la qualité d'ajustement du modèle aux données, si l'ajustement est correct, les valeurs prédites seront proches des valeurs observées. On calcule pour chaque observation la probabilité prédite $\widehat{p}(Y = 1/X = x_i)$. On classe les observations par déciles de probabilités prédites. On compare dans chaque classe les effectifs observés et les effectifs théoriques.
   - Si dans chaque classe ces deux effectifs sont proches alors le modèle est calibré
   - S'il existe des classes dans lesquelles les effectifs sont trop différents, alors le modèle est mal calibré
     H_0 : les probabilités théoriques sont proches de celles observées (modèle calibrée)
     H_1 : les probabilités théoriques sont différentes des observées (modèle non calibrée)
     Pour chaque observation x_i,on calcule la probabilité prédite par le modèle $\widehat{p}(Y = 1/X = x_i)$, que l'on classe par ordre croissant ;
   - Ensuite on subdivise ces probabilités en G groupes de tailles égales. En général pour un échantillon très grand, on attribue à G la valeur 10. Le dernier groupe, celui des $\widehat{p}(Y = 1/X = x_i)$, les plus grands, possède un effectif supérieur aux autres. Par la suite, on note m_g l'effectif du groupe g, f_g l'effectif des cas $(Y = 1)$

dans le groupe g, μ_g la moyenne des $\widehat{p}(Y = 1/X = xi)$, dans le groupe g.
La statistique de test d'Hosmer-Lemeshow est la suivante :

$$\widehat{C} = \sum_{g=0}^{G} \frac{(f_g - \mu_g m_g)^2}{\mu_g m_g (1 - m_g)} sous H_0 \widehat{C} \rightsquigarrow \chi^2_{(G-2)}$$

2. **Validation croisée**
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 suit Pour $i = 1$ à n, on construit la règle de décision sur la base privée de son i^{ime} 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 mal classés de la sorte. L'estimation du taux d'erreur ainsi obtenu est pratiquement sans biais, 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 des observations communes ont tendance à se ressemble.

Conclution

Le crédit scoring est généralement considéré comme une méthode d'évaluation du niveau du risque associé à un dossier de crédit potentiel. Cette méthode implique l'utilisation de différentes techniques statistiques comme l'analyse discriminante de Fisher et la régression logistique pour aboutir à un modèle de scoring basé sur les caractéristiques du client, il sera classé par le modèle comme : Bon Payeur ou Mauvais Payeur.

3

Implémentation et discussion des résultats

Introduction

Ce chapitre est essentiellement concentré sur l'élaboration d'une fonction score, grâce aux deux méthodes vu au chapitre précédent, à savoir la méthode de Fisher, et la régression logistique. Le but essentiel est de choisir la méthode la plus optimale qui présente moins d'erreur par rapport aux données.

3.1 Constitution de l'échantillon

C'est étape consiste à définir une représentativité statistique et homogénéité des échantillons. Il faut disposer de deux sous-échantillons :

1. un échantillon d'entreprise ayant connu l'événement à détecter (défaut, faillite),
2. un échantillon d'entreprise ne l'ayant pas connu, réputées saines.

Notre basse de donnée représente un échantillon de 46 entreprises, ils se composent de 23 entreprises jugées comme défaillantes et 23 entreprises saines. Nous notons ici que les entreprises sont représentées par la variable Y, les entreprises défaillantes prennent la valeur 0, alors que les entreprises non défaillantes prennent la valeur 1.

Le tableau suivant résume la classification des entreprises saines et défaillantes dans les deux secteurs d'activités, Commerce et industrie.

Principales caractéristiques	Entreprises saines	Entreprises défaillantes
Secteurs d'activités :		
Commerce :	10	15
Industries :	13	08

TABLE 3.1 – Tableau de donnée

3.2 Sélection des variables

Il s'agit ici de sélectionner les variables importantes et qui ont le pouvoir de discrimination importante et d'éviter de répéter les variables.

– **(A) :Choix des ratios**

On a distingué trois grandes catégories de ratios représenté sur le tableau suivant :

Aspect	Ratios	Intitule
Ratios de structure	R_1	Ratio d'autonomie financière
	R_2	Ratio de trésorerie immédiate
	R_3	Ratio d'équilibre financier
Ratios d'activité	R_4	Part des frais financiers dans la valeur ajoutée
	R_5	Part des frais financiers dans la valeur ajoutée
	R_6	Ratio crédit client
Ratios de rentabilité	R_7	Rentabilité financière

TABLE 3.2 – Tableau des ratios

Ratios	Valeur minimal	Valeur maximal	Moyenne	Ecart-type
R_1	0, 42	1, 00	0.71	0.41
R_2	0, 00	0, 54	0.27	0.38
R_3	0, 45	12, 91	6.725	8.75
R_4	0, 01	0, 44	0.22	0.31
R_5	0, 12	3, 29	1.705	2.24
R_6	0, 01	2, 84	1.42	2.00
R_7	0, 00	0, 21	0.105	0.15

TABLE 3.3 – Échantillon des valeurs des ratios

Ce tableau montre que les valeurs prises par les sept ratios retenus sont dispersées. Elles diffèrent fortement d'une entreprise à une autre.

Pour avoir une idée préliminaire sur le pouvoir de discrimination de chaque ratio, nous utilisons le test de différence de moyennes de student relatives à chaque ratio, entre les entreprises défaillantes, et les entreprises saintes. Les résultats de ce test se résument dans le tableau suivent :

Ratios	Entreprises saines	Entreprises défaillantes	Ecart	Test-t	signification
R_1	0, 9617	0, 7478	0, 2139	5, 019	$0,000^{*}$
R_2	0, 1396	7, 783	7.64	2, 538	$0,015^{**}$
R_3	3, 5343	1, 2561	2, 2783	3, 498	$0,01^{*}$
R_4	0, 11813	0, 1030	0, 0151	2, 579	$0,013^{**}$
R_5	1, 9548	1, 1913	0, 7635	3, 234	$0,02^{*}$
R_6	0, 4383	1, 3722	0, 9339	-4, 928	$0,00^{*}$
R_7	8, 913	5, 522	3, 391	2,256	$0,29^{*}$

TABLE 3.4 – Tableau des moyennes comparées des ratios retenus

Remarque :
$*$: signifé que le ratio et significative à 5% ;
$**$: signifé que le ratio et significative à 10%

Les premiers résultats de notre étude montrent que la moyenne relative au ratio 1 (ratio d'indépendance financière) est plus élevée chez les entreprises saines (0, 96) que chez les entreprises défaillantes (0, 75). La différence entre ces deux moyennes est positive (+0, 21), et statistiquement significative. Ce ratio est discriminant selon le test de student. Cette situation s'applique également pour les ratios R2 (Ratio de trésorerie immédiate), R3 (ratio de l'équilibre financier), R5 (ratio fournisseurs) et R7 (Rentabilité financière). Par contre, la moyenne relative au ratio 6 (ratio clients) est plus élevée chez les entreprises défaillantes (1, 37) que chez les entreprises saintes (0, 44). Le délai client est plus long chez les entreprises défaillantes que chez les entreprises saintes. Or, ces remarques élémentaires ne nous

permettent pas de trancher définitivement sur les variables les plus discriminantes.

– **(B)Construction de la fonction discriminante**

1. **Par la méthode de l'analyse discriminante de Fisher**

Le traitement de notre base des données nous a permis d'identifier la fonction score suivante :

fonction	
R_1	2, 071
R_2	-0, 036
R_3	0, 070
R_4	1, 662
R_5	0, 706
R_6	-1, 219
R_7	8, 224
constante	-2, 772

TABLE 3.5 – Résulta de traitement

Donc notre fonction score peut s'écrire ainsi :

$$\begin{aligned} S_1(X) &= 2,071R_1 - 0,036R_2 + 0,070R_3 + 1,662R_4 \\ &+0,706R_5 - 1,219R_6 + 8,224R_7 - 2,772 \end{aligned}$$

L'affectation aux groupes se fera en fonction des centroïdes de ces derniers, c'est-à-dire par comparaison avec un score discriminant "moyen" pour chaque groupe. Ce score moyen est calculé à partir de la fonction discriminante, où l'on remplace les valeurs individuelles par les moyens des variables indépendantes pour le groupe dont on s'occupe. Les scores discriminants moyens pour les deux groupes sont donnés ainsi :

fonction 1	
Appartenance	Scores moyens
0(entreprises défaillantes)	-1, 343
1(entreprises saines)	+1, 343

TABLE 3.6 – Tableau de valeur des Scores moyens

Chaque score individuel discriminant est ensuite comparé aux deux scores moyens et affecté au groupe dont-il est le plus proche.

Généralement, on teste la capacité prédictive de la fonction score soit par des tests statistiques faisant appel à des hypothèses probabilistes, soit par un test pragmatique par le biais de la matrice de confusion. Concernant les premiers tests, nous utilisons la corrélation canonique et Lambda de Wilks.

Fonction	Valeur propre	Pourcentage de la variance	pourcentage cumulé	Corrélation canonique
1	1.886	100.0	100.0	0.808

TABLE 3.7 – Résulta du test de corrélation canonique

Plus la corrélation canonique est proche de 1, plus le modèle soit meilleur . Dans notre cas, la corrélation canonique est égale à 0, 808. Ce résultat est très encourageant parce que cette valeur confirme un pouvoir discriminant assez important de la fonction discriminante extraite.

Le tableau suivent représente le résultat obtenu a partir de test de la fonction discriminante

Test de la/ou des fonctions	Lambda de Wilks	khi-deux	ddl	Signification
1	0, 346	42, 929	7	00.0

TABLE 3.8 – Résultat de test de la fonction discriminante

La valeur de Lambda de Wilks étant faible, et égale à 0, 346, et donc plus proche de 0 que de 1, avec un khi-deux ayant un degré de signification nul. Cela veut dire qu'au niveau global, la différence des moyennes des groupes est significative. Pour s'assurer que la fonction discriminante classifie bien les entreprises en sous-groupes, on analyse la matrice de confusion qui regroupe les entreprises bien classées et les mal classées. C'est le moyen le plus utilisé.

La matrice de confusion de notre fonction score se présente comme suit :

Appartenance			Classe d'affectation prévu		Total
			0	1	
Originale	Effectif	0	19	04	23
		1	03	20	23
	Pourcentage	0	82.6	17.4	100
		1	13.0	87.0	100

TABLE 3.9 – Matrice de confusion

Cette matrice fait ressortir que la fonction score extraite ci-dessus permet de classer $84,78\%(\frac{19+20}{46})$ des entreprises correctement un ans avant l'occurrence de la défaillance de c'est entreprises. Ce taux peut se décortiquer ainsi :

- Le pourcentage des bien classées pour les entreprises saines est égal à $(\frac{20}{23}) = 87\%$.
- Le pourcentage des bien classées pour les entreprises défaillantes est égal à $(\frac{19}{23}) = 82.6\%$.

Par contre, le taux d'erreurs(entreprises mal classées) est égal seulement $(\frac{7}{46}) = 15.21\%$.

Toute fois, on distingue pour ce taux entre : l'erreur du premier type (classer une entreprise défaillante par l'utilisation de la fonction score parmi les entreprises saintes) : ce taux est égal à $(\frac{4}{23}) = 17,4\%$, et l'erreur du second type (classer une entreprise sainte comme une entreprise défaillante par le modèle) : ce taux est égal à $(\frac{3}{23}) = 13\%$.

- **Par la méthode régression logistique**

Pour construire notre fonction nous utilisons python ou nous faisons appelle a la fonction fit() qui nous permet d'obtenir le tableau suivant :

Ratios	Coef	Errere de std	valeur de z	Pr(>|z|)	$IC_{90\%}$
R1	42,85	28, 15	1, 52	0, 04	[0, 45 ; 5, 30]
R2	-3, 02	2, 27	-1, 33	0, 10	[-27, 02 ; -2, 10]
R3	0, 40	28, 15	0, 01	0, 14	[-1, 30 ; 7, 04]
R4	-7, 49	5, 90	-1, 32	0, 17	[0, 45 ; 5, 30]
R5	0, 53	0, 26	2, 03	0, 03	[0, 30 ; 3, 96]
R6	-2, 72	2, 04	-1, 33	0, 13	[-12, 40 ; 0, 30]
R7	10, 49	5, 86	1, 88	0, 05	[0, 51 ; 9, 26]
Constance	17, 43	9, 73	1, 79	0, 07	[0, 21 ; 5, 60]

TABLE 3.10 – Résultat du modèle de régression logistique

Apartire des différentes valeurs de $Pr(z)$, nous constatons que les ratios retenus pour notre fonction sont R_1, R_5 et R_7, on obtient la fonction score suivante :

$$S_2(X) = 42,85R_1 + 0,53R_5 + 10,49R_7 + 17,43$$

Notre fonction étant élaborer il nous reste juste a validé le modèle avec le test d'Homser Lemeshow en Python, ou nous été assez compliqué car Python ne dispose pas de bibliothèque capable pour le calcul de la formule Homser Lemeshow, nous avons utilisé la fonction hoslem. Test de la bibliothèque Resource Selection du logiciel R qui nous permis d'effectuer ce test et nous avons eu le résultat suivant :

Au seuil de significativité de 5%, l'ajustement du modèle est bon car la p-value de la statistique chi-deux à 8 degrés de liberté est supérieure à 5%. Par conséquent l'hypothèse H_1 est rejetée, on conclut donc que le modèle est calibré donc valide.

Chix deux	ddl	p-value
36	8	$8,9710^{-5}$

TABLE 3.11 – Résulta de test

Conclusion

Aprér avoire présenter et calculer la fonction de score pour les deux technique a savoire l'analyse discriminante de Fisher, et la régression logistique, nous pouvons dire que l'analyse discriminante de Fisher est le meilleur modèle pour notre base de données, car elle a considéré tous les ratios sont discriminant et permettent d'expliquer l'appartenance d'un client à une modalité Y (saine ou défaillante).

Conclusion générale

A travers ce mémoire, réaliser durant mon stage au sein de la banque BNA Agence 586 de Tazmalt, Ce mémoire ma offert l'opportunité de concrétiser mes connaissances théorique acquises et de les confronter à la pratique dans le service de crédit de la banque BNA.

Ce mémoire m'a permis d'élaborer deux techniques statistiques émanant de la méthode de scoring, ces deux techniques qui sont l'analyse discriminante de Fisher, et la régression logistique. Nous avons présenté la démarche théorique et pratique de la construction de la fonction score pour chaque technique. La validation de l'analyse discriminante de Fisher s'est faite grâce à la corrélation canonique et du test Lambda de Wilks, qui m'ont permis d'obtenir la fonction $S_1(X)$.

Pour ce qui est de la régression logistique, le modèle a été valide grâce au test d'Homser Lemeshow, et nous avons obtenu la fonction $S_2(X)$.

On peut dire que l'analyse discriminante de Fisher est le meilleur modèle pour notre base de données, car elle considère tous les ratios en discriminant et permettant d'expliquer l'appartenance d'un client à une modalité Y.

Il faut noter que la taille de l'échantillon de mon travail était réduit, ce qui ma empêché de définir un seuil, et ma fournis des résultats acceptables, ceci est dû au manque d'accès aux données réelles.

Comme perspectives de recherche on peut citer :

1. Le présent travail peut être étendu en tenant compte d'un plus grand nombre et d'une plus grande variété de variables, notamment, celles qualitatives.

2. Bien que ces deux méthodes soient classiques dans les recherches nous pourrons établir une comparaison avec des nouvelles méthodes comme le réseaux neurone artificiel.

Bibliographie

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

[2] ALTMAN.E.I, S. *Credit risk measurement.* Developments over the last 20 years-(1998), p.1721-1742.

[3] Chavent, M. *Analyse discriminante linéaire et quadratique.* 2015.

[4] CHIBEL.Z, BAMOUSSE.Z, E. *Prévision du risque de crédit : ambition du scoring analyse comparative des pratiques de crédit scoring.* Article International Journal of Management et Marketing Research (MMR), Université Hassan Premier.

[5] COHEN.E. *Analyse financière.* édition économico, paris 1990.

[6] COUSSERGUES.S., BOURDEAUX.G, P. *Gestion de la banque-8e ed. : Normes et règlementation à jour.* Nouvelles stratégies bancaires , Vol. 1. Dunod, 2017.

[7] DIETSCH.M, P. *Mesure et gestion du risque de crédit dans les institutions financières.* Édition Revue Banque, Paris 2003.

[8] DUTAILLIS.G.P, Hamel.J, R. L. L. *Le risque du crédit bancaire.* Éditions Riber, 1967, pp 45.46.

[9] FERRONIERE.J, C. *Les opérations de banque.* Dunod, 1963, page 187/190-192/193-196.

[10] GONZALEZ.P.L. *Calcul d'un score (scoring) Application de techniques de discrimination.* 2019.

[11] Guigal.M.M. *Le guide de la banque, Comptes, carte bancaire, services.* Edition ComprendreChoisir.com, paris 2011.

[12] GUIZANI.A. *Traitement des dossiers refusés dans le processus d'octroi de crédit aux particuliers.* Thèse de doctorat en Sciences de Gestion, L'institut Supérieur de Gestion, Université Sousse, Tunisie, 2014.

[13] GUIZANI.A. *Traitement des dossiers refusés dans le processus d ?octroi de crédit aux particuliers.* Thèse de doctorat en Sciences de Gestion, L'institut Supérieur de Gestion, Université Sousse, Tunisie,, 2014.

[14] Hassen, M. *Cours de scoring.* 2013,2014.

[15] Jimbo, H. *modelling Risk of Non-Repayment of Bank Credit by Method of Scoring.* Journal of Advances Statistics 4.

[16] KHAROUBI.C, T. *Analyse du risque de crédit.* Banque et Marchés , RB édition, 2016.

[17] MEYSSONNIER.L. *banque : mode d ?emploi.* édition EYRROLLLES, 1992.

[18] ROUACH.M, NAULLEAU.G, T. *le contrôle de gestion bancaire et financière.* revue de banque, 1994, p 310.

[19] SAPORTA, G. *Sensibilité, spécificité, courbe ROC.* Conservatoire des arts et métiers, 2012.

[20] VIDAL.M.F, AND BARBON.F. *Credit scoring in financial ilnclusion.* 2019.

Résumé

La crise financière qui secoue le monde actuellement, notamment les défaillances successives des grandes banques, qu'ont remis sur le devant de la scène la problématique des risques bancaires, dont le risque crédit. Ce risque doit être géré actuellement par des méthodes plus sophistiquées. Dans ce mémoire nous avons présenté deux méthodes qui nous ont permet d'établir deux fonctions, à savoir l'analyse discriminante de Fisher, et la régression logistique. Ces deux fonctions nous ont permet d'évaluer les risques de non remboursement encourus par une banque au vu de nos données. Il en ressort que l'analyse discriminante de Fisher est plus efficace par rapport à la régression logistique pour l'évaluation du risque de non remboursement de crédit.

Mots clefs : Banques, ratios, risques, analyse discriminante de Fisher et régression logistique.

Abstract

The financial crisis that is currently shaking the world, particularly the successive failures of the major banks, has brought the issue of banking risks, including credit risk, back to the forefront. This risk must now be managed by more sophisticate methods.

In this work we present two methods that allow us to establish two functions, Fisher discriminant analysis, and logistic regression. Both functions allow us to evaluate the risk of non payment incurred by a bank in view of our data, it appears that Fisher discriminant analysis is more efficient than logistic regression for the evaluation risk of credit non payed.

Keys-words : Banks, ratios, risk, Fisher discriminant analysis and regression logistics.