République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la
Recherche Scientifique

المدرسة الوطنية المتعددة التقنيات
Ecole Nationale Polytechnique

Ecole Nationale Polytechnique

Département de Génie Industriel

Mémoire de Projet de Fin d'Etudes d'Ingénieur

Thème

Modélisation des résiliations des contrats d'assurance vie basé sur le Data Mining et élaboration d'un score d'attrition.

Application : Cardif El Djazair.

Présenté par :

M. Yanice AMBES.

M. Rabah GACEM.

Dirigé par :

Mme. Amel KASMI.

Mlle. Hanane SLOUGUI.

Promotion : juin 2012

ملخص

المؤسسات اليوم تواجه عددا من المشاكل الناتجة عن المنافسة وتغير الأسواق . ذهاب العملاء هو المشكل الحقيقي حاليا في مختلف المجالات لأن الزبون هو القاعدة الأساسية لتواجد المؤسسة. البحوث والدراسات الموجودة التي تتناول هذا الموضوع هي جديدة نسبيا, فإنها تحاول أن تشرح أسباب مغادرة المتعاملين باستخدام تقنيات التنبؤ لتنقيب البيانات .بناء على هذه الدراسات التي أجريت مؤخرا وعلى طريقة جديدة نقترحها لشرح المشكلة من خلال التعاريف و طرق و نموذج تصميم و تسليط الضوء عل قاط التشابه و الاختلاف من حيث اختيار الهدف من المتغيرات و التقنيات في مجال التنبؤ بتنقيب البيانات سنحاول ايجاد حل للحد من ذهاب لعملاء .بالإضافة إلى ذلك ركزت هذه الدراسة على وجود علاقة حقيقية بين استخراج البيانات وإدارة علاقات العملاء

الكلمات الرئيسية :

ذهاب العملاء, تنقيب البيانات, ادارة علاقات العملاء.

Résumé

Les organisations sont confrontées de nos jours à plusieurs problèmes qui résultent de la concurrence et de l'évolution des marchés. Les résiliations des contrats d'assurances vies constituent une vraie problématique dans différents secteurs d'assurances, car le client est l'une des raisons d'être de l'organisation. L'état de l'art qui traite ce sujet est relativement récent, l'un des principaux objectifs réside en la prédiction des départs des clients en utilisant des techniques de prévisions de data mining. Ce travail de recherche s'inscrit dans le cadre de la fouille de données. Basé sur des études récentes, il se propose d'expliquer la problématique à travers des définitions, des méthodes, et à travers la conception de modèles. Il souligne les points de similitudes et de différences en termes de choix de variables cibles et prédictives, et en termes de techniques data mining. De plus, cette étude met l'accent sur l'existence d'une réelle relation entre le data mining et la gestion de la relation client.

Mots clefs :

Data mining, résiliation, score d'attrition, gestion de la relation client.

Abstract

The organizations are nowadays confronted to serval problems resuling from competition and market evolution. Contract churn is a real problem in life insurance because customer i one of the reasons for the organization, it attempts to explain the reasons for churn and tries to predict the departures of customers using forecasting techniques of data mining. This research is part of the search data. Based on recent studies on the churn, it proposed to explain the problem through the definitions, methods, design models and highlights points of similarity and difference in terms of choice of target variables and predictive techniques in terms data mining. In addition, this study focused on the existence of a real relationship between data mining and customer relationship management.

Keywords:

Data mining, chun, churn prediction, client relationship management.

Dédicaces

Je dédie ce modeste travail aux personnes qui me sont les plus chers au monde, que j'aime et que j'adore, à mes chers parents.

A mon frère Ilyes et à ma famille.

A mes meilleurs amis qui m'ont aidé et encouragé, et tous ceux qui se reconnaitrons, et qui, au fil des années, sont devenus des ami(e)s ; je vous dis merci.

Enfin, à toutes les personnes que j'estime et que je respecte.

Yanice

Dédicaces

JE dédie ce modeste travail à :
Mon très cher père à qui m'adresse au ciel les voeux les plus ardents pour la conservation de sa santé et de sa vie.

A celle qui m'a transmis la vie, l'amour, le courage, à toi chère maman toutes mes joies, mon amour et ma reconnaissance.

Pour mes chers frères : Adel et Hamza .
Pour mes chères soeurs : Nawel et Nassima.

A mes meilleurs amis qui m'ont aidé et encouragé, et tous ceux qui se reconnaitrons, et qui, au fil des années, sont devenus des ami(e)s ; je vous dis merci.

Que toute personne m'ayant aidé de près ou de loin, trouve ici l'expression de ma reconnaissance.

Rabah

Remerciements

Le présent travail sanctionne la fin de notre formation de graduat en génie industriel à l'école nationale polytechnique d'Alger. Bien que ce travail soit la conjugaison de nos efforts, sa réussite et sa mise sur pied sont le fruit de plusieurs personnes qu'à divers titres et échelons ont accepté de nous porter aide.

Nos remerciements s'adressent tout d'abord à Mme. KASMI Amel, notre promotrice et enseignante d'analyse de données au Département Génie Industriel de l'Ecole Nationale Polytechnique d'Alger. Tout au long de ce travail, elle a su nous apporter un soutien constant, une disponibilité, une écoute, une confiance et des conseils précieux.

Nous tenons à remercier en tout premier lieu M. François-Xavier HUSSENOT pour nous avoir accepté au sein de son entreprise, Mlle. SLOUGUI Hanane pour nous avoir guidé durant notre stage, ses remarques, orientations et conseils nous ont prêté main forte pour arriver au bout de ce travail. Nos profondes reconnaissances vont également aux collaborateurs de Cardif-El-Djazair pour leur remarquable contribution à la réalisation de ce travail.

Il s'adresse aussi aux enseignants du Département Génie Industriel l'Ecole Nationale Supérieure Polytechnique Alger, à leur tête Mlle ABOUN, qui nous ont soutenus et encouragés lors des phases critiques et auxquels nous devons notre formation ingénieur.

Rabah et Yanice

Table des matières

Table des figures

Liste des tableaux

Introduction Générale

Au cours de ces dernières années, avec la concurrence rude, les bases de données et les résultats issus du Data Mining sont devenus la principale source d'information des décideurs, ce développement a entraîné une croissance rapide au niveau de la gestion des sources données et la manière dont elles sont traitées.

La priorité des assureurs est de posséder un fichier clients opérationnel, avec des coordonnées complètes et actualisées, enrichies par des informations économiques et comportementales. L'enjeu n'est donc pas d'accumuler les données, mais de les sélectionner, de les exploiter et pourquoi pas, de les enrichir. De ce fait les efforts sont concentrés sur les études de traçabilités, segmentations et fidélisations des clients, ainsi, un passage obligatoire par le Data Mining s'impose, car c'est le processus le plus puissant existant actuellement pour la gestion d'un nombre phénoménal de données, cette démarche permet d'extraire les informations pertinentes et très fines se trouvant dans un énorme champs de données (un diamant dans une mine) pour une seule finalité et une seule vision, satisfaire le client, essayer d'en acquérir le maximum possible et de minimiser le nombre de départ.

L'objet de ce travail s'insère dans cette préoccupation en apportant un outil d'aide à la décision relatif à la gestion du nombre de départ (résiliation ou bien attrition) des clients qui constitue une vraie problématique pour les organisations dans différents secteurs d'activités. Notre étude consiste alors à traiter et expliquer les raisons des résiliations et de prédire le départ des clients en utilisant des techniques prévisionnelles de Data Mining. Ce travail s'inscrit dans le cadre de la fouille des données et de la gestion de la relation client. En se basant sur l'état de l'art existant et relatif au Data Mining et la relation client ainsi que le contexte algérien de l'assurance vie, nous proposons un outil d'aide à la décision issu du Data Mining ayant pour objectif de lutter contre l'attrition en prévoyant les événements qui risqueront d'avoir un impact sur la relation client.

Notre travail est structuré comme suit :
Dans le premier chapitre, on présentera le contexte de l'étude, à savoir l'assurance vie et plus particulièrement le marché de l'assurance vie en Algérie, on enchainera par un chapitre " Data Mining " on présentera le concept ainsi que toute la méthodologie, on va compléter par une présentation des méthodes de data mining, cette étape importante à décrire lorsqu'on veut s'initier à la modélisation des phénomènes via le Data Mining ; et enfin, on abordera la modélisation dont le but sera de construire un outil de scoring " score d'attrition " et on montrera l'importance de chaque chapitre cité précédemment.

Chapitre 1

Présentation du contexte de l'étude

Sommaire

L'apparition de l'assurance est considérée comme un phénomène relativement récent même dans les pays développés. Dès lors, l'on peut comprendre que cette notion soit restée longtemps inconnue dans des pays en voie de développement comme l'Algérie, dans la mesure où les facteurs économiques et sociaux qui sont à la base du développement de l'assurance n'ont pas connu partout la même évolution. Le présent chapitre, après avoir exposé les généralités, nous nous pencherons sur le marché Algérien des assurances vies, ensuite nous présenterons la structure dans la quelle nous avons effectué notre étude, et finirons par présenter le produit autour du quel la problématique c'est posée.

1.1 Généralités sur l'assurance vie

1.1.1 Historique de l'assurance vie :

L'assurance sur la vie a eu des débuts plus difficiles puisqu'elle passait pour immorale dans la mesure où le décès de l'assuré était susceptible de procurer un avantage matériel à un tiers. Elle apparaissait également dangereuse pour l'assuré, dans la mesure où elle pouvait donner un intérêt au bénéficiaire de l'indemnité de "hâter le trépas de l'assuré". [fdsd]

Mais c'est encore dans le domaine maritime qu'elle se développa puisqu'il devint l'usage d'assurer les cargaisons d'esclaves comme marchandises à transporter, puis le capitaine et l'équipage, et enfin, au XVIe siècle, des Compagnie d'Anvers l'appliquèrent aux passagers. Par ailleurs, en 1653, un banquier napolitain a suggéré à Mazarin la création d'association dont les membres verseraient des cotisations dans une caisse commune, dont le contenu serait réparti, entre les membres survivants, à la fin d'une période déterminée (10 à 15 ans) : les Tontines.
La Révolution marque un coup d'arrêt au développement des assurances en France.

La loi Le Chapelier prohibe tout groupement ayant pour but la défense de "prétendus intérêts communs", et c'est ainsi qu'un décret du 24 Août 1793 a supprimé les Compagnies pratiquant des opérations d'assurance vie. Ceci n'empêchait pas Napoléon lui même de souscrire une assurance vie auprès du Lloyd's de Londres en 1813. Si la nécessité de protéger les patrimoines a donné lieu à l'invention du mécanisme contractuel de l'assurance, celui-ci a "débordé" sa vocation initiale. Mais surtout, en garantissant la solvabilité de l'assuré, et en le mettant à l'abri d'une dette de responsabilité, elle a permis de développer le domaine de la responsabilité civile, dans le but d'indemniser les victimes de dommages. Le développement de l'assurance est la pierre angulaire des systèmes d'indemnisation des sociétés modernes, laquelle a donné lieu à des assurances obligatoires dans les domaines de risques les plus importants, parmi lesquels :

- La Loi Badinter du 5 juillet 1985, qui organise l'indemnisation automatique des victimes d'accidents de la route directement par les assureurs.
- La loi Spinetta, de 1978, qui organise l'indemnisation automatique par les assureurs, et dans le cadre de l'assurance dommages-ouvrage, des désordres de construction relevant de la garantie décennale.
- La loi de 1982 concernant l'indemnisation des Catastrophes Naturelles, garantie dommages, voire, la mise en place de garanties obligatoires contre les attentats, ou les infractions, soit par un système d'assurance obligatoire, soit par des Fonds de garantie.
- La loi en préparation sur l'indemnisation de l'aléa thérapeutique et les accidents médicaux.

C'est ainsi qu'un mécanisme purement contractuel, purement soumis au principe de l'effet relatif des Conventions de l'article 1165 du Code Civil, se trouve conférer des effets bénéfiques pour des tiers, notamment en leur permettant d'agir directement contre l'assureur du responsable pour obtenir l'indemnisation directe de leur préjudice. [fdsd]

1.1.2 Définition :

La vie d'une personne peut être assurée par elle-même ou par un tiers **" article 67 du code des assurances "**. L'assurance sur la vie est un contrat par lequel en échange d'une prime, l'assureur s'engage à verser au souscripteur ou au tiers par lui désigner, une somme déterminée en cas de mort de la personne assurée ou sa survie à une époque déterminée. Ainsi donc elle couvre deux risques distincts : Le risque de décès prématuré et le risque de survie prolongée.

Cette définition qui englobe toutes les variétés d'assurance sur la vie fait apparaître en plus de l'assureur, trois personnes susceptibles d'être intéressées par le contrat : le souscripteur, l'assuré sur la tête de qui l'assurance est contractée, et le bénéficiaire appelé à recueillir le profit du contrat à la réalisation du risque. Les assurances sur la vie comprennent plusieurs catégories, qui sur le plan de l'exécution du contrat présentent quelques particularités. On classe habituellement les assurances-vie en assurances en cas de mort et assurances en cas de vie ; à ces deux catégories s'ajoutent les assurances mixtes, composées de garanties de chacune des autres. L'intérêt d'une assurance-vie, même temporaire, est sa grande crédibilité à l'égard des bailleurs de fonds : elle produit ses effets que le décès de l'assuré soit dû à une cause naturelle ou à un accident. D'autre part, les sommes stipulées payables au décès de l'assuré à un bénéficiaire désigné n'entrent pas dans la succession. Une telle disposition permet de compenser les inégalités de partage résultant des règles successorales en vigueur ou d'avantager une personne physique ou morale étrangère à la succession (personne défavorisée, entreprise ...).

Enfin, le capital assuré au profit d'un bénéficiaire déterminé ne peut être saisi par les créanciers de l'assuré. [dP]

1.1.3 Les différents produits d'assurance vie :

Les assurances de personnes garantissent l'individu contre les événements qui touchent à son existence et à sa santé : il s'agit principalement des assurances sur la vie, ou encore de celles qui couvrent les risques liés à la maladie ou aux accidents

1.1.3.1 L'assurance en cas de vie

Elle est définit comme étant un contrat en vertu duquel l'assureur s'engage à payer au bénéficiaire la somme stipulée si l'assuré est vivant à l'époque fixée par la police. Lorsqu'une assurance de capital différé est souscrite, souscripteur et bénéficiaire se confondent en une seule et même personne. Si cette personne est toujours en vie à une date précisée dans le contrat, l'assureur lui verse un capital d'un montant déterminé.

1.1.3.2 L'assurance en cas de décès :

Conformément à ce contrat, l'assureur s'engage à payer au bénéficiaire la somme stipulée si l'assuré décède au cours de l'assurance. Cependant, selon **l'Art.68 du code des assurances** : " l'assurance en cas de décès contractée par un tiers sur la tête de l'assuré est nulle, si ce dernier n'y a pas donné son consentement par écrit avec indication de la somme assurée.

1.1.3.3 L'assurance mixte :

C'est un contrat en vertu duquel l'assureur s'engage à payer au bénéficiaire le capital stipulé si le décès de l'assuré survient dans un délai déterminé et à défaut de ce décès, de payer au terme du contrat à l'assuré ou à la personne désignée, la somme stipulée.

La grande diversité des assurances actuellement disponibles facilite l'élaboration de polices personnalisées, plus à même de répondre aux besoins des individus. C'est, par exemple, le cas des assurances de revenu familial ou des assurances-vie de crédit qui combinent sur un même contrat différentes techniques d'assurance permettant de prendre en charge des risques de nature diverses.[dmf]

1.2 L'assurance vie en Algérie

Ces dernières années, notamment depuis 2002, une des évolutions majeures de l'assurance est le dynamisme de l'assurance vie comme instrument d'épargne privilégié. En effet, l'épargne de l'assurance vie dans les pays développés est supérieure aux autres modes de l'épargne.

En Algérie, le marché de l'assurance vie connaît une croissance continue mais reste très en retard par rapport aux autres pays du Maghreb. En outre, l'épargne générée à travers ce type d'assurance reste très modique, malgré le riche potentiel économique du pays. Par ailleurs, le niveau d'épargne actuel généré peut évoluer considérablement si les facteurs incitateurs à son développement sont réunis.

L'étude du marché de l'assurance de personnes notamment l'assurance vie durant la période 1995-2008 s'inscrit dans le cadre des changements adoptés par les autorités en matière de fonctionnement de l'économie. Parmi les événements majeurs qui ont caractérisé le début de cette période, on relève la promulgation de la loi relative aux assurances économiques n°95/07 du 25 Janvier 1995.

Néanmoins, depuis cette date, l'évolution du marché algérien de l'assurance a été marquée par plusieurs faits importants comme la levée du monopole, l'ouverture du marché dont l'exploitation avait été interdite aux investisseurs nationaux et étrangers depuis trente ans, etc. [NAO00]

1.2.1 Le marché algérien de l'assurance vie :

En Algérie, depuis la fin des années 80, depuis les réformes économiques nationales en général (1988) et le secteur des assurances en particulier (1990), les compagnies d'assurances sont en face d'un nouvel environnement contraignant, où la performance et la compétitivité deviennent le but de toute compagnie voulant garder sa part du marché.

Dans ce contexte, la question de l'évaluation du système assurantiel national revêtait une importance essentielle. Parmi les événements majeurs ayant caractérisé le début de cette période, on cite le programme d'ajustement structurel qui a engendré des changements, permettant le rétablissement des équilibres macroéconomiques.

Pour ce qui est du système assurantiel, la promulgation de la loi relative aux assurances économiques n°95/07 du 25 Janvier 1995 a généré des changements à savoir : la levée du monopole, l'ouverture du marché dont l'exploitation avait été interdite aux investisseurs nationaux et étrangers depuis trente ans, etc.[NAO00] [BEN00]

1.2.2 Cadre réglementaire de l'assurance en Algérie :

La transition de l'économie algérienne d'une économie planifiée vers une économie ouverte sur le marché remonte à l'année 1995 durant laquelle a été promulguée le 25 Janvier 1995 l'ordonnance N°95/07 relative aux assurances qui a mis fin au monopole exercé par l'Etat que se soit dans la production ou dans la distribution, avec l'instauration d'un système basé sur la concurrence à travers la réouverture du marché aux sociétés

constituées par des capitaux publics ou privés d'origine nationale ou étrangère depuis les nationalisations de 1966.

Dans ce contexte, les objectifs visés par le ministère des finances par le biais de la loi N° 95/07 peuvent être évoqués en quatre points essentiels :

- Promotion et développement du marché des assurances ;
- Une meilleure utilisation de l'épargne drainée ;
- Une meilleure prise en charge de l'assuré et les bénéficiaires des contrats d'assurance en protégeant leurs droits ;
- Et enfin, l'amélioration de la prestation de service rendu en matière d'assurance.

En 2006, les pouvoirs publics ont révisé cette ordonnance qui est considéré comme un déverrouillage réglementaire. C'est ainsi que la nouvelle loi n°06/04 a vu le jour le 20 février 2006, et a mis l'accent sur cette catégorie d'assurance pourvoyeur d'épargne longue après avoir été négligée par les compagnies d'assurance et l'ordonnance n°95/07.

La loi N°06/04 introduit une seconde vague de libéralisation, mais cette dernière n'est pas l'apport essentiel de cette loi. Étant donné que l'implication du secteur privé dans le domaine des assurances se fait progressivement, jusque là et dans tous les autres secteurs, la libéralisation signifiait simplement la possibilité, à des capitaux étrangers, de pouvoir créer des sociétés de droit algérien et de pouvoir exercer leur activité dans le pays.

Ainsi, les pouvoirs publics focalisent cette nouvelle loi autour de trois axes, mettant en exergue la stimulation de l'activité, l'amélioration de la gouvernance et de la sécurité financière des sociétés d'assurances ainsi que la réorganisation de la supervision. [1506][16][17]

1.2.3 Structure du marché des assurances de personnes en Algérie :

Le marché algérien de l'assurance, tout en étant largement ouvert depuis 1995, est dominé par les entreprises publiques nationales, puisque pas moins de 75% du marché est encore détenue par la SAA, la CAAR et la CAAT. Sur ce marché, les assurances de personnes restent très marginales et représentent en moyenne 5% du marché durant la période 1995-2006. Pour 2006, le chiffre d'affaires réalisé est de 2,8 milliards de dinars, soit 6,2% du portefeuille global du secteur, alors que dans les pays développés, les assurances de personnes représentent plus des 2/3 du volume des primes générées par l'industrie des assurances. [NAO00]

1.2.4 Entraves au développement des assurances vie en Algérie :

Le faible développement des assurances vie en Algérie, implique un faible développement de l'épargne au travers de ce type d'assurance, puisqu'il est considéré comme un vecteur essentiel dans la mobilisation de l'épargne.

Généralement, le retard constaté dans le développement des assurances vie en Algérie est expliqué par le facteur religieux et à l'existence d'un système de sécurité sociale généreux. Cependant, l'analyse comparée réalisée par le Conseil National des Assurances algérien avec d'autres pays musulmans a infirmé cette conclusion. De ce fait, les entraves au développement de cette catégorie d'assurance en Algérie ne se réduit pas à ces deux facteurs, mais elles relèvent à la fois de l'environnement économique et socioculturel.[BEN00] [LAT02]

1.2.4.1 Les principaux facteurs socioculturels :

Nous retiendrons ci-après les principaux facteurs socioculturels qui peuvent freiner le développement des assurances de personnes :

- **Un système de sécurité social généreux :** Les assurances de personnes ont des substituts qui se présentent sous forme d'assurances sociales, tel que les pensions de fin de carrière, la gratuité de l'éducation, le remboursement des frais médicaux, etc. Les citoyens algériens voient que les indemnités qu'ils perçoivent dans le cadre de la sécurité sociale sont suffisantes pour demander les assurances privées.
  Alors qu'en France, le système de sécurité sociale et de retraite obligatoire sont performants, les primes émises au titre des assurances de personnes représentent 63,48% du marché, sur les 435 milliards de francs de primes émises en 1998 par l'assurance vie et capitalisation, 427 milliards ont été émis au titre des assurances en cas de vie. Comparée à l'Algérie, la France produit 5696 fois plus de primes d'assurance vie que l'Algérie avec une densité de 1257,2 USD.
- **La religion :** L'existence dans la société d'une perception négative de l'assurance et de l'assurance vie en particulier, laquelle est assimilée à l'usure et aux jeux du hasard. Elle est perçue comme un moyen de contrarier la volonté de dieu. Ce jugement est fondé sur des arguments. Le premier, c'est que le contrat d'assurance vie contient l'intérêt qui est considéré par la religion islamique comme l'usure, Le second, c'est que le contrat d'assurance vie est un contrat aléatoire, alors que l'Islam a interdit la vente aléatoire qui est une vente à risque, il comprend une sorte d'ignorance et de jeu de hasard.
- **Le manque de confiance et l'absence de culture d'assurance auprès du public :** Les difficultés rencontrées par le marché des assurances vie sont dues principalement à l'absence de culture auprès de la population algérienne. D'une part, les algériens ont une mauvaise perception de l'assurance, en la considérant comme une lourde taxe imposée dont ils sont obligés de supporter, plutôt qu'une sorte de couverture. D'autre part, les assureurs, depuis des années, n'ont pas réussi à construire une relation de confiance avec les assurés car cela dépend de la capacité des assureurs à tenir leurs engagements. Par ailleurs, le manque, voire l'absence de communication et de vulgarisation des produits d'assurance offerts part les compagnies d'assurance envers les assurés potentiels, l'assurance de personnes est restée presque inconnue de la part des algériens malgré les diverses formules offertes sur le marché, puisque cette consommation reste inutile à leur regard et n'est pas considérée comme étant un produit de consommation.
- **Le caractère traditionnel de la société algérienne en termes de solidarité en cas de malheur :** Dans le cadre de vie traditionnel, le citoyen algérien n'était jamais isolé et abandonné à lui même tant matériellement, moralement qu'en matière de sécurité. Il faisait partie d'un tout. Engagé dans un réseau serré de relations et de solidarités, ses besoins peuvent être satisfaits au sein de la famille, groupe et communauté locale, et les jeunes jouent un rôle de sécurité face aux obstacles et difficultés qu'un individu peut rencontrer à un âge avancé. La solidarité est une des valeurs de base de la culture algérienne et que, dans la pratique, la vie moderne a quelque peu disloqué ce système ancestral de solidarité. C'est pourquoi des sociétés comme les mutuelles sociales, de même que les coopératives de production ou les mutuelles de prévoyance, sont des structures adaptées pour un pays comme l'Algérie

où elles prendraient le relais de la solidarité tribale et familiale. [BEN00] [LAT02]

1.2.4.2 Les facteurs économiques :

Les obstacles économiques auxquels est confrontée la branche de l'assurance de personnes sont assez nombreux. Nous retiendrons entre autres :

- **Le faible niveau des revenus disponibles des ménages :** L'application du programme de stabilisation des années quatre vingt dix a engendré l'apparition d'un double phénomène, d'une part, l'augmentation des prix, et d'autre part, une légère augmentation des revenus des ménages. Par conséquent, une détérioration évidente du pouvoir d'achat ainsi que le niveau de vie d'une grande partie de la population algérienne. Le niveau de la pauvreté à l'échelle nationale est estimé à 42,4% en 1995, dont 5,7% représente la pauvreté extrême, contre 14,1% de très pauvreté. Ce pourcentage est engendré par la diminution du pouvoir d'achat des salariés qui a été constatée entre 1986 et 1994, 7% et 3,6% respectivement pour l'année 1995 et 1996. La baisse des revenus de la population a influencé le portefeuille des compagnies d'assurances en ce qui concerne les contrats d'assurances de personnes, par contre, une amélioration des indicateurs économiques à la fin de la décennie a entraîné une augmentation réelle de 20% des revenus entre 1995 et 2000, et de 15% entre 2000 et 2002, cette amélioration se traduit par un accroissement de l'épargne.
- **Absence d'avantages fiscaux :** L'assurance vie constitue pour l'Etat une source de drainage de l'épargne. Elle est considérée comme un sacrifice d'une consommation immédiate, susceptible de financer les investissements en général d'intérêt public. Cette épargne devra être récompensée par des encouragements fiscaux, tant sur les primes que sur le revenu de leur épargne.
  Les produits algériens d'assurance vie étaient soumis d'une part à la TVA ce qui avait fait chuter le chiffre d'affaires de certaines compagnies, d'autre part, aux droits de succession. Jusqu'à présent l'assurance vie n'a pas bénéficié de l'avantage de déductibilité de la prime d'assurance à caractère d'épargne du revenu imposable de l'assuré. Par conséquent, pénalisé l'épargne par l'assurance au lieu de l'encourager. Actuellement, le régime fiscal ne prévoit que l'exonération des primes d'assurances de personnes à la TVA et un abattement de 25% avec un plafond de 20.000 DA sur le montant de la prime nette annuelle soumise à l'impôt sur le revenu global, bénéficié par les personnes souscrivant volontairement des contrats d'assurance de personnes, d'une durée minimale de huit ans. Ces deux incitations ne peuvent pas à elles seules militer pour le développement de cette branche d'assurance en Algérie.
- **L'instabilité monétaire :** Pour redresser les déséquilibres et relancer la croissance économique l'Algérie a décidé de dévaluer sa monnaie, qui est souvent inefficace car elle peut relancer l'inflation sans améliorer la position des paiements extérieure, ce qui amène à une nouvelle dévaluation et entraîne dans le pays un cercle vicieux d'inflation et dévaluation.
  La dévaluation du dinar algérien qui était d'environ 40% et l'aggravation de l'inflation qui a atteint 3,5% pour l'année 2002, ont entraîné une forte détérioration du pouvoir d'achat des ménages. Par conséquent, elle a découragé les clients à s'assurer par peur de voir la valeur de leur épargne érodée par l'inflation.
- **La distribution des produits d'assurance :** L'expérience vécue par les entreprises d'assurance algérienne a démontré les limites des réseaux traditionnels dans

la vente des produits d'assurance vie. Cela est expliqué par les professionnels du secteur, d'une part, par l'insuffisance en termes de communication entre les distributeurs d'assurance et leurs clients qui ignorent l'existence de l'assurance vie et ses spécificités. D'autre part, par le fait que les représentants ne sont pas spécialisés dans la commercialisation des assurances vie et ce, due principalement au :
  - Manque de formation dans le secteur.
  - Taux de commission accordés aux intermédiaires pour certains contrats d'assurance vie inférieurs à ceux attribués en assurance non vie. C'est pour cette raison, les intermédiaires ne se forcent pas à vendre les produits d'assurances vie.
- **La faiblesse du marché financier algérien :** La faiblesse du marché financier algérien avait un effet négatif sur le développement de l'assurance vie. Sachant que les produits modernes de ce type d'assurance ont connus un développement considérable dans les pays développés. Le succès de ces produits dépend de la performance des bourses. Alors que le marché financier algérien qui a été mis en place depuis quelques années, son développement demeure actuellement limité. Par conséquent, le marché financier n'a pas offert les instruments financiers utilisés comme support aux produits d'assurance algérienne (SICAV, FCP,...) qui s'est limité généralement aux assurances temporaires décès à capitaux décroissants, souscrites par une banque à ses débiteurs dans le cadre du remboursement d'un crédit bancaire.
- **Les prix des produits d'assurances vie :** La question de la fiabilité des bases techniques de tarification de l'assurance vie en Algérie est posée par référence à la mortalité, du fait qu'à ce jour, il n'existe pas de tables de mortalité spécifiques reflétant la mortalité de la population algérienne, cela est dû à l'absence des statistiques sur la population algérienne permettant de calculer les primes adéquates. Les assureurs du marché algérien utilisent des tables de mortalité françaises (TD - TV 1960/64, 1973/77,...). Pour l'instant, aucun texte ne réglemente l'utilisation des tables de mortalité, ni leur taux technique. [BEN00] [LAT02]

1.3 Présentation Cardif

CARDIF est une des Compagnies d'assurance vie, filiale de BNP Paribas Assurance á 100 % Elle emploie plus de 8000 collaborateurs, conçoit et commercialise ses produits et services dans 42 pays, et assure plus de 35 millions d'assurés dans le monde. Numéro 1 de l'assurance des emprunteurs, elle se vit attribuée la notation AA par « Standard & Poor's » et sa gestion de qualité est certifiée ISO 9000 dans plusieurs pays. Elle compte parmi ses partenaires 35 des 100 plus grandes banques dans le monde. Elle propose des produits d'épargne individuelle, d'épargne retraite collective ainsi qu'une large gamme de produits de prévoyance.

1.3.1 Produits de prévoyance

- Assurance des emprunteurs.
- Assurance des factures.
- Assistance voyage.
- Assurance des cartes de crédit.

Comme on a aussi des produits qui assure les différents dommages tel que :

- Multirisques habitation.
- Automobile.

- Protection juridique.
- Extension de garantie.
- Assurance GAP.

Ce qui différencie la compagnie Cardif des assureurs traditionnels, c'est qu'elle crée et gère un centre de profit avec ses partenaires, alors que les autres assureurs se contentent de concevoir et de délivrer des produits. Il existe donc une culture partenariale forte basée sur intérêt commun avec le partenaire ainsi qu'une considération particulière pour la satisfaction du client. Elle met également un point d'honneur à constamment innover afin de rester en avant de la courbe, et ce, en accordant un intérêt accru à l'activité de l'emprunteur.

1.3.2 Historique

CARDIF a démarré son activité en France en 1973, avec la commercialisation de produits d'assurance vie dans le réseau CETELEM, société spécialisée dans les crédits à la consommation du groupe Compagnie Bancaire. Elle a été la première compagnie d'assurance à distribuer, il y a plus de trente ans en France, ses produits d'assurance par l'intermédiaire de banques, de sociétés de crédit et de sociétés de grande distribution. En effet, cette filiale, qui commercialise des produits de services dans le domaine de l'épargne et de la prévoyance par l'intermédiaire de multiples canaux de distribution. Très vite, des accords de distribution ont été signés avec les autres sociétés du groupe Compagnie Bancaire, puis avec Paribas et le Crédit du Nord. Par la suite, des partenariats bancaires ont été noués en dehors du groupe d'origine avec des banques, des sociétés de crédit et des sociétés de grande distribution. En 1989, l'activité s'est développée hors de France et ce n'est qu'en 1999 lors de la fusion entre BNP et Paribas que CARDIF est entré dans le giron de BNP Paribas. en 2006, CARDIF s'implante en Algérie, Mexique, Pérou, Bulgarie et Roumanie ; après avoirs conquit 30 pays dans le monde : Afrique du Sud, Allemagne, Autriche, Argentine, Belgique, Brésil, Chili, Chine, Corée du Sud, Espagne, Etats-Unis, France, Hongrie, Inde, Irlande, Italie, Japon, Luxembourg, Pays-Bas, Pologne, Portugal, Royaume-Uni, République tchèque, Russie, Slovaquie, Suède, Suisse, Taiwan, Thaïlande et Vietnam. Puis en 2007 elle s'implante au Canada, Croatie, Danemark, Norvège et la Turquie. Actuellement Cardif SPA est présente dans plus de 40 pays, dont cinq en Amérique latine et sept en Asie. Elle assure plus de 50 millions de personnes dans le monde. Dans les prochaines années, la part de l'international dans le chiffre d'affaire global devrait continuer à croître.

1.3.3 L'implantation de CARDIF en Algérie

L'Algérie est le second pays Africain après l'Afrique du Sud dans lequel la compagnie d'assurance s'implante. Cette dernière est par ailleurs la première compagnie d'assurance étrangère qui s'établit sur le marché algérien. CARDIF EL DJAZAIR est une société de droit algérien de 28 collaborateurs, spécialisée dans les assurances de personnes. Elle répond ainsi aux nouvelles exigences législatives qui prévoient la séparation juridique entre les compagnies pratiquant les branches d'assurance de personnes des compagnies pratiquant les assurances de dommage. Elle été agrée par le ministère des Finances en Octobre 2006 en vertu des dispositions de l'Ordonnance relative aux assurances n°95/07 du 25 Janvier 1995 complétée et modifiée par la loi de Février 2006. Cardif devint ainsi le seul opérateur en bancassurance agréé par les autorités financières du pays ainsi que

c

FIGURE 1.1 – Implantation internationale

Partenaire	Produits
BNP Paribas El Djazair	Assurance Des Emprunteurs
	Prévoyance Individuelle
Cetelem	Assurance Des Emprunteurs
CNEP	Assurance Des Emprunteurs
	Total Prévoyance

TABLE 1.1 – Partenaires de Cardif el Djazair et produits.

l'unique entreprise spécialisée dans les assurances de personnes à ce jour. C'est en Février 2007 qu'elle reçoit un visa du Ministère des Finances algérien pour la commercialisation de son premier produit assurance ADE (Assurance Des Emprunteur) avec Cetelem. La forme juridique de CARDIF EL DJAZAIR est du type (SPA) société par actions. Son capital social est de 1.000.000.000,00 DZA. Filiale à 100% de BNP Paribas Assurance. Elle a réalisé en 2007 un chiffre d'affaire de 16 839 733 DZA, et ce d'Avril à Décembre. CARDIF EL DJAZAIR s'appuie également sur le réseau des 39 agences de BNP Paribas El Djazaïr, et avec la signature en mois de mars 2008 d'un accord partenariat avec la Caisse Nationale d'Epargne et de Prévoyance (CNEP-Banque).

1.4 Le produit CNEP Totale Prévoyance

C'est un produit d'assurance vie qui couvre les risques liés au décès et à l'invalidité, et qui garantie le paiement d'un capital au bénéficiaire ou ses ayants droits. Baptisé " Prévoyance individuel ce produit disponibles en deux formules Toutes causes offre une couverture quelle que soit la cause et Accidentels c'est-à-dire le capital garanti est doublé en cas de décès accidentels. Nous expliquons en détails tous qui concernent ce produit

ci-dessous :

- **L'objet du contrat**
  L'objet du contrat est de faire bénéficier l'adhérent des garanties Décès ou d'invalidité absolue et définitive par le versement d'un capital aux bénéficiaires désignés, sous réserves des exclusions.
- **Conditions d'admission**
  Est admissible au présent contrat, toute personne physique répondant aux conditions suivantes :
  - Être âgé de 19 à 60 ans ;
  - Être détenteur d'un compte chèque chez la CNEP-Banque ;
  - Donner son consentement écrit à l'assurance ;
  - S'acquitter de la Déclaration de Bonne Santé ou du questionnaire médical pour la couverture en Formule 1.
- **Risques garantis**
  Le Produit CNEP Totale Prévoyance peut garantir des risques sous une protection totale ou bien partielle et cela en souscrivant dans l'une des formules suivantes :
  1. Formule 1 : Décès, Invalidité Absolue et Définitive Toutes causes ;
  2. Formule 2 : Décès, Invalidité Absolue et Définitive Accidentels.
- **Conditions d'admission**
  Avant de souscrire a un contrat d'assurance vie, des conditions doivent être vérifié afin d'être admissible au présent contrat, toute personne physique répondant aux conditions suivantes est admis à souscrire à un contrat d'assurance vie CTP :
  - Être âgé de 19 à 60 ans ;
  - Être détenteur d'un compte chèque chez CNEP banque ;
  - Donner son consentement écrit à l'assurance ;
  - S'acquitter de la déclaration de bonne santé, ou du questionnaire médical pour la couverture en formule 1.
- **Limites d'âge de couverture** La limite d'âge de couverture est de :
  - 70 ans pour la garantie Décès ;
  - 60 ans pour la garantie Invalidité absolue et définitive.
- **Cumul de garantie**
  Le capital souscrit par l'Adhérent pour l'ensemble de ses adhésions ne peut excéder les montants suivants :
  - Formule 1 : 5 millions de dinars,
  - Formule 2 : 10 millions de dinars.
- **Prime d'assurance**
  La prime d'assurance est mensuelle et indiquée sur le bulletin d'adhésion. Une réduction de 15 % est accordée à l'adhérent-conjoint sous réserve que les garanties souscrites par ce dernier soient inférieures ou égales à celles choisies par l'adhérent. Cette réduction n'est plus applicable lorsque l'adhésion principale est résiliée.
- **Territorialité**
  Les garanties sont acquises dans le monde entier à condition que les séjours à l'étranger ne dépassent pas trois mois consécutifs.

Conclusion

Véritable outil d'épargne et de transmission de capital, l'assurance vie s'impose comme un placement qu'il faut détenir pour sa stratégie patrimoniale. En Algérie malgré l'ouverture du secteur de l'assurance depuis 1995, le marché de l'assurance de personnes reste toujours à l'état embryonnaire. Son rôle est encore très limité dans la mesure où il ne représente qu'une part très faible de la création de la valeur économique nationale, comparativement à nos voisins, aux pays développés et à la moyenne mondiale. Cela se vérifie par sa part dans le portefeuille global du secteur qui n'a jamais dépassé les 06%.

Le législateur algérien, conscient de ces insuffisances, a renforcé le marché par de nouvelles règles afin de l'adapter à l'ouverture du secteur, les premiers effets sont notés à travers l'agrément récent d'une compagnie première, à se constituer à partir de capitaux européens (CARDIF EL DJAZAIR), spécialisée dans les assurances de personnes en distribuant ses produits à travers le réseau bancaire (la bancassurance).

Dans le cadre de notre problématique, concernant les résiliations anticipées des contrats d'assurance vie au sein de CARDIF EL DJAZAIR, nous avons eu à exploiter des bases de données, afin d'en tirer des causes potentiels de résiliation des contrats d'assurance vie, en se basant sur le Data Mining et le scoring, qui feront, de ce fait, l'objet du chapitre suivant.

Chapitre 2

DATA Mining et Scoring

Sommaire

Le Data Mining et la statistique, envahissent aujourd'hui de nombreux domaines, qui vont de l'infini petit (génomique) à l'infiniment grand (astrophysique), du plus quotidien (gestion de la relation client) au moins quotidien (aide au pilotage aéronautique), du plus ouvert (e-commerce) au plus sécuritaire (prévention du terrorisme, détection de la fraude dans l'utilisation des cartes bancaires), du plus industriel (contrôle qualité, pilotage de la production) au plus théorique (enquêtes en science humaine, études biologiques, médicales et pharmaceutique). A cette simple énumération on devine que le spectre des applications du Data Mining et de la statistique est très large. Les plus concernés sont les secteurs ou d'importants volumes de données doivent être analysés, parfois en vue de prendre des décisions rapides comme le montrent certains des exemples précédents.

Ce chapitre définit la notion de Data Mining, et en décrit les principales applications et les apports au marketing de bases de données, à la gestion de la relation client et à d'autres domaines financiers, industriel et scientifiques. Il décrit les différentes phases d'une étude ou d'un projet de Data Mining ainsi qu'une application courante "Le scoring".

2.1 Le Data Mining

Les données brutes, malgré leur quantité qui augmente d'une façon exponentielle, n'ont presque aucune valeur, ce qui est le plus important en fait c'est les connaissances pour lesquelles nous sommes tous assoiffés et qui sont obtenus par la compréhension de ces données, mais plus on a de données plus ce processus devient difficile. De nos jours, les changements de notre environnement sont dénotés par des capteurs qui sont devenus de plus en plus nombreux. Par conséquent, la compréhension de ces données est très importante. Et comme il est dit par Piatestky-Shapiro, " [...] as long as the world keeps producing data of all kinds [...] at an ever increasing rate, the demand for data mining will continue to grow". D'où la fouille de données devient une nécessité. Voilà donc en quoi consiste le Data Mining :

2.1.1 Qu'est-ce que le Data Mining ?

Le Data Mining, ou fouille de données, est l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de (souvent grandes) bases de données informatique, de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données. En bref, le Data Mining est l'art d'extraire des informations, voire des connaissances, à partir des données. Le Data Mining est soit descriptif, soit prédictif : les techniques descriptives (ou exploratoires) visent à mettre en évidence des informations présentes mais enfouies sous le volume des données ; les techniques prédictives (ou explicatives) visent à extrapoler de nouvelles informations à partir des informations présentes, ces nouvelles informations pouvant être qualitatives (classement ou scoring) ou quantitatives (prédiction).[LIA10]

David Hand (1998) en donne la définition suivante : " Data Mining consists in the discovery of interesting, unexpected, or valuable structures in large data sets". Les règles à trouver sont du genre :

- Les clients ayant tel profil acquièrent plus souvent tel type de produit ;
- Les clients ayant tel profil arrivent plus souvent au contentieux ;
- Les clients ayant acquis le produit A et le produit B acquièrent plus souvent le produit C, en même temps ou n mois plus tard ;
- Les clients ayant eu tel comportement, clos tels et tels produits dans tel intervalle de temps, risquent de nous quitter pour la concurrence.

2.1.2 A quoi sert le Data Mining ?

Les avantages procurés par l'utilisation des règles et des modèles découverts à l'aide du Data Mining sont multiples, dans de nombreux domaines.

- **Secteur bancaire :**
  De multiples techniques de Data Mining (scoring, classification, associations de produits...) ont envahi la banque du fait du grand nombre de dossiers et de leur caractère relativement standard. Cet essor du Data Mining dans l'activité bancaire s'explique par la conjonction de plusieurs éléments : développement des nouvelles technologies de communication (internet, téléphonie mobile...) et de traitement de l'information (entrepôt de données), les attentes accrues de qualité de service des clients, la concurrence exercée sur les banques à réseau par les sociétés de crédit et les " nouveaux entrants " (banques étrangères, grande distribution et compagnies

d'assurance, lesquelles développent parfois une activité bancaire au travers d'un partenariat avec une banque traditionnelle), la pression économique internationale pour une plus grande rentabilité et productivité des banques, sans oublier l'aspect règlementaire.

- **La grande distribution :** Ce secteur développe ses cartes de crédit privatives, qui lui permettent de se constituer de grandes bases de données (parfois de plusieurs millions de porteurs) enrichies par les informations comportementales provenant des tickets de caisses, et lui permettent de concurrencer les banques dans la connaissance du client. En outre, les services associés à ces cartes (caisses réservées, promotions exclusives ...) sont facteurs de fidélisation. La détection des associations de produits sur les tickets de caisse permet d'identifier les profils de clients, de mieux choisir les produits et de mieux les disposer dans les rayons, en tenant compte du facteur " régional " dans les analyses.
- **Les assurances de biens et de personnes :** les études de ventes croisées (cross-selling), de montées en gamme (up-selling) et d'attrition sont, avec l'adaptation de la tarification aux risques encourus, les sujets dominants dans un secteur ou l'appétence ne se pose pas dans les même termes qu'ailleurs, puisque certains produits (assurance automobile) sont obligatoire, et qu'il s'agit, à l'exception des jeunes, soit de prendre un client à un concurrent, soit de faire monter en gamme un client que l'on détient déjà, en lui vendant par exemple des garanties optionnelles supplémentaires. Le besoin de Data Mining dans ce secteur s'est exacerbé avec le développement de la concurrence des nouveaux entrants que sont les banques qui, pratiquant ce que l'on nomme la bancassurance, possèdent l'avantage de réseaux étendus, de contacts fréquents avec le client et de bases de données riches. Cet avantage est surtout tangible face aux assureurs " traditionnels " non mutualistes, qui éprouvent parfois des difficultés à fédérer dans des bases de données marketing des informations disséminées et jalousement détenues par leurs agents généraux. De surcroit, les bases clients de ces assureurs, quand elles ne sont pas compartimentées par agent général, sont encore souvent structurées par contrat et non par client. Pourtant, ces réseaux, avec leurs taux de fidélisation inférieurs à ceux des mutuelles, ont bien besoin d'améliorer leur gestion de la relation client, et donc leur connaissance globale de leur clientèle. Si les études d'appétence de l'assurance ressemblent à celles de la banque, les études de sinistralité présentent quelques particularités, avec l'intervention de la loi de Poisson dans le modèle linéaire généralisé pour modéliser le nombre de sinistre. Les assureurs disposent d'un atout, avec la détention d'informations assez nombreuses sur leurs clients, notamment par le truchement des contrats d'assurance habitation et responsabilité civile qui fournissent des informations assez précises sur la famille et son cadre de vie.
- **Secteur de la téléphonie :** L'ouverture à la concurrence du marché de la téléphonie fixe et mobile, ont avivé les problèmes de churn (départ pour la concurrence) des clients, particuliers, professionnels ou entreprises. On imagine l'importance de la fidélisation dans ce secteur, quand on sait que les couts d'acquisition d'un client en téléphonie mobile dépassent en moyenne deux cent euros et que plus d'un million d'utilisateurs changent chaque année d'opérateur dans certains pays. C'est dont tout naturellement le score de churn qui tient la vedette du Data Mining dans la téléphonie. Pour les mêmes raisons, des opérateurs utilisent des outils de Text Mining

afin d'analyser automatiquement le contenu des lettres de réclamation des clients. Les autres sujets d'étude dans la téléphonie sont le score d'impayés, l'optimisation des campagnes marketing direct, l'analyse des comportements des internautes et le dimensionnement des centres d'appels. On s'intéresse aussi à la probabilité qu'un client change de téléphone mobile.

- **L'industrie automobile :** elle utilise assez couramment le Data Mining. Un thème classique est le score de rééachat d'un véhicule de la marque. Renault a ainsi construit un modèle prédisant des clients susceptibles d'acheter un nouveau véhicule Renault dans les six mois à venir. Ces clients sont identifiés à partir des données des concessionnaires, lesquels reçoivent en retour une liste de clients au score élevé, qu'ils peuvent alors contacter. Dans le domaine de la production, le Data Mining est utilisé pour rechercher l'origine des défauts de construction, de façon à pouvoir les minimiser.
- **Le secteur médical :**
  un secteur friand de statistique. Le Data Mining y est donc naturellement répandu, tant dans les applications descriptives que prédictives. Parmi les premières, on rencontre la détermination de groupes de patients susceptibles d'être soumis à des protocoles thérapeutiques déterminés, chaque groupe rassemblant tous les médicaments, en vue notamment de détecter des anomalies de prescription. Parmi les applications prédictives, on trouve la recherche des facteurs de décès ou de survie dans certaines pathologies (infarctus, cancers...), à partir des données recueillies lors des essais cliniques, afin de choisir le traitement le plus approprié en fonction de la pathologie et de l'individu.
- **L'industrie agroalimentaire :**
  grande consommatrice de statistique, elle est utilisée dans les " analyse sensorielles ", qui croisent les données sensorielles (gout, saveur, texture...) perçues par les consommateurs avec les mesures instrumentales physico-chimiques, ainsi qu'avec les préférences en faveur de tel ou tel produit. Ce sont par ailleurs des modèles prédictifs d'analyse discriminante et de régression logistique qui ont permis de distinguer des spiritueux de leurs contrefaçons, à partir de l'analyse d'une dizaine de molécules présentes dans le breuvage. Aussi de comprendre et maitriser l'évolution des micro-organismes, et prévenir les risques liés à leur développement dans les industries agroalimentaires, et de gérer la date limite de conservation.
- **La biologie :**
  de façon générale, la biologie utilise beaucoup la statistique. On la rencontre depuis longtemps dans la classification des espèces vivantes et l'exemple classique du classement de trois espèces d'iris par Fisher grâce à son analyse discriminante linéaire. L'agronomie demande à la statistique d'évaluer l'effet d'engrais ou de pesticides. Autre utilisation du Data Mining à la mode : la détection des facteurs expliquant la pollution de l'air.[TUF10]

Un sondage effectué en juillet 2005 sur le portait web www.kdnuggets.com révélait les principaux secteurs utilisant le Data Mining : la banque (12%), la gestions de la relation client (12%), le marketing direct (8%), la détection de la fraude (7%), les assurance (6%), la distribution(6%), les télécommunication (5%), les études scientifiques (4%) et la santé (4%). [TUF03]

2.1.3 Mise en oeuvre du Data Mining

Les principaux facteurs de succès d'un projet sont :
- Des objectifs précis, importants et réalistes ;
- La richesse, et surtout la qualité, des informations collectées ;
- La collaboration des compétences métiers et statistiques de l'établissement ;
- La pertinence des techniques de Data Mining utilisées ;
- Une bonne restitution des informations générées et une bonne intégration le cas échéant dans le système d'information ;
- L'analyse des résultats et des retours d'expérience de chaque utilisation du Data Mining pour l'utilisation suivante.

Dans une entreprise, la mise en oeuvre des techniques de Data Mining peut revêtir plusieurs formes.

Soit l'entreprise externalise totalement l'activité de Data Mining, comme de l'infogérance, en fournissant chaque fois qu'il le faut des fichiers commerciaux bruts à des prestataires spécialisés, les prestataires lui restituant les fichiers commerciaux enrichis avec des informations comme le score du client, son segment comportementale, etc.

Soit l'entreprise sous traite l'essentiel de l'activité de Data Mining, en confiant à des prestataires le soin d'élaborer les modèles de Data Mining dont elle a besoin, mais elle se fait livrer ces modèles, afin de pouvoir les appliquer elle-même à ses fichiers, voir les retoucher légèrement.

Soit l'entreprise développe elle-même ses modèles de Data Mining, en utilisant les logiciels du marché, éventuellement avec l'aide de consultants spécialisés. [TUF10]

2.2 Le déroulement d'une étude de Data Mining :

A l'origine le Data Mining était vue comme un procédé automatique ou semi-automatique. Aujourd'hui, on est revenu de cette illusion. Le Data Mining n'est pas un produit qui peut être acheté, mais bien une discipline qui doit être maitrisée.

Avant d'appliquer automatiquement des algorithmes de calculs sur les données, il faut passer par une phase d'exploration et d'analyse qui ne saurait être automatisée : elle fait intervenir le bon sens et la connaissance du contexte (culture générale). Quand on veut produire de la connaissance, le problème ne se limite pas à répondre à des questions. Il faut d'abord poser les questions. C'est cette première étape qui pour l'essentiel fait que le Data Mining est une discipline et pas un simple produit. [LIA10]

De bonnes pratiques ont émergé au fil du temps pour améliorer la qualité des projets. Parmi celles-ci, les méthodologies aident les équipes à organiser les projets en processus. Au nombre des méthodes les plus utilisées se trouvent la CRISP-DM qui est la méthode la plus employée dans les années 2010.

2.2.1 CRISP-DM le processus de Data Mining

L'exploration de données se propose d'utiliser un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, l'intelligence artificielle ou l'informatique, pour construire des modèles à partir des données, c'est-à-dire trouver des structures intéressantes ou des motifs selon des critères fixés au préalable, et d'en extraire un maximum de connaissances utiles à l'entreprise. La méthode CRISP-DM découpe le processus de fouille de données en six étapes permettant de structurer la technique et de l'ancrer dans un processus industriel. Plus qu'une théorie normalisée, c'est un processus

d'extraction des connaissances métiers. En tant que modèle de processus, CRISP-DM offre un aperçu du cycle de vie du Data Mining.

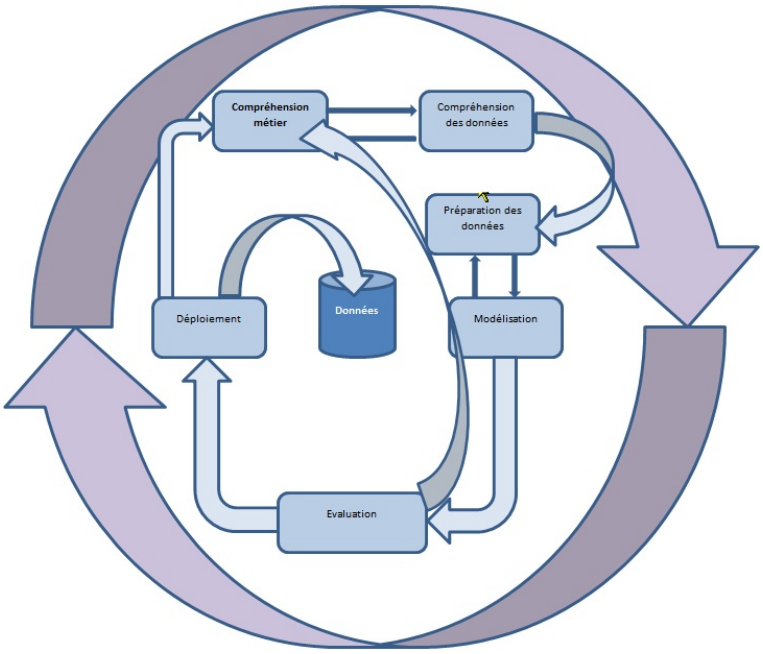

FIGURE 2.1 – CRISP-DM Processus [BER10]

Le modèle de cycle de vie comporte six phases dotées de flèches indiquant les dépendances les plus importantes et les plus fréquentes entre les phases. La séquence des phases n'est pas strictement établie. De fait, les projets, pour la plupart, passent d'une phase à l'autre en fonction des besoins.

– **Présentation de la compréhension du problème**
  Avant d'entamer la modélisation et les calculs, n'importe qu'elle entreprise doit comprendre la nature du problème et définir les attentes, il faut se pencher sur les bénéfices que la société souhaite en tirer du Data Mining. Cette consultation doit englober le plus grand nombre de personne possible. L'étape finale de cette phase CRISP-DM concerne la production d'un plan de projet à l'aide d'informations ainsi recueillies. Bien que cette étude puisse paraitre superflue, elle s'avère au contraire indispensable. La compréhension des objectifs de la société en matière de Data Mining garantit une approche homogène indispensable avant la mise en oeuvre de ressources précieuses. [LAR05] [IAN05][BER10]

  1. **Définition des objectifs :**
     il faut commencer par choisir le sujet, définir la population cible, définir l'entité statistique étudié, définir certains critères essentiels et en particulier le phénomène à prédire, planifier le projet, prévoir l'utilisation opérationnelle des informations extraites et des modèles produits, et spécifier les résultats attendus. Les objectifs doivent être précis et conduire à des actions concrètes, comme l'affinement d'un ciblage pour une compagne de marketing direct. Dans le domaine commercial, les objectifs doivent aussi être réalistes et tenir compte des

réalités économiques, des actions marketing déjà menées, du taux de pénétration, de la saturation du marché.

2. **Evaluation de la situation :**
   Maintenant que l'objectif est clairement établi, passons à l'évaluation de la situation actuelle. Cette étape soulève des questions telles que :
   (a) Quels types de données sont disponibles pour l'analyse ?
   (b) Le personnel nécessaire à la réalisation du projet est-il disponible ?
   (c) Quels sont les plus grands facteurs de risque en jeu ?
   (d) Existe-t-il un plan de secours pour chaque risque ?

3. **Production du plan du projet**
   A présent, on est prêt à créer le plan du projet de Data Mining. Les questions posées jusqu'ici, ainsi que les objectifs de Data Mining et les objectifs commerciaux formulés, formeront la base de ce plan. Le plan du projet est le document principal régissant tout le travail de Data Mining. S'il est bien crée, il permettra d'informer toutes les personnes associées au projet des objectifs, des ressources, des risques et du programme de toutes les phases du Data Mining. On peut publier ce plan, ainsi que la documentation recueillie lors de cette phase, sur le réseau interne de la société.

– **Compréhension et préparation des données :**
La phase de compréhension des données implique l'étude des données disponibles pour le Data Mining. Cette étape revêt une importance vitale, car elle permet d'éviter les problèmes inattendus au cours de la phase suivante, la préparation des données, phase généralement la plus longue d'un projet. Cette phase implique l'accès aux données et leur exploration à l'aide de tables et de graphiques pouvant être organisés, on peut ainsi déterminer la qualité des données et décrire les résultats de ces étapes dans la documentation du projet. [LAR05] [IAN05]

1. **Collecte des données initiales :**
   Durant cette étape est réalisé le recensement des données utiles, accessibles (internes ou externes à l'entreprise ou à l'organisation), légalement et techniquement exploitables, fiables et suffisamment à jour sur les caractéristiques et le comportement des individus étudiés : clients, patients, usagers... Ces données proviennent du système d'information de l'entreprise, ou alors sont stockées dans l'entreprise hors du système d'information centralisé (fichier Excel, Access...), ou bien sont achetées ou récupérées à l'extérieur de l'entreprise, ou encore sont calculées à partir des données précédentes. [TUF10]

2. **Exploration et préparation des données :**
   Cette étape a trait à l'exploration et à la mise en forme des données, elle se déroule en trois opérations : La première opération consiste à fiabiliser, remplacer, ou supprimer les données incorrectes, soit qu'elles aient trop de valeurs manquantes, de valeurs aberrantes, ou qu'elles aient trop de valeurs extrêmes (" outliers ") s'écartant trop des valeurs habituellement admises. La deuxième opération est la création d'indicateurs pertinents à partir des données brutes contrôlées et le cas échéant corrigées, par exemple en modifiant des unités de

mesure, en remplacent les dates par des durées, des anciennetés ou des âges. La troisième opération est la réduction du nombre de dimensions du problème : réduction du nombre d'individus, du nombre de variables, du nombre de modalités des variables. La réduction du nombre d'individus consiste, comme indiqué plus haut, à éliminer certains individus hors norme de la population, la réduction du nombre de variables à ignorer certaines variables trop corrélées entre elles et la réduction du nombre de modalités au moyen par exemple de regroupement des modalités qui sont trop nombreuses ou dont les effectifs sont trop petits. [TUF10]

3. **Sélection des données :**
   En fonction de la collecte initiale de données réalisée dans la phase précédente, on peut commencer par choisir les données pertinentes pour les objectifs de Data Mining. En général, les données peuvent être sélectionnées de deux manières :
   - Sélection des enregistrements (lignes) : implique des décisions concernant les comptes, les produits ou les clients à inclure ;
   - Sélection des attributs ou des caractéristiques (colonnes) : implique des décisions concernant l'utilisation de caractéristiques telles que le montant de transactions ou le revenu des ménages.
4. **Nettoyage des données :** Lorsqu'on nettoie les données, on peut examiner en profondeur les problèmes des données que nous avons choisi d'inclure dans l'analyse. Parmi les problèmes posés par les données :
   - Problèmes des données manquantes, qu'il faudra remplacer par des valeurs estimées ;
   - Les erreurs dans les données, qu'il faudra découvrir et corriger ou bien exclure ;
   - Métadonnées erronées ou manquantes, qu'il faudra examiner manuellement et rechercher la signification correcte.

– **Modélisation :**
Cette étape constitue le coeur de l'activité de Data Mining, c'est à ce stade que les efforts commencent à être récompensés. Les données que nous avons mis du temps à préparer sont importées dans les outils d'analyse et résultats commencent à éclaircir le problème posé lors de la compréhension du problème. La modélisation est généralement effectuée en utilisant plusieurs itérations, Généralement, les data miners exécutent plusieurs modèles en utilisant les paramètres par défaut, puis affinent ces derniers ou reviennent à la phase de préparation des données pour effectuer les manipulations requises par le modèle de leur choix. Il est rare qu'une question de Data Mining soit résolue de façon satisfaisante avec un seul modèle et une seul exécution. C'est pourquoi le Data Mining est si intéressant. [LAR05] [IAN05][BER10]

1. **Choix des techniques de modélisation appropriées :**
   Il arrive souvent que les data miners utilisent plusieurs techniques pour traiter un problème à partir de perspectives différentes. Pour savoir si les modèles à utiliser ont une incidence sur notre choix, il nous faut étudier les points suivant :
   - Le modèle exige-t-il que les données soient divisées en ensembles de test et d'apprentissage ?

- Avons-nous suffisamment de données pour produire des résultats fiables avec un modèle donné ?
- Le modèle exige-t-il un certain niveau de qualité des données ? Nos données actuelles répondent-elles à ce niveau ?
- Le type de données est-il approprié au modèle ? si ce n'est pas le cas, pouvons-nous effectuer les conversions nécessaires en utilisant des noeuds de manipulation de données ?

2. **Création des modèles :** A ce stade, il faut être bien préparé pour créer les modèles qu'on a étudiés pendant si longtemps. Il faut prendre le temps de tester plusieurs modèles avant de tirer des conclusions fermes et définitives. La plupart des data miners créent plusieurs modèles et comparent les résultats avant de les déployer ou de les intégrer. Garder une trace des données et des paramètres utilisés pour chaque modèle afin de suivre l'évolution des opérations que nous effectuerons avec les différents modèles. Ceci nous aidera à discuter des résultats avec d'autres personnes et à retrouver la trace des opérations effectuées, le cas échéant. A la fin du processus de création des modèles, on dispose de trois types d'informations à utiliser dans les décisions de Data Mining :
   - Les valeurs des paramètres, qui comprennent les notes que nous avons pris concernant les paramètres aboutissant aux meilleurs résultats ;
   - Les modèles réels produits ;
   - Les descriptions des résultats du modèle, qui incluent les problèmes de performances et de données rencontrés lors de l'exécution du modèle et de l'exploration de ses résultats.
3. **Evaluation du modèle :** A présent on dispose d'un ensemble de modèles initiaux, on les analyses en détail pour déterminer ceux qui sont suffisamment précis ou efficaces pour être dits finaux. Un modèle final peut désigner un modèle " prêt pour le déploiement " ou un modèle " illustrant des motifs intéressants ". Outre les qualités de précision du modèle, d'autres critères de choix sont parfois à prendre en compte, comme la lisibilité du modèle du point de vue de l'utilisateur, et sa facilité d'implémentation dans l'informatique de production. [LAR05] [IAN05][TUF10]
4. **Processus de révision :** Les méthodologies efficaces prévoient généralement du temps pour réfléchir sur les points positifs et négatifs du processus qui vient de se terminer. Le Data Mining fonctionne de la même manière. Une partie du processus CRISP-DM consiste à tirer des leçons de notre expérience de façon à ce que les futurs projets de Data Mining soient plus efficaces. Il faut d'abord récapituler les activités et les décisions pour chaque phase, en incluant les étapes de préparation des données, la création de modèles, etc. Ensuite, pour chaque phase, il faut tenir compte des questions suivantes et emmètre des propositions d'amélioration :
   - Cette étape a-t-elle contribué à la valeur des résultats finaux ?
   - Existe-t-il des moyens de simplifier ou d'améliorer cette étape ou opération particulière ?
   - Quelles ont été les erreurs ou les échecs rencontrés au cours de cette phase ? Comment peuvent-ils être évités la prochaine fois ?

- Avons-nous eu des surprises (bonnes ou mauvaises) pendant cette phase ? Avec du recul, existe-t-il un moyen de prédire ces évènements ?
- Des décisions ou des stratégies alternatives auraient-elles pu être utilisées lors d'une phase donnée ?

5. **Déploiement des modèles :**

   Ce déploiement passe par l'implémentation informatique des modèles de Data Mining, préalable à l'utilisation des résultats pour l'action (adaptation des procédures, ciblages...) et à la mise à disposition des utilisateurs (informations sur le poste de travail...). [TUF10]
   Ce processus a pour but l'utilisation des nouvelles connaissances pour apporter des améliorations au sein de l'entreprise. Par exemple, découvrir des motifs alarmants dans les données indiquant un changement de comportement des clients âgés de plus de 30 ans. Ces informations seront sans aucun doute utiles pour la planification et la prise de décisions marketing. [LAR05] [IAN05]
   De façon générale, la phase de déploiement de CRISP-DM comprend deux types d'activité :
   - Planification et surveillance du déploiement des résultats.
   - Exécution de tâches de synthèse, telles que la production d'un rapport final et la révision du projet. [BER10]

6. **Production du rapport final :**

   L'élaboration d'un rapport final permet non seulement de compléter les points manquants de la documentation antérieure mais également de communiquer les résultats. Même si cette tâche peut paraître simple, il est important de présenter nos résultats aux différentes personnes ayant un intérêt à les connaitre. Il peut s'agir non seulement des administrateurs techniques responsables de la mise en oeuvre des résultats de la modélisation mais aussi des commanditaires (marketing et gestion) qui prendront des décisions en fonction de nos résultats. Nous commencerons par tenir compte des personnes qui liront notre rapport. S'agit-il de développeurs techniques ou de responsables intéressés par le marché ? Nous devrons peut être créé des rapports distincts en fonction de chaque type de personne si leurs exigences différent. Dans les deux cas, notre rapport doit inclure la majorité des points suivants :

   - Une description complète du problème initial ;
   - Le processus utilisé pour effectuer le Data Mining ;
   - Les coûts du projet ;
   - Des remarques sur tout écart par rapport au plan de projet initial ;
   - Un récapitulatif des résultats du Data Mining (modèles et constations) ;
   - Une présentation du plan proposé pour le déploiement ;
   - Des recommandations pour tout travail de Data Mining ultérieur, incluant des pistes intéressantes issues de l'exploration de la modélisation.

2.3 Une application du data mining : le scoring

Pour illustrer concrètement l'apport et la mise en oeuvre des techniques et méthodes de Data Mining dans le monde de l'entreprise, nous allons parler d'une importante branche du Data Mining : le calcul de score, ou " scoring ". Il s'agit de l'application à l'entreprise des techniques de classement. Par son ancienneté et son universalité, on peut voir le scoring comme l'archétype des applications du Data Mining en entreprise.

2.3.1 Les différents types de scores

Les principaux types de scores utilisés dans la banque sont :

- Le score d'appétence (synonyme : score de propension à consommer, score d'affinité),
- Le score de (comportement) risque,
- Le score d'octroi,
- Le score de recouvrement,
- Le score d'attrition.

Nous les définissons comme suit.

1. **Le score d'appétence** mesure la probabilité d'un client d'être intéressé par un produit ou un service donné. Il est calculé pour un individu ou un foyer client de la banque depuis quelque mois, sur la base des données caractérisant le fonctionnement de ses comptes et de ses produits bancaires pendant cette période, ainsi que de ses caractéristiques sociodémographiques.

2. **Le score de (comportement)** risque mesure la probabilité d'un client avec un compte courant, une carte bancaire, une autorisation de découvert ou un crédit, de rencontrer un incident de paiement ou de remboursement. Il est calculé pour un individu client de la banque depuis quelques mois, sur la base de ses caractéristiques sociodémographiques, et des données caractérisant le fonctionnement de ses comptes et produits bancaires pendant cette période. Il s'agit d'un score comportemental de risque.

3. **Le score d'octroi (score d'acceptation)** est un score de risque calculé pour un client qui est nouveau ou a une faible activité avec la banque. On ne dispose pas de données historisées (ou pas suffisamment) pour ce nouveau client, et le risque est calculé en temps (quasi) réel, au moment ou le client sollicite la banque sur la base de données déclaratives (notamment socioprofessionnelles) fournies par le client, croisées avec des données de géomarketing fournissant le niveau de vie et les habitudes de consommation dans la zone d'habitation du client. On peut aussi calculer un score d'octroi pour un client déjà connu, si l'on veut intégrer au calcul des éléments propres à la demande.
4. **Le score de recouvrement**
   évalue le montant susceptible d'être récupéré sur un compte ou un crédit au contentieux, et peut suggérer les actions de recouvrement les plus efficaces, en évitant des actions disproportionnées pour des clients fidèles, rentables et sans véritable risque.

5. **Le score d'attrition**
   mesure la probabilité d'un client de quitter la banque. Il est calculé pour un individu client de la banque depuis au moins plusieurs mois, sur la base des données caractérisant le fonctionnement de ses comptes et de ses produits bancaires pendant cette période, de ses relations avec la banque, ainsi que de ses caractéristiques sociodémographique. L'attrition est plus complexe à calculer pour un client que pour un produit, pour deux raisons. La première est qu'il y a plusieurs façons de partir : en diminuant les avoirs, ou en conservant les avoirs et en diminuant les flux. La seconde est que le score est moins fiable si l'on veut suffisamment tôt le départ du client, et qu'il ne devient parfois très fiable qu'à un moment tellement proche du départ du client que celui-ci est inéluctable. L'utilisation d'un score sur le lieu de vente exige la restitution de l'information sous une forme simple et immédiatement compréhensible. Une concertation avec les utilisateurs finaux aboutit fréquemment au principe d'un découpage des notes de score en trois classes de valeur : faibles, moyennes, fortes. [BAR01]

2.4 Déploiment du score

2.4.1 Score stratégique :

Un score stratégique est un score utilisé dans une démarche proactive de ciblage de clients pour des actions commerciales plus ou moins centralisées.

2.4.2 Score opérationnel :

Un score opérationnel est un score, de comportement risque (pour des clients) ou d'octroi (pour les clients et les prospects), destiné à l'aide à la décision des chargé de clientèle.

La différence entre les deux types de scores, stratégique et opérationnel, est que le second est plutôt utilisé de façon " réactive " (initiative du client) et lors d'un contact (en agence, au téléphone ou sur le site internet de l'établissement), tandis que le premier est plutôt utilisé de façon " proactive " (initiative de la banque) et en marketing direct. [TUF10]

Conclusion

L'avenir de l'exploration des données dépend de celui des données numériques. Avec l'apparition du Web 2.0, des blogs,..., il y a une explosion du volume des données, et les gisements de matière première pour la fouille de données sont donc importants. De nombreux domaines exploitent encore peu la fouille de données, pour leurs besoins propres. Pour que les problèmes liés à la vie privée des personnes soient réglés, la fouille de données peut aider à traiter des questions dans plusieurs domaines. Enfin, avec l'apparition de nouvelles données et de nouveaux domaines, les techniques continuent à se développer, nous allons en développer quelques unes dans le chapitre qui suit.

Chapitre 3

Les techniques du DATA MINING

Sommaire

POUR effectuer les taches du Data Mining il existe plusieurs techniques issus de disciplines scientifiques diverses (statistiques, intelligence artificielle...), afin de faire apparaître des corrélations cachées dans des gisements de données pour construire des modèles à partir de ces données. Dans ce chapitre, nous présentons les techniques du Data Mining les plus connues, nous parlerons brièvement de certaines d'entre elles, et nous détaillerons concernant les techniques utilisées dans l'étude de cas, à savoir, l'analyse factorielle des correspondances multiples et la régression logistiques.

3.1 Classement des techniques du Data Mining

Le Data Mining permet d'accomplir les six types d'analyses suivantes :

1. Description ;
2. Classification ;
3. Association ;
4. Estimation ;
5. Segmentation ;
6. Prévision.

Ces types d'analyses se répartissent dans les techniques descriptives et prédictives. Nous affinons cette classification dans le tableau suivant :

Techniques descriptives		**Techniques prédictives**		
Corrélation simple	**Corrélation complexe**	**Présent**		**Future**
Description	Classification Association	Estimation	Segmentation	Prévision

TABLE 3.1 – Les 6 grand types de techniques du Data Mining

Notre approche, pour la présentation des principales techniques utilisées dans le Data Mining a été de les diviser en méthodes descriptives et méthodes prédictives.

3.2 Les techniques descriptives

La description consiste à mettre au jour pour une variable donnée, la répartition de ses valeurs (tri, histogramme, moyenne, minimum, maximum, etc.) Pour deux ou trois variables données, faire ressortir des liens entre les répartitions des valeurs des variables. Ces liens s'appellent " tendances ". L'intérêt est de favoriser la connaissance et la compréhension des données, parmi les différentes techniques utilisées dans la description, nous avons :

3.2.1 La classification

La classification automatique ou plus simplement classification, est la plus répandue des techniques descriptives d'analyse de données et de Data Mining. Elle est utilisée quand on dispose d'un grand volume de données au sein duquel on cherche à distinguer des sous-ensembles homogènes, susceptibles de traitements et d'analyses différenciés.

3.2.1.1 Principe :

La classification est l'opération statistique qui consiste à regrouper des objets (individus ou variables) en nombre limité de groupes, les classes (ou segments, ou clusters), qui ont deux propriétés. D'une part, ils ne sont pas prédéfinis par l'analyste mais découverts au cours de l'opération, contrairement aux classes du classement. D'autre part, les classes de la classification regroupent les objets ayant des caractéristiques similaires et séparent les objets ayant des caractéristiques différentes (homogénéité interne et hétérogénéité externe), ce qui peut être mesuré par des critères tels l'inertie interclasse. Comme le classement, la classification consiste à répartir les objets en groupes. Toutefois cette

répartition n'est pas effectuée en fonction d'un critère prédéfini, et ne vise pas à rassembler les objets possédant la même valeur pour ce critère. Autrement dit, on ne sait pas à l'avance la classe à laquelle chaque objet appartient, contrairement au classement. Même le nombre de classes n'est pas toujours fixé à l'avance. Cela vient de ce qu'il n'y a pas de variable à expliquer : la classification est descriptive et non prédictive. Elle est beaucoup utilisée en marketing, médecine, sciences humaines... En marketing, on lui donne souvent le nom de segmentation ou typologie ou analyse typologique. Les Anglo-Saxons parlent de clustering. Enfin, les spécialistes de réseaux de neurones parlent de reconnaissance de forme non supervisée. [TUF10]

3.2.1.2 Intérêt

1. Favoriser, grâce à la Meta typologie, la compréhension et la prédiction.
2. Fixer des segments qui serviront d'ensemble de départ pour des analyses approfondies.
3. Réduire les dimensions, c'est-à-dire le nombre d'attributs, quand il y'en a trop au départ.

3.2.1.3 Méthodes

1. Classification hiérarchique ;
2. Classification des K moyennes ;
3. Réseaux de kohonen.

3.2.2 La recherche d'associations

La détection des règles d'association est une autre technique descriptive, l'une des plus populaires du Data Mining, surtout dans la grande distribution ou elle permet d'analyser les produits simultanément acheté par un client. Ceci explique les surnoms parfois données à cette technique : Analyse du panier de la ménagère (" market basket analysis "), ou analyse du ticket de caisse.

3.2.2.1 Principe

Rechercher des règles d'associations consiste à rechercher les règles du type : " Si pour un individu, la variable $A = X_a$, la variable $B = X_b$, etc, alors, dans 80% des cas, la variable $Z = X_z$,cette configuration se rencontrant pour 20% des individus. " Autrement dit, on recherche les valeurs conjointes les plus fréquentes d'un ensemble de variable d'une base. La valeur 80% est appelée indice de confiance et la valeur de 20% est appelée indice de support de la règle $(A = X_a, B = X_b, ...) \longrightarrow (Z = X_z)$.

La première partie de la règle est appelée " antécédent " ou " condition " ; la seconde " conséquent " ou " résultat " ; les expressions de la forme $A = X_a$ sont appelées " items ". Dans une règle d'associations un item n'est jamais à la fois dans la condition et le résultat. Une règle est donc une expression de la forme : Si condition alors résultats. [NC03]

3.2.2.2 Intérêt

Mieux connaitre les comportements.

3.2.2.3 Méthodes

1. Algorithme a priori ;
2. Algorithme du GRI (induction de règle généralisée).

3.2.3 L'analyse factorielle

- **Principe**
  Pour les analyses multivariées, les techniques factorielles sont très appréciées des statisticiens, auxquels elles permettent, à la fois de représenter en deux ou trois dimensions, le plus fidèlement possible, les individus d'une population, et aussi de détecter les liaisons entre les variables ainsi que les variables séparant le mieux les individus. Elles font appel à l'algèbre linéaire et à un outil très bien adapté à la classification et à la reconnaissance des formes : l'oeil. Un simple coup d'oeil permet de repérer d'éventuels groupes isolés d'individus. Les techniques factorielles sont aussi un puissant outil de réduction de dimensions d'un problème, qui permet de diminuer le nombre de variables étudiées en perdant le moins possible d'information. Cela est parfois fort utile comme traitement préalable à la mise en oeuvre de certains algorithmes sensibles au nombre de variables en entrée, tels que les réseaux de neurones, mais sert aussi parfois avant une classification. [TUF10]

- **Méthodes**

3.2.3.1 Analyse en composante principale

L'analyse en composantes principales est une technique d'analyse des données qui permet, à partir de p variables numériques analysées, de construire $m(\leq p)$autres variables, appelées composantes principales ou facteurs, qui sont des combinaisons linéaires des variables analysées, et qui présentent d'intéressantes caractéristiques :
- Les composantes principales sont ordonnées selon l'information qu'elles restituent, la première étant celle qui restitue le plus d'information ;
- On sait quelle part d'information restitue chaque composantes principales, et des critères permettent de décider combien de composantes principales il est judicieux de conserver ;
- Les composantes principales sont des vecteurs indépendants, c'est-à-dire des variables non corrélées linéairement entre elles (l'ACP n'est donc pas affectée par la présence de données corrélées) ;
- On a une inégalité stricte m < p s'il existe des relations linéaires entre les variables analysées ;
- Les composantes principales (en tout cas, les premières) sont moins sujettes aux fluctuations aléatoires que les variables analysées. [AGR07]

3.2.3.2 Analyse factorielle des correspondances

Rappelons que le tableau de contingence de deux variables qualitatives (ou discrétisées) A et B, de modalités $(a_k)_k$ et $(b_l)_l$ est le tableau (x_{ij}) dans lequel :
La valeur x_{ij} = le nombre d'individus x tels que $A(x) = a_i$ et $B(x) = b_j$.
Le test du χ^2 permet de détecter une dépendance entre les deux variables. Les effectifs et la contribution au χ^2 de chaque cellule du tableau de contingence montrent les liaisons entre modalités des deux variables : soit sureffectif (forte liaison positive), soit sous-effectif (forte liaison négative), soit équilibre (faible liaison). La lecture de ce tableau renseigne bien sur les rapports entre les deux variables, mais, s'il y a de nombreuses modalités, il est fastidieux de parcourir toutes les cellules. Et s'il y avait plus de deux variables à croiser, la lecture serait encore plus difficile. L'analyse

factorielle des correspondances (AFC), ou analyse des correspondances binaires, remédie à cette difficulté en offrant une visualisation en deux dimensions des tableaux de contingence :
– Deux modalités de A et B liées positivement (sureffectif) sont proches ;
– Deux modalités de A et B liées négativement (sous-effectif) sont opposées ;
– Les plus fortes oppositions sont sur le premier axe (horizontal) ;
– Les modalités non liées aux autres sont au centre.
On voit que le principe de l'AFC est séduisant, ce qui explique sa popularité dans l'analyse des données " à la française ". [AGR07] [BEN82]

3.2.3.3 Analyse des correspondances multiples

L'analyse des correspondances multiples (ACM) est une extension de l'analyse factorielle des correspondances qui s'applique à plus de deux variables qualitatives. Dans ce cas, on ne peut plus se contenter du tableau de contingence simple, et il faut ou bien le généraliser, ou bien lui substituer le tableau disjonctif complet. [TUF10]
La généralisation du tableau de contingence simple est le tableau de contingence multiple, également dit tableau de Burt. Dans ce tableau carré symétrique, au lieu que les lignes correspondent aux modalités d'une seule variable, et les colonnes aux modalités d'une autre, ce sont à la fois les lignes et les colonnes qui correspondent aux modalités de l'ensemble des variables X_1,X_2,...,X_p. Au croisement des deux modalités, on trouve le nombre d'individus possédant cette modalité. Au croisement de deux modalités différentes, appartenant à des variables différentes, on trouve le nombre d'individus possédant à la fois la première et la seconde modalité : au croisement de la Kème modalité et x_{ik} de X_i et de la Ième modalité x_{jl} de X_j, on a le nombre ? d'individus x tels que $X_i\ (x) = x_{ik}$ et $X_j\ (x) = x_{jl}$. Le tableau de contingence multiple est composé de blocs, dans lesquels les blocs diagonaux sont des tableaux de contingence simples des variables avec elles-mêmes, et les autres blocs sont les tableaux de contingence simples des variables prises deux à deux.

$$\left|\begin{array}{ccccccc} & \vdots & & x_{jl} & & \vdots & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{jk} & \vdots & \cdots & \alpha & & \vdots & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & \vdots & & \vdots & & \vdots & \end{array}\right|$$

Quant au tableau disjonctif complet, il n'est rempli que de 0 et de 1, il a une ligne par individu, et un nombre de colonnes égal à la somme du nombre de modalités $\sum_{k=1}^{p} m_k$ de l'ensemble des variables X_1,X_2,...,X_p. Le croisement de la i^e ligne et de la colonne correspondant à la modalité x_{jk} de la variable X_j vaut 1 si X_j (i^e individu) $=x_{jk}$, et 0 sinon. C'est ce tableau qui permet d'obtenir des résultats sur les individus et pas seulement sur les variables ; il est donc fort utile.

$$\left|\begin{array}{ccccc} & \vdots & & X_J & & \vdots & \\ & \vdots & & x_j & & \vdots & \\ & \vdots & & \vdots & & \vdots & \\ i & \vdots & 0 & 1 & 0 & \vdots & \\ & \vdots & & \vdots & & \vdots & \end{array}\right|$$

La possibilité de réaliser une ACM à partir du tableau disjonctif complet se fonde sur le fait qu'une AFC fournit les mêmes axes factoriels, qu'elle soit calculée sur le tableau de contingence ou sur le tableau disjonctif complet. En revanche, les valeurs propres sont très différentes selon le tableau à partir duquel elles sont calculées (si λ est valeur propre du tableau disjonctif complet, alors λ^2 est valeur propre du tableau de Burt), de même que la somme de ces valeurs propres, qui est, égale à l'inertie totale. Cette inertie totale est égale au χ^2 divisé par le nombre total d'individus dans la population, lorsque le tableau de contingence est utilisé, et avec le tableau disjonctif complet elle vaut :

$$\frac{m_1 + m_2}{2} - 1$$

Ou m_1 et m_2 sont les nombres de modalités des deux variables X_1 et X_2 Dans une ACM sur le tableau disjonctif complet, on obtient une inertie totale égale à :

$$\frac{1}{p}(\sum_{i=1}^{p} m_i) - 1$$

Ou p est le nombre de variables et m_i est le nombre de modalités de la i^{eme} variable. Cette somme de valeurs propres ne dépend pas de la structure des données c'est-à-dire qu'elle ne dépend pas des liaisons entre variables, et n'a donc pas de signification statistique particulière. L'inertie d'une modalité d'effectif n_j, c'est-à-dire sa contribution à l'inertie totale, est :

$$\frac{1}{p}(1 - \frac{n_j}{n}),$$

Ce qui montre au passage la nécessité d'éviter d'avoir des modalités d'effectifs trop petits, pour ne pas déséquilibrer les résultats. L'inertie d'une variable à m, modalités vaut donc :

$$\sum_{j=1}^{m_i} \frac{1}{p}(1 - \frac{n_j}{n}) = \frac{m_i - 1}{p},$$

Et, comme elle dépend de son nombre de modalités, on voit qu'il y a intérêt à éviter des disparités entre les nombres de modalités des différents variables. Ces points sont à prendre en compte lors de la phase préparation des données développé au chapitre précédent. Bien sûr, se pose la question du nombre d'axes factoriels à retenir et de l'interprétation à en donner. Il faut tout d'abord savoir que le nombre de valeurs propres non trivialement égales à 0 ou 1, c'est-à-dire le nombre d'axes factoriels, est :

$$\sum_{i=1}^{p} m_i - p,$$

Ce qui entraine, compte tenu de la valeur de la somme des valeurs propres indiquée plus haut pour le tableau disjonctif complet, que la valeur moyenne des valeurs propres faut $1/p$. l'analogue du critère de Kaiser de l'ACP consiste donc à ne retenir que les axes dont les valeurs propres sont supérieures à $1/p$. Un second critère est, comme pour l'ACP, l'existence d'un coude dans le diagramme en bâtons des valeurs propres. En revanche, contrairement à l'ACP, le pourcentage d'inertie totale expliquée par les premiers axes n'est pas forcément significatif ; il est souvent assez faible, à cause du grand nombre de modalités rencontrées. Le cas d'une valeur propre égale à 1, parfois rencontré en AFC, est ici exceptionnel. Le pourcentage d'inertie expliquée est encore plus petit quand l'ACM est effectuée sur le tableau disjonctif complet, en raison du nombre de colonnes créées par le codage disjonctif. Dans ce cas, une solution a été proposée pour remédier à ce pessimisme par J. -P. Nakache et al. Dans un article des cahiers de l'analyse des données. Elle consiste à considérer, non les valeurs propres, mais leurs carrés ou d'autres fonctions particulières. Ces transformations assurent que les premières valeurs propres représentent un plus grand pourcentage d'inertie que les premières valeurs propres du tableau disjonctif complet. En raison du faible pourcentage d'inertie expliquée et sa dépendance à la méthode employée, les valeurs propres et les pourcentages d'inertie sont rarement importants dans l'interprétation d'une ACM. Ce sont des mesures pessimistes de la qualité d'une ACM et il est abusif de parler de part d'information restituée au sujet des pourcentages d'inertie. En pratique, on dépasse rarement les cinq premiers axes.[AGR07][AND03]
Une fois que sont déterminés les axes factoriels à conserver, comment les interpréter ? Le plus intéressant est de repérer les modalités apportant la plus forte contribution à chaque axe factoriel. Cette contribution vaut :

$$\frac{1}{\lambda}\frac{n_j}{n.p}(v_j)^2$$

Ou λ est la valeur propre de l'axe, v_j est la coordonnée de la modalité sur cet axe, et les autres notations sont comme ci-dessus. On s'intéressera généralement aux modalités dont la contribution est supérieure au poids :

$$\frac{n_j}{n.p},$$

C'est-à-dire dont la coordonnée v_j est supérieure à $\sqrt{\lambda}$.Un axe est expliqué par les modalités à forte contributions.
Quand on regarde la représentation d'une modalité sur un axe, il faut s'assurer de la qualité de cette représentation. Elle est mesurée par le cosinus carré de l'angle de la modalité avec l'axe. Ce cos^2 est le pourcentage pris par l'axe dans la dispersion de la modalité, et plus il est proche de 1, meilleure est la représentation de la modalité. Il faut n'apprécier la proximité de deux modalités sur un axe que si elles ont toutes deux un cos^2 assez grand sur cet axe. Quand, sur un même axe, deux modalités ont des coordonnées élevées (donc sont éloignées du centre) mais que l'une a un cos^2 plus élevé que l'autre, cela signifie que toutes deux sont différentes du profil moyen (représenté par le centre) mais que la différence à la moyenne est, pour l'une plus que pour l'autre, expliquée par cet axe et non par d'autres caractéristiques. Notons que la somme des cos^2 sur l'ensemble des axes vaut 1. Quant aux variables

supplémentaires, même si elles ne contribuent pas aux axes, leurs cos^2 peuvent être analysés.

Récapitulatif de l'ACM

Les atouts de l'ACM sont multiples :

1. L'ACM permet d'appréhender les liaisons non linéaires (de degré > 1) entre des variables continues préalablement discrétisées, et détecter des dépendances entre variables dont le coefficient de corrélation linéaire est pourtant proche de 0 ;
2. Elle permet de représenter simultanément individus et modalités sur un même plan (en utilisant les cosinus carrés pour s'assurer de la qualité de la projection) ;
3. Permet de visualiser les variables supplémentaires sans les prendre en compte dans le calcul des correspondances.

Dans la représentation graphique d'une ACM :

1. Deux individus sont proches s'ils ont à peu près les mêmes modalités ;
2. Deux modalités de deux variables différentes sont proches si ce sont presque les mêmes individus qui possèdent ces modalités (fort sureffectif dans le tableau de contingence) ; en particulier, elles sont confondues si elles sont possédées par exactement les mêmes individus ;
3. Deux modalités d'une même variable sont proches si les deux groupes d'individus qui les possèdent se ressemblent vis-à-vis des autres variables.

De plus, une modalité est d'autant plus éloignée du centre que son effectif est petit puisque le carré d^2 de la distance au centre est inversement proportionnel à l'effectif. On a $d^2 = (\frac{n}{n_j}) - 1$. De telles modalités peuvent à elle seules suffire à déterminer presque exclusivement les premiers axes factoriels, et à escompter complètement les phénomènes généraux intéressants derrière des phénomènes particuliers ne concernant qu'une poignée d'individus. D'où la nécessité d'éviter les modalités d'effectifs trop petits.

3.3 Les techniques prédictives

Les techniques prédictives de statistique et de Data mining, sont les plus anciennement, les plus souvent et les plus utilement mises en oeuvre. Les techniques prédictives sont nombreuses, en progrès constant et s'appliquent à des problèmes variés. Parmi ces techniques, on distingue deux grandes opérations : le classement (ou discrimination) et la prédiction (ou régression). Ces deux opérations visent à estimer la valeur d'une variable (dite variable " à expliquer ", " cible ", " réponse ", " dépendante " ou " endogène ") d'un individu ou d'un objet en fonction de la valeur d'un certain nombre d'autres variables du même individu, indiquées comme variables explicatives (dites encore variables " indépendantes ", " de contrôle " ou " exogènes "). Ce qui les distingue est la nature de la variable à expliquer : qualitative dans le cas du classement, continue dans le cas de la prédiction. [TUF10]

3.3.1 Les arbres de décision

La technique de l'arbre de décision est employée en classement pour détecter des critères permettant de répartir les individus d'une population en n classes (souvent n=2) prédéfinies. On commence par choisir la variable qui, par ses modalités, sépare le mieux les individus de chaque classe, de façon à avoir des sous-populations, que l'on appelle noeuds, contenant chacune le plus possible d'individus d'une seule classe, puis on réitère la même opération sur chaque noeud obtenu jusqu'à ce que la séparation des individus ne soit plus possible ou plus souhaitable (au vu de certains critères dépendant du type d'arbre). Par construction, les noeuds terminaux (les feuilles) sont tous majoritairement constitués d'individus d'une seule classe. Un individu est affecté à une feuille, et donc à une certaine classe avec une assez forte probabilité, quand il satisfait l'ensemble des règles permettant d'arriver à cette feuille. L'ensemble des règles de toutes les feuilles constitue le modèle de classement.

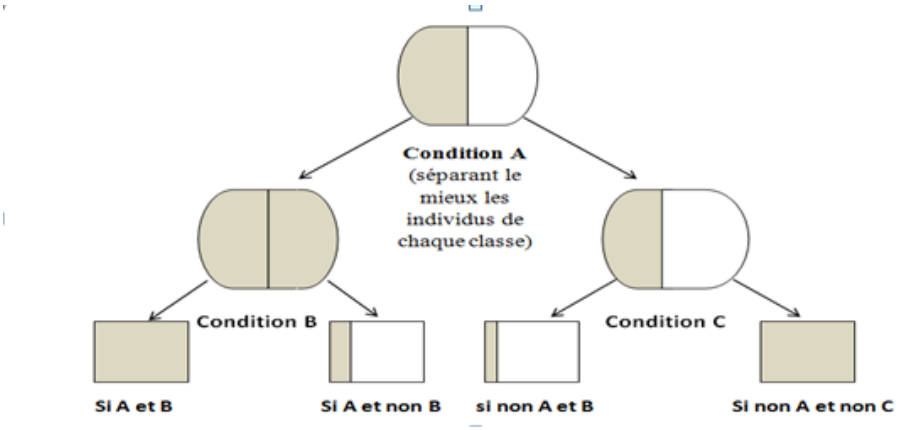

FIGURE 3.1 – Schéma d'un arbre de décision

Les arbres de décision conçus pour le classement permettent généralement d'effectuer une prédiction. Le principe est que la variable à expliquer doit avoir une variance plus faible dans les noeuds-fils que dans les noeuds-père, et aussi avoir une moyenne la plus distincte possible d'un noeud-fils à un autre.[LAM06]

3.3.2 Régression logistique

Si ses racines plongent assez loin dans l'histoire de l'analyse des données (Verhulst 1838), la régression logistique a été introduite plus récemment dans les logiciels, et donc dans la pratique quotidienne de la plupart des statisticiens. La régression logistique se définit selon Desjardins (2005) comme une technique permettant d'ajuster une surface de régression à des données lorsque la variable dépendante est dichotomique. Cette technique est utilisée pour des études ayant pour but de vérifier si des variables indépendantes peuvent prédire une variable dépendante dichotomique. En outre, la régression logistique peut correspondre à une technique statistique dont l'objet est, à partir d'un fichier d'observations, de produire un modèle permettant de prédire les valeurs prises par une variable catégorielle, le plus souvent binaire, en se basant sur une série de variables explicatives continues et/ou binaires. Contrairement à la régression multiple et l'analyse discriminante,

la régression logistique n'exige pas une distribution normale des prédicteurs ni l'homogénéité des variances. Par ses nombreuses qualités donc, cette technique est de plus en plus préférée à l'analyse discriminante par les statisticiens et les spécialistes du scoring. [DES05]

La régression logistique devient universelle, puisqu'elle peut traiter des variables à prédire à 2 valeurs (sans faire d'hypothèse aussi restrictives que l'analyse discriminante), à $k \geq 3$ valeurs ordonnées, à $k \geq 3$ valeurs nominales, et que les variables explicatives peuvent être quantitatives ou qualitatives. Qui plus est, ses résultats sont très explicites, surtout dans sa version logit avec les odds-ratios très populaire en médecine et en épidémiologie. La régression logistique est enfin largement répandue dans des domaines nombreux et divers. D'abord utilisée dans la médecine (caractérisation des sujets malades par rapport aux sujets sains par exemple), cette technique de classement et de prédiction s'est rependue dans la banque assurance (détection des groupes à risque), la science politique (explication des intentions de vote), le marketing (fidélisation des clients)... [TUF10]

C'est pour cette raison que cette méthode a été utilisé dans la modélisation, dans notre étude de cas.

3.3.2.1 Principe de la régression logistique binaire :

Dans la régression logistique binaire, on considère une variable à expliquer (variable " cible ") binaire Y=0 ou 1, et p variables explicatives X_j continues, binaires ou qualitatives (dont les indicatrices ramènent au cas d'une variable binaire). Dans les notations qui suivent, on rassemble les p variables x_j dans un vecteur $X = (X_1,X_2,...X_p)$. Les variables cibles qualitatives à $k \geq 2$ modalités sont traitées par ce que l'on appelle la régression logistique polytomique, les k modalités pouvant etre ordonnées (régression logistique ordinale) ou non (régression logistique multinomiale = régression logistique nominale).

Dans tout problème de régression, on cherche à écrire l'espérance conditionnelle de la variable à expliquer Y comme combinaison linéaire de répresseurs X. On regarde l'espérance car ce n'est qu'en moyenne, calculée pour chaque p-uplets de valeurs des p régresseurs, que la variable à expliquer est linéaire par rapport aux régresseurs : la variable à expliquer peut fluctuer autour de sa moyenne. L'objectif de la régression logistique est donc celui de toute régression : modéliser l'espérance conditionnelle $E(Y|X = x)$. On veut connaitre la valeur moyenne de Y pour toute valeur de X. pour une valeur Y valant 0 ou 1 (loi de Bernoulli), cette valeur moyenne est la probabilité que Y=1. On a donc :

$$E(Y/X = x) = Prob(Y = 1/X = x)$$

.

En régression linaire, on cherche à faire passer un hyperplan au milieu du nuage des points$(x_1,x_2,...x_p,y)$, de sorte que l'ensemble des valeurs moyennes de Y pour toutes les valeurs de X est approché par cet hyperplan, d'équation :

$$E(Y/X = x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

.

Cette approximation ne convient évidemment plus lorsque $Y = 0$ ou 1, puisque le terme $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$ est non borné alors que Prob $(Y = 1/X = x)$ est dans l'intervalle $[0, 1]$.

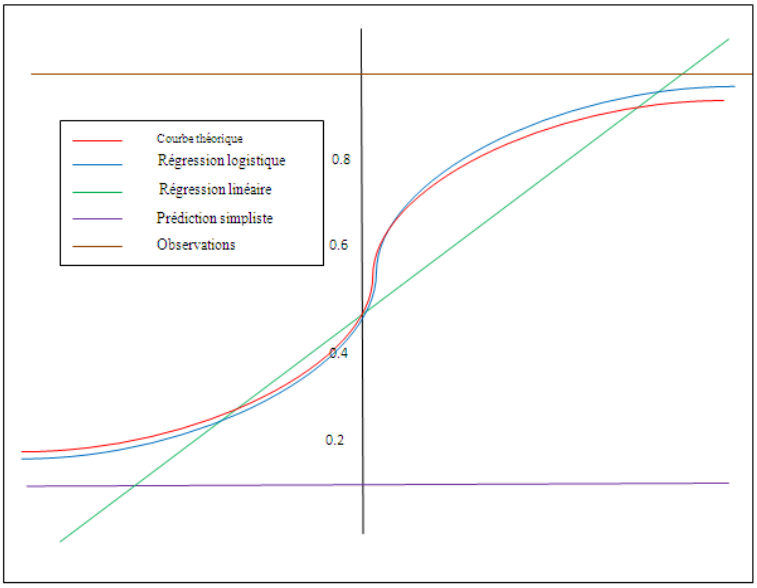

FIGURE 3.2 – Comparaison des régressions linéaires et logistique

En réalité, dans la situation favorable (celle qui nous intéresse) ou X parvient à discriminer les valeurs de Y, le nuage des points (x, y) a une allure dont (ou $p = 1$) : quand x est petit, on a plus souvent $y = 0$, et quand x est grand, on a plus souvent $y = 1$. Pour simplifier, nous nous sommes placés ici dans le cas très classique d'une variable x multinormale et homoscédastique : $x \sim N(0, 1)$sur l'ensemble des points tels que $y = 0$ et $x \sim N(1, 1)$ sur l'ensemble des points tels que $y = 1$.

Le cas extrême dit de la " séparation complète ", ou existe un x_0 tel que $y = 0$ pour tout $x \leq x_0$et $y = 1$ pour tout $x \geq x_0$: paradoxalement, la régression logistique ne sait pas trouver de solution à cette situation si simple.

Revenons au cas de la Figure : visiblement, les valeurs $Prob(Y = 1/X = x)$ quand x varie suivent la courbe théorique représentée par les croix " + " sur la figure. C'est une courbe en S et non une droite. Cette forme de courbe est appelé courbe logistique. Si l'on suit l'expression de cette courbe, on peut écrire $\pi(x) = Prob(Y = 1/X = x)$ sous la forme :

$$\pi(x) = \left(\frac{e^{\beta_0 + \sum_j \beta_j x^i}}{1 + e^{\beta_0 + \sum_j \beta_j x^j}}\right)$$

Équation équivalente à :

$$\log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

.

La fonction $f(p) = log(\frac{p}{1-p})$ est appelée logit. C'est un cas particulier des fonctions de lien rencontrées dans les régressions logistiques et les modèles linéaire généralisés. Dans ce type de modèle, ce n'est donc pas l'espérance $\pi(x) = E(Y/X = x)$ qui est écrite comme combinaison linéaire des variables explicatives, mais $f(\pi(x))$, f étant la fonction de lien. Dans la régression logistique la plus courante, on modélise le logit de l'espérance conditionnelle comme combinaison linéaire des variables explicatives. Cette écriture est cohérente avec la règle bayésienne de l'analyse discriminante et le calcul de la probabilité a posteriori qui en découle dans le cas d'une distribution normale de X/Y avec égalité des variances et égalité des probabilités a priori. Mais il existe d'autres courbes en S, car si la variante logit de la régression logistique apparait comme la plus naturelle, ne serait-ce qu'en raison de son lien avec l'analyse discriminante, il en existe d'autres que nous présentons dans un tableau, avec leur fonction de lien et leur fonction de transfert (inverse de la précédente) correspondantes.

Modéle	**Fonction de lien**	**Fonction de transfert**
Logit	$Log(\frac{\mu}{[1-\mu]}$	$\frac{e^t}{1+e^t}$
Probit(normit)	Fonction inverse de la fonction de répartition d'une loi normale centrée réduite	$s(t) = \int_{-\infty}^{t} \frac{(e^t - z^2)}{\sqrt{2\pi}}$
Log-log(complémentaire)	$Log[-Log(1-\mu)]$	$1 - e^{-e^t}$

TABLE 3.2 – Les différentes fonctions de liens

Le modèle log-log (aussi appelé gombit en référence à Gompertz) a une courbe en S dissymétrique mais très proche de celle du logit quand t est petit et la probabilité est inférieure à 0,1.

En raison de son lien avec la loi normale, le probit est parfois appelé normit. Il est aujourd'hui beaucoup moins prisé que le logit. Bien que sa fonction de transfert ait un tracé en S rassemblant à celui du logit, le probit offre moins d'avantages que le logit, notamment dans l'interprétation des coefficients. Le probit est donc déconseillé quand de nombreux cas ont une forte ou une faible probabilité, c'est-à-dire quand la queue de la distribution est importante. [TUF10]

3.3.2.2 Les odds-ratios

L'odds-ratio d'une variable explicative mesure l'évolution du rapport des probabilités d'apparition de l'événement $Y = 1$ contre $Y = 0$ lorsque $X_i$passe de x à $x + 1$. Dans ce cas, logit $(\pi(x))$ augmente du coefficient β_i de $X_i$et donc la cote $\frac{\pi(x)}{[1-\pi(x)]}$ est multipliée par $\exp(\beta_i)$.

Ceci s'écrit :

$OR = \frac{\pi(x+1)/[1-\pi(x+1)]}{\pi(x)/[1-\pi(x)]} = e^{\beta_i}$

attention à une simplification abusive fréquente (mais parfois volontaire) : l'odds-ratio est différent du risque relatif $\pi(x + 1)/\pi(x)$, sauf quand $\pi(x)$ est petit (détection de

phénomène rare).

Si X_i est binaire 0/1, la formule de l'odds-ratio devient :

$$OR = \frac{prob(Y = 1/X_i = 1)/prob(Y = 0/X_i = 1)}{prob(Y = 1/X_i = 0)/Prob(Y = 0/X_i = 0)} = e_i^{\beta}$$

Une variable X binaire a un seul odds-ratio. Si l'on s'intéresse à l'apparition d'une maladie (Y=1 pour un malade), un odds-ratio de 1,5 pour la variable " sexe " (=1 pour homme et 0 pour femme) signifiera que le rapport des malades aux bien-portants est 1,5 plus important pour les hommes que pour les femmes.

Quant au variables qualitatives, elles ont autant d'odds-ratios que de modalités moins une, l'une des modalités étant prise pour et son coefficient étant généralement posé égal à 0 (c'est la convention la plus fréquente et la plus commode, mais ce coefficient peut aussi être l'opposé de la somme de tous les autres coefficients).

Un odds-ratio < 1 (un coefficient < 0) indique une influence négative de la variable explicative sur la variable à prédire, et un odds-ratio > 1 (un coefficient > 0) indique une influence positive. Pour la modalité de référence, les logiciels proposent souvent par défaut la dernière modalité, mais ils permettent parfois le choix. Dans ce cas, on choisit parfois la modalité la plus fréquente pour référence, mais on peut préférer choisir une modalité de référence qui soit extrême du point de vue de la cible (plus faible ou plus fort risque, par exemple), afin que les coefficients des modalités de la variable aient tous de même signe. On peut gagner en lisibilité, surtout quand le nombre de modalités dépasse trois.

Quand toutes les variables explicatives sont qualitatives, l'ensemble des modalités de référence est représenté par la constante β_0 : un individu " moyen " dont toutes les modalités sont celles de référence a pour probabilité $\pi(x) = Prob(Y = 1/X = x) = exp(\beta_0)/[1 + exp(\beta_0)]$. On voit donc un intérêt qu'il y a à choisir pour toutes les variables la modalité la plus fréquente pour référence : les diverses modalités de référence risquent moins d'être incompatibles et l'individu " moyen " a plus de chances d'exister. [TUF10]

3.3.2.3 Estimation des paramètres

Les paramètres à estimer dans un modèle logistique logit sont les coefficients β_i de la combinaison linéaire exprimant le logit de la probabilité Prob $(Y = 1/X = x)$. La régression logistique, et plus généralement le modèle linéaire généralisé, diffère du modèle linéaire simple en ce que l'estimation des paramètres ne se fait pas par la méthode des moindres carrés mais par celle du maximum de vraisemblance. Il nous faut parler ici un peu de cette méthode.

La méthode du maximum de vraisemblance consiste à estimer un paramètre β de la loi d'une variable aléatoire X au vu d'un certain nombre d'observations indépendantes, en écrivant une fonction de vraisemblance, fonction de β dont il s'agit de trouver le maximum.

Si la loi est discrète, la fonction de vraisemblance s'écrit par définition :

$$L(\beta, x^1, x^2, ..., x^n) = Prob\beta(X = x^2) \times Prob(X = x^n) \times ... \times Prob\beta(X = x^n)$$

.

Si la loi est continue de densité $f\beta$, la fonction de vraisemblance s'écrit par définition :

$$L(\beta, x^1, x^2, ..., x^n) = f\beta(x^1) \times f\beta(x^2) \times ? \times f\beta(x^n)$$

.

La valeur de β qui maximise $L(\beta, x^1, x^2, ..., x^n)$ est celle qui maximise la probabilité des réalisations observées. La situation est inverse de celle ou l'on connait le paramètre de la loi de X et ou l'on calcule la probabilité d'observer les réalisations x^i : ici, on a observé les x^i et on cherche le paramètre qui maximise la probabilité de les observer. $L(\beta, x^1, x^2, ?, x^n)$ est une fonction de densité si on la voit comme fonction de $(x^1, x^2, ..., x^n)$ est une fonction de vraisemblance si on la voit comme fonction de β. L'indépendance des observations est indispensable pour pouvoir écrire la vraisemblance en $(x^1, x^2, ..., x^n)$ comme produit des vraisemblances de chaque observation x^i. il est noter que cette méthode se généralise au cas de données tronquées, dont on calcule la vraisemblance conditionnelle en divisant la fonction de densité par Prob $(X > s)$, ou s est le seul de troncature.

La recherche d'un maximum se fait en recherchant une valeur pour laquelle la dérivée première s'annule et la dérivée seconde est négative (nous supposons que ces dérivées existent). On cherche le plus souvent à maximiser le logarithme de la vraisemblance, problème équivalent mais plus simple à résoudre, le logarithme transformant les produits en sommes. Cette recherche de maximum est un problème d'optimisation qui se traite par des algorithmes tel celui de Newton-Raphson.

L'estimation du maximum de vraisemblance permet d'estimer le paramètre λ d'une loi de poisson, le paramètre α d'une loi exponentielle, les paramètres μ et σ d'une loi normale, etc. De façon générale, l'estimateur du maximum de vraisemblance peut exister et être unique, ne pas être unique, ou ne pas exister. Dans un problème de régression, nous n'avons plus des lois simples mais des lois conditionnelles et il s'agit de maximiser l'une des fonctions suivantes :

$$Prob\beta(Y = y^1/X = x^1) \times Prob\beta(Y = y^2/X = x^2) \times ...Prob\beta(Y = y^n/X = x^n)$$

. Ou

$$f\beta(Y = y^1/X = x^1) \times f\beta(Y = y^2/X = x^2) \times ...f\beta(Y = y^n/X = x^n)$$

.

$f\beta$ étant une fonction de densité conditionnelle.

Dans le cas de la régression logistique binaire, nous observons des données $[(x^1, y^1), (x^2, y^2), ..., (x^n, y^n)$ dans lesquelles chaque y^i vaut 0 ou 1, et x^i est le vecteur de variables explicatives de la $i^{ème}$ observation. Si $y^1 = 1$, la probabilité d'obtenir (x^i, y^i) sera par définition Prob $(Y = 1/X = x^1) = \pi(x^i)$. Si $y^i = 0$ probabilité d'obtenir (x^i, y^i) sera Prob $(Y = 0/X = x^i) = 1 - \pi(x^i)$. on peut unifier ces deux cas en écrivant que la probabilité d'obtenir (x^i, y^i) est :

$$\pi(x^i)y^i(1 - \pi x^i)(1 - y^i)$$

quel que soit y^i. Maintenant, afin de poursuivre les calculs et d'appliquer la formule précédente, il faut faire intervenir une hypothèse fondamentale au cadre de la régression logistique : les observations (x^i, y^i) sont indépendantes. Cette hypothèse ne pourra être levée que dans une régression logistique sur données corrélées. Cette hypothèse d'indépendance permet d'écrire la fonction de vraisemblance comme produit des probabilités :

$$\prod_{i=1}^{n} \pi(x^i) y^i(x^i)(1-y^i)$$

En remplaçant $\pi(x^i)$ par l'expression qui est la sienne dans le modèle logistique, on voit que la fonction de vraisemblance vaut :

$$\prod_{i=1}^{n} (\frac{e^{\beta_0+\sum_j \beta j x^i}}{1+e^{\beta_0+\sum_j \beta j x^i}})^{y^i} (1-\frac{e^{\beta_0+\sum_j \beta j x^i}}{1+e^{\beta_0+\sum_j \beta j x^i}})^{1-y^i}$$

C'est cette fonction $L(\beta, x^1, x^2, ..., x^n)$ des coefficients β_j qui doit être maximisée : il faut trouver les coefficients tels que $L(\beta, x^1, x^2, ..., x^n)$ soit le plus proche possible de 1, car cela signifiera que le modèle s'ajuste le mieux possible aux données observées. En réalité, la vraisemblance ne peut valoir 1, et le modèle contient autant de coefficients qu'il ya d'observations (x^i, y^i) distinctes. Un tel modèle est qualifié de saturé. Pour prendre une analogie, le modèle linéaire simple est saturé quand le nuage de points se réduit à deux points, et la droite ajuste parfaitement le nuage formé des deux points se réduit à deux points. La détermination du meilleur modèle logistique passera donc par une recherche des coefficients qui maximisent la vraisemblance. C'est un problème qui n'a pas de solution analytique, c'est-à-dire de solution s'exprimant directement à partir des données initiales à l'instar des fonctions discriminantes de l'analyse discriminante que l'on obtient en inversant la matrice des covariances. Ici la solution optimale $(\beta_0, \beta_1, ..., \beta_p)$ sera trouvée par une méthode numérique itérative, les plus répandues de ces méthodes numérique étant les algorithmes de Newton-Raphson et de Fisher. Cette absence de solution analytique peut-être le principal inconvénient de la régression logistique, qui la rend plus difficile à programmer pour un éditeur de logiciel, en rend le calcul plus long que celui d'une analyse discriminante, et peut rendre dans certains cas exceptionnels impossible l'obtention d'une solution fiable. L'algorithme ne converge pas lorsque les groupes sont complètement séparés, alors que l'analyse discriminante reste efficiente dans ce cas.[TUF10] Par ses multiples avantages et sa facilité d'application, nous utiliserons cette méthode dans la modélisation.

3.3.2.4 Tests statistiques de la régression logistique

Sans anticiper sur les tests généraux sur les modèles prédictifs que sont la matrice de confusion, la courbe ROC et l'indice de Gini, on peut récapituler les tests spécifiques à la régression logistique. Les quatre premiers viennent d'être vus :

- Le test du χ^2 sur les indicateurs de Wald, qui doivent être < 3.84 ;
- Les intervalles de confiance à 95% des odds-ratios ne doivent pas contenir 1
- La valeur de $-2LogL(\beta_k)$ doit être la plus basse possible, ou on effectue un test du χ^2 sur la modification du -2Log-vraisemblance quand on enléve un coefficient β_k(hypothèse nulle :β_k) ;
- Les critères d'Akaiké (AIC° et de Schwartz(BIC), qui doivent être le plus bas possible ;
- Le R^2 de Cox-Snell et le R^2 ajusté de Nagelkerke ;
- Le test de Hosmer et Lemeshow sur la comparaison des proportions observées et théoriques ;
- Le test de la déviance normalisée et du χ^2 de Pearson normalisé ;
- Les tests de concordance (liés à l'aire sous la courbe ROC et l'indice de Gini).

Le R^2 de cox-Snell est un équivalent du R^2 de la régression linéaire, défini à l'aide des vraisemblances par :

$$R^2 = 1 - [\frac{L(\beta_0)}{L(\beta_k)}]^{\frac{2}{n}}$$

Et qui ne peut donc dépasser (pour un modèle saturé) :

$${R^2}_{max} = 1 - [L(\beta_0)]^{\frac{2}{n}}$$

s Le R^2 de Nagelkerke (" max-rescaled R-square ") est le quotient $\frac{R^2}{{R^2}_{max}}$ qui varie entre 0 et 1. [TUF10]

3.4 Les réseaux de neurones

Il est difficile de parler du Data Mining sans parler des réseaux de neurones, qui sont à la base à la fois de certaines techniques descriptives et de certaines techniques prédictives de Data Mining. Ils sont largement répandus grâce à leur puissance de modélisation (ils peuvent approcher n'importe quelle fonction suffisamment régulière), qui fait merveille dans une grande variété de problèmes, face à des phénomènes complexes, des formes irrégulières, des données difficiles à appréhender et ne suivant pas de loi probabiliste particulière. Cependant leur utilisation est parfois freinée par les difficultés qu'elle présente : le côté " boite noire " des réseaux, la délicatesse des réglages à effectuer, la puissance informatique requise et surtout les risques de sur-apprentissage et de convergence vers une solution globalement non optimale.

3.4.1 Principe

Les réseaux de neurones sont des modèles représentants le fonctionnement du système nerveux. Les unités de base sont les neurones. Ils sont généralement organisés en couches, comme l'illustre la figure ci-dessous.

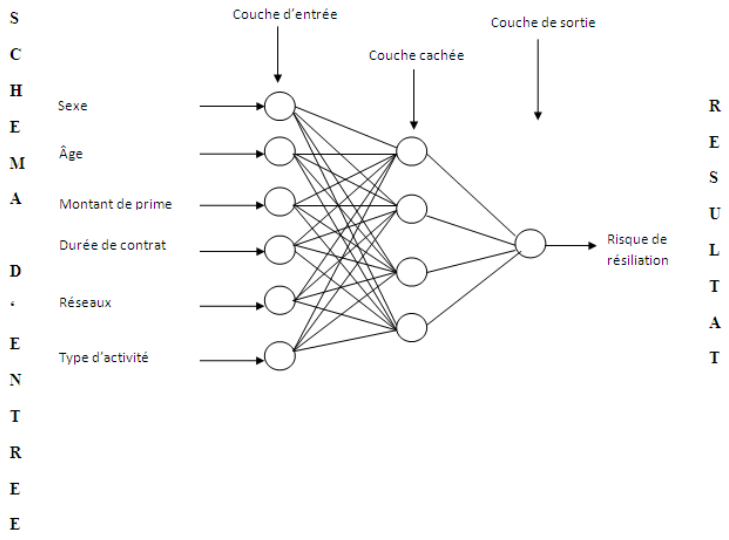

FIGURE 3.3 – Shéma d'un réseau de neurones

Un réseau de neurones, également appelé perceptron multicouche, est un modèle simplifié de la façon dont le cerveau humain traite les informations. Le fonctionnement de ce modèle repose sur la simulation d'un grand nombre d'unités de traitement simples interconnectées, qui sont en quelque sorte des versions abstraites de nos neurones. Ces unités de traitement sont organisées en couches. Il existe généralement trois types de couche dans un réseau de neurones : une couche d'entrée dans laquelle les unités représentent les champs d'entrée, une ou plusieurs couches cachées, ainsi qu'une couche de sortie dans laquelle des unités représentent les champs de sortie. Les unités sont reliées entre elles par des connexions de puissance (ou de pondération) différentes. Les données d'entrée sont présentées dans la première couche et les valeurs transmises entre les neurones d'une couche à l'autre. Le résultat final est obtenu à partir de la couche de sortie.
Lors de son apprentissage, le réseau procède à l'examen de tous les enregistrements afin de générer des prévisions et modifie les pondérations lorsque l'une de ses prévision s'avère incorrecte. Ce processus se répète plusieurs fois et le réseau continue d'améliorer ses prévisions jusqu'à ce que l'un des critères d'arrêt soit atteint.
Au début, tous les coefficients de pondération sont aléatoires et les réponses en provenance du réseau risquent de ne pas avoir de sens. Le réseau apprend à travers l'apprentissage. Les exemples dont le résultat est connu sont présentés à plusieurs reprises au réseau et les réponses qu'il donne sont comparées aux résultats connus. Les informations de cette comparaison sont réacheminées via le réseau, modifiant progressivement les coefficients de pondération. Au fur et à mesure de l'apprentissage, les résultats connus répliqués par le réseau sont à chaque fois plus précis. Lorsque l'apprentissage est terminé, le réseau peut être appliqué à d'autres observations pour lesquelles le résultat est inconnu. [TUF10] [WIS09] [PAR04].

De façon générale, les étapes dans la mise en oeuvre d'un réseau de neurones pour la prédiction ou le classement sont :

1. L'identification des données en entrée et en sortie,
2. La normalisation de ces données,
3. La constitution d'un réseau avec une structure adaptée,
4. L'apprentissage du réseau,
5. Le test du réseau,
6. L'application du modèle généré par l'apprentissage,
7. La dénormalisation des données en sortie.

3.4.2 Les principaux réseaux de neurones

Il existe différents modèles de réseaux de neurones. Les principaux, le perceptron multicouches (PMC), le réseau à fonction radiale de base (RBF : Radial Basis Function) et le réseau de Kohonen. Plus récents, les réseaux par estimation de densité de Speckt (1990) sont utilisés, soit pour le classement (réseau PNN : probabilistic neural networks), soit pour la prédiction (réseau GRNN : general regression neural networks). Il existe aussi des réseaux analogues aux RBF mais basés sur la théorie mathématique des ondelettes. Le

réseau de Kohonen est un réseau à apprentissage non supervisé, utilisé pour la classification automatique, tandis que les autres réseaux cités, PMC, RBF... sont des réseaux à apprentissage supervisé, utilisés avec en sortie une ou plusieurs variables à expliquer.

3.5 L'analyse discriminante

Avant la diffusion de la régression, logistique, l'analyse discriminante de Fisher fut longtemps LA grande méthode de classement, utilisée dans de nombreux contextes allant de la biologie, avec les travaux fondateurs de Fisher en 1936, jusqu'au crédit scoring. Aujourd'hui encore, cette méthode est celle qui est privilégiée par la banque de France pour scorer, cette méthode, même si elle est limitée, fournit des prédictions explicites, précises et robustes, pourvu que l'on ait bien préparé les données. Par ailleurs, un prolongement inventé par Gilbert Saporta, sous le nom de méthode DISQUAL, a permis à l'analyse discriminante d'étendre son cadre hors des variables explicatives quantitatives pour traiter aussi les variables qualitatives. Cette méthode se trouve au carrefour des méthodes paramétriques, semi-paramétriques (régression logistique) et non-paramétriques (estimation de densité de probabilité), et a aussi à voir avec l'analyse en composantes principales. [TUF10]

3.5.1 Principe

Voici la situation-type traitée par l'analyse discriminante : on a un ensemble d'individus appartenant chacun à un groupe, le nombre de groupes étant fini et supérieur à 1. Deux problèmes se posent à nous : trouver une représentation des individus qui sépare le mieux les groupes (analyse discriminante descriptive) ou trouver des règles d'affectation des individus à leur groupe (analyse discriminante prédictive). Une autre formulation est la suivante : on a un ensemble d'individus caractérisés par une variable à expliquer Y qualitative et des variables explicatives Xi quantitatives .on peut vouloir trouver une représentation des liaisons entre Y et les X_i (analyse discriminante descriptive), ou vouloir trouver des règles de prédiction des modalités de Y à partir des valeurs des X_i (analyse discriminante prédictive). L'analyse discriminante offre plusieurs approches à cette double problématique. [TUF10]

3.5.2 L'analyse discriminante géométrique descriptive (analyse factorielle discriminante)

On a une variable cible (à expliquer) Y qualitative à K modalités, correspondant à K groupes G_i dont on note n_i les effectifs. L'analyse factorielle discriminante consiste à remplacer les X_j par des axes discriminants, c'est-à-dire des combinaisons linéaires des X_j prenant les valeurs les plus différentes possibles pour des individus différant sur les variables cible. On reconnaitra dans ce mécanisme une analyse en composantes principales du nuage des K centres de gravité des classes (pondérés par ni/n). Les axes sont au nombre de K-1 ou p, le plus petit des deux. On peut illustrer simplement l'approche géométrique descriptive :

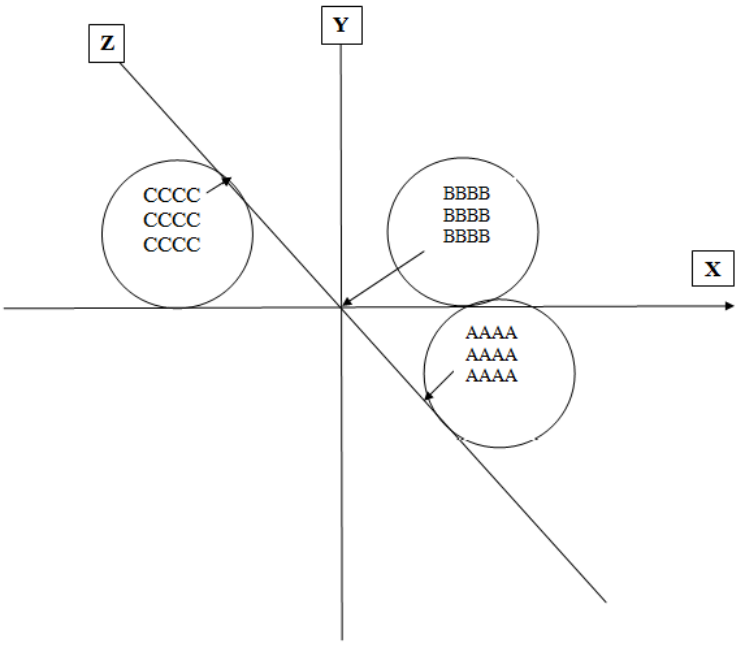

FIGURE 3.4 – Analyse factorielle discriminante

Dans cet exemple, on voit que :
- L'axe " x " sépare bien les groupes " B " et " C " mais non les groupes " A " et " B ".
- L'axe " y " sépare bien les groupes " A " et " B " mais non les groupes " B " et " C ".
- Et en revanche l'axe " z ", combinaison linéaire de " x " et " y ", sépare bien les trois groupes.
- La droite d'équation z=1 sépare les " B " et les " C ", tandis que la droite d'équation z= -1 sépare les " A " et les " B " : donc " z " est une fonction de score.[TUF10]

3.5.3 L'analyse discriminante géométrique prédictive

Il s'agit dans ce cas de construire une fonction de classement (règle d'affectation, ...) qui permet de prédire le groupe d'appartenance d'un individu à partir des valeurs prises par les variables prédictives. En ce sens, cette technique se rapproche des techniques supervisées en apprentissage automatique telles que les arbres de décision, les réseaux de neurones, ...Elle repose sur un cadre probabiliste. Le plus connu est certainement l'hypothèse de distribution multinormale (loi normale). Additionnée à l'hypothèse d'homoscédasticité, les nuages de points conditionnels ont la même forme, nous aboutissons à l'analyse discriminante linéaire. Elle est très séduisante dans la pratique car la fonction de classement s'exprime comme une combinaison linéaire des variables prédictives, facile à analyser et à interpréter. Cette technique est, avec la régression logistique, très utilisée dans le scoring, lorsque nous voulons par exemple caractériser l'appétence - la propension à acheter - d'un client face à un nouveau produit. [TUF10]

3.5.4 L'analyse discriminante sur variables qualitatives (méthode DISQUAL)

Les variables quantitatives suffisent à décrire le fonctionnement financier d'une entreprise et à construire un score de risque par une analyse discriminante linéaire. Quand on a voulu appliquer cette méthode aux scores d'octroi aux particuliers, mis en oeuvre lorsqu'un organisme spécialisé de crédit à la consommation veut examiner le risque attaché à la demande d'un particulier , par exemple un client achetant un meuble à crédit dans un grand magasin, le problème s'est posé de devoir prendre en compte des variables qualitatives propres aux personnes physiques : sexe, situation de famille, catégorie socio-professionnelle, etc. C'est pour traiter ce genre de situation que Gilbert Saporta inventa en 1975 la méthode DISQUAL (Discrimination sur variables Qualitatives).

Elle consiste à partir des variables qualitatives, à découper en classes toutes les variables quantitatives (plutôt avec le même nombre de modalités et des effectifs proches), à effectuer une analyse des correspondances multiples sur le tableau disjonctif complet de ces variables, à récupérer les coordonnées (continues) des individus sur les axes factoriels les plus discriminantes (les autres axes représentent du " bruit statistique ", puis à injecter ces coordonnées en entrée d'une analyse discriminante linéaire classique. On obtient alors une fonction de score de Fisher combinaison linéaire des axes factoriels. Or, comme ces axes sont eux-mêmes des combinaisons linéaires des indicatrices des modalités des variables initiales (qualitatives), la fonction de Fisher peut s'exprimer comme une combinaison linéaire d'indicatrices de modalités, ce qui revient à attribuer une note à chacune de ces modalités. Outre l'intérêt de pouvoir traiter des variables qualitatives, et d'éviter la plupart des inconvénients de l'analyse discriminante énumérés un peu plus bas , on voit que la méthode DISQUAL fournit ses résultats sous une forme très pratique : les coefficients (les " notes ") de deux modalités sont comparables, puisque il s'agit d'indicatrices et non de variables quantitatives de grandeurs éventuellement très différentes, et on peut d'ailleurs normaliser les coefficients de façon à avoir un score compris entre 0 et 100, par exemple. [TUF10]

3.5.5 Inconvénients de l'analyse discriminante :

L'analyse discriminante linéaire classique, quand elle ne bénéficie pas du perfectionnement apporté par la méthode DISQUAL, présente plusieurs inconvénients.

1. Elle ne s'applique en principe qu'aux variables explicatives continues sans valeurs manquantes, même si on peut tolérer les variables explicatives discrètes. Pour les variables discrètes ou qualitatives, ou en cas de phénomènes non linéaires, on peut recourir à la méthode DISQUAL.
2. Elle est sensible aux individus hors norme.
3. Elle requiert théoriquement les hypothèses de multinormalité, d'homoscédasticité et d'indépendance linéaire des variables explicatives. L'hétéroscédasticité (non homoscédasticité) peut être due à la présence d'individus hors norme. Quant à l'existence de relations linéaires entre les variables explicatives, ou colinéarité, elle entraine une moindre stabilité des résultats et éventuellement des aberrations dans le signe des paramètres.

Chapitre 4

Modélisation des résiliations des contrats d'assurance vie et élaboration d'un score d'attrition

Sommaire

Dans cette partie nous allons mettre en oeuvre deux techniques exposées précédemment afin de construire un outil de scoring "'score d'attrition"' qui pourra être utilisé dans une démarche proactive de ciblage clients pour des actions commerciales. Afin que le travail soit bien organisé, la méthode CRISP-DM a été suivie.

4.1 Compréhension du problème

L'attrition, c'est l'usure, c'est-à-dire le fait de rompre ou pas le contrat. Le score d'attrition est un cas d'étude de data mining très important. La problématique qui se pose est de pouvoir minimiser le nombre d'attrition (résiliation) avec un processus data mining qui permettra un suivi continue et un meilleur ciblage qui va servir dans les actions et compagnes marketing.

4.2 Compréhension des données

Dans cette phase, nous avons défini les données dont on a besoin au niveau opérationnel (Marketing), vérifié leur disponibilité au niveau de la base de donnée, nous avons ensuite défini l'historique à considérer et tous les indicateurs à prendre en compte durant l'étude.

4.3 Exploration et Préparation des données

Nous disposerons de variables explicatives qualitatives ou discrètes pour la plupart, continues pour quelques autres. Après discrétisation des variables continues, nous pourrons utiliser l'ensemble des variables dans une analyse des correspondances multiples, laquelle nous servira à explorer les données et nous assurer de leur cohérence. Puis nous passerons à la phase de modélisation, et nous calculerons un modèle de score en employant la régression logistique binaire. La base de données contient 35937 dossiers d'assurance vie " CNEP Totale Prévoyance " pour une période allant de juin 2009 à décembre 2011, dont 9707 résiliations, soit un taux de 27%. Cette base se compose de la variable à expliquer " cible ", avec ces modalités OUI (résiliation) et NON (pas de résiliation), et de 10 variables explicatives (nous en avons omise quelques-unes, présentant des valeurs manquantes ou bien pour cause de redondance).

- Trois variables numériques continues : l'âge en années, le montant de la prime mensuelle et la durée du contrat.
- Sept variables qualitatives : le type de contrat, le type d'assuré, l'adresse de l'assuré, le type d'activité, le code agence ou le contrat a pris lieu, le sexe et le code produit.

Certaines variables portes sur le produit lui-même (durée, montant, type...), d'autres sur le profil personnel du contractant (âge, adresse, sexe). Toutes ces données sont bien sûr celles mesurées au moment de la signature du contrat d'assurance. Nous avons ajoutés une variables CLE qui est un numéro de client allant de C1 à C35937, et avons construit une nouvelle base contenant que les variables qui ont été retenues.
Constatant que seules trois variables ne sont pas découpées en classes, nous allons commencer par les discrétiser, ce qui permettra ultérieurement de les utiliser conjointement aux autres, à l'aide de mêmes techniques. Ceci procurera plus de simplicité et de lisibilité. Nous avons décidé de diviser la variable âge en 5 modalités : [19-24], [25-34], [35-44], [45-54] et [55,64], l'âge présentant des limites naturelles entre 19 et 64 ans. Puis nous croisant les modalités de l'âge avec la variable à expliquer.

	[19-24]	[25-34]	[35-44]	[45-54]	[55-64]	Total
R-Oui	19.708%	27.267%	27.335%	28.330%	28.347%	27.011%
R-Non	80.292%	72.733%	72.665%	71.670%	71.635%	71.989%

TABLE 4.1 – Tableau de contingence croisant la variable « Tranche d'âge » à la variable « Résiliation ».

Le tableau de contingence montre que la première modalité de l'âge correspond à un taux de résiliation nettement inférieure à celui des autres. Il y'a donc un seuil à 24 ans. De plus nous remarquons un taux presque égal de 25 à 44 ans et de 45 à 64 ans. Le découpage de l'âge en 3 tranches a donc été décidé. Cette démarche a été réitérer pour le montant de la prime, que nous avons vu utile de regrouper avec le type de contrat pour créer une variable " Montant de prime par type de contrat ". Cette variable a été découpée en 7 modalités, nous croisons ces dernières avec la variable résiliation.

	F-A [0 ;265[	F-A [265 ;530[	F-A [530 ;795[	F-T [0 ;530[	F-T [530 ;795[	F-T [795 ;1060[	F-T [1060 ;1590[	Total
R-Oui	17.151%	**14.854%**	**13,645%**	29.647%	34.431%	**38.972%**	**40.281%**	27.011%
R-Non	82.849%	85.146%	86.355%	70.353%	65.569%	61.028%	59.719%	72.989%
Total	100%	100%	100%	100%	100%	100%	100%	100%

TABLE 4.2 – Tableau de contingence croisant la variable « Type de contrat par montant de prime » à la variable « Résiliation ».

Ce tableau montre que pour la formule accidentelle, un regroupement peut être fait entre 265 et 795Da du fait de leur ressemblance en terme de résiliation, de même en ce qui concerne la formule " toutes causes ", pour les montant de prime allant de 795 à 1590Da, qui présente un taux de résiliation atteignant les 40%. Le découpage de cette variable se fera en 5 modalités.
En ce qui concerne les durées de contrats, cette variable présente un seuil naturel de 31 mois, la production ayant débuté en juin 2009. Le choix a été fait de découper cette variable en 5 modalités. Passons à l'examen du tableau de contingence :

	Durée des Contrats [0,4]	Durée des Contrats [5,8]	Durée des Contrats [9,12]	Durée des Contrats [13,20]	Durée des Contrats [21,31]	Total
R-Oui	25.102%	42.978%	**28.385%**	**27.070%**	5.552%	27.011%
R-Non	74.898%	57.022%	71.615%	72.930%	94.448%	72.989%
Total	100%	100%	100%	100%	100%	100%

TABLE 4.3 – Tableau de contingence croisant la variable « Durée des contrats » à la variable « Résiliation ».

Il s'avère utile de regrouper les durées de contrat de 9 à 20 mois, de ce fait nous aurons une variable à 4 modalités.

Ayant noté, des périodes d'animation commerciale, dont le but a été de motiver les différents intervenants au niveau de la banque (chargé de clientèle, directeur régional,...) par un système simple de récompense des intervenants en fonction du nombre de contrats signés.
Nous avons donc décidé de crée une variable " challenge " représentant ces périodes, le but étant de voir l'effet de cette dernière sur la variable cible " résiliation ", et de ce fait de pouvoir répondre à la question " Les challenges influent ils sur les résiliations ? ". Durant la période de production, cinq animations commerciales ont été réalisées, de ce fait, la variable " challenge " sera découpé en 6 modalités, une des modalités représentant les périodes d'hors challenge.

Les autres variables ont été prises comme trouvé dans la base de données sans aucun changement.

Un tableau récapitulatif des différentes variables explicatives ainsi que leurs codifications est donné comme suit :

Variables	Modalités	Codification
Sexe	Homme	M
	Femme	F
Réseaux	Alger centre	ALG C
	Alger est	ALG E
	Alger ouest	ALG O
	Annaba	ANB
	Bejaia	BEJ
	Blida	BDA
	Chleff	CHL
	Constantine	CST
	Ghardaïa	GRD
	Oran centre	ORN C
	Oran est	ORN E
	Sétif	STF
	Tizi-Ouzou	T-O
	Tlemcen	TMN
Challenge	challenge n°1 (sept09-mars10)	C1
	challenge n°2 (oct10-déc10)	C2
	challenge n°3 (fév11-avril11)	C3
	challenge n°4 (mai11-aout11)	C4
	challenge n°5 (sep11-déc11)	C5
	Hors challenge	HC
Type d'activité	Profession libérale	P-lib
	employé secteur public	E-pub
	employé secteur privé	E-pri
	Retraité	Ret
	Sans profession	S-P
	Autres catégories	Atrs
Tranche d'âge	Entre 19 et 24 ans	[19-24]
	Entre 25 et 44 ans	[25-44]
	Plus de 45 ans	[45-64]

Variables	Modalités	Codification
Montant de prime par type de contrat (T-C/M-P)	Formule toutes causes, prime inférieur à 530 Da	F-T [0 ; 530[
	Formule toutes causes, prime entre 530 et 1060 Da	F-T [530 ; 1060[
	Formule toutes causes, prime supérieure à 1060 Da	F-T [1060 ; 1590[
	Formule accidentelle, prime inférieur à 265 Da	F-A [0 ; 265[
	Formule accidentelle, prime entre 265 et 795 Da	F-A [265 ; 795[
Durée de contrat	4 mois et moins	[0,4]
	Entre 5 et 8 mois	[5,8]
	Entre 9 et 20 mois	[9,20]
	21 mois et plus	[21,31]

TABLE 4.4 – Codification des variables

4.3.1 Analyse factorielle des correspondances multiples

Toutes les variables étant à présent disponible sous forme de classe, nous pouvons effectuer une analyse des correspondances multiples.

Pour ce faire, nous recourons au logiciel Xlstat [1] , avec le quel ont été réalisé aussi les différents tableaux de contingence et tests d'indépendance.

Il faut effectuer un petit travail de transformation de la variable réseaux, qui présente à elle seul 14 modalités, car il est nécessaire d'éviter des disparités entre les nombres de modalités des différentes variables, pour ce faire nous avons réalisé une ACM avec les variables prises comme présenté précédemment, suite à cette étape, nous avons décidé de regrouper les réseaux présentant une ressemblance (proximité des projections des modalités sur les axes factorielles), le regroupement se présente comme suit :

- Alger centre, Alger ouest ;
- Alger est, Bejaia, Tizi-Ouzou ;
- Chleff, Oran est ;
- Oran centre, Tlemcen ;
- Constantine, Blida.

Les autres réseaux n'ont pas été regroupés du fait de leur distinction dans la projection réalisé. Une nouvelle ACM a été réalisé sur ce nouveau découpage de la variable réseaux. La variable à expliquer ainsi que " sexe " ont été spécifiés comme des variables " supplémentaires " dans L'ACM, ce qui permettrait de les représenter dans le plan des variables, mais sans les faire intervenir dans les calculs d'axes factoriels.

Le résultat de la projection des différentes modalités sur le plan constitué par les 2 premiers axes factorielles est le suivant :

La légende de la figure suivante montre que les modalités présentant un risque de

1. Xlstat est un outil d'analyse de données et de statistique pour Microsoft Excel, il offre de très nombreuses fonctionnalités qui font d'Excel un outil performant et facile d'accés pour répondre à la majorité des besoins en analyse de données.

résiliation sont visiblement concentrées vers le haut du plan factoriel, et précisément en haut à gauche. En bas, elles sont beaucoup plus rares.

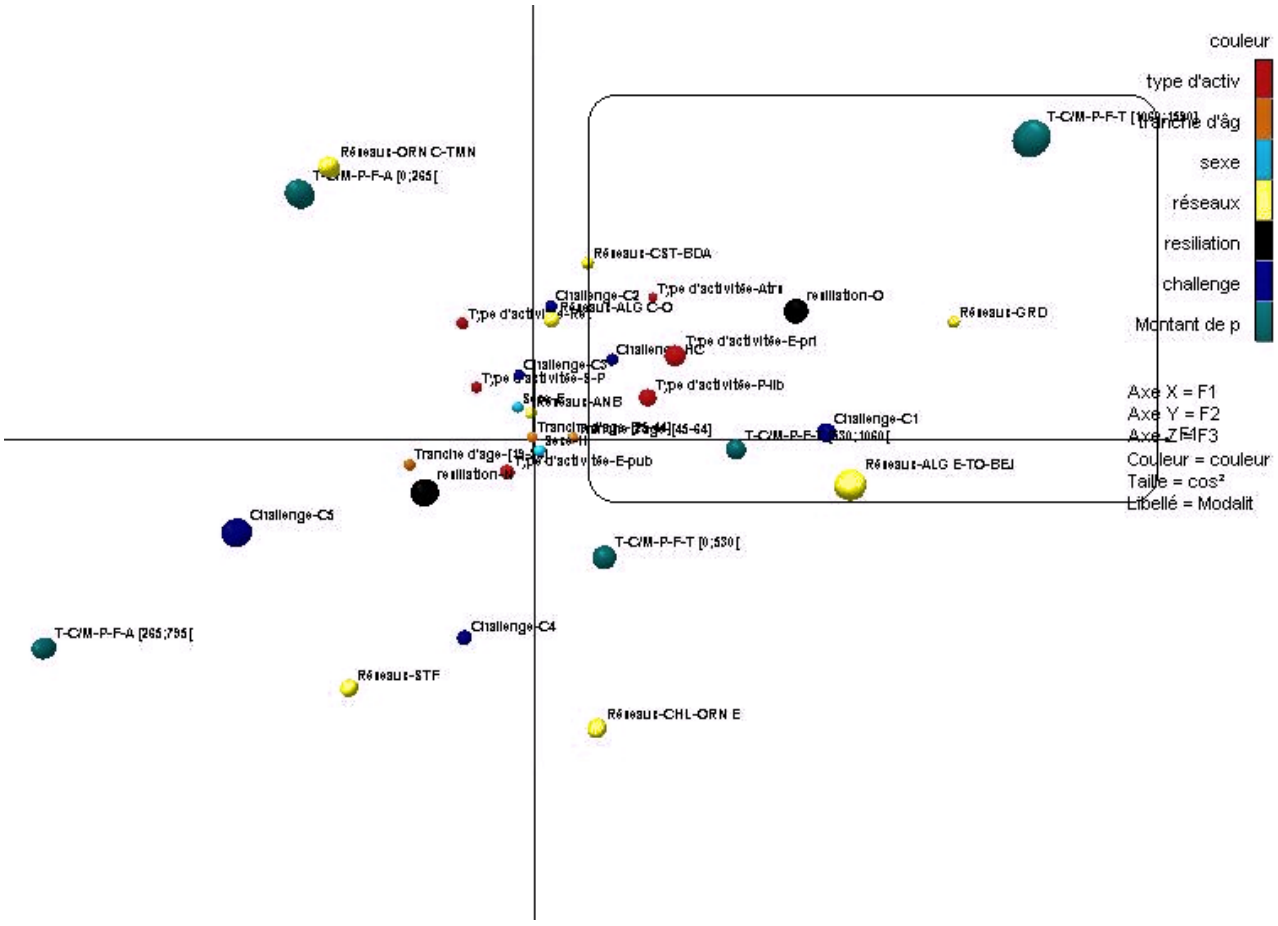

FIGURE 4.1 – Analyse des correspondances multiples sur la base de données.

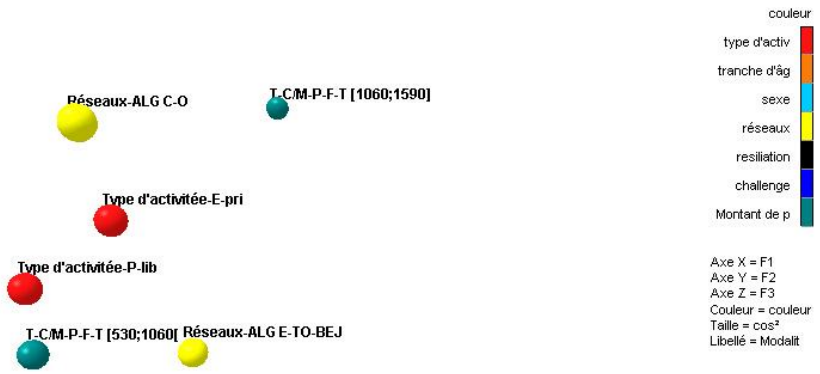

FIGURE 4.2 – Zoom sur la projection.

On remarque que les modalités sont représentées par des boules de dimensions différentes. Ces dimensions sont proportionnelles à la qualité de représentation des différentes modalités sur le plan factorielle " cosinus carrés ". Nous parlerons que des modalités bien représentées sur le plan factorielle et dont le positionnement dans le plan peut légitiment être commenté, en excluant les modalités dont aucun des cosinus carrés n'est élevé. Nous pourrons constater plus loin, dans les tableaux croisés des variables explicatives avec la variable à expliquer, que les modalités non commenté sont aussi celles qui sont les plus rares.

Le plan factoriel permet de vérifier la cohérence des modalités entre elles. Par exemple, les modalités " femme " et " sans emploi " sont proches en haut au milieu.

Nous constatons a première vue que la variable " Challenge " est ordonnée suivant l'axe F1, de plus c'est celle qui a le plus contribué à la construction de cette axe, nous pourrons conclure, que ce dernier représente un axe " période de production (challenge) " allant des plus anciennes période d'animation commerciale " C1 " , complètement à droite, au plus récente " C5 ", complètement à gauche. De ce fait, le positionnement de la modalité " résiliation Oui " n'est pas fortuit, mais il représente le fait que ce sont les contrats anciens qui présente un fort taux de résiliation, logique en terme d'assurance vie. Nous parlerons en détail des résiliations dans les croisements de chacune des variables explicatives avec la variable à expliquer.

De plus, l'ACM permet de déceler des structures particulières non apparentes au départ. Nous remarquons par exemple une séparation entre les modalités représentant la formule " accidentelle ", et ceux de la " toutes causes ", les 1eres ayant des coordonnées négatives par rapport à l'axe F1, contrairement aux autres, dont les coordonnées sont positives. Cette séparation est due au fait que l'axe F1 représente l'axe " période de production ". Après croisement de la variable " montant de prime par type de contrat " avec la variable " challenge ", nous remarquons que plus de 60% des ventes en accidentelle se sont déroulé durant les deux dernier challenge " C4, C5 ".
Des groupements de modalités sont apparents aussi.

En zoomant sur le graphique ont pourra remarquer le groupement suivant :

En examinant de plus prés ce rapprochement, on pourra noter que les réseaux Alger

centre- ouest-Est, Tizi-Ouzou et Bejaia, présentent un profil particulier de clientèle (employé du secteur privé et Profession libérale) qui, prennent des contrats " toutes causes " à montant de prime important allant de 530 à 1590 Da. En effet en regardant de plus prés les tableaux de contingence croisant ces variables, on s'aperçoit que 35% des clients de ces réseaux sont " employé secteur privé " et " profession libérale ", ce qui est loin d'être le cas pour les autres réseaux, ou sa ne dépasse pas les 20%. De plus, prés de 50% des ventes en " Toutes causes, prime allant de 1060 à 1590Da " ainsi que 36% des ventes en " Toutes causes, prime allant de 530 à 1060Da " se sont faites dans ces réseaux.

Nous notons aussi le rapprochement suivant :

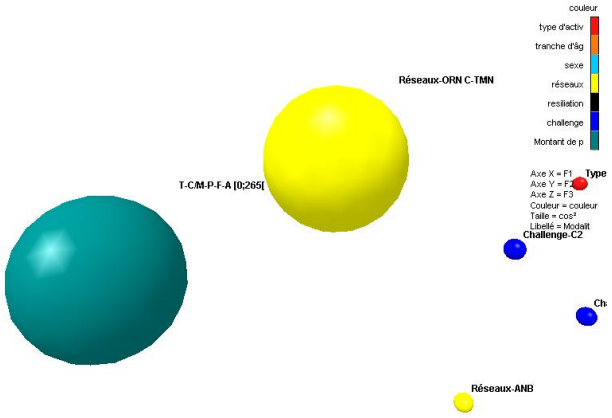

FIGURE 4.3 – Zoom sur la projection.

Après croisement de la variable " réseaux " avec la variable " Montant de prime par type de contrat ", nous constatons que ce rapprochement n'est pas fortuit, mais montre le fait que " 53% " c'est-à-dire plus de la moitié des ventes des réseaux Oran centre et Tlemcen sont de l'accidentelle à montant de prime inférieure à 265 Da. Les commerciaux devront se pencher sur la question.

Avant d'examiner le croisement de chacune des variables explicatives avec la variable à expliquer, nous allons procéder à un classement des ces dernières selon leur pouvoir discriminant, mesuré par la valeur du Khi^2 de Wald développé au chapitre précédent, le quel est fourni par l'analyse de type 3. Ces valeurs sont restituées dans le tableau suivant :

Source	**DDL**	$Khi^2(Wald)$	$Pr > Wald$
Durée des contrats	3	1753.477	< 0.0001
T-C/M-P	4	958.615	< 0.0001
Réseaux	7	801.726	< 0.0001
Type d'activité	5	386.401	< 0.0001
Tranche d'âge	2	81.930	< 0.0001
Sexe	1	25.202	< 0.0001

TABLE 4.5 – Les résultats de l'analyse de type 3

On observe une chute du khi^2 de Wald après la 4eme variable, c'est-à-dire que les variables " tranche d'âge " et " sexe " se révèlent peu prédictif du risque de résiliation, par contre elles seront prises en compte dans la modélisation, car, dans une démarche de scoring. Les variables qui caractérisent le profil des clients sont toujours gardées du fait qu'elles sont moins corrélées aux autres variables et apportent une information différente, intrinsèque et non liée à la relation assurance. Logiquement la variable " durée des contrats " est celle qui discrétise le plus la variable à expliquer, chose que nous avions constaté durant l'analyse factorielle.

Nous noterons que la variable " Challenge " n'a pas été prise en compte dans ce tableau, car, elle ne sera pas prise en compte dans la modélisation, du fait de la forte liaison qui existe entre cette variable et la variable " durée des contrats ".

4.3.2 Tri croisé

Passons maintenant à l'examen des tableaux de contingences croisant les différentes variables explicatives avec la variable résiliation. Nous commençons par la variable la plus liées à la cible, " la durée de contrats ". Pour des durées de contrats dépassant les 21

mois, le risque de résiliation est très faible (moins de 6% de résiliations), c'est-à-dire que, dépassant ce seuil, nous pouvons considérer que les clients sont fidèles à leur contrat. Le pic de résiliation est observé pour les contrats ayant une durée entre 5 et 8 mois, ou on note 43% de résiliations, remarque qui devrait être prise en considération au futur, car dans les pays ou l'assurance vie est développée, la durée moyenne des contrats est de 2 ans.

	Durée des contrats-[0,4]	Durée des contrats-[5,8]	Durée des contrats-[9,20]	Durée des contrats-[21,31]	Total
R-Oui	25.102%	**42.978%**	27.735%	5.552%	27.011%
R-Non	74.898%	57.022%	72.265%	94.448%	72.989%
Total	100%	100%	100%	100%	100%

TABLE 4.6 – Tableau de contingence croisant la variable « Durée des contrats » à la variable « Résiliation ».

	F-A [0 ;265[	F-A [265 ;795[	F-T [0 ;530[	F-T [530 ;1060[	F-T [1060 ;1590[	Total
R-Oui	17.151%	14.365%	29.647%	35.390%	40.281%	27.011%
R-Non	82.849%	85.635%	70.353%	64.610%	59.719%	72.989%
Total	100%	100%	100%	100%	100%	100%

TABLE 4.7 – Tableau de contingence croisant la variable « Montant de prime par type de contrat » à la variable « Résiliation ».

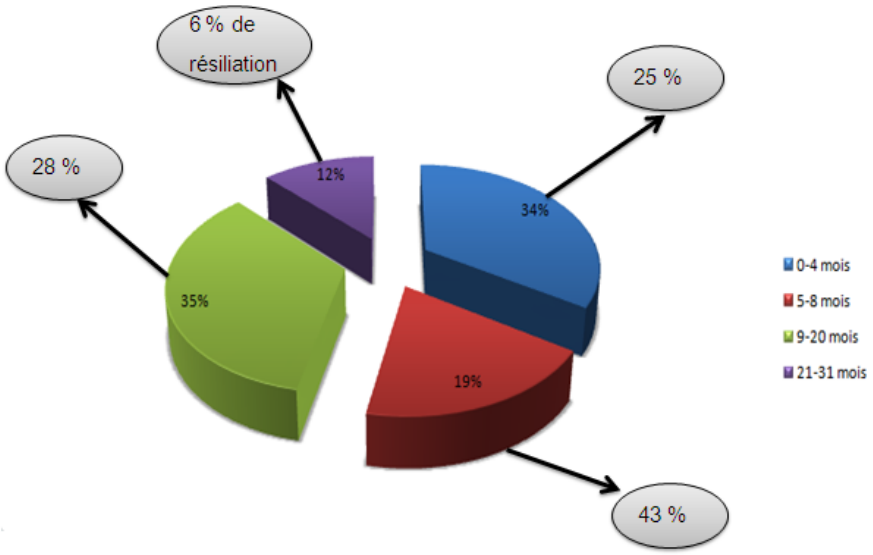

FIGURE 4.4 – Diagramme en camembert du tableau de contingence " Durée des contrats-Résiliation"

Les résiliations sont très logiquement liées aux différents montants de prime. La gradation est très nette depuis les montants de primes inferieure à 530 Da jusqu'aux montants supérieur à 1060Da allant de 29% à 40% de résiliation. En ce qui concerne la formule accidentelle, nous avons au taux de résiliation inferieure à la moyenne (15%), ceci s'explique par le fait que 88% des contrats " Formule accidentelle " sont des contrats à 130Da de prime mensuelle.

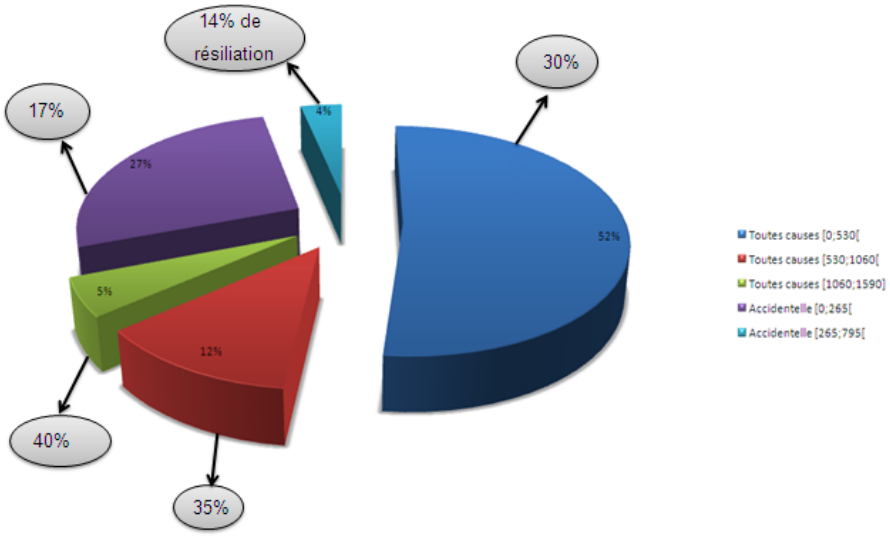

FIGURE 4.5 – Diagramme en camembert du tableau de contingence " Durée des contrats-Résiliation"

De même pour les réseaux, nous remarquerons que ce sont les réseaux qui sont située à droite du plan factoriel qui présentent de fort taux de résiliations, allant de Ghardaïa à 53% de résiliations, Constantine et Blida à 37% et Alger est, Tizi-Ouzou et Bejaia à 31%. Un taux faible par rapport à la moyenne à Sétif (17%), les autres réseaux fluctuent entre 24 et 26%.

	Rsx-ALGE- TO-BEJ	Rsx-ALGC-O	Rsx-ANB	Rsx-CST-BDA	Rsx-CHL-ORN E	Rsx-GRD	Rsx-STF	Rsx-ORN C-TMN	Total
R-Oui	**31.4%**	24.0%	26.0%	**36.6%**	26.9%	**52.6**	17.2%	24.6%	27.0%
R-Non	68.6%	76.0%	74.0%	63.4%	73.8%	47.4%	82.8%	75.4%	73.0%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%

TABLE 4.8 – Tableau de contingence croisant la variable « Réseaux » à la variable « Résiliation »

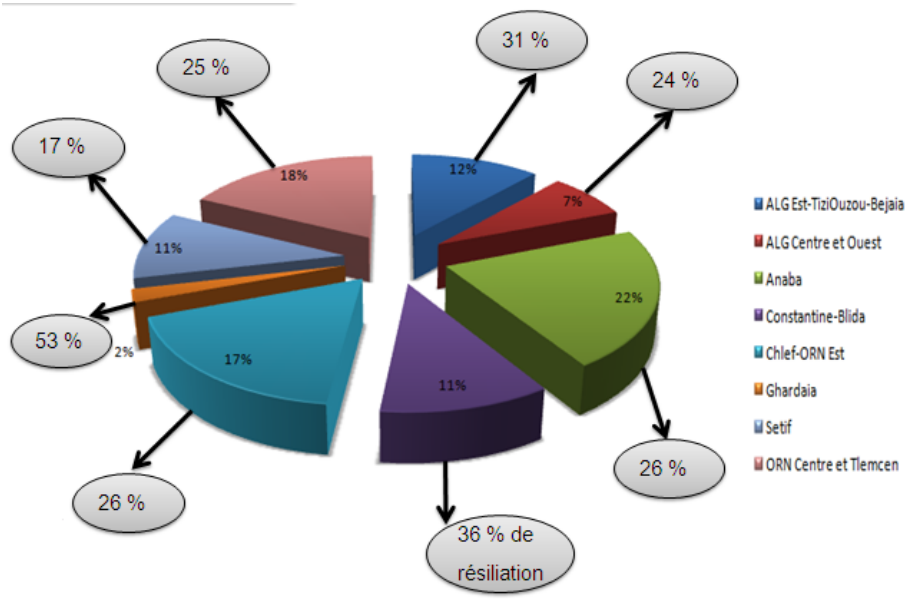

FIGURE 4.6 – Diagramme en camembert du tableau de contingence " Réseaux-Résiliation"

Pour la variable " type d'activité ", " autres catégories " présente un taux très supérieure à la moyenne, sauf que cette modalité ne représente que 1% du porte feuille clients. Bizarrement, 34% de résiliations chez les employés du secteur privé, chose qui était inattendue au départ. Un taux de résiliation inferieur à la moyenne chez les employés du secteur public (23%), qui représentent à eux seuls 62% des ventes.

	Type d'activité P-lib	Type d'activité E-pub	Type d'activité E-pri	Type d'activité Ret	Type d'activité S-P	Type d'activité Atrs	Total
R-Oui	**33.374%**	23.817%	**34.638%**	26.198%	28.427%	**40.421%**	27.011%
R-Non	66.626%	76.183%	65.362%	73.802%	71.573%	59.579%	72.989%
Total	100%	100%	100%	100%	100%	100%	100%

TABLE 4.9 – Tableau de contingence croisant la variable « Type d'activité » à la variable « Résiliation ».

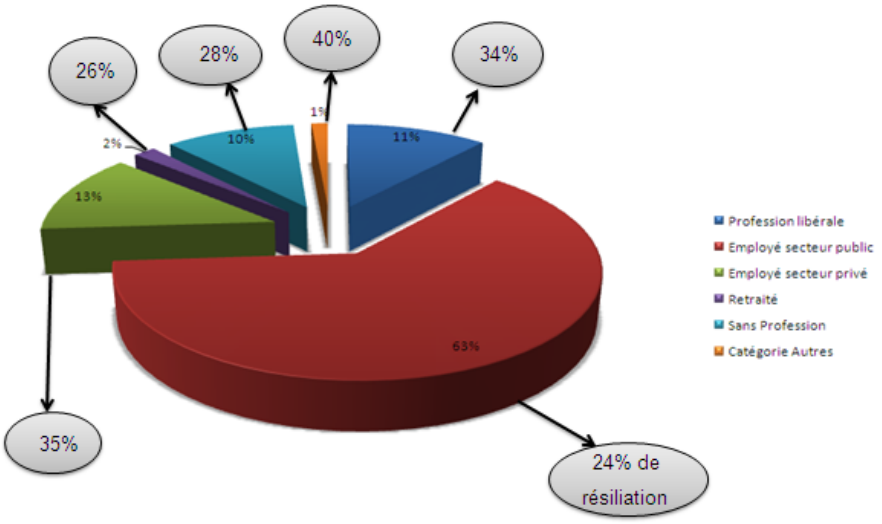

FIGURE 4.7 – Diagramme en camembert du tableau de contingence " Type d'activité-Résiliation"

Les tranches d'âges présentent une tendance un peu particulière. On se serait attendu à un taux de résiliations faible chez les seniors et fort chez les jeunes, dans notre cas, le phénomène est tout autre. Un taux de résiliations de 20% chez les moins de 24 ans, et de 28% chez les plus de 45 ans. De ce fait, il s'avère que moins le client est à un âge avancé et plus il recherche la sécurisation de ses opérations financières.

	Tranche d'âge [19-24]	Tranche d'âge [25-44]	Tranche d'âge [45-64]	Total
R-Oui	19.708%	27.301%	28.333%	27.011%
R-Non	80.292%	72.699%	71.667%	72.989%
Total	100%	100%	100%	100%

TABLE 4.10 – Tableau de contingence croisant la variable « Tranche d'âge » à la variable « Résiliation ».

	Sexe-H	Sexe-F	Total
R-Oui	26.313%	29.076%	27.011%
R-Non	73.687%	70.924%	72.989%
Total	100%	100%	100%

TABLE 4.11 – Tableau de contingence croisant la variable « Sexe » à la variable « Résiliation »

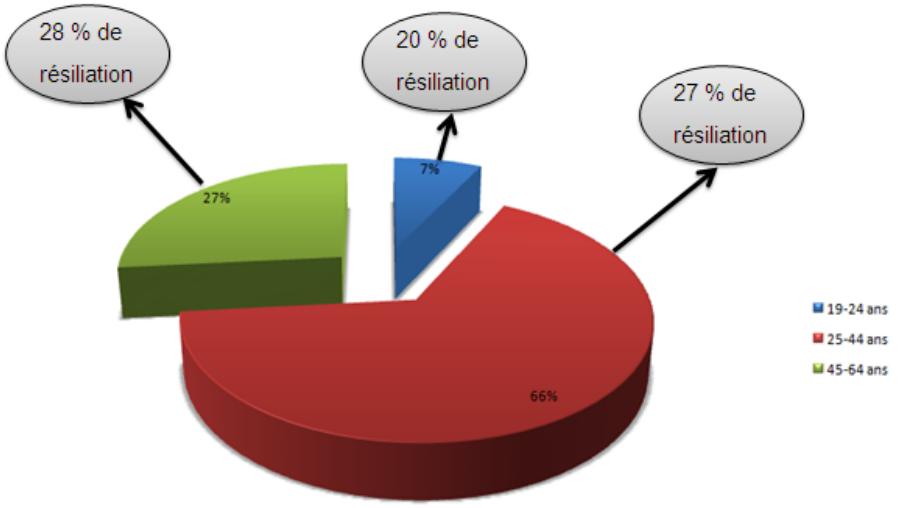

FIGURE 4.8 – Diagramme en camembert du tableau de contingence " Tranche d'âge-Résiliation"

Un taux de résiliation légèrement plus important chez les femmes (29%). Nous remarquerons aussi un taux de 21% de sans emploi chez les femmes, contre 7% chez les hommes, d'où la légère différence dans le taux de résiliation.

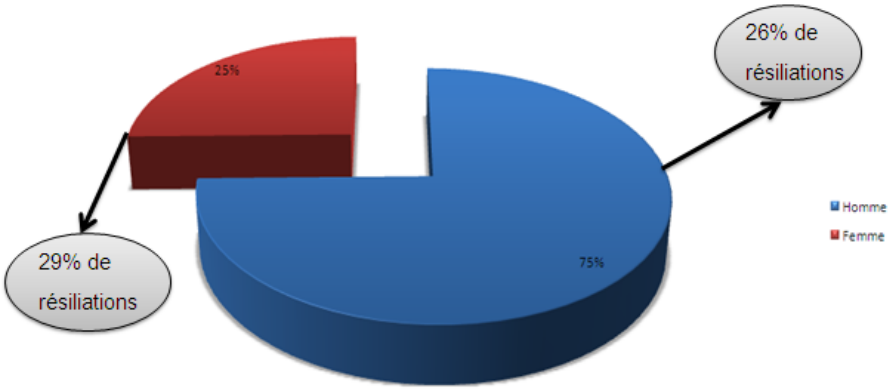

FIGURE 4.9 – Diagramme en camembert du tableau de contingence " Sexe-Résiliation"

4.4 Modélisation

La méthode de modélisation employée dans cette étude de cas est la régression logistique, la plus classique dans ce type de situation, pour les raisons déjà exposé au chapitre précédent. On modélise la modalité " Oui " de la variable cible, c'est-à-dire la résiliation, et ce, à l'aide des variables présélectionnées sur la base des premiers tests statistiques.

L'analyse des effets de type 3 à déjà été vue dans le chapitre précédent. Elle est effectuée, pour chaque variable, en comparant le sous-modèle excluant la variable au modèle incluant cette variable et les autres, afin de tester l'hypothèse nulle que cette variable est sans effet dans le modèle pourvu que les autres variables y soient. Voir tableau 4.3.1.

Il faut ensuite lire attentivement le tableau des paramètres du modèle logistique. Il faut notamment repérer les statistiques de Wald inférieures à 3.84 (seuil de significativité à 5%) et les coefficients incohérents avec les taux de résiliations constatés.
Il n'y a pas de problème pour les variables " durée des contrats ", " tranche d'âge " et " sexe ".

Pour la variable " réseaux ", la modalité " ALG E-TO-BEJ " a un coefficient anormalement bas puisque il est inférieure à 0, hors, c'est une modalité qui influe positivement sur la variable cible. D'ailleurs la statistique de Wald de ce coefficient est trop basse. De plus le coefficient 0.004 de la modalité " ORN C-TMN " est incohérent avec le fait que cette modalité présente un risque moindre par rapport à la variable de référence choisie. D'ailleurs sa statistique de Wald est inférieure à 3.84. Nous regrouperons cette modalité avec " ALG C-O " dont le taux de résiliation est quasiment égale. De même pour la première modalité " ALG E-TO-BEJ " qui sera regroupé avec " CST-BDA " avec le même raisonnement.

Pour les types d'activités, la modalité " Atrs " a un coefficient incohérent avec le fait qu'elle est la plus risqué en terme de résiliation. Nous regrouperons cette modalité avec " E-pri ". Pour la variable " montants de primes par type de contrat ", les coefficients des modalités représentant la formule accidentelle ne sont pas cohérant avec les taux de résiliations observés auparavant. Ces deux modalités devraient être regroupées et cela passerait par la création d'une modalité " formule accidentelle ".

Nous effectuons les regroupements cités précédemment, avant de relancer la régression logistique sur les différentes variables.
Voici le modèle que nous obtenons alors :

Source	Value	Standard error	Wald chi-Square	Pr > Chi²
Intercept	-0.925	0.036	663.481	<0.001
Sexe-H	0.000	0.000		
Sexe-F	0.164	0.030	30.089	<0.001
Tranche d'âge [25-44]	0.000	0.000		
Tranche d'âge [45-64]	0.152	0.029	26.557	<0.001
Tranche d'âge [19-24]	-0.324	0.054	36.194	<0.001
Durée des contrats [9,20]	0.000	0.000		
Durée des contrats [5,8]	0.712	0.033	458.674	<0.001
Durée des contrats [21,31]	-2.185	0.070	969.205	<0.0001
Durée des contrats [0,4]	-0.071	0.030	5.453	<0.020
T-C/M-P-F-T [0 ;530[	0.000	0.000		
T-C/M-P-F-T [530 ;1060[	0.153	0.038	15.975	<0.0001
T-C/M-P-F-T [1060 ;1590]	0.262	0.054	23.323	<0.0001
T-C/M-P-F-A [0 ;795[	-0.832	0.033	655.147	<0.0001
Type d'activité-E-pub	0.000	0.000		
Type d'activité-P-lib	0.474	0.040	141.724	<0.0001
Type d'activité-E-pri-Atrs	0.582	0.036	255.886	<0.0001
Type d'activité-Ret	0.073	0.098	0.557	<0.455
Type d'activité-S-P	0.281	0.043	42.885	<0.0001
Réseaux-ANB	0.000	0.000		
Réseaux-ALGE-TO-BEJ-CST-BDA	0.147	0.038	15.199	<0.0001
Réseaux-ALG C-O ORN C-TMN	-0.129	0.038	11.365	0.001
Réseaux-CHL-ORN E	-0.219	0.041	28.024	<0.0001
Réseaux-GRD	1.243	0.090	192.539	<0.0001
Réseaux-STF	-0.875	0.053	277.132	<0.0001

TABLE 4.12 – Les paramètres du modèle logistique.

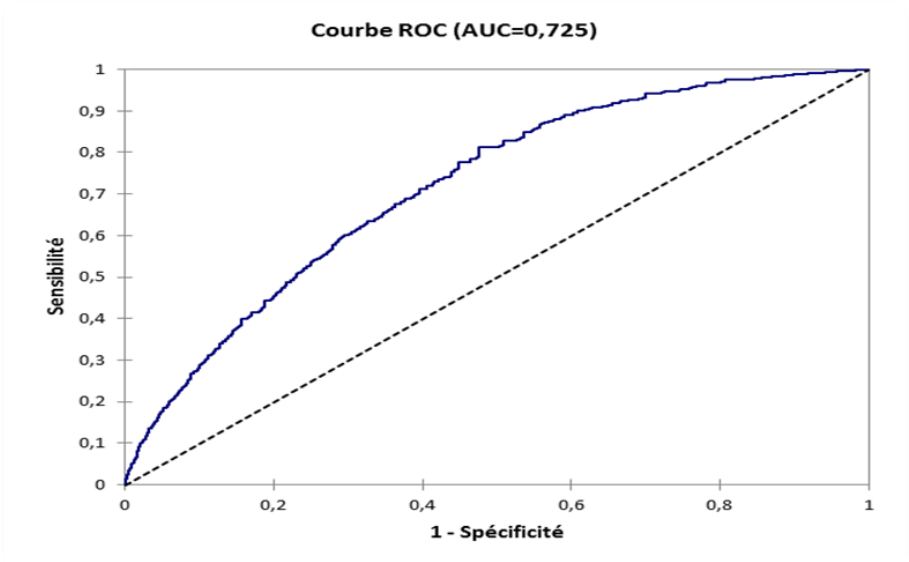

FIGURE 4.10 – Courbe ROC

L'air sous la courbe ROC vaut 0.725 sur l'échantillon d'apprentissage, sélectionné aléatoirement par le logiciel Xlstat.

Seule la modalité " type d'activité-retraite " a une statistique de Wald légèrement inférieure au seuil critique de 3.84. Toutes fois son coefficient est cohérent. Nous avons testé le regroupement de cette modalité avec la modalité " type d'activité-S-P ", mais avons abouti à une baisse de performance (AUC=0.718). Nous conserverons donc le modèle ci-dessus tel quel.

Une fois le modèle logistique retenu, avec ses variables et leur découpage en modalités, il n'est plus nécessaire de disposer d'un échantillon d'apprentissage, et il vaut mieux ajuster les coefficients du tableau précédent en les réévaluant sur l'ensemble de la population. Cela permet d'obtenir les coefficients les moins biaisés possibles. Cela est effectuer automatiquement par le logiciel. Après, nous pouvons transformer le tableau précédent des coefficients en une grille score. Dans le cas d'une régression logistique sur des variables qualitatives ou discrétisées, avec par conséquent un coefficient par modalité de variable, il suffit de :

1. Substituer le logit (combinaison linéaire des indicatrices des modalités) à la probabilité EXP(logit) /(1+EXP(logit)) comme valeur du score ;
2. Puis normaliser le logit en sorte qu'il soit compris entre 0 et 100.

Dans cette normalisation du logit, les coefficients de la régression logistique sont remplacés par de nouveaux coefficients, appelés " nombre de points ", associés chacun à une modalité. Par exemple au lieu d'associer le coefficient 0.164 à la modalité " Sexe-F ", on lui associe 3 points. Le nombre de points associé à chaque modalité est déterminé en sorte que tout individu ait un nombre total de points compris entre 0 et 100, ces deux bornes étant atteignables, au moins en principe. Ce nombre de points est le score de l'individu.

Ce nombre de points est linéairement parfaitement corrélé au logit, mais non au " vrai " score logistique c'est-à-dire la probabilité EXP (logit) / 1+EXP(logit). En revanche, il est

parfaitement corrélé au score logistique en termes de rangs, et son pouvoir discriminant est exactement le même, puisque les rangs, et donc le classement des individus, sont conservés par la fonction croissante EXP(x)/ 1+EXP(x). En conséquence, l'aire sous la courbe ROC de la grille de score est égale à celle du score logistique (aux arrondis prés).[TUF03]

On note c (j, i) le coefficient du modèle associé à la modalité i de la variable j. Pour chaque variable j, on recherche le coefficient c (j, i) le plus petit, noté min(j), le coefficient c (j, k) le plus grand, et l'on calcule Deltamax (j) la différence des deux : le plus grand écart entre deux coefficients d'une même variable. Puis on calcule poids_total la somme sur j de tous les Deltamax.

Enfin, à chaque modalité i de la variable j est associé un nombre de points :

$$N(j,i) = 100 * \frac{c(j,i) - min(j)}{Poids\_total}$$

En résumé, on a :

$$N(j,i) = 100 * \frac{c(j,i) - min(c(j,k))}{\sum_i [max_m c(l,m) - min_m c(l,m)]}$$

[TUF03]

Ce calcul peut être effectué simplement avec un tableur de type Excel. Dans notre cas on obtient la grille suivante :

Source	score
Sexe-H	0
Sexe-F	**3**
Tranche d'âge-[25-44]	4
Tranche d'âge-[45-64]	**7**
Tranche d'âge-[19-24]	0
Durée des contrats-[9,20]	31
Durée des contrats-[5,8]	**43**
Durée des contrats-[21,31]	0
Durée des contrats-[0,4]	30
T-C/M-P-F-T [0 ;530[	17
T-C/M-P-F-T [530 ;1060[	19
T-C/M-P-F-T [1060 ;1590]	**19**
T-C/M-P-F-A [0 ; ;795]	0
Type d'activité-E-pub	0
Type d'activité-P-lib	7
Type d'activité-E-pri-Autrs	**9**
Type d'activité-Ret	0
Type d'activité-S-P	4
Réseaux-ANB	12
Réseaux-ALG E-TO-BEJ-CST-BDA	14
Réseaux-ALG C-O ORN C-TMN	9
Réseaux-CHL-ORN E	8
Réseaux-GRD	**19**
Réseaux-STF	0

TABLE 4.13 – Grille de score

L'intérêt d'une telle grille est bien sur sa lisibilité. Nous y avons surligné pour chaque variable sa modalité de poids maximal. La somme des nombres de points surlignés vaut 100. Des analystes qui ne sont pas statisticiens peuvent très facilement comprendre et commenter une telle grille. Leur appropriation de l'outil de scoring en sera facilitée et ils pourront confronter les nombres de points avec leur intuition métier. Celle-ci peut d'ailleurs leur suggérer que tel découpage n'est pas approprié ou que telle variable a trop de poids, et inciter le statisticien à réexaminer son modèle.

Dans notre étude de cas, les critères personnels (sexe, âge) ont moins de poids. A l'opposé, la durée de contrat est un facteur prédominant le risque de résiliation, ce qui est classique. A titre d'exemple, nous calculons le nombre de points de deux demandeurs. Un homme âgé de plus de 45 ans employé secteur public, qui possède un contrat d'assurance toute cause d'un montant de prime inférieur à 530 Da, d'une durée inférieur à 4 mois, contracté à Ghardaïa, aura une note égale à :

0+7+17+0+30+19=73 points.

Une femme sans profession âgé entre 25 et 44ans, possédant un contrat d'assurance accidentelle, d'une durée dépassant les 9 mois contracté dans l'un de ces réseaux (Alger centre-ouest Oran centre ou Tlemcen) aura une note égale à :

3+4+4+0+31+9=53 points.

Nous verrons plus loin quel est le jugement porté sur leur risque. Nous somme donc en mesure d'attribuer un nombre de points à tous les clients. Mais ce calcul ne constitue pas encore un outil d'aide à la décision, car il ne permet pas directement à l'analyste de se forger une opinion sur les clients qu'ils examinent. Un nombre de points ne lui indique pas s'il vaut plutôt mieux cibler ou pas le client. La dernière étape de la constitution de l'outil de scoring consiste à découper en tranches le nombre de points. On constitue généralement trois tranches de score :

- La moins risquée ;
- la tranche intermédiaire ;
- la plus risquée.

Nous commencerons le découpage du nombre points en décile. Nous croisons ensuite les déciles avec la variable cible.

Le tableau croisé met en évidence que le taux de résiliation ne croît pas linéairement. On voit ainsi que ce taux reste très bas et faiblement croissant dans les 5 premiers déciles, connaît un saut à 26.19% dans le 6eme décile, puis un deuxième à 48.71% dans le 8eme. Le taux de résiliation présente ensuite un fort accroissement dans les 2 derniers déciles.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	Total
R-Oui	3.8%	3.9%	4.0%	7.8%	14.4%	26.2%	36.4%	48.7%	59.1%	85.7%	27.0%
R-Non	96.2%	96.1%	96.0%	92.2%	85.6%	73.8%	63.6%	51.3%	40.9%	14.3%	73.0%

TABLE 4.14 – Tableau de contingence croisant les déciles à la variable « Résiliation ».

Rang pour la variable nombre de points	Nombre d'observation	Minimum	Maximum
0	125	4	13
1	404	14	23
2	1453	24	34
3	4467	34	43
4	**6429**	**44**	**53**
5	8138	54	63
6	**9457**	**64**	**73**
7	4455	74	83
8	988	84	93
9	21	94	98

TABLE 4.15 – Analyse de la variable nombre de points.

Les seuils de tranches de nombres de points sont : 53 et 7, en gras dans le tableau précédent. Nous avons obtenus une tranche de 35.83% de contrats dont le risque est très faible, puisque le taux de résiliation ressort à 6.78%, très loin du taux moyen de 27%. Nous avons ensuite une tranche de 48.96% des contrats dont le taux de résiliation est légèrement supérieur à la moyenne.

Nous avons enfin une tranche très risquée dont prés de deux tiers des contrats sont résiliés. Ils représentent environ 15% des contrats. Avec 73 points notre contractant de tout à l'heure présente encore un risque moyen, mais à la limite du risque fort. L'autre contractante a un risque faible avec 53 points.

Nombre de points	Cible	Cible	
Pourcentage en ligne	**Résiliation-O**	**Résiliation-N**	**Total**
Risque faible	6.78%	93.22%	35.83%
Risque moyen	31.31%	68.69%	48.69%
Risque fort	64.51%	35.49%	15.20%
Total	27.011 %	72.989%	100%

TABLE 4.16 – Nombre de points

Nous disposons donc d'un outil opérationnel, qui pourra être déployé et considérer comme point de vue du pilotage, par la quantification du taux de risque de la production dans les comptes des exercices futurs, et de façon plus générale, l'ajustement fin de la politique de l'entreprise.

4.5 Fiabilité et stabilité du modèle

Lorsqu'un modèle est déployé, en aucun cas il est définitif, il peut changer et faiblir au cours du temps, c'est pour cela qu'une modélisation est faite régulièrement pour conserver toute la robustesse, mais on peut au préalable vérifier après chaque nouveau déploiement l'efficacité et la stabilité grâce aux outils de Data Mining, tels que la matrice d'affectation

qui donne un pourcentage sur la qualité de la prédiction sur une période étudiée, si ce dernier commence à diminuer alors on dit qu'il y a une certaine instabilité cela, est lié au comportement client qui change au fil du temps ainsi on peut songer déjà à établir une autre modélisation. Dans notre cas, d'après la matrice de confusion (d'affectation) on peut constater que notre modèle prédit la résiliation avec un taux d'environ 74% contre un taux de 95.39% pour les non résiliations, d'ici on peut dire qu'on tient un bon modèle.

from/to	**0**	**1**	**Total**	**% correct**
0	25022	1209	26230	95.39%
1	2524	7184	9708	74.00%
Total	27546	8392	35938	84.70%

TABLE 4.17 – Matrice de confusion

4.6 Évaluation des résultats globaux

L'évaluation est l'étape d'interprétation des résultats du modèle obtenu, l'objectif est de voir est ce que grâce à notre étude nous avons amélioré l'existant ou pas.

D'après les résultats de la partie modélisation, on peut dire que nous avons apporté des améliorations significativement positifs et cela se traduit par :

- La visualisation sur les clients à risque (cela permet de mener à bien les actions et compagne marketing) ;
- La réduction du nombre de clients à risque ;
- Augmenter la durée de vie d'un client dans le système.

4.7 Déploiement final

On met en oeuvre le score après l'avoir présenté à ses futurs utilisateurs et leur avoir éventuellement demandé leur validation, et après une période de test " en grandeur nature ". On effectue le suivi de l'utilisation et de la qualité du score. Ce suivi porte :

- Le nombre de clients par valeur ou tranche de score ;
- Le nombre de clients passant d'une valeur à une autre du score entre deux dates de calcul restitué sous la forme d'une matrice de transition ;
- Le taux de souscription et les montants souscrits par valeur ou tranche de score.

On comparera les résultats des différents segments de clientèle, des différents produits, des différents canaux de vente (réseaux et agences) et des différents mois (évolution au cours du temps).

L'objectif est donc de déployé le modèle sur la base afin de vérifier la pertinence des variables définies dans les modèles auprès de clients nouveaux, ainsi que de fournir un outil utile aux commerciaux, pour lesquels il constitue un tableau de bord et une mesure du potentiel de leur clientèle. Le suivi est important car les scores, comme on l'a vu précédemment, ont une durée de vie limitée. Ils vieillissent notamment sous l'influence de l'évolution de l'environnement économique, juridique ou sociodémographique, de l'évolution de l'environnement concurrentiel et de l'offre commerciale.

Ainsi un bon déploiement via le marketing (équipe data mining) permettra le monitoring des scores générés, de mener des actions plus rapidement ainsi que de faire face

aux changements de comportement des clients rapidement. Néanmoins, c'est un travail couteux en terme de temps du fait de l'absence d'infrastructure de déploiement (outils de transformation des données, logiciels...), et l'immobilisation de toute une équipe pour la gestion des bases scorées.

Conclusion

Comme nous avons pu le voir, ce dernier chapitre a porté sur la modélisation des résiliations des contrats d'assurance vie (attrition), on a pu apprécier la résolution d'un problème réel qui préoccupe la majorité des décideurs dans une entreprise (particulièrement en Assurance). Après une bonne étude de la problématique et la mise en oeuvre des moyens pour la résoudre (outils d'aide à la décision : le Data Mining), nous avons estimé un modèle avec une bonne précision et stabilité, ce dernier sera utilisé pour la détection des profils à risques pour les mois a venir et il va contribuer à renforcer les actions et compagnes de fidélisation avec un bon ciblage (choisir les vrais clients et non aléatoirement).

Conclusion Générale

Développer une stratégie CRM est devenu un objectif majeur des entreprises actuelles. Or, la réalité nous a montré que ces projets sont des projets risqués et très couteux à mettre en oeuvre. Une des conditions nécessaires pour réussir l'implémentation d'une stratégie de gestion de la relation client est la disponibilité de données "Clients" fiables, pérennes, précises et répondant aux besoins des décideurs permettant ainsi une gestion efficace. Pour cette condition, les applications CRM doivent être supportées par des processus de Data Mining conçus autour d'objectif CRM qui ne se limitent pas à recueillir que les données comportementales. Cette vision Data Mining intègre un nouvel objectif clair : Maximiser l'efficacité de la gestion de la relation client. Notre travail consistait à voir l'impact du Data Mining sur la gestion de la relation client et comment faire un bon ciblage pour mieux fidéliser.

En premier lieu nous avons fourni un état de l'art sur l'assurance vie, son évolution, ainsi que le contexte de l'étude, nous avons montré aussi le cheminement pour acquérir l'information utile et fine pour une robuste modélisation, ils seront suivi par un chapitre décrivant les méthodes les plus efficaces pour réaliser un bon Data Mining, et enfin le dernier a porté sur la modélisation des résiliations (attrition), dans ce dernier on peut apprécier la résolution d'un problème réel qui préoccupe la majorité des décideurs dans une entreprise (particulièrement en Assurance), après une bonne étude de la problématique et la mise en oeuvre des moyens pour la résoudre (outils d'aide à la décision : le Data Mining), nous avons estimé un modèle avec une bonne précision et stabilité, ce dernier sera utilisé pour la détection des profils à risques pour les mois a venir et il va contribuer à renforcer les actions et compagnes de fidélisation avec un bon ciblage (choisir les vrais clients et non aléatoirement).

Ainsi et avec le déploiement de ce modèle des entreprises peuvent atteindre leurs objectifs en termes de business, il permet aussi aux data mineurs et chefs de projet d'atteindre leurs objectifs en termes de Data Mining et essentiellement d'obtenir la satisfaction de la clientèle.

Perspectives futures : Pour les perspectives futures, pour enrichir la partie Data Mining déjà réalisée, on envisage d'intégrer la partie techniques, nécessaire à l'homogénéisation des données : Le datawarehouse ainsi que les datamarts, et des enquêtes auprès des clients, pour pouvoir prendre en considération d'autres variables qui peuvent influer sur les résiliations.

Bibliographie

[1506] Points économique n°50. 2006.

[16] La loi 06/04 modifiant et complétent l'ordonnance 95/07 relatives aux assurances.

[17] L'ordonnance n° 95/07 relatives aux assurances.

[AGR07] A. AGRESTI. *An introduction to categorical data analysus*. Wiley, 2007.

[AND03] T.W. ANDERSON. *An introduction to multivariate statistical analysis*. Wiley, 2003.

[BAR01] M. BARDOS. *Analyse discriminante, application au risque et scoring financier*. DUNOD, 2001.

[BEN82] J-P. BENZEKRI. *Histoire et préhistoire de l'analyse des données*. DUNOD, 1982.

[BEN00] M. BENLALAM. L'approche de la caar pour le developpement du marché des assurances de personnes en algérie. 2000.

[BER10] I. BERADA. *SPSS clémentine*. SPSS Maghreb, 2010.

[DES05] J. DESJARDINS. *L'analyse de regression logistique*. TECHNIP, 2005.

[dmf] Autorité des marchés financiers. Les differents produits d'assurance vie. *www.l'autorite.qc.ca*.

[dP] Bourse de Paris. Généralité sur l'assurance vie. *www.trader-finance.fr*.

[fdsd] Fédération française des sociétés d'assurance. Histoire de l'assurance vie. *www.assurance.info.com*.

[IAN05] H. IAN. *Data Mining : practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[LAM06] A. LEBART, MORINEAU. A., and PIRON. M. *Statistique exploratoire multidimentionnelle : visualisation et inférences en douille de données*. DUNOD, 2006.

[LAR05] D. LAROSE. *Discovering knowledge in data : an introduction to data mining*. John Wiley & sons, Inc., Hoboken, New Jersey, 2005.

[LAT02] A. LATROUS. Le rôle de l'assurance dans la collécte de l'epargne et les perspéctives de son évolution. journée d'étude sur "le rôle de l'épargne dans l'économie nationale". 2002.

[LIA10] B. LIAUDET. *Cours de data mining.* IAP (ingénierie d'affaires et de projets-finance), 2010.

[NAO00] M. NAOURI. Les assurances de personnes : évolution et spécificité. 2000.

[NC03] J. NAKACHE and J. CONFAIS. *statistique explicative appliquée.* TECHNIP, 2003.

[PAR04] P. PARIZEAU. *Réseaux de neurones.* Université de Lavale, 2004.

[TUF03] S. TUFFERY. *Data Mining et scoring : bases de données et gestion de la relation client.* DUNOD, 2003.

[TUF10] S. TUFFERY. *Data Mining et statistique décisionnelle.* TECHNIP, 2010.

[WIS09] P. WISRA. *Réseaux de neurones artificiels : architecture et application.* Université de Haute Alsace, 2009.