République Algérienne Démocratique et Populaire

Ministères de l'Enseignement Supérieur et de la Recherche Scientifique

المدرسة الوطنية المتعددة التقنيات
Ecole Nationale Polytechnique

École Nationale Polytechnique

Département de Génie Industriel

Mémoire de Magister

Présenté par :

Redouane HALOUANE

Ingénieur d'état en statistique (USTHB)

Thème :

Élaboration d'un modèle de Gestion de la Relation Client basé sur le Data Mining et le Datawarehousing

Application : Prédiction du Churn à WTA

Directeur de mémoire : BELMOKHTAR Oum hani Professeur ENP

Président de juré : BOUBAKEUR Ahmed Professeur ENP

Examinateur : BABA ALI Riadh Maitre de conférence USTHB

2010/2011

ملخص

آلمؤسسات اليوم تواجه عددا من المشاكل الناتجة من المنافسة والأسواق المتغيرة. ذهاب العملاء هو المشكلة الحقيقية حاليا في مختلف المجالات لأن الزبون هو القاعدة الاساسية لتواجد المؤسسة. البحوث و الدراسات 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 وطرق ونماذج تصميم وتسليط الضوء على نقاط التشابه والاختلاف من حيث اختيار الهدف من المتغيرات والتقنيات في مجال التنبؤ بتنقيب البيانات سنحاول ايجاد حل للحد من ذهاب العملاء. بالإضافة إلى ذلك ركزت هذه الدراسة على وجود علاقة حقيقية بين استخراج البيانات وإدارة علاقات العملاء.
الكلمات الرئيسية:
ذهاب العملاء

Résumé

Les organisations sont confrontées de nos jours à plusieurs problèmes qui résultent de la concurrence et de l'évolution des marchés. **Le churn** des clients constitue une vraie problématique dans différents secteurs d'activité car les clients sont l'une des raisons d'être de l'organisation. L'état de l'art qui traite ce sujet est relativement récent et tente d'expliquer les raisons du churn, et essaye de prédire les départs des clients en utilisant des techniques de prévisions de data mining. Ce travail de recherche s'inscrit dans le cadre de la fouille des données et des méthodes de traitement d'informations de l'entreprise. Basé sur des études récentes sur le churn, 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, churn, prévision du churn, turnover, prévision du turnover, gestion de la relation client.

Abstract

Nowadays organizations are faced to several problems resulting from competition and market evolution. Customer churn is a real problem in different industries because customers are one of the reasons for the organization evolution, 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 and methods of information processing company. Based on recent studies on the churn, it proposed to explain the problem through 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, churn, churn prediction, turnover, turnover prediction, customer relationship management.

Table des matières

Liste des figures :

Liste des flux :

Introduction générale et problématique

Au cours de ces dernières années, avec les exigences de la clientèle et la concurrence rude les bases des 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 de données et la manière dont ils sont traitées. ***(BERADA, 2010)***

La satisfaction du parc des abonnés dans le secteur des télécommunications et l'une des priorités de chaque entreprise, 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 donnés(un diamant dans une mine) pour une seule finalité et une seule vision satisfaire le client et essayer d'en acquérir le maximum possible et de minimiser le nombre de départ. ***(GUILLERON, 2000)***

L'objet de ce travail de magister s'insère dans cette préoccupation en apportant un outil d'aide à la décision relatif à la gestion du **churn** (nombre de départ 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 du churn et de prédire le départ des clients en utilisant des techniques prévisionnelles de data mining. Ce travail de recherche 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 aux churn, la relation client et le data mining, nous proposons un outil d'aide à la décision issu du data mining ayant pour but la construction d'une relation durable et fidèle avec sa clientèle et pour en assurer sa pérennité.
Notre travail est structuré comme suit :
Dans le premier chapitre on donnera un état de l'art sur la gestion de la relation client et son évolution depuis sa création, on enchainera par un autre chapitre « la collecte et le traitement des données « le datawarehouse » où on expliquera l'utilité et l'architecture générale de ce dernier après on expliquera le fonctionnement et les bonnes méthodes pour faire un excellent datawarehoussing, le troisième chapitre est le cœur de ce mémoire à savoir le datamining on présentera le concept ainsi que toute la méthodologie. Le quatrième chapitre est dédié aux méthodes de classification et de segmentation cette étape importante à décrire lorsqu'on veut s'initier à la modélisation des phénomènes via le data mining. Dans le dernier chapitre on

abordera la modélisation d'un cas télécom et on montrera l'importance de chaque chapitre cité précédemment.

Chapitre 1 : La gestion de la relation client

L'arrivée d'un nouveau sigle pose toujours problème: s'agit-il de quelque chose de totalement nouveau ou encore d'un habillage marketing d'un problème ancien. Ni tout à fait l'un, ni tout à fait l'autre, répondrons-nous. Non, le CRM n'est pas un nouveau concept car les entreprises ont toujours cherché à satisfaire le client. Non, le CRM est plus qu'une création marketing des cabinets de consultants car il propose une vision totalement nouvelle de la gestion de la relation client. Il faut constater que la technologie permet de traiter dans une approche unifiée et décloisonnée des problématiques qui ont été séparées pendant de nombreuses années: stratégie marketing, gestion de la force de vente, service client, réingénierie des processus, rentabilité des clients, conception assistée des produits par les clients, etc. Le CRM est un terme fédérateur pour définir un objectif commun à des fonctions encore trop souvent cloisonnées. Afin d'établir la vision fédératrice du concept CRM, il nous a semblé important de faire un travail préliminaire d'inventaire des théories et des composants du CRM.

Ce chapitre se propose tout d'abord d'expliciter le contenu de ces trois lettres et la convergence des huit leviers du CRM. Nous présenterons ensuite les avantages escomptés lors de la mise en œuvre de la gestion de la relation client. ***(Lefébure et Venturi, 2005)***

I. Histoire tourmentée de la relation client

L'émergence du concept de gestion de la relation client est le résultat d'une lente évolution de la mentalité des entreprises. Il est toujours difficile de construire une approche simplificatrice des concepts marketing mais un historique rapide montre qu'un nouveau concept apparait tous les dix ans pour modeler les orientations stratégiques.

I.I. D'une orientation produit à une orientation client

- ***L'ère préindustrielle : relation de proximité***

L'ère préindustrielle s'est terminée plus ou moins récemment selon les secteurs. Pour prendre l'exemple du commerce l'apparition des grandes surfaces les concentrations des centrales d'achat et les pressions concurrentielles sur les petits commerces ont débuté il y a quelques dizaines d'années. Auparavant, le commerce à destination du grand public était avant tout bâti sur un modèle de valeurs de proximité de fonds de commerce à taille humaine et de relations personnelles pour ne pas dire de voisinage. ***(Lefébure et Venturi, 2005)***

- ***Les fifties et sixties : reconstruction et push marketing***

Les années 1950 et 1960 furent les années de la production de masse. Il fallait proposer des produits aux consommateurs pour répondre à une demande explosive. La demande était simple l'offre devait l'être également.

Pendant cette période les entreprises se sont essentiellement concentrées sur la création de nouveaux produits et l'élargissement de l'offre. ***(Lefébure et Venturi, 2005)***

- ***Les seventies : segmentation de marchés et mass markets***

Les années 1970 furent les années de la rationalisation. L'optimisation de la production visait à baisser les coûts de fabrication. Il fallait par la combinaison d'une baisse des coûts, d'une amélioration des processus de vente et de la création de nouveaux moyens de toucher la clientèle, élargir la taille de leurs marchés potentiels. Les entreprises ont commencé à segmenter leurs clientèles et ont élargi leurs gammes de produits. ***(Lefébure et Venturi, 2005)***

- ***Les eighties : « consommateur » et one to many :***

Les années 1980 furent les années de la qualité. Les exigences des consommateurs commençaient à se faire sentir. Il fallait pour satisfaire ceux-ci améliorer la qualité des produits. Les entreprises se sont lancées dans la mesure de la qualité des produits et dans le développement des services aux clients.

Pendant plus de trente ans, les entreprises ont perfectionné leurs techniques de production et de gestion pour mieux connaître et maîtriser les produits. Dans la même période elles ont évidemment développé des approches du client mais celles-ci sont restées épisodiques et peu industrielles. ***(Lefébure et Venturi, 2005)***

- ***Les nineties : l'orientation client et le one to some***

Depuis le début des années 1990, le marché connaît une profonde modification avec l'inversion du paradigme marketing : passage d'une orientation produit à une orientation client.

Les années 1990 marquent le début de l'ère du client. Les bases de données client se multiplient. L'essor du marketing direct permet de mettre en avant les avantages de la relation directe. Les canaux d'accès et d'information prolifèrent. ***(Lefébure et Venturi, 2005)***

- ***Début 2000 : l'inversion des relations client-fournisseur et le one to one***

Sans aucun doute, les années 2000 marquent l'intensification de cette tendance client avec l'émergence du concept de marketing one to one une offre spécifique pour chaque client possible essentiellement grâce à l'avènement de l'Internet. Les entreprises quels que soient leurs secteurs d'activité concentrent leurs efforts sur le service et la gestion de la relation client. En parallèle, les nouveaux horizons ouverts par les technologies de communication et de l'information dessinent également une inversion des rôles : le consommateur joue un rôle de plus en plus actif jusqu'à se substituer aux distributeurs à s'auto-conseiller et à assurer lui-même son propre service client. ***(Lefébure et Venturi, 2005)***

- ***L'explosion de la bulle: l'heure des bilans et de la raison***

Après avoir cédé à l'euphorie générale et lancé sans compter des projets parfois pharaoniques, les entreprises marquent une pause dans leurs investissements technologiques et notamment dans le CRM. Cette pause est l'occasion de tirer un premier bilan des retours sur investissements, bilan parfois mitigé avec de réels succès mais aussi de véritables flops, certains allant jusqu'à l'abandon pur et simple du projet. À la lumière de ce bilan, les entreprises reconfigurent leurs attentes en matière de CRM, ce qui a conduit à une évolution dans la nature de la demande et donc des solutions proposées par le marché. Après une période de déraison les projets sont évalués sur leurs perspectives de retour sur investissement à court terme. ***(Lefébure et Venturi, 2005)***

II. Le CRM qu'est-ce que c'est ?

II.1. Définition

Toutes les acceptions s'accordent malgré ces nuances pour dire que le CRM permet de développer et d'améliorer les relations avec les clients. *Le CRM est une démarche qui doit permettre d'identifier, d'attirer et de fidéliser les meilleurs clients, en générant plus de chiffre d'affaires et de bénéfices*. ***(Lefébure et Venturi, 2005)***

Cette définition met en avant le souhait de construire une relation choisie, et non subie, et souligne le souci de rentabilité.

Sous-tendant cette définition, trois dimensions sont implicites dans le CRM :

• une dimension temporelle avec la nécessaire construction d'une relation profitable sur le long terme

• une dimension relationnelle avec le souhait d'être le plus proche possible du client, quel que soient le point de contact et le moment choisi par ce dernier

• une dimension opérationnelle avec le besoin de gérer la complexité de la combinaison clients-offres-canaux avec des outils dédiés. Pour tenir compte de toutes ces dimensions, nous proposons de définir le CRM de la façon suivante :

Le CRM est la capacité à bâtir une relation profitable sur le long terme avec les meilleurs clients en capitalisant sur l'ensemble des points de contacts par une allocation optimale des ressources. ***(Lefébure et Venturi, 2005)***

Le CRM vise à développer une proximité et une relation continue avec les clients. Pour cela, l'entreprise cherche en permanence à mieux comprendre les besoins présents et futurs de chacun d'eux. Grâce à cette connaissance elle peut ensuite ajuster de la manière la plus

économique possible les canaux de distribution de contact, les options sur les produits, les conditions de livraison et la communication de son offre aux besoins.

Le CRM est le moyen d'assurer une cohérence globale entre :

• des clients aux enjeux et aux attentes très différents

• des offres de plus en plus personnalisées

• des canaux de contacts de plus en plus nombreux

II.2. Les huit leviers du CRM

- ***La réingénierie des processus***: les entreprises sont conduites à revoir l'organisation de leurs processus. Elles doivent déterminer comment les simplifier, les recomposer et les optimiser pour faciliter la fabrication et la fourniture de produits et services au client. ***(SMITH, WHEELER, 2002) (Lefébure et Venturi, 2005)***
- ***La réactivité :*** cette nouvelle tendance est mise en avant par Michael Porter. Après le management stratégique des années 1970, le management de la qualité en 1980, le *speed management* s'impose. Le management de la vitesse signifie que les entreprises compressent le temps de conception des produits. Il faut savoir affronter les évolutions de plus en plus rapides des comportements, ainsi que les ruptures technologiques introduites par les concurrents. Par exemple, Dell construit un ordinateur en moins de vingt-quatre heures après avoir reçu une commande. ***(SMITH, WHEELER, 2002) (Lefébure et Venturi, 2005)***
- ***La personnalisation de masse :*** cette tendance est décrite par Joseph Pine dans son ouvrage ***(J. Pine The New Frontier in Business Competition. 1999)***. La personnalisation de masse combine les économies d'échelles par une organisation optimale des processus et la personnalisation du produit et du service au goût du client : la combinaison du sur-mesure et du prix standard. Les logiciels de CRM assemblent et collectent les informations sur les goûts et préférences du client pour permettre aux équipes de production l'organisation des processus.
- ***Le marketing relationnel :*** il s'agit certainement de la révolution la plus importante pour le marketing. Le marketing relationnel nécessite de créer des relations au travers de l'ensemble des canaux de distribution, au niveau des partenaires, des fournisseurs, de l'utilisateur de produits et services. ***(SMITH, WHEELER, 2002) (Lefébure et Venturi, 2005)***

- ***L'amélioration de la satisfaction client :*** un nombre croissant d'entreprises se tournent vers la satisfaction et le service client pour conserver leurs clients et se différencier des concurrents. Le développement des serveurs vocaux, des centres d'appels et des sites Internet informatifs a permis aux clients de contacter directement les entreprises. La réception des réclamations clients offre des possibilités importantes d'améliorer les produits et permet d'apporter une compensation aux clients insatisfaits. De nombreuses études montrent qu'un litige traité rapidement et de manière efficace est un élément important de fidélisation. ***(SMITH, WHEELER, 2002) (Lefébure et Venturi, 2005)***
- ***Le one to one marketing:*** ce concept, développé par Don Peppers et Martha Rogers, **(PRG *Le Marketing one to one. 1999)***, suggère que les entreprises peuvent segmenter leur marché de manière individuelle. Cette approche intellectuellement séduisante a connu des difficultés de mise en œuvre. La rupture qu'implique le passage d'une optique « produit » à une optique « client individuel » est- certainement trop difficile à réaliser. Il semble plus honnête aujourd'hui d'utiliser la notion de *one to few* pour exprimer les enjeux de la différenciation clients. La mise en œuvre de quelques processus différenciés de traitement représente des enjeux quantifiables, mesurables et rentables, ce qui est plus difficile avec le one to one.
- ***La modification du mix marketing :*** les éléments traditionnels du mix marketing connaissent une évolution profonde :

– une augmentation des services périphériques au produit (le pack auto Groupama ou MAAF enveloppe l'assurance automobile dans un ensemble de produits complémentaires comme le crédit automobile, par exemple)

– une segmentation de plus en plus fine de la clientèle avec des notions de potentiel, de cycle de vie, de vitesse de développement, de potentiel d'innovation, etc.

– une stratégie de distribution multi canal permettant d'allier des canaux réactifs comme le SMS ou l'e-mail, des canaux plus conviviaux comme le téléphone ou la force de vente, et des canaux informatifs comme le mailing ou les sites Internet

– une politique de prix basée sur la valeur du client, en complément de la valeur intrinsèque de la transaction. Ils imposeront une flexibilité, tant dans la mise en œuvre que dans le paramétrage des logiciels de CRM. Ces derniers devront être ouverts et modulaires pour s'intégrer et se compléter comme les éléments d'un Lego. Il est évident que l'urbanisation des

applications et des échanges de flux s'impose aux architectes des systèmes d'information. ***(SMITH, WHEELER, 2002) (Lefébure et Venturi, 2005)***

- ***L'intelligence des clients et du personnel :*** un accès de plus en plus large à l'information est la caractéristique du monde actuel. Des clients et des collaborateurs toujours mieux formés et informés sont la contrepartie d'un client qui exige plus de professionnalisme et plus de conseils de ses fournisseurs. Cette tendance signifie que le personnel de vente n'attend plus les directives du management, mais qu'il est prêt à utiliser cette connaissance accumulée, de manière à s'adapter parfaitement au marché. La sophistication des outils et l'amélioration du niveau de formation sont des leviers importants pour l'ajustement au marché. Nous espérons avoir pu mettre en évidence que le CRM ne saurait être réduit à la simple sélection d'une offre de logiciels intégrés. Les enjeux et les impacts potentiels sont beaucoup plus larges. Malgré les déboires de certaines mises en œuvre, il nous semble évident que le CRM est et reste une source importante de différenciation pour une entreprise qui sait en interpréter tous les enjeux et les difficultés. Il ne faut pas occulter les multiples impacts de ce type de projet, sous peine de ne comptabiliser que les dépenses, sans perspectives de retour sur investissement. ***(SMITH, WHEELER, 2002) (Lefébure et Venturi, 2005)***

III. Le positionnement du CRM

Aujourd'hui les entreprises réalisent que l'augmentation de l'efficacité des vendeurs n'est plus suffisante. Il ne s'agit plus seulement d'améliorer leur productivité. Il faut également donner plus de latitude aux personnes au contact du client ou au personnel administratif dans l'optique d'améliorer la qualité du service au client. Il faut à la fois assurer une proximité avec le client tout en assurant un contrôle des coûts de service. L'approche unifiée du client intègre dans un tout cohérent :

• le marketing stratégique, qui doit être plus terre à terre

• le marketing études et le marketing opérationnel, qui doivent savoir assurer une meilleure liaison entre concept et mise en œuvre

• le service après-vente, qui doit concilier respect des normes de productivité et reconnaissance des meilleurs clients

• la gestion de la force de vente qui doit accepter de vivre avec l'évidence qu'elle ne maîtrise plus l'ensemble de la relation client. Le partage est nécessaire et profitable pour tous.

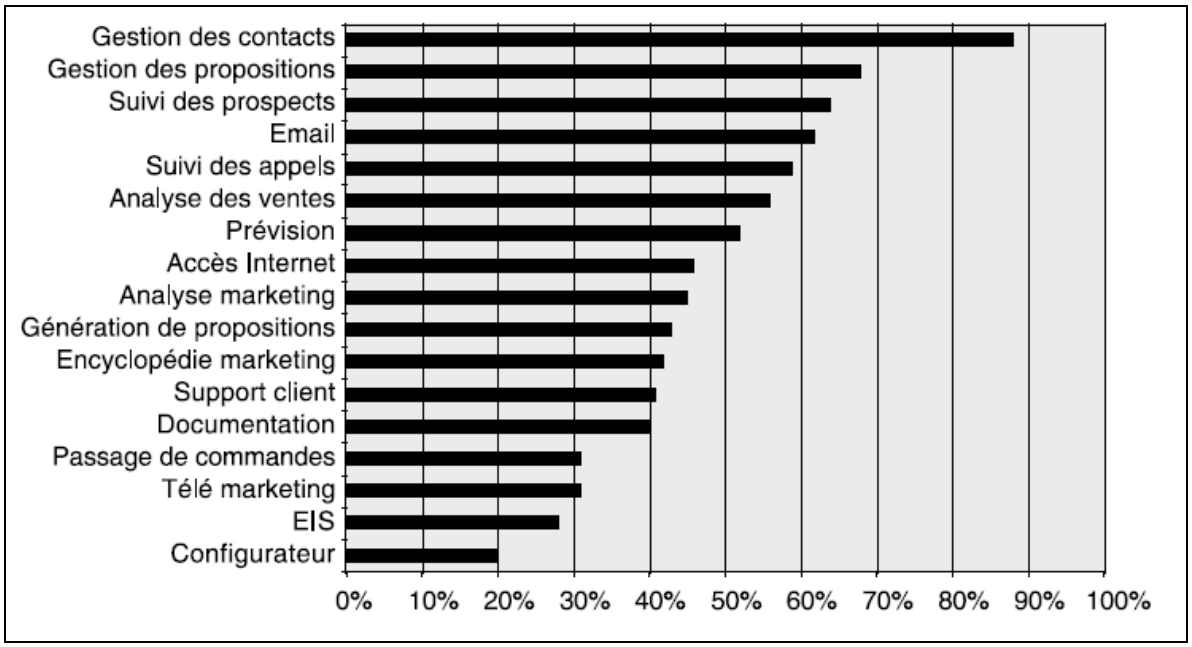


Figure 1 : Les fonctions cibles du CRM ***(Lefébure et Venturi, 2005)***

IV. Les avantages du CRM

- **Pour la force de vente**

Aujourd'hui, les forces de vente qu'elles soient itinérantes ou fixes en face à face ou par téléphone proposent des solutions qui dépassent souvent leurs capacités techniques (complexité des produits et manque de formation). La mise en place d'un outil de CRM leur permet d'accéder à des aides pour les grilles tarifaires, la lecture des stocks et la configuration des produits. Cette assistance leur permet de construire une offre cohérente et de minimiser les risques d'erreurs (factures, conceptions, …).

Accélérer l'intégration des nouveaux vendeurs

Les logiciels de CRM apportent une aide méthodologique pour l'application des méthodes de vente éprouvées. Ils guident les nouveaux commerciaux à travers le cycle de vente (proposition, relance, etc.). Ils réduisent les coûts de formation et d'information et permettent d'identifier le comportement purement opportuniste de certains clients qui recherchent l'avantage immédiat. La capacité d'accéder à une connaissance globale par des interfaces conviviales améliore considérablement l'efficacité du vendeur. ***(Lefébure et Venturi, 2005)***

Accélérer les cycles de vente

La troisième fonction des logiciels de CRM est d'améliorer la productivité et partant d'accélérer la vente. Ils assurent pour cela un support informatique pour les fonctions administratives ou répétitives dans la vente : élaboration des devis et propositions aide à la

configuration des produits, accès on-line aux grilles tarifaires, suivi des coûts de vente. Ces outils diminuent donc les tâches administratives en automatisant certains processus manuels et récurrents pour les commerciaux. Ceux-ci passent ainsi plus de temps à la vente et les éléments de reporting nécessaires au contrôle de gestion sont fiabilisés (clés de répartition des coûts). ***(Lefébure et Venturi, 2005)***

Augmenter les taux de transformation

Le but majeur du CRM est d'augmenter l'efficacité commerciale, c'est-à dire le rapport entre le temps et les moyens investis sur un client et la marge générée par celui-ci. Le principe général consiste à centraliser un maximum d'informations structurées sur le client pour mieux anticiper des événements et trouver le bon moment, le canal optimal et le bon prétexte pour le prochain contact ou la prochaine action. Cette approche générale appliquée au prospect mais aussi au client doit développer les offres complémentaires et donc le chiffre d'affaires unitaire par client. ***(Lefébure et Venturi, 2005)***

- **Pour l'entreprise**

L'affectation des investissements marketing sur des segments plus petits, le *narrow casting* entraîne une diminution des coûts de marketing direct sous réserve d'une industrialisation des coûts de conception. L'efficacité des propositions se traduit par une amélioration du taux de transformation des propositions en vente et de la fidélité du client. Cette capacité de mieux le cibler se traduit immédiatement par une amélioration des rendements de la fonction commerciale. La mise en œuvre d'instruments de mesure et d'évaluation des actions facilite la prolifération de tests et l'optimisation des actions sur la base d'un apprentissage progressif. L'entreprise redécouvre une certaine forme de créativité au moyen des tests et acquiert une courbe d'expériences. Certaines fonctions typiquement « centres de coûts » comme le service client sont partiellement transformées en « centres de profit » grâce aux outils de CRM, un appel au service client peut se transformer en occasion de ventes additionnelles. Ainsi, les opérateurs de téléphonie mobile et les organismes financiers ont tous à des degrés différents insérés des opérations de rattrapage sur les appels entrants des clients identifiés comme fragiles ou à fort potentiel. ***(Lefébure et Venturi, 2005)***

Augmenter le résultat

L'orchestration efficace des différents canaux de recrutement et leur optimisation permanente génèrent plus de prospects et moins de perte de clients. Ces prospects mieux renseignés dès l'amont sont plus rapidement et plus efficacement transformés en clients. Les clients qui présentent certains signes prédictifs d'attrition se voient allouer des efforts spécifiques (offres

spéciales, prise de contact, entretien découverte, etc.) afin d'essayer de modifier leurs comportements. Une meilleure connaissance de la valeur économique des clients permet d'attribuer les ressources financières en priorité aux clients ou prospects ayant le plus fort potentiel. Les politiques de communication ou de promotion peuvent être modifiées pour attirer les meilleurs profils de clients et éviter de développer des tendances opportunistes axées sur les prix ou les remises chez les clients. Les techniques de segmentation offrent la possibilité de construire des offres plus adaptées avec un meilleur mixe des offres et des canaux elles améliorent la part de marché par client et elles diminuent l'attrition. ***(Lefébure et Venturi, 2005)***

Réduire l'attrition

L'attrition aussi dénommée *churn* (pour *change and turn*) dans le secteur des télécommunications, exprime la désaffection des clients. Elle se mesure en taux en prenant sur une cohorte de clients arrivés dans la même période. Le fait de disposer d'informations riches et nombreuses sur les clients peut contribuer à réduire ce taux d'attrition : par une plus grande personnalisation des offres par l'anticipation des tendances au churn grâce à des analyses statistiques par un partage des informations et des clignotants entre tous les canaux et les acteurs en relation avec le client. La détection de l'attrition n'est toutefois que le dernier élément de la chaîne. Un score d'attrition ne fait qu'évaluer les facteurs prédictifs. Il est souvent difficile de rattraper un client qui a décidé de vous quitter. Il est par contre important d'identifier les causes. Le bon sens d'un consultant en restauration illustre bien ce problème : la personne la plus importante dans un restaurant n'est ni le cuisinier, ni le serveur, mais le plongeur…lequel est capable de vous dire ce qui n'est pas consommé ! ***(Lefébure et Venturi, 2005)***

Améliorer la qualité de l'information

Le partage des informations entre un nombre important d'utilisateurs bien encadré par des procédures organisationnelles assure une meilleure intégrité des données. Les incohérences de données ou les informations obsolètes ont plus de chances d'être détectées et corrigées avec un système partagé et unifié. L'objectif même du CRM est le partage de l'information entre les canaux d'interactions : le mailing, le télémarketing, les centres de réception d'appels, la force de vente, les services administratifs, le service après-vente, le Minitel, le serveur vocal interactif ou Internet. Cette homogénéité par les systèmes améliore globalement la perception du client et permet à l'entreprise d'être plus efficace dans sa gestion de la relation lorsqu'elle choisit de favoriser l'interactivité avec le client. ***(Lefébure et Venturi, 2005)***

Augmenter la valeur de l'entreprise

Le CRM a un impact important sur l'augmentation de la valeur à vie des clients, ce que les Anglo-Saxons appellent *Lifetime Value* ou LTV en capitalisant sur les informations acquises lors de chaque interaction. En améliorant les taux de transformation lors de l'acquisition, les ventes croisées et la rétention des clients fidèles, une entreprise accroît de facto sa capitalisation boursière. Le CRM contribue à créer de la valeur sur chaque client de l'entreprise et par conséquent sur l'entreprise elle-même (notion de « capital client »). Ce potentiel de différenciation est bien perçu par les analystes financiers qui considèrent que les entreprises équipées de logiciel CRM ont plus de facilités de communication avec des partenaires et sont donc plus faciles à fusionner. Cette capacité de communication est tant dirigée vers l'amont (architecture en flux tendus avec des systèmes d'EAI) que vers l'aval. ***(Lefébure et Venturi, 2005)***

- **Pour le client**

Améliorer la qualité des contacts

Grâce aux outils de CRM, le client est globalement mieux accueilli, orienté et conseillé lorsqu'il entre en relation avec l'entreprise. À l'accueil, il est reconnu par son nom et les informations sur les relations précédentes peuvent être mises à profit pour orienter et personnaliser le dialogue. En cas d'orientation entre différents départements l'intégration de l'informatique et du téléphone permet de transmettre l'appel au bon interlocuteur en même temps que le dossier informatique suit : le client n'a pas à raconter son histoire encore et encore à chaque nouvel interlocuteur. ***(Lefébure et Venturi, 2005)***

Améliorer la fidélisation

Grâce aux fonctions de conseil et d'aide à la vente qu'offrent les outils de CRM, le client se voit proposer des offres sur-mesure en fonction de son profil ou de son comportement lors de l'entretien. Cette personnalisation si elle est correctement paramétrée par l'entreprise se traduit naturellement par une intensification de la relation avec les clients et un développement du taux de multi vente (ventes de plusieurs produits sur un contact). ***(Lefébure et Venturi, 2005)***

Faire du client un ambassadeur

La confiance développée doit se traduire par des recommandations auprès des prospects. La recommandation reste le stade ultime de la satisfaction : le client se transforme en ambassadeur de l'entreprise. Cette reconnaissance peut se traduire de différentes façons : obligations de passer par un fournisseur en B to B, communication de coordonnées clients ou

parrainage en B to C (Business to Consumer). Ce mode de recrutement par le bouche à oreille ou par des formes plus structurées de parrainage reste de loin le mode d'acquisition le moins coûteux, le plus efficace et le plus fidélisant. ***(Lefébure et Venturi, 2005)***

Avez-vous réellement besoin du CRM ?

Vous faites peut-être partie des d'entreprises chanceuses qui sont satisfaites par leur système d'information :

• pas d'information manquante ou dispersée
• cohérence et fiabilité des données
• valorisation satisfaisante des données
• outils adaptés aux profils des utilisateurs
• peu de litiges ou réclamations client
• connaissance parfaite des attentes des clients
• délais satisfaisants de réactivité face aux incidents
• sémantique commune entre tous les départements de l'entreprise
• accord et partage sur les indicateurs de pilotage de l'entreprise

Dans le cas contraire, vous pensez qu'il est possible d'améliorer votre efficacité en améliorant vos systèmes d'information. Mais, pour autant, le CRM vous concerne-t-il ? Voici quelques questions pour vous aider dans votre autodiagnostic : ***(Lefébure et Venturi, 2005)***

Question	**Oui**	**Non**
Avez-vous perdu des clients suite à une réactivité plus forte de vos concurrents ?		
Avez-vous des difficultés à déterminer vos investissements marketing par cible et par canaux ?		
Avez-vous des difficultés à identifier les clients qui présentent la plus forte valeur ?		
Avez-vous des processus commerciaux qui ne sont pas homogènes selon les canaux ?		
Avez-vous des difficultés à coordonner les actions entre les différents canaux de distribution ?		
Avez-vous une augmentation des réclamations client ?		
Avez-vous le sentiment de ne pas être capable d'anticiper et de réagir à des modifications de comportement de votre client ?		

Avez-vous des problèmes pour traiter les problèmes clients en un temps aux points de contact ?		
Avez-vous des difficultés pour accéder à l'ensemble de l'historique des relations entre votre entreprise et un client ?		
Avez-vous des difficultés pour coordonner les campagnes, la gestion événementielle du client, les séquences d'actions et la communication institutionnelle ?		
Avez-vous la perception que vos investissements commerciaux pourraient être optimisés ?		
Aimeriez-vous avoir une méthode d'évaluation de votre capital client ?		
Pensez-vous pouvoir améliorer l'efficacité de vos campagnes ?		
Pensez-vous qu'il est de plus en plus difficile de coordonner les acteurs de votre entreprise ?		

V. Les composants de l'offre CRM

- **Acheter et intégrer ou développer ?**

La plupart des entreprises ont déjà et souvent depuis longtemps initialisé ne serait-ce que de manière parcellaire une démarche de gestion de la relation client. La tendance actuelle est de s'appuyer davantage sur des progiciels. Ils présentent par opposition aux développements spécifiques purs de multiples avantages :

• l'accumulation d'une multitude d'expériences d'utilisateurs dans des domaines où les processus sont encore peu normalisés comme le déroulement des programmes de marketing spécifiques par client et de la gestion des campagnes

• une cohérence de tous les outils servant à la relation client et donc un partage des informations depuis le ciblage marketing jusqu'au service après-vente

• une meilleure capacité d'intégration avec des environnements existants ou à venir tant en termes de technologies (par exemple, la migration progressive sur XML ou Java, l'intégration des plates-formes EAI (*Enterprise Application Integration)* pour assurer les liens entre le back et le front office) qu'en termes de systèmes fonctionnels comme les interfaces avec les serveurs de commerce électronique ou avec les progiciels générateurs de portails

• des charges de tests et de maintenance notoirement inférieures à celles nécessaires dans le cas de développements internes. ***(Lefébure et Venturi, 2005)***

- **Les pièges à éviter**

La clé du succès commence par la compréhension de la problématique du métier. Il faut faire un état des lieux du système actuel et se définir une cible. Le projet CRM doit réduire le gap entre les besoins (souvent important) et l'existant (souvent préoccupant) en sachant distinguer la part du rêve (ce que l'on voudrait faire) de ce qu'il est souhaitable de faire (ce qu'attendent les clients). Il est naturel de se détourner des difficultés actuelles depuis longtemps dénoncées mais non traitées comme les erreurs de facturation pour imaginer un monde idéal mais non réalisable. L'utopie ne fait pas bon ménage avec la gestion d'un projet CRM ! Il faut faire preuve de beaucoup de réalisme pour contourner les obstacles classiques :

• Complexité des processus : les données doivent être capturées à partir d'une multitude de systèmes opérationnels internes et externes. Il faut les intégrer, les vérifier, les corriger, les modéliser et les restituer dans un mode qui reflète les préoccupations de métier de l'utilisateur.

• Changement de point de vue : les systèmes opérationnels actuels sont construits pour suivre des produits et des services et non pour gérer le client. Pour construire un système de CRM les informations dispersées dans une multitude de systèmes opérationnels dans et au-dehors de l'entreprise, doivent être intégrées dans une vision unifiée du client.

• Désynchronisation : les systèmes opérationnels sont maintenus de façon séparée, il faut harmoniser les termes et les définitions de manière à unifier la sémantique du contenu informationnel de l'entreprise.

• Évolutivité : les systèmes de CRM manipulent un volume important de données et de processus. Les technologies doivent être flexibles pour supporter les multiples composants d'un CRM et permettre les évolutions.

• Émergence d'une nouvelle vision client : les études sur le comportement des clients mettent en évidence une certaine irrationalité dans les pratiques commerciales. Il faut d'abord accepter le constat avant d'accepter le changement. Un travail de préparation et d'éducation est souvent nécessaire en amont pour faire prendre conscience du besoin de changer l'existant.

• Résistance au changement : la mise en évidence des problèmes n'est pas suffisante pour espérer une évolution du comportement de l'entreprise. Une politique de formation est nécessaire pour préparer et accompagner le changement.

• Conduite du changement : les solutions de CRM imposent d'une part un décloisonnement de certains départements voire quelques réorganisations et d'autre part la réingénierie de certains processus. Or, tout changement d'organisation ou de méthode nécessite un minimum de

préparation pour obtenir l'adhésion des futurs utilisateurs. Il est donc vital pour la réussite d'un projet CRM d'anticiper ces risques et de mettre en place une approche spécifique sur la conduite du changement.

• Gestion de projets : sur un domaine traditionnellement peu enclin à préciser ses spécifications, les projets CRM sont généralement très difficiles à maîtriser sur le plan des budgets et des délais. ***(Lefébure et Venturi, 2005)***

- **Les composants CRM**

Une solution de CRM se construit autour des éléments suivants :

• les systèmes et les données de back office : *supply-chain*, ressources humaines, comptabilité, finance…

• des bases de données client qui capturent l'ensemble des informations reliées aux clients, éventuellement unifiées sous la forme d'un datawarehouse

• des canaux de relation qui permettent d'interagir avec les clients ou les fournisseurs

• des accès à des bases de données externes pour enrichir le système d'information

• des outils de gestion des données qui permettent d'assurer les fonctions stratégiques de pilotage et les fonctions tactiques pour réaliser les actions commerciales

• des outils de gestion de la connaissance pour transformer la donnée en information.

(Lefébure et Venturi, 2005)

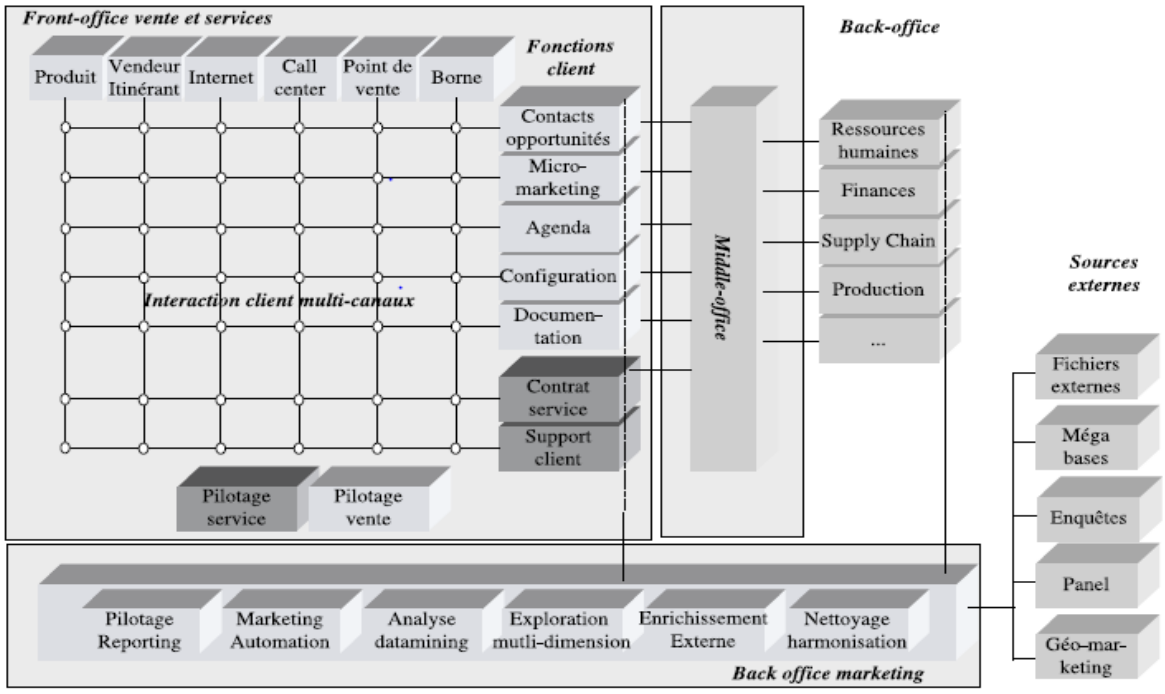


Figure 2 : La cartographie globale du CRM ***(Lefébure et Venturi, 2005)***

VI. Un peu de sentiment

Jusqu'à présent la présentation du CRM a été essentiellement technique : des outils des fonctions élargies et des processus optimisés. Il serait pourtant dangereux de croire que le CRM se limite à la sélection et à la mise en œuvre des bons outils. Il faut aussi positionner le CRM en termes d'avantages pour le client. En ce sens, la définition des gourous du one to one **Peppers et Rogers (PRG *Le Marketing one to one, 1999)*** a le mérite de repositionner le client au cœur du CRM : « Le CRM est une stratégie qui vise à améliorer le taux de rétention des clients en rendant pour un client la fidélité plus avantageuse que l'infidélité. »

Cette valeur ne peut pas être estimée qu'en argent, il faut aussi recréer un sentiment de proximité, d'affinité, de connivence avec le client qui existe dans la relation face à face. Il faut créer des émotions, du vécu, du sentiment pour différencier l'entreprise. Cette maîtrise des coûts de la relation client est également vitale : les entreprises qui ne les maîtriseront pas disparaîtront. Les travaux de Porter sur l'analyse concurrentielle des entreprises ont mis en évidence qu'un leader se doit de produire moins cher mais aussi de construire de la différenciation. Le CRM permet de baisser les coûts mais ses outils sont publics. Toutes les entreprises peuvent les acquérir. La différenciation ne peut se construire durablement sur les prix du service. Il faut donner au CRM une dimension sociale pour insuffler dans l'entreprise une dynamique de culture client. ***(PEPPERS et ROGERS, 1999)***

Fil directeur

La méthodologie IDIC développée par Peppers et Rogers se place comme la plus complète possible :

- Identifier : il faut d'abord identifier les clients avec le niveau de détail le plus fin possible. Il ne faut pas se contenter des noms et des adresses ou d'agrégats sur le CA (chiffre d'affaire) et les visites mais il s'agit de repérer les habitudes, les préférences, etc.

- Différencier : les clients sont différents sur deux axes, leur valeur pour l'entreprise et leurs attentes. Les degrés et les types de différenciation doivent permettre de décider quelle stratégie de one to one est la plus appropriée.

- Interagir : il faut interagir et faciliter cette interaction avec les clients. Il faut parallèlement veiller à l'efficience de cette interaction en orientant les clients vers les canaux les moins coûteux, dans le respect de la valeur de chaque client. La logique économique pousse aux glissements de la force de vente vers le centre d'appels et de celui-ci vers le Web. Pour améliorer la qualité de l'interaction, il faut rassembler et traiter l'information pour identifier la valeur et les besoins clients.

- Customiser : vous devez personnaliser la relation en fonction des besoins et de la valeur du client. Pour inscrire le client dans une relation de fidélité, l'entreprise doit adapter son comportement pour éviter de tomber dans une relation anonyme fondée, soit sur les produits, soit sur des logiques de campagnes massives. Elle doit s'adapter pour produire des biens personnalisés et mieux adaptés à des coûts standards (personnalisation de masse) mais aussi changer des éléments de la chaîne de valeur du client : facture, emballage, manuel de prise en main, etc.

Il y a donc trois visions concomitantes du CRM. L'une technique, met l'accent sur la capture et le partage d'informations dans une optique d'optimisation des processus. La deuxième, plus qualitative voit dans les outils de CRM de simples moyens pour améliorer la qualité globale de la relation client en la rendant moins anonyme. La dernière, plus pragmatique est focalisée sur les résultats tangibles et le retour sur investissement du CRM. ***(PEPPERS et ROGERS, 1999)***

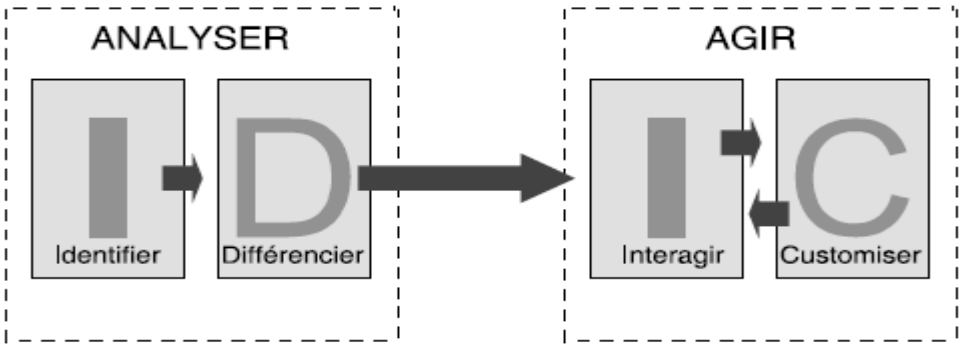


Figure 3 : Le développement d'une stratégie client

Dans cette optique, le CRM doit apporter un support à chaque étape du cycle d'achat : dans la formation d'opinion, lors de la considération d'achat et de l'acquisition, pendant la possession et lors de la reconsidération. Cette proximité sur toute la ligne est le gage de la fidélité.

Ce marketing relationnel délivre de la valeur maximale aux relations client par l'exploitation de la connaissance du client afin de dégager des opportunités de vente de produits. La vente n'est plus une activité limitée aux seuls hommes de marketing ou aux forces de vente traditionnelles. Elle est intégrée dans l'ensemble des processus et des systèmes qui délivrent des services aux clients. La création de la relation n'est cependant justifiée que si elle se traduit par une amélioration de la position concurrentielle et de la profitabilité de l'entreprise… sinon le risque de sur-qualité relationnelle est évident.

(PEPPERS et ROGERS, 1999)

Chapitre 2 : Collecte et traitement des données le DATAWAREHOUSE

On l'a vu pour fidéliser son client il faut le connaître. Pour maîtriser les coûts il faut identifier les postes de dépenses. Cette connaissance passe par la compilation des informations internes et externes disponibles sur le client et sur les niveaux d'utilisation des canaux et des offres de l'entreprise. Plus d'informations, c'est plus de connaissance et donc plus d'efficacité dans la relation. Il est donc essentiel d'accéder dans des conditions correctes à toutes les données disponibles dans l'entreprise et qui se rapportent au client. Or, les données relatives aux clients sont généralement éparpillées dans les bases de données de différents systèmes opérationnels il n'est pas rare de devoir compiler plus de deux cents sources d'information pour obtenir une vision complète de la relation. Donc plutôt que de remettre à plat l'ensemble de ces systèmes d'information pour les centrer sur le client projet titanesque s'il en est, la plupart des entreprises décident de faire du neuf avec du vieux: elles interfacent leurs différents systèmes opérationnels pour en compiler le contenu dans une base de données à part **l'entrepôt de données**. Cette base de données n'ayant pas de vocation opérationnelle particulière elle dispose d'une structure mieux adaptée pour effectuer des traitements lourds et transversaux. Elle dispose également le plus souvent d'un historique bien supérieur à ce que les systèmes opérationnels gèrent. De ce fait, elle est d'une taille souvent très largement au-delà des bases de données opérationnelles taille que seules les récentes innovations technologiques permettent d'absorber correctement.

Ce chapitre présentera dans un premier temps cette évolution de l'infocentre vers les entrepôts de données en mettant en évidence les avantages recherchés au moyen des datawarehouses. Ensuite, nous aborderons la méthodologie de mise en œuvre de ces entrepôts de données et le choix des informations utiles pour l'installation d'un projet CRM. ***(MEIER, 2008) (Lefébure et Venturi, 2005)***

I. Explosion du datawarehouse

Aujourd'hui, l'informatique a la possibilité de créer une véritable synergie entre les différentes approches marketing. En effet, l'augmentation de la puissance de traitement, les nouvelles capacités de communication et la baisse des coûts ouvrent de nouvelles perspectives de stocker et de gérer les informations pour connaître le client. **Moore**, le cofondateur d'Intel, leader mondial des microprocesseurs, prédisait, il y a près de vingt ans, que la puissance informatique augmenterait d'un facteur double, tous les deux ans, Pour illustrer ce point, la puissance de calcul du superordinateur de 1970 qui coûtait près d'une dizaine de millions d'euros est aujourd'hui disponible sur une console de jeu Sony Playstation pour moins de 100 euros. Grâce à cette baisse des coûts, sans équivalent dans d'autres secteurs économiques, la

technologie n'est plus un obstacle. Toutes les entreprises, grandes ou petites, peuvent accéder à des outils très sophistiqués.

L'accès aux données était auparavant limité au niveau des fonctionnalités disponibles et des postes autorisés. L'ergonomie et l'intégration de la bureautique ont permis de développer une maîtrise des interfaces graphiques par des interlocuteurs qui n'avaient connu auparavant que des grands systèmes. Aujourd'hui, le développement de l'informatique dans les entreprises est tel que les ordinateurs sont souvent plus nombreux que les salariés. Les postes de travail ont des possibilités de récupération et de traitement des données très évoluées. Le développement d'Internet et l'augmentation de la puissance des réseaux de communication offrent des capacités d'accès par des canaux de plus en plus nombreux. Le développement des postes nomades a conduit à décentraliser les données et les traitements sur des postes de plus en plus portatifs avec synchronisation régulière par rapport au site central. Le besoin de connectivité a conduit les directions de l'information dans le développement des projets d'EDI(Échange de données informatisées), de solutions de mobilité, et de workflow pour améliorer et accélérer les flux liés aux transactions vers les clients, ou la remontée des informations au plus près de l'intervention. Les premières formes d'accès par Minitel sont progressivement remplacées par des ordinateurs portables reliés par ligne fixe, des assistants personnels utilisant les technologies WiFi ou des Webphones, qui permettent aux itinérants d'être en lien perpétuel avec les informations de l'entreprise. Les possibilités d'être en contact n'auront bientôt plus de limites. ***(NISBET, ELDER, MINER, 2009) (Lefébure et Venturi, 2005)***

II. Entrepôt de données

II.1. Les premiers infocentres

Dans les années 1970, IBM a lancé le concept d'infocentre. Il s'agissait d'extraire des données des systèmes de production et de les rendre accessibles à l'utilisateur final autrement que par des langages de programmation destinés à des spécialistes. Véritable révolution si l'on se rappelle cette époque, l'informatique était encore une technique ésotérique, citadelle totalement hermétique à la compréhension des utilisateurs. L'infocentre comprenait des fichiers destinés à l'utilisateur final et un langage de requête évolué et convivial.

Les systèmes d'infocentre présentaient les caractéristiques suivantes :

• ***Administration***

Elle était la plupart du temps mise entre les mains des utilisateurs afin de respecter à la lettre la logique d'autonomie qui avait guidé la création de ce concept.

• ***Alimentation***

L'infocentre était souvent chargé par des mécanismes d'annule et remplace, par opposition à des mises à jour incrémentales, où seules les modifications sont chargées à chaque vacation.

- ***Contenu***

L'infocentre regroupait en général deux types de données : une photo instantanée d'un sous-ensemble jugé pertinent des données de production et pour justifier l'investissement consenti des agrégats de gestion, c'est-à-dire des données synthétiques pré calculées pour constituer les tableaux de bord des différentes directions.

- ***Structure***

Les premières bases de données relationnelles n'existaient pas encore et la faible puissance de calcul alors disponible ne permettait pas d'exploiter les structures alternatives de l'époque de manière efficace. L'infocentre était la plupart du temps basé sur des fichiers indexés ou des formats spécifiques aux outils utilisés. Pour ce qui est des outils d'interrogation, leur convivialité et leur caractère évolué nous laisseraient rêveurs aujourd'hui, à l'heure du tout Windows, de l'intranet, et de Business Objects, Brio ou Impromptu. Quoi qu'il en soit, pour l'époque, ils apportaient effectivement une amélioration indéniable par rapport à l'outil principal, pour ne pas dire unique, dont ne disposait toute personne désireuse d'accéder à une donnée. L'offre était relativement pléthorique et la plupart des fournisseurs proposaient un langage d'interrogation en mode commande comparable aujourd'hui à du SQL panaché avec du Basic. Un doux mélange qui conduisait souvent l'utilisateur final à devenir d'abord un spécialiste de ce langage puis souvent l'expert en programmation de requêtes pour les autres utilisateurs n'ayant pas acquis une maîtrise suffisante du langage.

En d'autres termes, l'infocentre qui aurait dû libérer l'utilisateur de sa dépendance vis-à-vis des professionnels de l'informatique a en fait simplement déplacé le problème. Il a créé une nouvelle caste de professionnels de l'infocentre pas tout à fait informaticiens et plus totalement utilisateurs non plus.

Plus d'un quart de siècle s'est écoulé depuis l'apparition du concept d'infocentre bien sûr les lacunes du passé ont été progressivement comblées. Les fournisseurs d'infocentre, pour conserver leur parc de clients ont cherché à faciliter l'utilisation de leurs outils en intégrant tant bien que mal de nouvelles technologies telles que le client-serveur, le tout Windows, le stockage en base de données relationnelle, Internet…

Aujourd'hui encore de nombreuses entreprises s'appuient totalement sur un infocentre pour leur pilotage, ce qui prouve que quoi qu'on en dise il apporte un premier niveau de solution pour désengorger le service informatique de demandes de requêtes ponctuelles et pour

apporter un peu plus d'autonomie aux utilisateurs. ***(Lefébure et Venturi, 2005) (LEBART, 2006)***

II.2. Industrialiser l'infocentre : les entrepôts de données

La décennie 1990 s'est caractérisée par l'émergence du concept d'entrepôt de données le datawarehouse pour les anglophones. De quoi s'agit-il ? Le pape du datawarehouse, Bill Inmon a proposé une définition qui quinze ans après fait toujours référence : « L'entrepôt de données est une collection de données orientées sujet intégrées, non volatiles et historiques, organisées pour le support du processus d'aide à la décision. » ***(Bill. I, Using the Datawarehouse)***

Les systèmes de production ont été développés au fil du temps. Ils sont donc nécessairement stratifiés et peu cohérents entre eux or, une refonte globale qui permettrait d'atteindre cette cohérence est généralement infaisable sur le plan économique. Il faut donc atteindre cette nécessaire cohérence en laissant les systèmes de production évoluer à leur rythme. L'entrepôt de données apporte une solution à cette problématique en proposant de mettre en place une base de données (l'entrepôt) dans laquelle sont déversées après nettoyage et homogénéisation des informations en provenance des différents systèmes opérationnels. Il s'agit donc de construire une vue d'ensemble cohérente des données de l'entreprise pour pallier la stratification et l'hétérogénéité historique des systèmes de production, sans pour autant remettre à plat ces derniers.

Le datawarehouse se positionne ainsi comme la nouvelle solution à un problème vieux comme l'informatique : comment sortir des informations d'un système optimisé pour l'entrée de données ? ***(Lefébure et Venturi, 2005) (LEBART, 2006)***

II.3. Faciliter l'utilisation de l'entrepôt de données : les datamarts

Les entrepôts de données contiennent des données pléthoriques tant en termes de profondeur (l'historique) qu'en termes de largeur (la richesse des informations stockées). Par construction leur structure est orientée stockage et non-traitement, puisque les utilisations qui en seront faites ne sont pas totalement prédéterminées lors de leur conception. Structuré de manière à pouvoir stocker des éléments de détails historisés, l'entrepôt de données est souvent complexe dans sa structure et difficilement exploitable par des utilisateurs finaux. De plus, le volume et les traitements de chargement sont souvent techniquement incompatibles avec des requêtes utilisateurs non déterministes, par exemple, des requêtes libres dont la charge de traitement ne peut pas être précisée a priori. Pour pallier ces limites fonctionnelles et techniques, il n'est pas rare que des datamarts soient créés en aval de l'entrepôt, voire parfois en parallèle à

l'entrepôt. Ces datamarts contiennent un sous-ensemble de données pertinentes pour une activité particulière :

• Sous-ensemble des instances : par exemple, seuls les clients actifs sont repris dans le datamart, voire un échantillon aléatoire représentatif pour faciliter les comptages.

• Sous-ensemble des attributs : seule les données élémentaires et les agrégats pertinents pour l'activité concernée sont transférés ou calculés au chargement du datamart depuis le datawarehouse. Un datawarehouse ou entrepôt de données est une collection de données structurées consolidant les informations en provenance des différents systèmes opérationnels et dédiée à l'aide à la décision. Par expérience, les datamarts suivants sont fréquemment présents dans les entreprises :

• ***Gestion de campagnes***: le datamart « gestion de campagnes » a pour objectif de simplifier l'utilisation des données à des fins de marketing opérationnel : comptage, ciblage et mesure des remontées d'opérations marketing adressées. Il contient la plupart du temps une vision simplifiée du client, des données de personnalisation, comme le nom, l'adresse, le nom du commercial affecté… et des agrégats permettant d'effectuer les ciblages les plus courants, par exemple, le chiffre d'affaires annuel, le fait que le client détient ou non tel produit ou service, la récence, la fréquence d'achat, l'utilisation du SAV…

• ***Études et data mining*****:** les entreprises matures dans le domaine du data mining constatent la plupart du temps une certaine récurrence dans les types d'études qu'elles mènent et donc des données généralement nécessaires pour ces études. La vocation du « datamart étude » est d'offrir un accès direct à des données pré calculées fréquemment utilisées pour des études répétitives.

• ***Pilotage*****:** la plupart des entreprises constituent non pas un mais plusieurs « datamarts pilotage » par exemple, un datamart concernera le pilotage des forces ou des réseaux de ventes, un autre celui des achats ou du contrôle de gestion. Tous ces datamarts pilotage auront comme caractéristique commune de pré calculer des indicateurs, tels que le chiffre d'affaires et le nombre de contacts et de les associer à des dimensions comme des périodes temporelles, des régions, des segments de clients ou des magasins. Les datamarts pilotage sont généralement structurés en étoiles, c'est-à-dire qu'une entité regroupe les indicateurs et que les différentes dimensions sont autant d'entités reliées aux indicateurs. Cette modélisation en étoile permet de proposer une navigation multidimensionnelle dans les données.

D'autres datamarts spécifiques peuvent exister pour couvrir des besoins métier spécifiques, comme le risque de crédit dans le domaine bancaire. Quelle qu'en soit la finalité, ces datamarts ont comme caractéristique commune de stocker des valeurs agrégées et pré

calculées afin d'en simplifier l'utilisation quotidienne. ***(Lefébure et Venturi, 2005) (LEBART, 2006)***

III. Entrepôt de données et CRM

L'entrepôt de données est souvent associé à une architecture de CRM. Cela étant et avant de présenter plus en détail son rôle potentiel dans le CRM, il faut souligner deux points essentiels:

• L'entrepôt de données et ses dérivés, les datamarts, ont par définition une couverture plus large que le CRM.

• L'entrepôt de données n'est pas systématiquement nécessaire pour mettre en œuvre une architecture de CRM.

L'entrepôt de données couvre d'autres domaines métier que le CRM

Le CRM est centré client. L'entrepôt de données se veut « orienté sujet », c'est-à-dire qu'il est structuré pour faciliter la manipulation d'entités métier telles que les clients ou les commandes. En ce sens, il est généralement le pivot d'une vision globale du client et l'une des entités métier essentielles de l'entrepôt de données est effectivement le client. Pourtant, dans son rôle de hangar à données, l'entrepôt manipule d'autres entités métier fondamentales pour l'entreprise et qui ne relèvent pas directement du CRM (voir figure 1) : la représentation des structures de l'entreprise, d'une part et la représentation des produits et de leur vie, d'autre part.

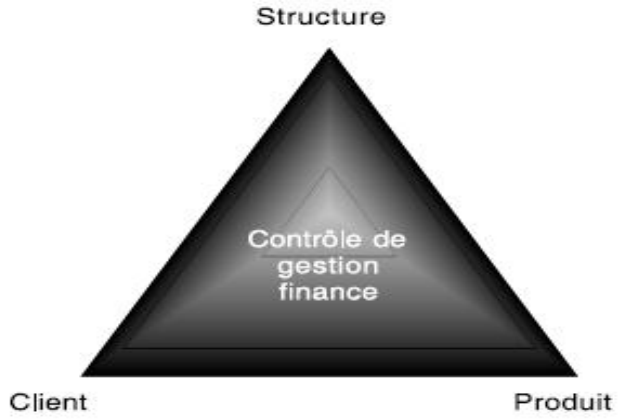


Figure 1 : Le datawarehouse couvre un champ plus large que le CRM ***(Mattison, 2007)***

Nous verrons plus en détail l'axe client, l'axe produit comprend une profondeur historique et des attributs produit utiles pour la logistique ou les stocks, mais sans intérêt du point de vue du client : encombrement du produit, fournisseurs, historique des commandes fournisseurs, historique des prix d'achat… De même, l'axe structure intègre une description détaillée de

l'organisation et des moyens ainsi que leur évolution dans le temps : par exemple, les horaires d'ouverture/fermeture des caisses, les remplacements de personnes, les niveaux de délégation dans la banque…

Pour illustrer ce propos, voici trois exemples d'utilisation de l'entrepôt de données :

• ***Axe client***

Quels sont les clients ayant acheté un micro-ordinateur entre janvier et février de l'année dernière, ayant ensuite acheté des cartouches d'encre couleur et n'ayant pas dans les douze mois précédant l'achat de leur micro-ordinateur acheté une imprimante ?

• ***Axe structure***

Quelle est l'évolution du chiffre d'affaires quotidien rapporté au nombre d'heures de présence des caissières depuis un mois ?

• ***Axe produit***

Quel est le délai de livraison moyen des salades fraîches par mes différents fournisseurs depuis un mois sur mes magasins de la région parisienne ?

Des données, telles que les horaires historiques des caissières ou le fournisseur de la salade acquise par un client donné la semaine précédente, n'ont en définitive que peu d'importance pour améliorer la relation avec le client. En ce sens, l'entrepôt de données comprend généralement un sur ensemble des données effectivement utiles pour le CRM. ***(SCHARFF,2004) (BERRY, 2004) (Lefébure et Venturi, 2005)***

IV. Une solution de CRM sans entrepôt de données

Nous avons insisté précédemment sur la nécessité de constituer une vision globale des informations disponibles sur chaque client nous en avons déduit l'importance d'un entrepôt de données en tant que colonne vertébrale d'une architecture de CRM. Cet entrepôt de données n'est finalement qu'un palliatif à la stratification des systèmes d'information opérationnels.

Dans les cas particuliers d'entreprises pour lesquelles la relation client est le cœur du métier, il est possible de considérer les bases de données des outils de CRM comme une alternative à la mise en place d'un entrepôt de données. En effet, dans cette situation, la très grande majorité des informations sur le client est en fait détenue par les outils de CRM. Ces outils proposent généralement une structuration des données orientée sur le client. Ainsi, la base de données des outils de CRM fait office d'entrepôt de données client parmi les entreprises qui ont ou pourraient prendre ce raccourci, citons notamment les sociétés d'assistance, les Vépécistes ou les sociétés de commerce électronique. Dans ce cas particulier, on aboutit donc

à une structure de base de données opérationnelle pour le CRM qui peut être directement utilisée pour le décisionnel. Pourtant, si la structure peut être reprise, il n'est néanmoins pas rare de voir les entreprises opter pour une duplication technique des bases opérationnelles. En effet, les traitements marketing sont transversaux et non déterministes. De ce fait, ils cohabitent habituellement très difficilement avec des applications transactionnelles optimisées. ***(SCHARFF, 2004) (BERRY, 2004) (Lefébure et Venturi, 2005)***

V. Qu'attendre d'un entrepôt de données ?

On peut approcher l'architecture d'un système de CRM par la métaphore de la tête et des jambes :

• D'un côté les jambes, des outils opérationnels pour gérer l'interaction en temps réel avec le client, l'automatisation des forces de vente et le service client, le tout dans une déclinaison par canal.

• De l'autre la tête, des outils analytiques pour connaître et orchestrer en back office les mouvements assurés par les jambes. L'entrepôt de données est le support de ces activités analytiques, qu'on peut regrouper sous le terme générique de marketing de base de données : « Le marketing de bases de données est défini comme le fait de gérer un système informatisé de bases de données relationnelles qui collecte des données pertinentes sur nos clients et nos prospects pour délivrer de meilleurs services et établir une relation à long terme avec eux. Une utilisation efficace de la base de données a pour effet de fidéliser, de réduire les pertes de clientèle et d'accroître la satisfaction des clients et les ventes. La base de données est utilisée pour cibler les offres sur les clients ou prospects, pour envoyer le bon message au bon moment et à la bonne personne – augmentant le taux de réponse par dollar attribué au marketing, diminuant les coûts par commande, fondant notre entreprise et augmentant nos profits. » (Extrait du site du National Center for Database Marketing).

Plus précisément, les principaux besoins des entreprises qui se lancent dans un projet d'entrepôt de données à vocation marketing sont traditionnellement :

• Piloter les opérations de marketing : ciblage, mailing, promotions, etc.

• Analyser les ventes.

• Définir la typologie client pour affiner l'offre.

• Suivre les évolutions des segments de clients.

• Étudier l'impact des promotions.

• Étudier la satisfaction des clients/produit/réseau de distribution.

• Étudier la concurrence.

• Suivre la performance des commerciaux (budget/réalisé).
• Établir un report des activités des filiales.
• Piloter les événements.
• Mesurer l'efficacité des canaux de distribution (ratio ventes/contacts).
• Évaluer les principaux postes de coûts liés à la vente.

Sur le plan fonctionnel, ces attentes peuvent se résumer en trois grands ensembles d'activités :

• le marketing opérationnel
• l'analyse
• le pilotage

La base de données s'inscrit donc dans le CRM comme un centre de profit par sa dimension opérationnelle.

VI. L'architecture générale d'un entrepôt de données

L'entrepôt de données, on l'a vu est plus large que ce que justifient les besoins du CRM. Nous allons maintenant nous focaliser sur les natures de données, les flux et les traitements de la partie marketing d'un entrepôt de données.

VI.1. Les fonctions

L'entrepôt de données doit, soit directement, soit par le biais de bases de données dérivées, supporter quatre grands ensembles de fonctions pour les utilisateurs du marketing :

• Le pilotage des ventes, des forces commerciales et des actions marketing.
• Le contrôle de gestion avec le suivi des différents postes de coûts de commercialisation, de service après-vente permettant d'évaluer la rentabilité globale des produits, des clients ou des activités de l'entreprise.
• L'analyse statistique des facteurs explicatifs de tel ou tel comportement ou la recherche de segmentation pertinente de clientèles.
• Le marketing opérationnel avec la gestion de campagnes depuis le ciblage jusqu'au suivi des remontées.

Cependant, au-delà de ces fonctions intrinsèques, l'entrepôt de données doit supporter également des fonctions plus techniques pour répondre complètement à la problématique du CRM :

• L'alimentation, qui englobe toute la gestion des flux depuis et vers les systèmes opérationnels de CRM.

• La production, qui recouvre des calculs plus ou moins sophistiqués de nouvelles informations à partir des données.

• La gestion transactionnelle, qui consiste à servir directement en données les canaux d'interactions. ***(Lefébure et Venturi, 2005)*** (***GOUARNE, 1999) (GRISLIN, 2006)***

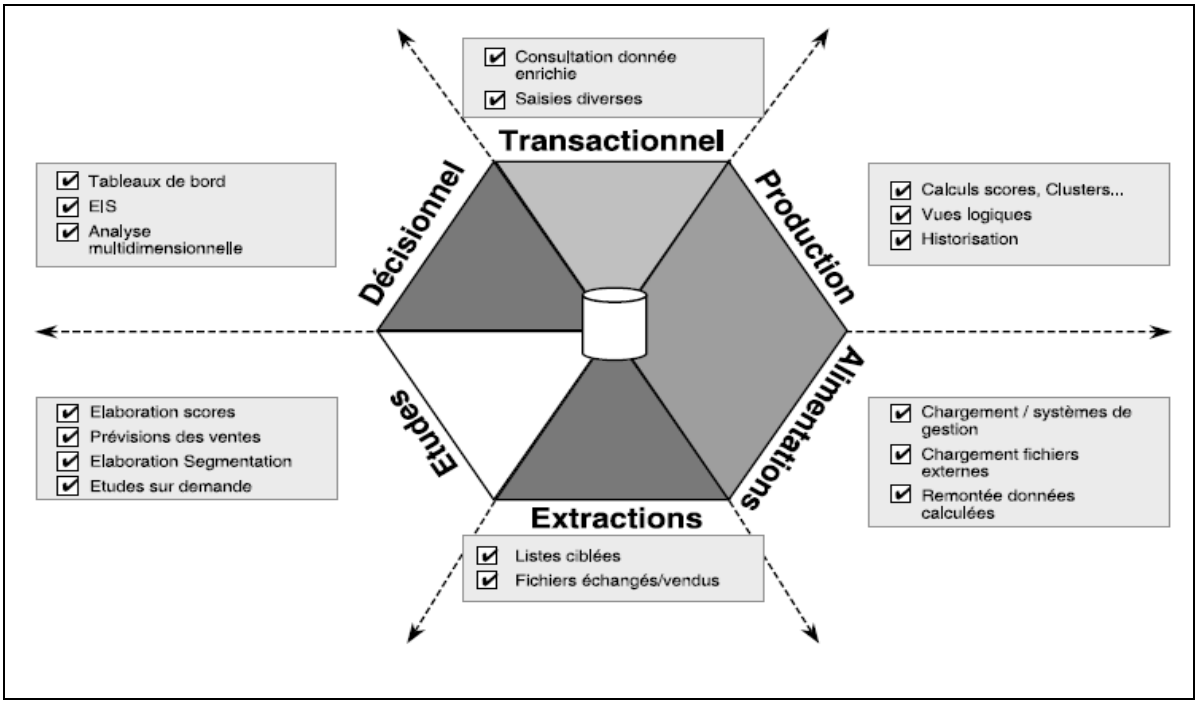


Figure 2 : Les fonctions du datawarehouse vis-à-vis du CRM ***(Lefébure et Venturi, 2005)***

VI.2. Les alimentations

Il s'agit des processus et des programmes qui vont, d'une part, transformer les informations des systèmes opérationnels avant de les intégrer dans l'entrepôt de données et d'autre part, assurer l'envoi des informations pertinentes vers les différents systèmes opérationnels que recouvre le CRM.

Pour le chargement des données, la complexité est très largement fonction du nombre de sources à intégrer dans l'entrepôt de données. Le nettoyage consiste à éliminer les données aberrantes et à harmoniser des codifications qui peuvent varier d'un système opérationnel à l'autre, par exemple, les codes décrivant la fin d'un appel téléphonique peuvent avoir des significations différentes selon qu'ils proviennent du système de gestion du service consommateur ou du plateau de gestion des appels dans le cadre du programme de fidélité.

L'appareillage et la hiérarchisation des sources sont des tâches relativement spécifiques à la gestion des clients. Il s'agit de définir et d'appliquer des règles de gestion permettant d'identifier que deux personnes décrites dans deux systèmes différents sont en fait une seule

et même personne et lorsque ces deux systèmes disposent d'une même information avec des valeurs différentes, d'arbitrer entre ces valeurs.

Dans le sens inverse, l'entrepôt de données est la seule source exhaustive concernant le client. C'est donc le seul endroit où il est possible de calculer certains indicateurs agrégés : qu'il s'agisse de simples agrégats, comme le nombre total de contacts entrants et sortants pour chaque client ou de calculs plus sophistiqués comme la probabilité d'attrition de chaque client dans le prochain mois. Ces informations nouvelles sont bien entendu utiles pour la gestion de campagnes (ciblage des clients ayant une probabilité élevée d'attrition), le pilotage (répartition des ventes par tranche de nombre de contacts), le contrôle de gestion (valorisation des coûts de service au client) ou les études (corrélation des contacts passés et des ventes futures). Elles ont également une valeur inestimable dans le cadre des outils de gestion de l'interaction avec le client et doivent donc être poussées par des interfaces vers les bases de données des différents canaux pour personnaliser la relation. Par exemple, un opérateur de télécommunication calcule systématiquement un score d'attrition pour tous les clients sur son entrepôt de données et interface celui-ci avec son logiciel de service client pour adapter les argumentaires commerciaux. Ainsi, lorsqu'un client appelle pour obtenir un renseignement ou tout autre service, l'opérateur dispose d'une probabilité d'attrition à l'écran et se voit proposer un script de rétention qu'il peut dérouler pendant que le client est en ligne. Ces fonctions d'échange peuvent sembler purement techniques. Elles contiennent en réalité une forte dose d'intelligence métier. ***(Lefébure et Venturi, 2005)*** (***GOUARNE, 1999) (GRISLIN, 2006)***

VI.3. La production

L'entrepôt de données se veut statique, du moins est-ce ce que nous en disent les gourous du domaine. Force est de constater que l'entrepôt de données vu du marketing a une dimension de plus en plus orientée production, en ce sens qu'il est le centre où sont produites de nouvelles informations nécessaires pour l'exécution de nombreux processus.

Exemple de score Figure 3 :

Ces formules doivent être programmées sur l'entrepôt, qui est la plupart du temps la seule source disposant de l'ensemble des données nécessaires pour ce calcul. Il devient ainsi possible de calculer cette probabilité de manière systématique et sur l'ensemble des clients. De plus, stockée dans l'entrepôt, elle devient une clé possible pour effectuer des tableaux de bord, des ciblages, des tests de performance ou d'autres études qui viennent ainsi s'ajouter à des agrégats plus simples mais tout aussi, voire plus prolifiques : des indicateurs de moyennes, de sommes ou de comptages directement calculés au niveau du client, dans

l'optique d'éviter de les calculer à la volée pour améliorer les temps de réponse et simplifier la vie de l'utilisateur.

Caractéristiques du compte	
Titulaire du compte	**1.454**
Non titulaire de compte	**0.000**
Le compte a été ouvert avant la création de la carte	**0.336**
Le compte a été ouvert après la création de la carte	**0.000**
Code option paiement de la carte	
Libre	**0.320**
Total	**0.198**
Minimum	**0.000**
Plafond inférieur à 10000 francs	**0.467**
Plafond compris entre 10000 et 15000 francs	**0.000**
Plafond supérieur à 15000 francs	**-0.212**
Caractéristiques du porteur le jour de la création de la carte	
Age supérieur à 60 ans	**-0.435**
Autres âges	**0.000**
Distance nulle	**0.260**
Distance non nulle et inférieure à 10 kilomètres	**0.000**
Distance supérieure à 10 kilomètres	**-0.295**
Monsieur	**-0.619**
Madame	**0.197**
Mademoiselle	**0.000**

Figure 3 : Illustre un exemple de score ***(Lefébure et Venturi, 2005)***

Le marketing opérationnel s'entend traditionnellement comme un moyen de gérer des campagnes que nous qualifierons de *batch*. L'informatisation aidant, il est aujourd'hui possible de passer à une gestion événementielle du client : à partir de la définition de règles comme l'envoi d'un *welcome pack* trois semaines après l'abonnement sauf si le client a appelé entretemps le service réclamation. Les outils d'automatisation du marketing appliquent systématiquement ces règles pour extraire dans le temps les cibles correspondant à cette description. ***(Lefébure et Venturi, 2005)*** (***GOUARNE, 1999) (GRISLIN, 2006)***

VI.4. Les transactions

Les cas où l'entrepôt de données sert également de support au transactionnel sont extrêmement rares. Néanmoins, il arrive parfois que certaines entreprises décident de faire fonctionner des applications transactionnelles de CRM comme les outils de services client ou de mesure de risque directement sur l'entrepôt de données. Cette approche présente de nombreux avantages : fonctionnement en temps réel sur des données à jour, interfaces moins nombreuses et plus simples, économies de stockage liées à la non-redondance des données… L'inconvénient majeur de cette architecture est de mêler sur une même base de données des traitements décisionnels transversaux lourds avec des applications transactionnelles, exigeantes en termes de qualité de service et de temps de réponse.

Les technologies évoluent elles sont de plus en plus à même de proposer des solutions de cohabitation : architectures massivement parallèles, *clustering* et redondance pure des systèmes disques… En parallèle, l'explosion d'Internet rend le CRM de plus en plus orienté sur le temps réel. Les besoins de réactivité et les progrès technologiques convergent donc progressivement pour apporter des informations de plus en plus actuelles et fiables aux points de communication entre les clients et l'entreprise. ***(Lefébure et Venturi, 2005)*** (***GOUARNE, 1999) (GRISLIN, 2006)***

VII. Quelques principes pour la collecte des informations

Les sources internes qui parlent du client sont multiples, comme l'illustre l'exemple de la figure 4. Ces informations internes doivent être captées via les processus d'alimentation en respectant autant que faire se peut quelques règles élémentaires.

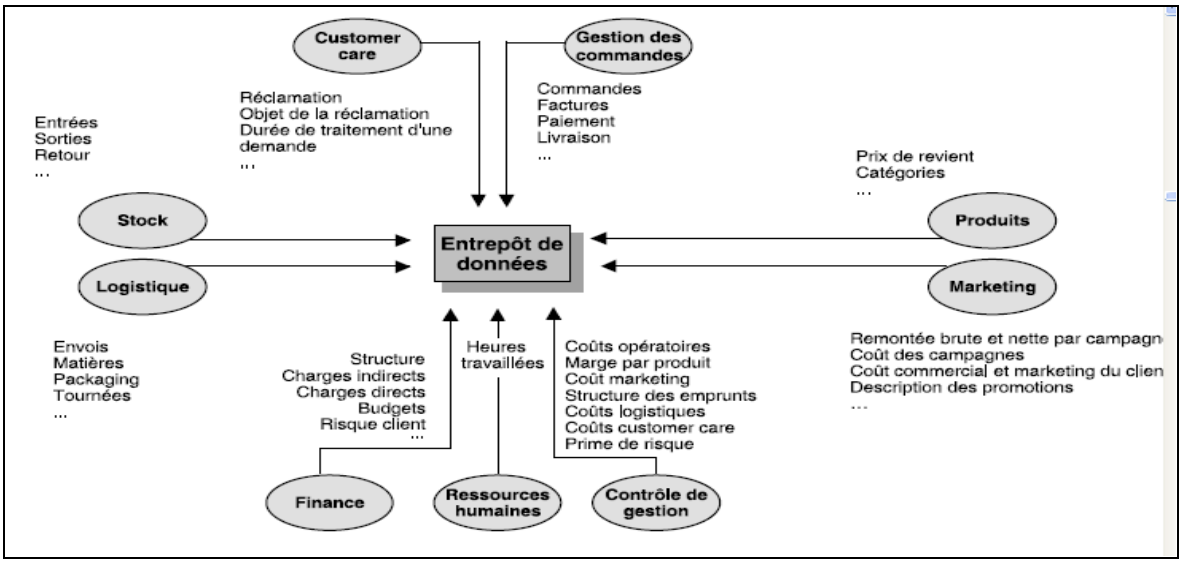


Figure 4 : les sources internes possibles ***(Lefébure et Venturi, 2005)***

- ***Historique des consommations* :** il est important de conserver l'historique des produits et services achetés. Il est en effet possible d'en déduire les centres d'intérêt des clients et surtout d'éviter de leur proposer des offres qu'ils auraient pu résilier ou décliner. De même, la conservation du mode d'entrée en relation permet souvent d'interpréter les motivations du client lors de l'entrée en relation.
- ***Collecte auprès de tiers* :** une prestation peut englober des composantes gérées par des tiers, par exemple, une assistance dépannage dans le cadre d'une carte de crédit. Les événements intervenant chez ces tiers ont une importance qui justifie de récupérer et

de stocker ces informations client externes. La gestion de la relation client doit être mise en œuvre dans une logique globale.

- ***Intérêt marketing* :** les systèmes de gestion n'ont besoin de la trace d'une commande qu'à partir du moment où elle est effective ; en marketing, il est primordial de conserver la trace des demandes d'informations ou des devis, ainsi que l'ensemble des sollicitations qui ont dû être mises en œuvre pour déclencher l'acte d'achat. Ils représentent la marque d'un centre d'intérêt du client qui doit être traité de manière prioritaire.
- ***Caractéristiques des produits achetés* :** il n'est pas rare que les systèmes de gestion soient faits de telle sorte que les caractéristiques volatiles des produits disparaissent des bases de données une fois qu'elles sont devenues obsolètes. Ainsi, le fait qu'un article ait été en promotion pendant quinze jours disparaîtra des bases de données après ces quinze jours. Il est important d'associer à l'achat de ce produit pendant la période, un indicateur permettant de repérer qu'il y a effectivement eu promotion. En effet, le comportement d'achat d'articles en promotion peut se révéler important pour l'animation marketing d'un client ou d'un groupe de clients.
- ***Historique de tous les contacts* :** l'historique des actions présente l'ensemble des moyens utilisés pour délivrer un message à un client. Il est utile de suivre les coûts associés à l'utilisation de ces vecteurs de communication. Les entreprises qui ont une large palette de moyens de communication pourront choisir d'agir sur des canaux moins coûteux pour limiter les dépenses. La conservation d'un historique complet des contacts est primordiale pour déterminer la réceptivité du client aux différents messages, et repérer les méthodes marketing qui donnent les meilleurs résultats.

Cet historique doit englober les différents types de contacts : appels entrants des prospects et des clients, courriers et réclamations, remontées de coupons ou autres remontées marketing direct (annonces presse, numéro vert ou indigo, fax et bus mailing). L'historisation de la navigation sur les serveurs vocaux ou les sites Internet (du moins les éléments les plus signifiants) doit être mise en place car elle constitue un moyen de comprendre les attentes des clients. ***(Lefébure et Venturi, 2005)*** (***GOUARNE, 1999) (GRISLIN, 2006) (MEYLAN, 2003)***

VIII. La construction d'un datawarehouse

VIII.1. La procédure idéale

La majorité des entrepôts de données aujourd'hui en exploitation ont accouché dans la douleur : retards importants de livraison, couverture fonctionnelle inférieure aux attentes d'origine, explosion des budgets initiaux, insatisfaction des utilisateurs, mauvaise qualité des informations…

Le coût d'implémentation d'un entrepôt de données, véritablement transversal et fédérateur de l'ensemble des systèmes d'une entreprise, dépasse la plupart du temps tout réalisme budgétaire. Il est donc souvent nécessaire de construire un plan d'urbanisation d'ensemble, une cible qui donne un fil directeur pour assurer une possibilité de convergence à terme, tout en développant l'entrepôt de données progressivement et par appartement. Il s'agit ici de respecter une maxime de bon sens : penser grand, mais commencer petit. La démarche standard se présente sous forme d'un processus répétitif chaque itération enrichit la solution tout en respectant plus ou moins six étapes principales (voir figure 5).

La première étape commence par la création d'un nouveau mode de compréhension des données avec l'objectif de mieux servir le client.

Transformer un système qui sait bien faire des transactions en un système orienté vers le client nécessite un travail de coordination important entre les différents départements impactés. Il est souvent indispensable de commencer le projet par des réunions de sensibilisation des acteurs pour clarifier les enjeux et préciser la cible. Un audit de l'existant permet d'établir une situation claire de la position actuelle de l'entreprise, afin d'apprécier ce qui doit changer dans les processus existants et les impacts sur l'organisation. L'objectif est de répertorier les besoins d'alignement fonctionnels et techniques par rapport aux objectifs stratégiques fixés.

Pour cela, il faut cerner les besoins d'évolution de l'existant, identifier les projets d'évolution en cours et les domaines d'évolution prioritaires en fonction des objectifs métiers. Il faut que les acteurs du projet soient conscients des opportunités et des menaces qui pèsent sur chacun des métiers.

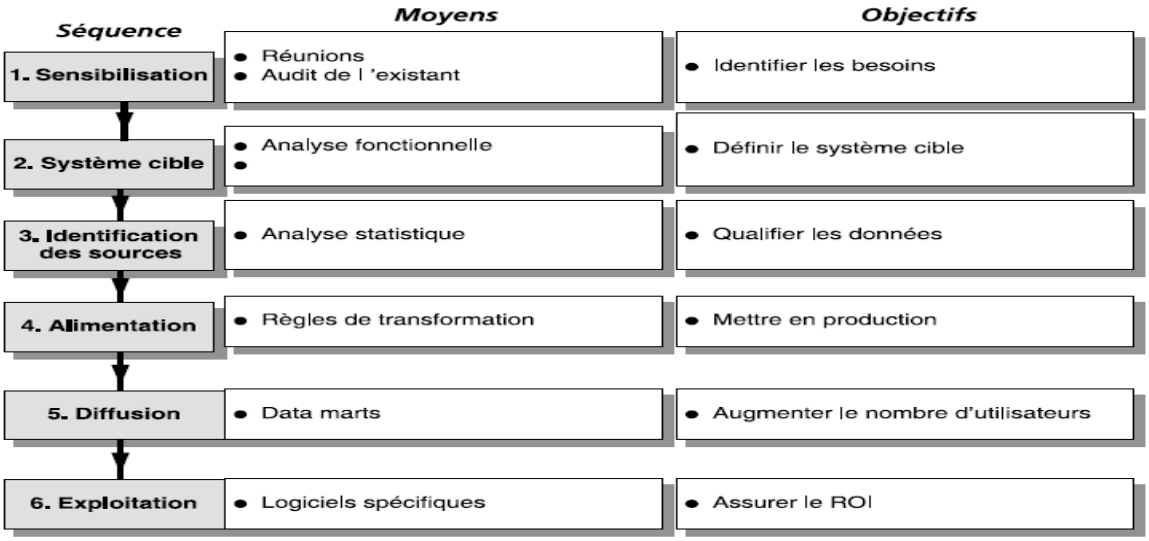


Figure 5 : étapes du projet ***(Lefébure et Venturi, 2005)***

La deuxième étape s'attache à identifier les pistes d'amélioration et à préciser les contours du système cible. Elle s'appuie sur l'audit de l'existant pour évaluer les forces et les faiblesses. Le principe consiste à partir des problématiques métier actuellement insolubles pour en déduire les fonctions et les données qui permettraient à l'entrepôt de données d'apporter une réponse. Cette phase d'analyse fonctionnelle s'appuie sur des entretiens avec les utilisateurs métier. Elle permet au groupe de projet de construire une vision et un consensus sur la cible. Elle doit aboutir à la mise en évidence du plan d'action et à la définition d'un lotissement dans le temps. Cette tâche de lotissement est particulièrement importante. Les lots doivent être cohérents et définis selon des critères de couverture : géographique, de sources de données considérées, fonctionnelle, organisationnelle, etc.

C'est à ce stade que les enveloppes budgétaires doivent être estimées et engagées. Il est essentiel de faire preuve de réalisme économique lors de cette étape en identifiant des indicateurs de mesure de succès de chaque projet identifié : un projet sans indicateur ne peut être piloté et donc géré. Au cœur de cette étape, il est nécessaire de construire un dictionnaire commun des informations qui contienne les termes qui définissent le client, les produits, les contacts ou les canaux. Cette deuxième étape aboutit à la construction d'un modèle de données **(MCD : Modèle conceptuel de données)**, cartographie théorique de l'organisation des données.

La troisième étape s'attache à identifier les sources et la disponibilité des données pour remplir le système cible. Il faut remonter aux sources des systèmes opérationnels. Il s'agit d'une tâche importante et laborieuse qui consiste à déterminer les processus de nettoyage nécessaires ainsi que les règles de transformation des données source vers la structure cible de

l'entrepôt de données. À ce stade, il faut choisir la donnée de référence parmi de multiples origines possibles : un client ne doit avoir qu'une seule adresse postale et il faudra choisir laquelle. Chaque système présente ses propres règles ou définitions pour décrire un client, une offre, un produit ou un marché. Il est également souvent nécessaire d'analyser et de valider les données : identification des valeurs aberrantes, des données manquantes, etc. L'objectif de cette troisième étape est de définir précisément comment transformer des données source éparpillées et hétéroclites, pour les restituer dans le cadre d'une structure de données orientée métier structurée et historique. La plupart du temps, il est nécessaire pour les évolutions ultérieures, de disposer d'un dictionnaire centralisé qui précise la signification des données, leurs origines, les changements effectués et les règles de constitution à partir des données source. Ces informations sont communément appelées les métadonnées. Ce dictionnaire est très utile pour l'utilisateur, qui sait ainsi quelles sont les données disponibles et pour l'administrateur, qui bénéficie d'une vision centralisée et unifiée de l'entreprise ou du métier. Celui-ci peut par la suite compléter, modifier cette représentation en veillant à préserver la cohérence du système.

La quatrième phase est l'alimentation du datawarehouse. Il s'agit de développer ou de paramétrer les règles de transformation qui constituent le processus de chargement de l'entrepôt de données. Cette phase est de plus en plus souvent effectuée avec des outils d'ETL (Extract Transform Loading) du type Data Stage ou Synopsis. Ils offrent une richesse d'interface qui facilite la création du code et la maintenance des procédures des chargements. Il est cependant souvent indispensable d'adjoindre des programmes en SQL pour traiter certaines données ou optimiser certains traitements. Certains outils permettent de créer de manière automatique le méta-dictionnaire. Pour la phase de chargement, il faut souvent avoir une approche modeste au démarrage, avec un chargement initial simple pour tenir compte des problèmes de volume et des temps de mise à jour des informations. La productivité de l'utilisateur final n'est pas seulement régie par la richesse fonctionnelle des données qui lui sont proposées, elle est également conditionnée par le confort d'utilisation et donc par les temps de réponse et de service. Le bon calibrage des investissements en termes de puissance machine, notamment est à ce stade un facteur clé pour que la solution soit acceptée par les utilisateurs.

La cinquième phase est celle de l'ouverture des accès aux utilisateurs. Jusqu'ici, les investissements ont été pour l'essentiel consacrés à stocker l'information sous une forme restructurée. Cette information se déprécie vite si elle n'est pas partagée. Pour gagner de la valeur elle doit être distribuée et transformée. Il faut déployer des outils conviviaux et rapides

chez les utilisateurs. Cette étape passe généralement par la création de datamarts, qui sont des visions des données limitées au métier de l'utilisateur (voir figure 6). Le responsable du ciblage des clients n'aura pas les mêmes besoins ni donc les mêmes données accessibles que le responsable de la plate-forme téléphonique. Ces datamarts sont en fait des informations pré calculées qui ont pour objectif de faciliter les travaux d'agrégation, de totalisation et de jointure pour l'utilisateur. Ils ont souvent pour objectif de garantir des temps de traitement ou de préparer des données sémantiquement correctes pour l'utilisateur. Cette phase d'ouverture doit être accompagnée d'une politique d'information sur les données et de formation aux outils. Une des causes d'échec dans l'utilisation des entrepôts se trouve dans le manque de formation des équipes techniques. La modification de l'univers des données et des langages représentée par le passage d'un infocentre à un entrepôt de données, se traduit pour les utilisateurs par la crainte de ne pas maîtriser le nouveau système et par une perte de productivité due à la maîtrise du nouveau langage. Dans ce contexte, il est difficile de changer et utopique de croire que la migration se fera car l'entrepôt est plus riche. L'entrepôt dans sa première version est un problème pour les utilisateurs ! La sixième étape consiste à exploiter les données, point qui est largement détaillé dans les deux chapitres suivants. ***(Lefébure et Venturi, 2005)*** (***GOUARNE, 1999) (GRISLIN, 2006) (MEYLAN, 2003)***

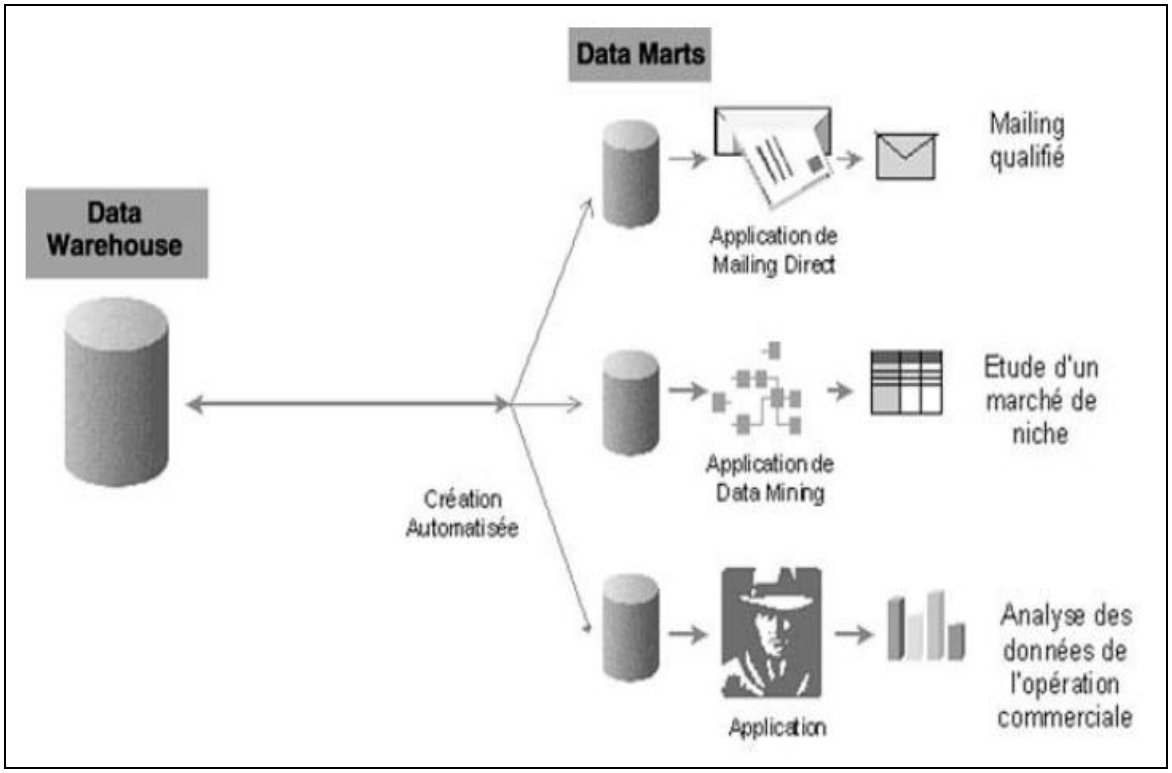


Figure 6 : les dartamarts ***(Lefébure et Venturi, 2005)***

IX. Modélisation de données

IX.1. La modélisation par sujet

Un entrepôt de données est généralement basé sur un SGBD (Système de Gestion des Bases de Données) relationnel.

La modélisation par sujet est une technique de conception logique qui vise à organiser et classifier les informations des bases légataires en données classées par sujet fonctionnel. Elle est basée sur la modélisation " Entité/Relation " et est préliminaire à la modélisation dimensionnelle. Chaque sujet correspond à une table gérée au sein de l'entrepôt. Il faut isoler les données stratégiques, déterminer les informations de détails nécessaires (profondeur, granularité) et conserver les métadonnées. ***(VANGENOT, 2005)***

XI.2. La modélisation dimensionnelle

La modélisation dimensionnelle (modèle multidimensionnel) souvent appelée modélisation OLAP (Online Analytical Processing) se présente comme une alternative au modèle relationnel. Elle correspond mieux aux besoins du décideur tout en intégrant la modélisation par sujet. C'est une méthode de conception logique qui vise à présenter les données sous une forme standardisée intuitive et qui permet des accès hautement performants. Elle aboutit à présenter les données non plus sous forme de tables ou cube (Une construction multidimensionnelle formée de la conjonction de plusieurs dimensions. Chaque cellule est définie par une seule valeur de chaque dimension) mais centré sur une activité. Un cube de dimension n (n > 3) est aussi dit **hypercube**. ***(TASLIMANKA, 2007) (RAPHALEN, 2002)***

Faits, indicateurs et dimensions

La table de faits est la clef de voûte du modèle dimensionnel où sont stockés les indicateurs de performance. Le concepteur s'efforce de considérer comme indicateurs les informations d'un processus d'entreprise dans un système d'information. Les indicateurs étant les données les plus volumineuses d'un système d'information, on ne peut se permettre de les dupliquer dans d'autres tables mais de les rationaliser au sein de la table de faits.

Table de faits des dépenses journalières
Clé date
Clé revenue
Clé Heure
Montant dépensé

Figure 7 : Modèle conceptuel d'une table de faits

Le terme de fait est utilisé pour représenter une mesure économique. Pour exemple, lors de la vente de produits sur un marché, on comptabilise les types de produits vendus, leur quantité et le montant de chaque vente au jour le jour et ce pour chaque produit et pour chaque magasin.
La mesure des quantités et des prix est réalisée à l'intersection de toutes les dimensions (produit, magasin, temps). Le nombre des dimensions détermine la finesse, la granularité de la table et indique la portée de l'indicateur.

Additivités des indicateurs

Les indicateurs les plus utiles d'une table de faits sont numériques et additifs. L'additivité des attributs d'une table de faits est cruciale pour les outils décisionnels. Les utilisateurs demandent rarement l'analyse d'une seule ligne. Dans notre exemple, constater les ventes de produits sur une année pour les magasins d'une région demande l'analyse de plusieurs milliers de lignes à la fois. Pour autant, tous les attributs utiles ne sont pas additifs. Certains sont semi additifs et ne peuvent être additionnés que pour certaines dimensions.
D'autres sont non additifs et ne peuvent pas être additionnés par dimensions. Pour cette dernière catégorie, on utilise des fonctions d'agrégations tel que, le calcul de moyenne, le ratio ou le comptage de lignes. ***(RAPHALEN, 2002)***

Les dimensions

Les tables de dimensions sont les entités complémentaires à la conception de la table de faits. Elles contiennent, autant que possible, des attributs sous forme de descriptions textuelles permettant de qualifier ou d'expliquer l'activité.
Des attributs de dimensions, nombreux, permettent de varier les possibilités d'analyse (par tranches ou en dés). Ces attributs rendent utilisables et intelligible les données de l'entrepôt de données. Ils établissent, en quelque sorte une interface homme/entrepôt de données.
En général, les tables de dimensions tendent à être peu profondes mais elles sont larges (l'inverse de la table de faits), en d'autres termes elles ont peu de lignes mais beaucoup de colonnes.

Tables de dimension "Produit"
Clé produit
Description du produit
Numéro (clé naturelle)
Tariff Plan
Tariff Plan Type
Rated duration

Figure 8 : Modèle conceptuel d'une table de dimension

XI.3. Structure de la base de données

Au sein de l'entrepôt de données les données sont redondantes et dé normalisées, nous sommes loin de la modélisation en troisième forme et pour cause, cela permet de faciliter l'utilisation et d'améliorer les performances lors de l'analyse des données.

Trois types de schémas sont fréquemment rencontrés, le schéma en étoile, le schéma en flocon de neige et le schéma en constellation de faits. ***(KUSIAK, AGARD, 2005)***

XI.3.1. Le schéma en étoile

Dans un schéma en étoile, une table centrale de faits contenant les faits à analyser, référence les tables de dimensions par des clefs étrangères. Chaque dimension est décrite par une seule table (feuille de l'arbre de tables) dont les attributs représentent les diverses granularités possibles.

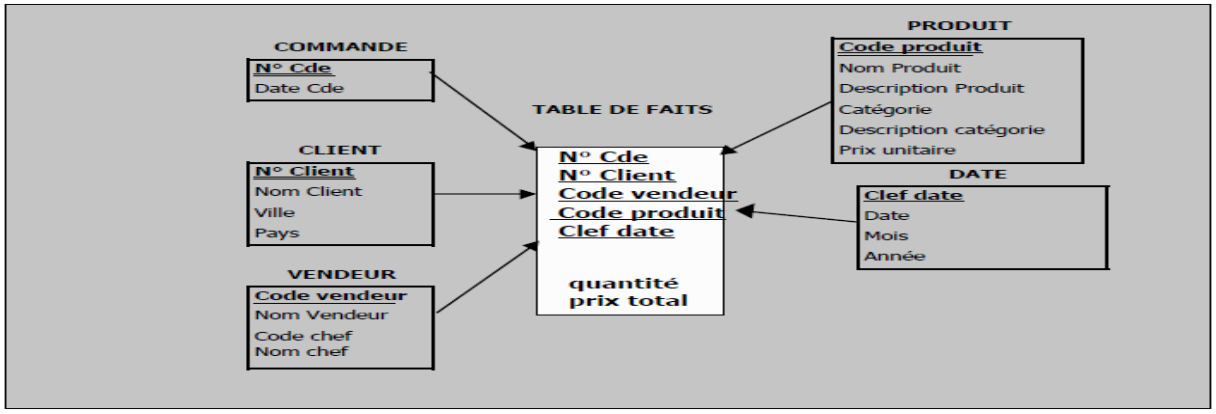


Figure 9 : Modèle en étoile ***(VANGENOT, 2005)***

XI.3.2. Le schéma en flocon

Dans un schéma en flocon, cette même table de faits, référence les tables de dimensions de premier niveau, au même titre que le schéma en étoile. La différence réside dans le fait que les dimensions sont décrites par une succession de tables (à l'aide de clefs étrangères) représentant la granularité de l'information. Ce schéma évite les redondances d'information mais nécessite des jointures lors des agrégats de ces dimensions. ***(VANGENOT, 2005)***

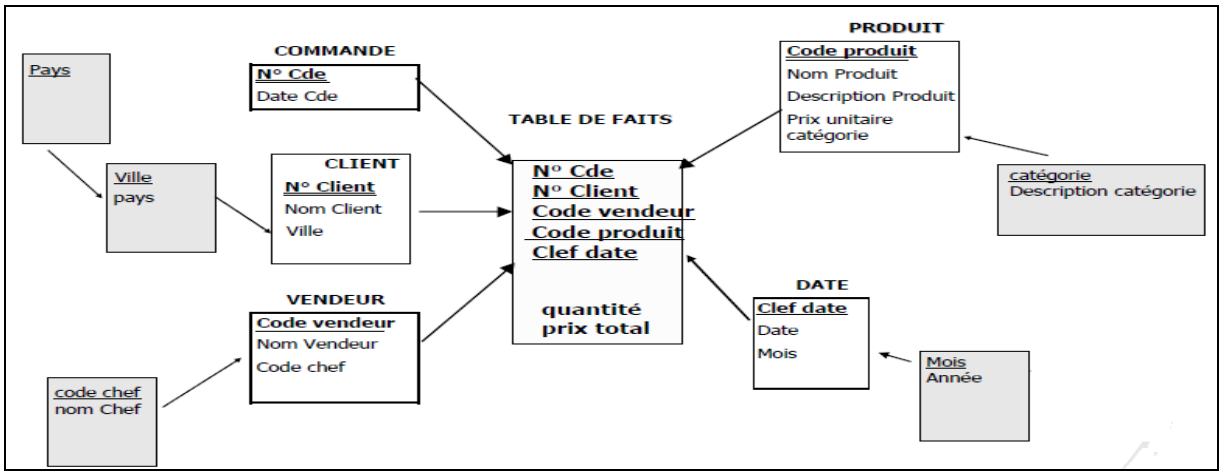


Figure 10 : Modèle en flocon ***(VANGENOT, 2005)***

XI.3.3. Les schémas en constellation de faits

Dans un schéma en constellation, plusieurs modèles dimensionnels se partagent les mêmes dimensions, c'est-à-dire, les tables de faits ont des tables de dimensions en commun.

Pour conclure, les différences entre ces trois modèles sont faibles et ne peuvent donner lieu à des comparaisons de performance. Ce sont des schémas issus de la modélisation dimensionnelle utilisés par les outils décisionnels. ***(VANGENOT, 2005)***

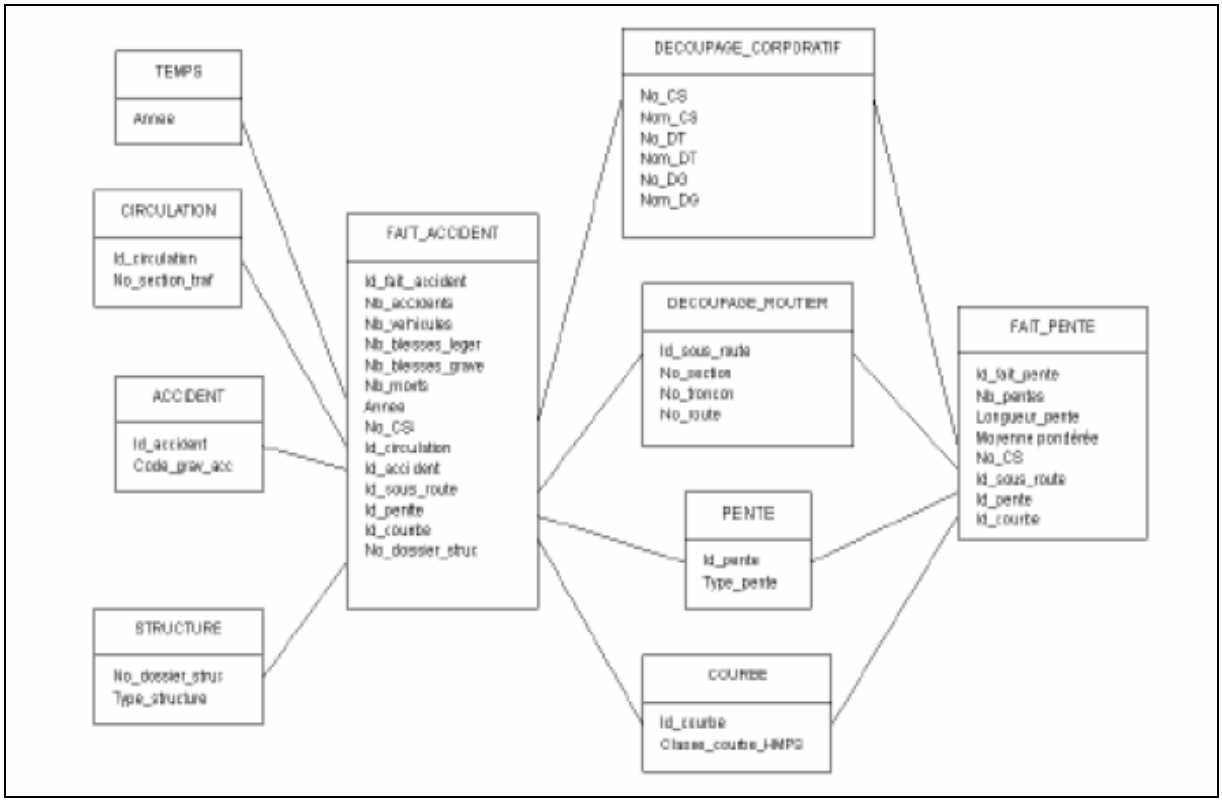


Figure 11 : Schéma en constellation de faits ***(VANGENOT, 2005)***

L'entreposage de données est la première technologie utilisée et la plus fiable aujourd'hui par les entreprises pour la planification, de prévision et de gestion pour exemple la planification des ressources, les prévisions financières et de contrôle, etc… Après l'évolution de la notion d'entreposage de données pendant le début des années 90, on pensait que cette technologie va croître à un rythme très rapide mais malheureusement ce n'est pas la réalité. On a beaucoup peine dans ce domaine concernant la conception et le développement, il reste encore beaucoup à faire mais c'est un domaine qui mérite une attention particulière de la communauté de recherche.

Chapitre 3 : Le Data Mining

L'exploration de données connue aussi sous l'expression de fouille de données, *data mining* (« forage de données »), ou encore extraction de connaissances à partir de données, « ECD », a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. L'utilisation industrielle ou opérationnelle de ce savoir dans le monde professionnel permet de résoudre des problèmes très divers, allant de la gestion de la relation client à la maintenance préventive, en passant par la détection de fraudes ou encore l'optimisation des sites web.

I. Qu'est-ce que le data mining

I.1. Définitions

Le data mining est un procédé d'exploration et d'analyse de grands volumes de données en vue d'une part de les rendre plus compréhensibles et d'autre part de découvrir des corrélations significatives, c'est-à-dire des règles de classement et de prédiction dont la finalité ultime la plus courante est l'aide à la décision. ***(LIAUDET, 2010)***

Le data mining est un procédé de production de connaissance. En terme de logique philosophique traditionnelle, le data mining consiste à produire des jugements (toutes les personnes sont x, la moyenne des y des personnes vaut tant, etc. : c'est l'étape de description et de compréhension des données) et des règles de raisonnements (si toutes les personnes sont « a » alors elles seront « b » : c'est l'étape modélisation qui permet la prédiction). ***(LIAUDET, 2010)***

Le data mining est un procédé qui permet de passer des données à la connaissance. ***(LIAUDET, 2010)***

Le data mining est un procédé qui permet de découvrir des « pépites » d'informations cachées dans la gangue des données. ***(LIAUDET, 2010)***

I.2. Pourquoi la naissance du data mining ?

• Augmentation des capacités de stockage des données.

• Augmentation des capacités de traitements des données

• Maturation des principes des bases de données (maturation des bases de données relationnelles).

• Croissance exponentielle de la collecte des données (scanners de supermarché, internet, etc.)

• Croissance exponentielle des bases de données : capacités atteignant le terabits (10 bits) et émergence des entrepôts de données : datawarehouse, rendant impossible l'exploitation manuelle des données.

• Plus grande disponibilité des données grâce aux réseaux (intranet et internet).

• Développement de logiciels de data mining. ***(LIAUDET, 2010)***

I.3. Intérêt du data mining

Les entreprises sont inondées de données (scanners des supermarchés, internet, bases de données, etc.). Ces données languissent dans des entrepôts de données (ou référentiels, ou datawarehouse).

• Le data mining permet d'exploiter ces données pour améliorer la rentabilité d'une activité.

• Le data mining permet ainsi d'augmenter le retour sur investissement des systèmes d'information. ***(LIAUDET, 2010)***

I.4. Finalités du data mining : comprendre et décider, savoir et prévoir (la raison et la volonté)

Le data mining est un outil qui permet de produire de la connaissance dans le but de comprendre les phénomènes dans un premier temps : savoir et de prendre des décisions dans un second temps : prévoir pour décider.

II. Le processus standard d'une étude de data mining

II.1. Une discipline et pas un produit

À 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 maîtrisée.

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. ***(LIAUDET, 2010)***

II.2. Comment faire du mauvais data mining ?

• En travaillant sans méthode

• En ne préparant pas correctement ses données.

• En appliquant des boîtes noires de calculs sans les comprendre.

Un mauvais data mining peut amener à des conclusions erronées et donc à des conséquences très coûteuses. ***(LIAUDET, 2010)***

II.3. Comment faire du bon data mining ?

• En suivant une méthode

• En préparant les données correctement

• En comprenant le principe des modes opératoires (des algorithmes de calculs). En étant capable de savoir pourquoi on en choisit un plutôt qu'un autre. Une compréhension des modèles statistiques appliqués par le logiciel est donc nécessaire. ***(LIAUDET, 2010)***

III. CRISP-DM le processus de Data Mining

CRISP-DM, qui signifie Cross-Industry Standard Process for Data Mining, est une méthode mise à l'épreuve sur le terrain permettant d'orienter vos travaux de Data mining.

• En tant que méthodologie, CRISP-DM comprend des descriptions des phases typiques d'un projet et des tâches comprises dans chaque phase, et une explication des relations entre ces tâches.

• En tant que modèle de processus, CRISP-DM offre un aperçu du cycle de vie du Data mining.

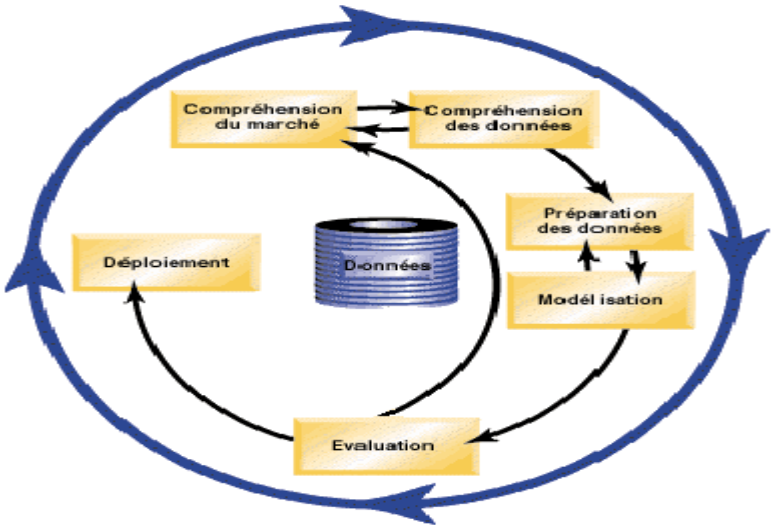


Figure 1 : CRISP-DM processus ***(SPSS Clémentine 11.1 Crisphelp, 2007)***

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. CRISP-DM vous permet de créer un modèle de Data mining adapté à vos besoins.

III.1. 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 tirer du Data mining. Cette consultation doit englober le plus grand nombre de personnes possible. L'étape finale de cette phase CRISP-DM concerne la production d'un plan de projet à l'aide des informations ainsi recueillies. Bien que cette étude puisse paraître 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 œuvre de ressources précieuses. ***(LAROSE, 2005) (IAN, 2005) (SPSS Clémentine 11.1 Crisphelp, 2007)***

- **Définition des objectifs**

Nous abordons à présent l'aspect le plus concret de la question. Suite aux recherches et aux études, il faut établir un objectif principal concret qui fasse l'unanimité des commanditaires du

projet et des autres unités de l'entreprise concernées par les résultats du projet. Cet objectif, dont la formulation peut être aussi confuse que « réduction du score d'attrition de la clientèle », se traduira par la suite par des objectifs de Data mining spécifiques qui guideront l'analyse.

- **Liste des tâches**

• Décrire le problème à résoudre à l'aide du Data mining.

• Énoncez toutes les questions le plus précisément possible.

• Déterminez tout autre impératif (par exemple, éviter de perdre les clients actuels tout en augmentant les possibilités de vente de produits associés).

• Précisez les bénéfices attendus (par exemple, réduction de 10 % du score d'attrition par les clients importants). ***(LAROSE, 2005) (IAN, 2005) (SPSS Clémentine 11.1 Crisphelp, 2007)***

- **Évaluation 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 :

• Quels types de données sont disponibles pour l'analyse ?

• Le personnel nécessaire à la réalisation du projet est-il disponible ?

• Quels sont les plus grands facteurs de risque en jeu ?

• Existe-t-il un plan de secours pour chaque risque ?

- **Risques et plans de secours**

La prudence veut aussi qu'on examine les risques encourus au cours du projet. Ces risques sont notamment liés aux domaines suivants :

• Programmation (Que se passe-t-il si le projet dure plus longtemps que prévu ?)

• Financement (Que se passe-t-il si le commanditaire du projet rencontre des difficultés budgétaires ?)

• Données (Que se passe-t-il si les données sont de mauvaise qualité ou peu représentatives ?)

• Résultats (Que se passe-t-il si les résultats initiaux sont moins spectaculaires que prévu ?)

Une fois ces différents risques évalués, on établit un plan de secours afin d'éviter les désastres.

Liste des tâches

• Évaluez tous les risques possibles avec précision.

• Établir un plan de secours pour chaque risque.

- **Analyse coût-bénéfice**

Cette étape permet de répondre à la question : Quel est le résultat net ? Dans le cadre de l'évaluation finale, il est essentiel de comparer les coûts du projet aux bénéfices rapportés en cas de succès.

Liste des tâches

Dans l'analyse, il faut inclure une estimation des coûts liés aux éléments suivants :

• Collecte des données et utilisation de données externes

• Déploiement des résultats

• Coûts d'exploitation

Ensuite, on prend en compte les bénéfices apportés par les éléments suivants :

• Réussite de l'objectif principal

• Connaissances supplémentaires engendrées par l'exploration des données

• Avantages éventuels issus d'une meilleure compréhension des données

- **Production d'un plan de projet**

À 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.

- **Élaboration du plan du projet**

Le plan du projet est le document principal régissant tout le travail de Data mining. S'il est bien créé, 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.

Liste des tâches

Lors de l'élaboration du plan, on vérifie qu'on a bien répondu aux questions suivantes :

• Est-ce que nous avons discuté des tâches du projet et du plan proposé avec toutes les personnes concernées ?

• Le plan comprend-il une estimation des dates pour toutes les phases ou tâches ?

• Est-ce que nous avons inclus dans le plan les efforts et les ressources nécessaires au déploiement des résultats ou de la solution commerciale ?

• Les demandes de révision et les points de décision sont-ils mis en évidence dans le plan ?

• Est-ce que nous avons signalé les phases comprenant généralement des itérations multiples, telles que la modélisation ? ***(LAROSE, 2005) (IAN, 2005) (SPSS Clémentine 11.1 Crisphelp, 2007)***

- **Exemple de plan de projet**

Le plan d'ensemble de l'étude se présente comme suit :

Phase	Temps	Ressources	Risques
Compréhension du problème	1 semaine	Tous les analystes	Changement économique
Compréhension des données	3 semaines	Tous les analystes	Problèmes de données, problèmes technologiques
Préparation des données	5 semaines	Consultant en Data mining, analyste en bases de données (quelques heures)	Problèmes de données, problèmes technologiques
Modélisation	2 semaines	Consultant en Data mining, analyste en bases de données (quelques heures)	Problèmes technologiques, incapacité à trouver un modèle adéquat
Évaluation	1 semaine	Tous les analystes	Changement économique, incapacité à mettre en œuvre les résultats
Déploiement	1 semaine	Consultant en Data mining, analyste en bases de données (quelques heures)	Changement économique, incapacité à mettre en œuvre les résultats

III.2. Présentation de la compréhension des données

La phase de compréhension des données de CRISP-DM 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.

La compréhension des données 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. ***(LAROSE, 2005) (IAN, 2005) (SPSS Clémentine 11.1 Crisphelp, 2007)***

- **Collecte des données initiales**

Cette phase de l'utilisation de CRISP-DM implique l'accès aux données. Les données proviennent de sources variées, telles que :

• Données existantes : Cette catégorie comprend plusieurs types de données, telles que les données transactionnelles, les données issues d'enquêtes, les logs Web, etc. Évaluez ces données existantes pour voir si elles suffisent à répondre aux besoins.

• Données acquises : la société utilise-t-elle des données d'appoint, telles que des données démographiques ? Si la réponse est non, peut-être faut-il envisager leur utilisation.

• Autres données : Si les sources ci-dessus ne répondent pas à aux besoins, il faut mener des enquêtes ou effectuer davantage de suivis afin de compléter les magasins de données existants.

Liste des tâches

• Quels sont les attributs (colonnes) de la base de données qui semblent les plus prometteurs ?

• Quels sont les attributs qui semblent sans intérêt et peuvent être exclus ?

• Le nombre de données permet-il de tirer des conclusions pouvant être généralisées ou d'effectuer des prévisions précises ?

• Les attributs sont-ils trop nombreux pour la méthode de modélisation choisie ?

• peut-on opérez la fusion de données issues de plusieurs sources ? Si oui, certains points risquent-ils de poser problème lors de la fusion ?

• Est-ce que nous avons envisagé le mode de traitement des valeurs manquantes dans chacune de nos sources de données ?

- **Élaboration d'un rapport sur l'exploration des données**

Liste des tâches

• Quels types d'hypothèse avons-nous formulés au sujet des données ?

• Quels attributs semblent prometteurs en vue d'une future analyse ?

• les explorations ont-elles révélé de nouvelles caractéristiques des données ?

• Comment ces explorations ont-elles modifié l'hypothèse initiale ?

• Pouvons-nous identifier des sous-ensembles de données particuliers à utiliser ultérieurement ?

- **Élaboration d'un rapport sur la qualité des données**

Suite à l'exploration et à la vérification de la qualité des données, maintenant on est prêt à élaborer un rapport qui orientera la prochaine phase de CRISP-DM.

Liste des tâches

• Est-ce qu'on a identifié des attributs manquants et des champs vides ? Si oui, ces valeurs manquantes ont-elles une signification ?

• L'orthographe présente-t-elle des incohérences pouvant engendrer des problèmes lors de fusions ou de transformations ultérieures ?

• Est-ce qu'on a exploré les écarts afin de déterminer s'il existe des « parasites » ou des phénomènes à analyser plus en profondeur ?

• Est-ce qu'on a vérifié la plausibilité des valeurs ? Relevez les conflits apparents (par exemple, des adolescents à revenus élevés).

• Est-ce qu'on a envisagé d'exclure les données sans impact sur vos hypothèses ?

• Les données sont-elles conservées dans des fichiers non hiérarchiques ? Si oui, les délimiteurs des différents fichiers sont-ils cohérents ? Chaque enregistrement contient-il le même nombre de champs ?

III.3. Présentation de la préparation des données

La préparation des données est l'un des aspects les plus importants et les plus coûteux en temps du Data mining. En fait, la préparation des données représente, selon les estimations, de 50 à 70 % du temps et des efforts consacrés à un projet. Le fait de consacrer une énergie suffisante aux phases initiales de compréhension du problème et de compréhension des données permet de réduire cette étape, mais la préparation et l'intégration des données en vue du Data mining requièrent encore beaucoup d'efforts. ***(LAROSE, 2005) (IAN, 2005) (SPSS Clémentine 11.1 Crisphelp, 2007)***

En fonction du type de société et de ses objectifs, la préparation des données comporte généralement les tâches suivantes :

• Fusion des ensembles et/ou des enregistrements de données

• Sélection d'un sous-ensemble de données exemple

• Agrégation des enregistrements

• Calcul de nouveaux attributs

• Tri des données en vue de la modélisation

• Suppression ou remplacement des blancs ou des valeurs manquantes

• Fractionnement en sous-ensembles d'apprentissage et de test

- **Sélection de données**

En fonction de la collecte initiale de données réalisée dans la phase CRISP-DM 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 des transactions ou le revenu des ménages.

- **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.

Problème posé par les données	**Solution possible**
Données manquantes	Excluez les lignes ou les caractéristiques, ou insérez une valeur estimée dans les blancs.
Erreurs dans les données	Procédez de manière logique pour découvrir manuellement les erreurs et les corriger, ou excluez les caractéristiques.
Codage des incohérences	Décidez d'une méthode de codage unique, puis convertissez et remplacez les valeurs.
Métadonnées erronées ou manquantes	Examinez manuellement les champs suspects et recherchez la signification correcte.

- **Prêt pour la modélisation ?**

Avant de créer des modèles, assurons-nous d'avoir répondu aux questions suivantes.

• Toutes les données sont-elles accessibles ?

• L'exploration et la compréhension initiale nous ont-elles permis de sélectionner des sous-ensembles de données pertinents ?

• Avons-nous nettoyé les données de manière efficace ou retiré les éléments irrécupérables ?
• Les ensembles de données multiples sont-ils correctement intégrés ? La fusion a-t-elle entraîné des problèmes nécessitant un complément d'informations ?

• Avons-nous étudié les impératifs des outils de modélisation qu'on prévoit d'utiliser ?

• Pouvons-nous résoudre certains problèmes de formatage avant la modélisation ?

III.4. Présentation de la modélisation

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 les 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 seule exécution. C'est pourquoi le Data mining est si intéressant. ***(LAROSE, 2005) (IAN, 2005) (SPSS Clémentine 11.1 Crisphelp, 2007)***

- **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.

Liste des tâches

Lorsqu'on décide des modèles à utiliser, on étudie les points suivants pour savoir s'ils ont une incidence sur notre choix :

• 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 nœuds de manipulation de données ?

- **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.

Gardez 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. À 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.

- **Évaluation du modèle**

À 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 ». ***(LAROSE, 2005) (IAN, 2005) (SPSS Clémentine 11.1 Crisphelp, 2007)***

III.5. Présentation de l'évaluation

A ce stade, nous avons réalisé la plus grande partie de votre projet de Data mining. Nous avons également déterminé, lors de l'étape de modélisation, que les modèles créés sont techniquement corrects et efficaces en fonction des critères de réussite du Data mining définis précédemment.

Néanmoins, avant de poursuivre, nous devons évaluer les résultats de nos efforts en utilisant les critères de réussite commerciale établis au début du projet. Cette étape est primordiale car elle permet de nous assurer que l'entreprise peut utiliser les résultats que nous avons obtenus. Le Data mining produit deux types de résultat :

• Les modèles finaux sélectionnés au cours de la phase précédente de CRISP-DM.

• Les conclusions ou déductions tirées des modèles eux-mêmes, ainsi que du processus de Data mining. Elles sont appelées constatations.

- **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.

Liste des tâches

Il faut d'abord récapituler les activités et les décisions pour chaque phase, en incluant les étapes de préparation des données, la création des modèles, etc. Ensuite, pour chaque phase, il faut tenir compte des questions suivantes et émettre 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 constaté que des modèles particuliers ne présentaient aucune perspective d'avenir ? Existe-t-il des moyens de prévoir ces impasses et par conséquent, de mieux concentrer les efforts ?

• Avons-nous eu des surprises (bonnes ou mauvaises) pendant cette phase ? Avec du recul, existe-t-il un moyen de prédire ces événements ?

• Des décisions ou des stratégies alternatives auraient-elles pu être utilisées lors d'une phase donnée ?

III.6. Présentation du déploiement

Le déploiement est le processus consistant à utiliser les nouvelles connaissances pour apporter des améliorations au sein de l'entreprise. Ceci peut se traduire par une intégration formelle telle que la mise en œuvre d'un modèle produisant des scores d'attrition qui sont ensuite lus dans un entrepôt de données. Le déploiement peut également signifier que vous utilisez les connaissances obtenues suite au Data mining pour provoquer un changement dans notre entreprise. Par exemple, vous avez peut-être découvert des motifs alarmants dans vos données indiquant un changement de comportement des clients âgés de plus de 30 ans. Ces résultats peuvent ne pas être intégrés formellement dans vos systèmes d'informations, mais ils seront sans aucun doute utiles pour la planification et la prise de décisions marketing. ***(LAROSE, 2005) (IAN, 2005) (SPSS Clémentine 11.1 Crisphelp, 2007)***

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

- **Production d'un 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 connaître. Il peut s'agir non seulement des administrateurs techniques responsables de la mise en œuvre des résultats de la modélisation mais aussi des commanditaires (marketing et gestion) qui prendront des décisions en fonction de nos résultats.

Liste des tâches

Commençons 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

créer des rapports distincts en fonction de chaque type de personne si leurs exigences diffèrent. 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 constatations)

• 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 et de la modélisation.

L'avenir de l'exploration de 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 numériques 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 de se développer.

Chapitre 4 : Les méthodes de classification et de segmentation

I. Classement des techniques du data mining

On distingue d'abord **deux grandes catégories de techniques** : les techniques descriptives et les techniques prédictives.

I.1. Les techniques descriptives (la classification)

- **Décrire**.
- Résumer, synthétiser, réduire, classer.
- Mettre en évidence des informations présentes mais cachées par le volume des données.
- Pas de variable cible à prédire.
- On les appelle aussi : **technique non supervisées.**
- Elles produisent des modèles de classement : typologie, méta-typologie. ***(LIAUDET, 2010)***

I.2. Les techniques prédictives (le scoring)

- **Prédire**.
- Extrapoler de nouvelles informations à partir des informations présentes.
- Les techniques prédictives présentent une variable cible à prédire.
- L'objectif est de prévoir la variable cible mais aussi de classer à partir de la variable cible.
- On les appelle aussi : **techniques supervisées.**
- Elles sont plus délicates à mettre en œuvre que les techniques descriptives.
- Elles demandent plus d'historique que les techniques descriptives.
- Elles produisent des modèles de prédiction. ***(LIAUDET, 2010)***

Deuxième distinction : variable numérique et variable catégorielle

Cette distinction est essentielle en statistique et en data mining.

Les **variables numériques** permettent de faire des résumés, des synthèses : moyenne, minimum, maximum, écart type, etc.

Les **variables catégorielles** permettent de faire des regroupements par catégories, c'est-à-dire des classements. ***(LIAUDET, 2010)***

II. Les 6 grands types de techniques du data mining

Le data mining permet d'accomplir les six types d'analyse suivants :

1 : Description - 2 : Classification - 3 : Association

4 : Estimation - 5 : Segmentation - 6 : Prévision.

Ces types d'analyse se répartissent dans les techniques descriptives et prédictives :

Techniques descriptives		Techniques prédictives		
Corrélation simple	Corrélation complexe	Présent		Futur
		Variable cible numérique	Variable cible catégorielle	
1 : Description	2 : Classification 3 : Association	4 : Estimation	5 : Segmentation	6 : Prévision

Distinction entre classification et classement

Dans un **classement**, on sait à l'avance à quelle classe l'individu appartient car on connaît à l'avance les classes. Le classement est un tri pour les variables numériques.

Dans une **classification**, on ne sait pas à l'avance à quelle classe un individu appartient car on ne connaît pas à l'avance les classes. La classification se fait en fonction de la population entière. ***(BACCINI, BESSE, 2005) (LAROCQUE, 2008) (LIAUDET, 2010)***

Classement	Classification
Ne crée pas nécessairement de nouvel attribut	Crée nécessairement un nouvel attribut
Les classes sont définies à partir d'un attribut unique ou d'un petit nombre d'attributs.	Les classes sont définies à partir d'un grand nombre d'attributs
Une classe est connue à partir d'un individu	Les classes sont connues à partir de la population
Les classes et leur nombre sont connus *a priori*.	Les classes et leur nombre sont connus *a posteriori*.
La classe d'appartenance d'un individu est définie par l'individu lui-même.	La classe d'appartenance d'un individu est défini par ses relations avec la population.

Classement	Classification
Plutôt prédictif. Les données des attributs de classement sont utilisés pour prédire une variable cible. Exemple : superposition du « churn » en fonction du choix de l'option internationale.	Plutôt descriptif. Le classification crée un attribut de classification qui est la variable cible de la classification elle-même.

Les techniques concrètes

Le data mining utilise des techniques concrètes qui peuvent être limitées à un type de technique spécifique ou être partagées par plusieurs types de techniques.

· Exemple de méthodes descriptives : la classification hiérarchique, la classification des K moyennes, les réseaux de Kohonen, les règles d'association.

· Exemples de méthodes prédictives : les méthodes de régression, les arbres de décision, les réseaux de neurones, les K plus proches voisins. ***(BACCINI, BESSE, 2005) (LAROCQUE, 2008) (LIAUDET, 2010)***

Les techniques du data mining

II.1. la description (technique descriptive)

Principe :

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 : des liens entre les répartitions des valeurs des variables. Ces liens s'appellent des « tendances ».

Intérêt :

· Favoriser la connaissance et la compréhension des données.

Méthode :

· Méthodes graphiques pour la clarté : analyse exploratoire des données.

Exemples :

· Répartition des clients par âge (lien entre les variables «client » et « âge »).

II.2. la classification (technique descriptive)

Principe

La **classification** (ou ***clustering*** ou **segmentation**) consiste à créer des classes (c'est-à-dire des sous-ensembles) de données similaires entre elles et différentes des données d'une autre classe (autrement dit, l'intersection des classes entre elles doit toujours être vide). Autrement dit, il s'agit pour n variables de créer des sous-ensembles disjoints de données. On dit aussi « segmente**r** » l'ensemble entier des données.

La classification définit les grands types de regroupement et de distinction : on parle de métatypologie (type de type). Elle permet une vision générale de l'ensemble (de la clientèle, par exemple).

Intérêt :

· Favoriser, grâce à la métatypologie, la compréhension et la prédiction.

· Fixer des segments qui serviront d'ensemble de départ pour des analyses approfondies.

· Réduire les dimensions, c'est-à-dire le nombre d'attributs, quand il y en a trop au départ.

Méthodes :

· Classification hiérarchique

· Classification des K moyennes

· Réseaux de Kohonen.

· Règles d'association.

Exemples :

Métatypologie d'une clientèle en fonction de l'âge, les revenus, le caractère urbain ou rural, la taille des villes, etc.

Pour un audit comptable, classer un comportement financier en catégorie normale et suspecte.

II.3. l'association (technique descriptive)

Principe :

L'association consiste à trouver quelles valeurs des variables vont ensemble. Par exemple, telle valeur d'une variable va avec telle valeur d'une autre variable.

Les règles d'association sont de la forme : si antécédent, alors conséquence.

L'association ne fixe pas de variable cible. Toutes les variables peuvent à la fois être prédicteurs et variable cible. On appelle aussi ce type d'analyse une « analyse d'affinité ».

Intérêt :

Mieux connaître les comportements.

Méthodes :

· Algorithme *a priori.*

· Algorithme du GRI (induction de règles généralisée).

Exemples :

· Analyse du panier de la ménagère (si j'achète des fraises, alors j'achète des cerises).

· Étudier quelle configuration contractuelle d'un abonné d'une compagnie de téléphone portable conduit plus facilement à un changement d'opérateur.

II.4. l'estimation (technique prédictive)

Principe :

L'estimation consiste à définir le lien entre un ensemble et une variable cible. Ce lien est défini à partir de données « complètes », c'est-à-dire dont les valeurs sont connues tant pour les prédicteurs que pour la variable cible. Ensuite, on peut déduire une variable cible inconnue de la connaissance des prédicteurs. À la différence de la segmentation (technique prédictive suivante) qui travaille sur une variable cible catégorielle, l'estimation travaille sur une variable cible numérique.

Intérêt :

· Permettre l'estimation de valeurs inconnues.

Méthodes :

· Analyse statistique classique : régression linéaire simple, corrélation, régression multiple, intervalle de confiance, estimation de points.

· Réseaux de neurones

Exemples :

· Estimer la pression sanguine à partir de l'âge, le sexe, le poids et le niveau de sodium dans le sang.

· Estimer les résultats dans les études supérieures en fonction de critères sociaux.

II.5. la segmentation (technique prédictive)

Principe :

La segmentation est une estimation qui travaille sur une variable cible catégorielle.

On parle de segmentation car chaque valeur possible pour la variable cible va définir un segment (ou type, ou classe, ou catégorie) de données. La segmentation peut être vue comme une classification supervisée.

Intérêt :

· Permettre l'estimation de valeurs inconnues.

Méthodes :

· Graphiques et nuages de points.

· Méthode des k plus proches voisins.

· Arbres de décision.

· Réseau de neurones.

Exemples :

· Segmentation par tranche de revenus : élevé, moyen et faible (3 segments). On cherche les caractéristiques qui conduisent à ces segments.

· Déterminer si un mode de remboursement présente un bon ou un mauvais niveau de risque crédit (deux segments).

II.6. la prévision (technique prédictive)

Principe :

La prévision est similaire à l'estimation et à la segmentation mise à part que pour la prévision, les résultats portent sur le futur.

Intérêt :

· Permettre l'estimation de valeurs inconnues.

Méthodes :

· Celles de l'estimation ou de la segmentation.

Exemples :

· Prévoir le prix d'action à trois mois dans le futur.

· Prévoir le temps qu'il va faire.

· Prévoir le gagnant du championnat de football, par rapport à une comparaison des résultats des équipes. ***(BACCINI, BESSE, 2005) (LAROCQUE, 2008) (LIAUDET, 2010)***

III. Fonctionnement général des méthodes de classification

- **Principe de la classification**

Une classe est un ensemble d'éléments qui sont semblables entre eux et qui sont dissemblables à ceux d'autres classes.

Classifier consistera à maximiser les similarités des éléments qui sont dans la même classe et à minimiser les similarités de ces éléments avec ceux des autres classes. Inversement, on peut dire que classifier consiste à minimiser la variation intra-classe et à maximiser la variation inter-classe. ***(LIAUDET, 2010)***

- **Classification et techniques supervisées**

Quand on part d'un volume de données très important, on a intérêt à faire une classification préalable pour réduire l'espace de recherche des algorithmes supervisés.

Comment mesurer la similarité ? Notion de distance entre les enregistrements

C'est le premier problème inhérent à la classification. La distance euclidienne entre deux enregistrements « x » et «y »est la suivante : d(x,y)= $\sqrt{\sum_i (Xi - Yi)\text{^}2}$

xi = x1, x2… xn représentent les valeurs des variables de « x ». De même pour « y ». Il existe d'autres calculs de distance. Pour que les distances soient comparables d'une variable à une autre, on va utiliser la technique des normalisations : normalisation « min-max » ou normalisation par le « test Z »

Normalisation « min - max » : x' = (x-min(x)) / amplitude(x)

Normalisation « test Z » : x'= (x - moy(x)) / écart type(x)

- **Comment mesurer les variables catégorielles ?**

C'est le second problème inhérent à la classification. Quand on a une variable booléenne, cela ne pose pas de difficulté. Faux vaut 0 et vrai vaut 1. Pour des variables énumérées, on

considérera que Si xi = yi alors xi-yi = 0 sinon xi-yi = 1 (c'est une sorte de généralisation du cas précédent).

IV. Fonctionnement général des méthodes supervisées

- **Rappels : variable cible et variables prédictives**

Variable cible

La variable cible est la variable dont on cherche à connaître la valeur. On parle aussi de variable à expliquer, réponse, variable dépendante, variable endogène. C'est la variable « en sortie ».

- **Variables explicatives**

Les variables explicatives sont les variables utilisées pour fabriquer le modèle. On parle aussi de variables prédictives ou de prédicteurs. Ce sont les variables « en entrée ».

- **Définition générale d'un modèle prédictif**

Un modèle prédictif est un ensemble de règles de découpage et d'association des variables explicatives. En appliquant ces règles à n'importe quel nouvel individu de la population, on pourra déterminer la valeur de l'individu pour la variable cible.
Les techniques prédictives sont nombreuses et leur domaine d'application tout autant. Elles servent aussi bien à calculer l'efficacité d'un traitement médical, à prévoir le temps en météorologie, qu'à prévoir le rendement d'une culture en agriculture. Ces techniques ont un cadre théorique précis qu'il faut connaître pour les appliquer correctement. ***(LIAUDET, 2010)***

- **Description intuitive d'un modèle prédictif**

Le but est de connaître une information qu'on ne connaît pas. Par exemple, on veut savoir si un client va rembourser le prêt qu'on lui fait.
Pour calculer cette information, on va s'intéresser aux clients qui ont déjà eu des prêts. Et on va chercher une corrélation générale entre les données économiques, sociales, géographiques et comportementales (le comportement des comptes) et le fait que ces clients aient ou n'aient pas remboursé leurs prêts. Cette corrélation, c'est le modèle prédictif. Une fois trouvée, on peut l'appliquer au client qui demande un prêt : c'est ce qu'on appelle une mesure de score de risque. ***(LIAUDET, 2010)***

- **Distinction entre les méthodes supervisées : classement et prédiction**

Le classement : variable cible catégorielle

Encore appelé « discrimination », le classement est une technique prédictive dont la variable cible est une variable catégorielle, le plus souvent booléenne. Le classement permet de placer chaque individu dans une classe correspondant à une catégorie de la variable cible.
À noter que le classement est aussi le nom donné à une technique de modélisation descriptive, par opposition à la classification. Il s'agit bien du même « classement » dans le sens où on connaît a priori les catégories de classement. Quand il s'oppose à la classification, le classement est descriptif, sans variable cible. Quand il s'oppose à la prédiction, le classement est prédictif, avec variable cible.
L'exemple type sera le classement prédictif par arbre de décision. ***(LIAUDET, 2010)***

La prédiction : variable cible continue

Encore appelé « régression », la prédiction est une technique prédictive dont la variable cible est une variable continue.

- **Deux grands types de technique : inductive et transductive**

Les techniques transductives
Elles ne présentent qu'une seule phase. Elles ne produisent pas de modèle.
C'est pendant la classification des individus connus que se fait la prédiction des données inconnues. Toute prédiction demande donc un accès à la population complète (ou à un échantillon) et demande une grande puissance de calcul et peut donc être assez longue. ***(LIAUDET, 2010)***

Les techniques inductives

1 : Elles présentent trois phases (parfois quatre) :

· Une phase d'apprentissage qui permet d'élaborer un modèle. C'est la phase inductive.

· Une phase de test pour vérifier le modèle obtenu (et éventuellement une phase de validation en plus).

· Une phase de prédiction ou de classement qui consiste à appliquer le modèle à de nouvelles données. C'est la phase déductive.

Les phases d'apprentissage, de test et de validation sont effectuées sur des échantillons distincts de la population.

2 : Elles produisent un modèle.

Les techniques inductives sont plus répandues car le modèle produit permet un contrôle du modèle (courbe de ROC et indice de Gini) et une application facilitée : une prédiction se fait à partir du modèle, sans retour à la population ou à un échantillon d'origine. C'est rapide et demande peu de puissance de calcul.

V. Les réseaux de neurones

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.

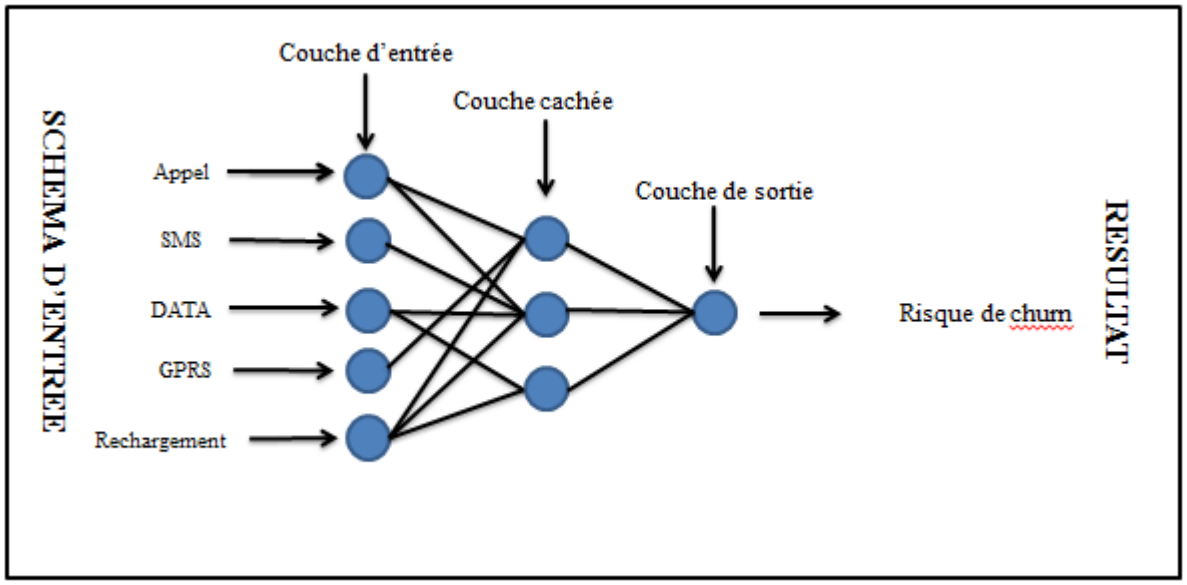


Figure 1 : Structure d'un réseau de neurones

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 prévisions 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. ***(Tuffery, 2010) (Wisra, 2009) (Parizeau, 2004)***

V.1. La notion neurone réel et formel

1- Neurone réel

Les neurones réels présentent trois régions principales : le corps cellulaire, les dendrites - prolongements relativement courts et arborescents du corps cellulaire - et l'axone, prolongement long et fibreux.

Un neurone utilise des dendrites pour rassembler des données d'entrée issues d'autres neurones. Ces données d'entrée sont combinées pour produire une réponse envoyée à d'autres neurones ou d'autres cellules. Les axones transportent les influx en provenance du corps cellulaire vers d'autres cellules (la longueur d'un axone est très variable ; elle peut atteindre 1 m chez l'homme et près de 10 m chez la girafe).

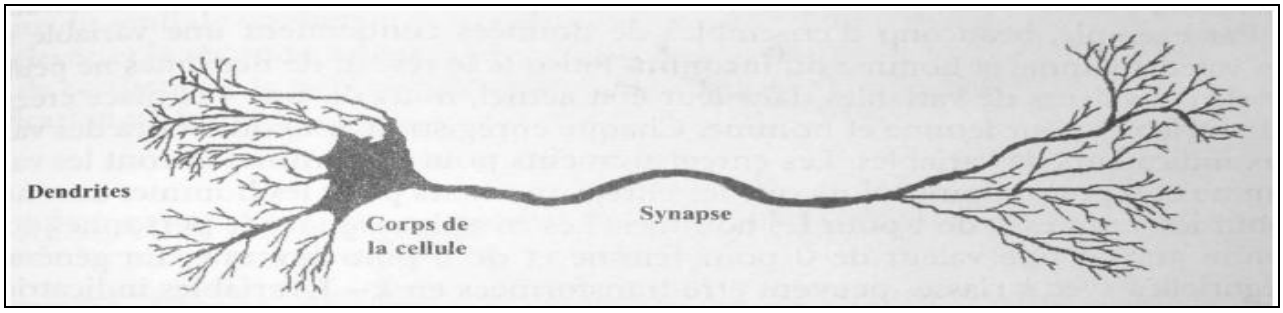


Figure 2 : Schéma d'un neurone réel ***(LIAUDET, 2010)***

2- Neurone formel

Les **données d'entrée** (xi) sont recueillies à partir des neurones du flux supérieur dans l'ensemble des données, et sont combinées dans une **fonction combinatoire** telle la somme. Cette fonction combinatoire est en entrée d'une **fonction d'activation** qui produit une réponse envoyée en entrée d'autres neurones.

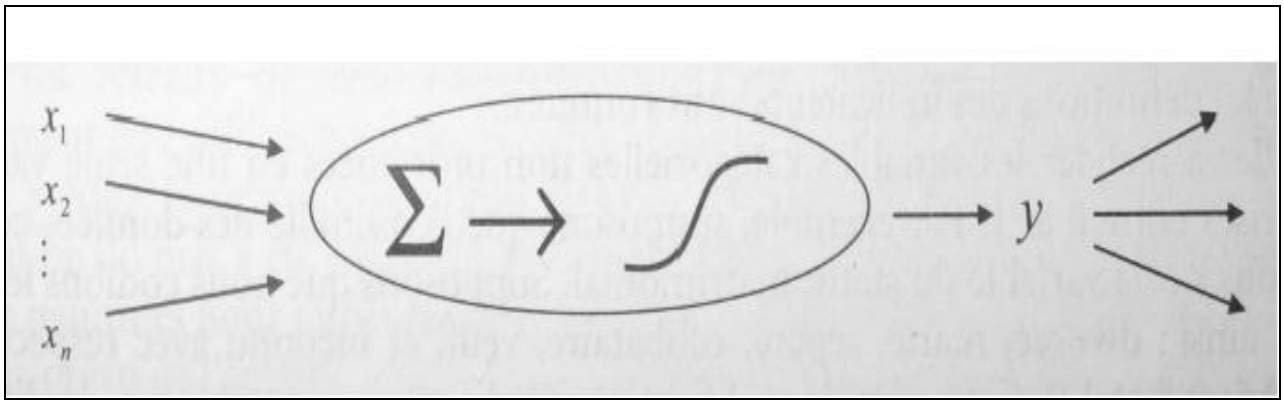


Figure 3 : Schéma d'un neurone formel ***(LIAUDET, 2010)***

Avantage des réseaux de neurones

Robuste aux données bruitées.

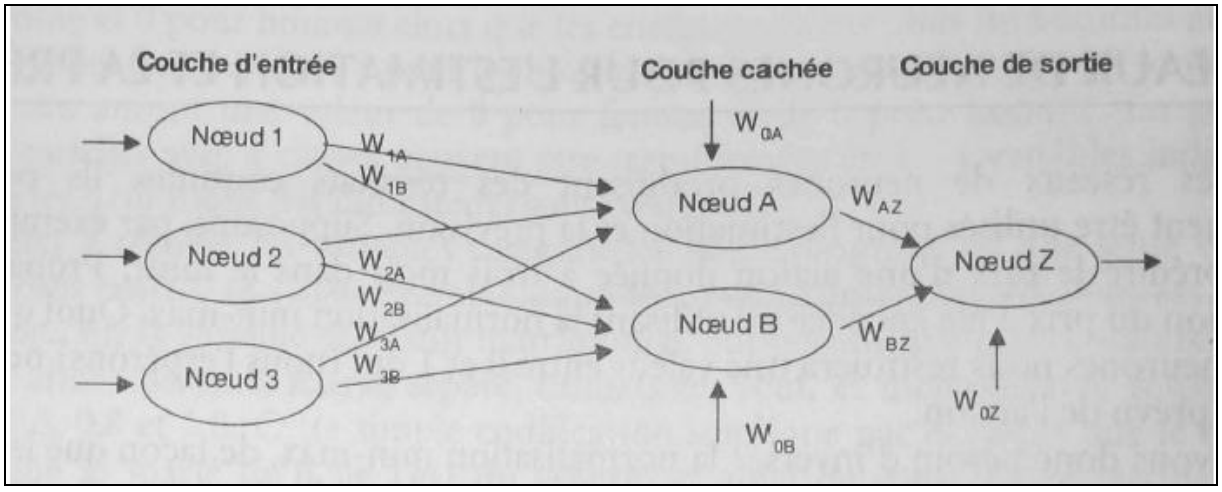
Permettent de modéliser de grandes variétés de comportements.

Inconvénients

Les résultats sont assez opaques, à la différence de la méthode des arbres de décision.

La mise en œuvre, qui passe par un apprentissage, peut être longue.

V.2. Architecture et principes de fonctionnement



Exemple d'un petit réseau de neurones

- Un réseau de neurones formels est disposé en couches de neurones formels.

  Les neurones sont appelés « nœuds ».

- La plupart des réseaux sont constitués de 3 couches successives : une **couche d'entrée**, une **couche cachée** et une **couche de sortie**. Toutefois, il peut y avoir 0 ou N couches cachées.
- D'une couche à l'autre, tous les nœuds de la première couche (nœuds « in ») sont reliés à tous les nœuds de la seconde (nœuds « out »).
- Chaque liaison a un poids : une valeur entre 0 et 1.
- Chaque nœuds des couches cachées et de sortie possède aussi un poids : une valeur entre 0 et 1.
- Le nombre de nœuds de la couche d'entrée dépend du nombre de variables prises en compte et de leur type. En simplifiant, on peut dire qu'on a un nœud par variable en entrée.
- Le nombre de couches cachées et le nombre de nœuds pour chaque couche cachée est paramétrable par l'utilisateur.
- En général, la couche de sortie ne contient qu'un nœud. Toutefois, elle peut en contenir plus. En simplifiant, on peut dire que ce nœud correspond à la variable de sortie.

Type des données en entrée

La valeur des données en entrée et en sortie doit être comprise entre 0 et 1.

Traitement des variables numériques

On applique une standardisation « min-max » aux données numériques :

$x' = (x - \text{moy}(X)) / (\max(X) - \min(X))$

Si on applique les résultats à une population dans laquelle le min et le max ont changé, on peut obtenir des résultats erronés.

Traitement des variables catégorielles :

Si elles sont ordonnées, on peut affecter à chaque catégorie une valeur comprise entre 0 et 1.

Si elles ne sont pas ordonnées, la méthode précédente risque de conduire à des résultats erronés du fait de la création de voisinages irréels.

Chaque catégorie peut être alors être traitée comme une variable booléenne.

Paramétrage de la couche cachée

On peut choisir le nombre de nœuds de la couche cachée et le nombre de couche cachée.

Plus le nombre de nœuds augmente, plus le réseau est apte à identifier des phénomènes complexes.

Toutefois, un trop grand nombre de nœuds conduit à un sur-apprentissage dans l'échantillon d'apprentissage finalement nuisible aux échantillons de test.

Valeurs des nœuds et des liaisons

Lors de l'initialisation, un poids est donné aléatoirement à chaque liaison et à chaque nœud des couches cachées et de sortie. **L'ajustement de ces poids représente la clé du mécanisme d'apprentissage par le réseau de neurones.**

· Pour un individu, les nœuds de la couche d'entrée prennent la valeur normalisée des variables d'entrée du modèle.

· Pour un individu, les nœuds des couches cachées et de sortie prennent une valeur qui est une combinaison (une somme le plus souvent) des combinaisons linéaires des nœuds « in » et des poids correspondants.

Pour un nœud j donné, on a donc :

NET j = Somme pour i de 0 à N (W i j * X i)

Avec :

NET : valeur du nœud dans le réseau.

i : allant de 0 à N, N étant le nombre de nœuds « in ».

W i j : poids de la liaison entre le nœud « i » qui est « in » et le nœud « j » qui est « out ».

X i : valeur du nœud « i », avec X 0 = 1.

V.3. La fonction sigmoïde

Dans un neurone réel, les signaux sont envoyés entre les neurones quand la combinaison des données d'entrée dépasse un certain seuil : le **seuil d'activation**. Le comportement n'est pas linéaire car la réponse ne dépend pas linéairement de l'incrément de la stimulation.

La fonction qui modélise ce comportement est appelée : **fonction d'activation**. C'est une fonction non linéaire. La fonction d'activation la plus commune est la **fonction sigmoïde** :

y = 1 / (1 + exp (- x))

Soit :

SIG (NET(n)) = 1 / (1 + exp (-NET(n)))

Avec :

exp : fonction exponentielle : exp(1) = 2,7.

NET(n) : NET du nœud « n ».

SIG (NET(n)) : sigmoïde du NET du nœud « n ».

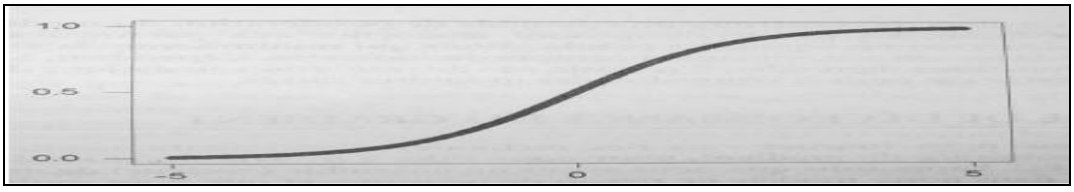


Figure 4 : Graphe de la fonction sigmoïde ***(LIAUDET, 2010)***

La fonction sigmoïde est telle que lorsque les données d'entrée sont proches du centre de l'intervalle, f(x) est linéaire. Lorsque les données d'entrée s'éloignent du centre, f(x) est curviligne. Lorsque les données sont très éloignées du centre, f(x) devient quasiment constante.

Donc les incréments du NET d'un nœud produisent des incréments variables de SIG(NET) : près du centre, un incrément du NET produit un incrément linéaire du SIG. Plus on s'écarte du centre, moins l'incrément du NET a d'effet sur le SIG. Loin du centre un incrément du NET ne produit pas d'incrément du SIG. La fonction sigmoïde est aussi appelée : **fonction d'écrasement** : elle écrase les extrêmes.

On va appliquer la fonction sigmoïde à la valeur NET de chaque nœud.

V.4. La SEC

Les réseaux de neurones sont une méthode supervisée : on choisit une variable cible. Chaque individu avec ses variables en entrée passe à travers le réseau et fournit un résultat dans le nœud de sortie. Cette valeur de sortie est comparée à celle de la variable cible.

Erreur de prévision = valeur de la donnée réelle - valeur de sortie

Cette erreur est analogue à celle des modèles de régression.

En général, les modèles à réseau de neurone calculent une somme des erreurs au carré (SEC) :

SEC = Somme pour tous les enregistrements (donnée réelle – donnée en sortie)

Le problème consiste donc à **minimiser la valeur de SEC en fonction de l'ensemble des valeurs de pondération** des nœuds et des liaisons.

V.5. La rétropropagation

En raison de la nature non linéaire de la fonction sigmoïde, il n'existe pas de résolution analytique de la minimisation de la SEC.

La rétropropagation met en œuvre des calculs mathématiques et algorithmiques complexes que nous ne présentons pas ici.

Nous présentons seulement les principaux concepts et paramètres qui entrent en jeu.

Méthode de décroissance du gradient de la SEC pour ajuster les pondérations

Pour minimiser la SEC, on utilise la « **méthode de décroissance du gradient** » qui donne la direction dans laquelle il faut ajuster la pondération pour faire décroître la SEC.

La courbe ci-dessus montre une évolution parabolique de la SEC en fonction d'une seule pondération. C'est une simplification qui permet de montrer que la dérivée de la courbe donne la pente et nous dit dans quel sens il faut ajuster le poids.

La rétropropagation consiste à **ajuster les poids des nœuds et des liaisons en remontant du**

Nœud de la couche de sortie aux nœuds de la couche d'entrée.

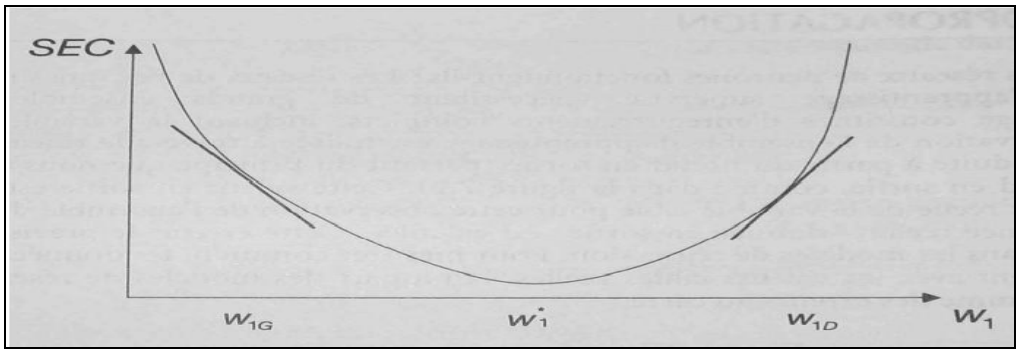


Figure 5 : Courbe de l'évolution du SEC en fonction du poids ***(LIAUDET, 2010)***

En général, les réseaux font la mise à jour après chaque calcul de la valeur de sortie d'un enregistrement. Cet ajustement sera fonction de :

· L'erreur de prévision

· Le taux d'apprentissage : valeur comprise entre 0 et 1.

Le taux d'apprentissage (Eta)

Le taux d'apprentissage est un paramètre qui favorise l'évolution de la SEC vers le minimum. Quand le taux d'apprentissage est faible, les ajustements sont faibles.

Quand le taux d'apprentissage est fort, les ajustements sont forts. Mais un taux d'apprentissage trop fort fait dépasser la SEC optimum.

Le taux d'apprentissage peut évoluer au cours de l'apprentissage. Au début, il est élevé pour s'approcher rapidement de la solution. Quand le réseau commence à converger, le taux est graduellement réduit pour ne pas dépasser la SEC optimum.

Le terme de moment (alpha)

Le terme de moment est un paramètre supplémentaire qui favorise l'évolution de la SEC vers le minimum. Intuitivement, on peut comprendre son fonctionnement ainsi : la courbe d'évolution de la SEC en fonction des poids n'est pas une simple parabole. Elle contient plusieurs minimums ou « paliers ». Le terme de moment permet d'éviter que la recherche du meilleur minimum s'arrête à un palier intermédiaire ou qu'il se trouve avant ou après le meilleur palier.

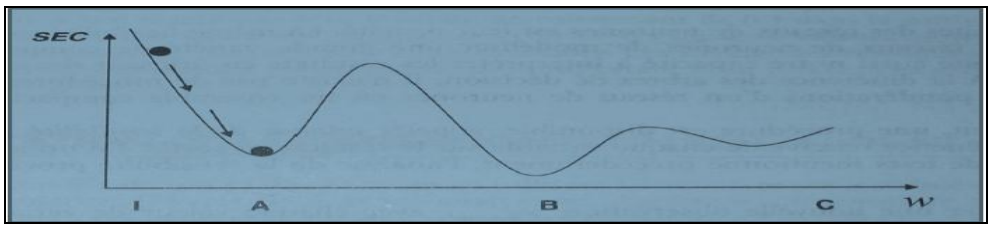


Figure 6 : Courbe de l'évolution du SEC en fonction du poids ***(LIAUDET, 2010)***

On peut interpréter cette courbe en disant que le terme de moment favorise le fait de ne pas s'arrêter au palier A, le fait de ne pas aller au palier C, le fait de s'arrêter au palier B.

Critères d'arrêt

L'algorithme peut traiter tous les enregistrements de l'ensemble des données d'apprentissage un nombre de fois indéterminé. Il faut donc déterminer un critère d'arrêt.

Le temps : à éviter

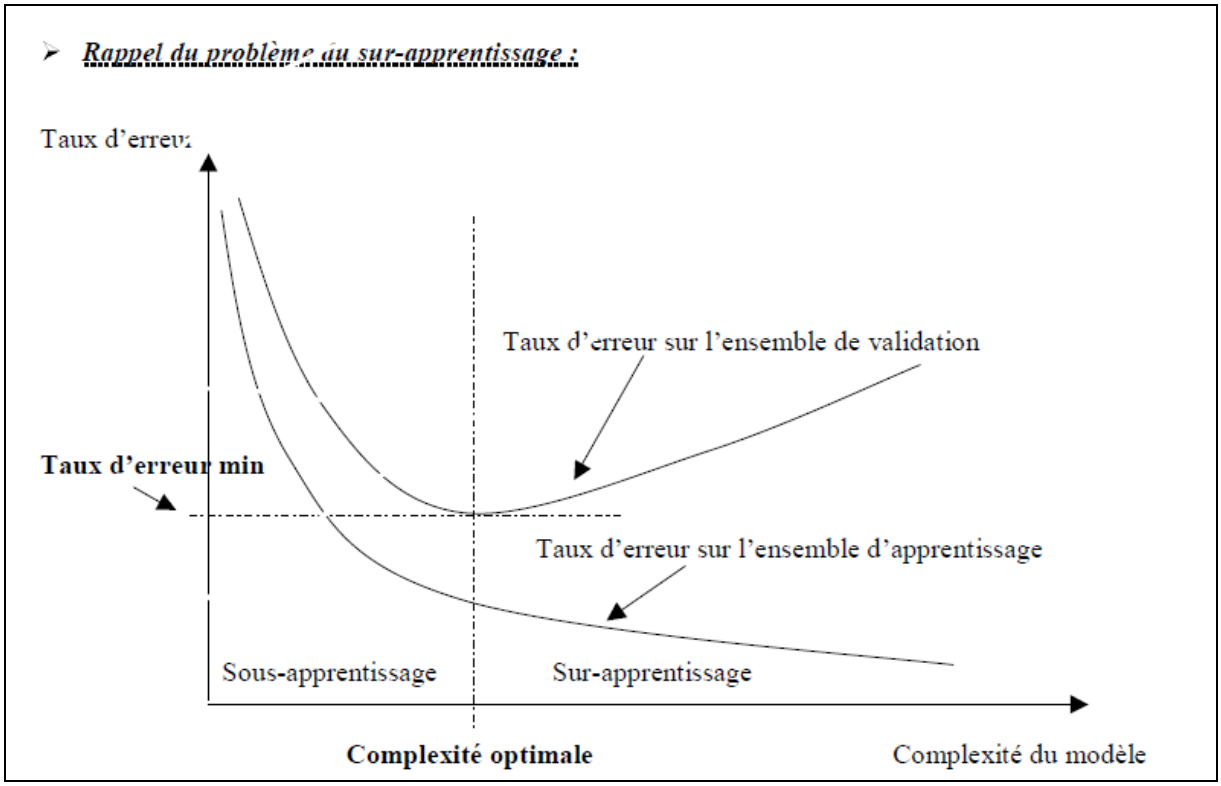
La modélisation par réseau de neurones peut prendre plusieurs heures ! Le temps peut donc un critère d'arrêt si on est pressé, mais il risque de conduire à un modèle peu efficace.

Minimiser la SEC dans l'ensemble d'apprentissage : à éviter !

La SEC peut être un critère d'arrêt mais elle risque de conduire à un sur-apprentissage mémorisant des caractéristiques idiosyncrasiques (propres aux individus indépendamment de leur groupe).

Minimiser la SEC dans l'ensemble de validation

En même temps qu'on minimise la SEC dans l'ensemble d'apprentissage, on vérifie qu'on la minimise aussi dans l'ensemble de validation. Quand elle croit dans l'ensemble de validation, c'est qu'on commence à entrer en phase de sur-apprentissage. C'est un bon critère d'arrêt. Le modèle doit atteindre le taux d'erreurs minimum pour l'ensemble de validation.



Chapitre 5 : Étude d’un cas télécom : La modélisation du churn

Dans ce chapitre nous fournirons une méthodologie complète d'une étude data mining, on discernera l'importance et l'utilité de cette démarche sur un cas WTA (Wataniya Télécom Algérie) qui nous permettra de donner une vision réelle et palpable.

En premier lieu on développera la problématique Churn et son positionnement chez WTA, ensuite on travaillera sur les bases de données et on montrera le cheminement complet de l'information.

On passera par la suite à la partie modélisation où on travaillera sur le ciblage et la segmentation de la base de données, le résultat de cette étape est primordial pour la phase finale qui sera dédiée à la conception d'une action de rétention client comme ça peut être une compagne de fidélisation.

I. La vision du client/data mining chez WTA

WTA fait partie des opérateurs téléphoniques qui se soucient de sa clientèles et qui pensent à leurs satisfactions en permanence, cette vision responsable et respectueuse à l'égard du client demande beaucoup de finesse et de technicité pour exploiter toute la richesse des données disponibles sur la base client, en faisant appel à trois disciplines complémentaires : les technologies de gestion de BDD (base de données), les méthodes statistiques d'analyse des données (segmentation, prédiction…), les techniques d'intelligence artificielle (réseaux de neurones, Arbres de décision, logique floue…).

II. Apports du DATA MINING pour WTA

- Analyser les comportements clients.
- Segmenter les clients.
- Prédire le risque d'attrition (churn)
- Up selling/Cross selling : ajuster/concevoir des offres en fonction des profils clients.
- Détecter des comportements anormaux (abus, fraudes, impayés…).
- Simuler l'impact de scénarios d'actions en fonction de l'analyse de sensitivité (changement tarifaire, bonus…).

III. Objectifs pour WTA

- **Disposer d'un système d'information efficace**

Grace au datamart marketing on peut avoir plus de réactivité et d'autonomie dans le reporting cela va permettre aux utilisateurs d'avoir plusieurs possibilités et angles d'analyse avec un système d'information fiable et rapide.

- **Exploiter toutes les données clients**

Grace à la diversité et le nombre de tables et variables qui existent sur le datamart, chaque *Data Miner* possède une visibilité à 360° sur le client, à partir d'ici on peut analyser son comportement et sa façon de réagir.

Cette disponibilité d'information peut nous permettre aussi de faire une segmentation globale de la base de données et cela avec les critères qu'on veut, soit par rapport à la valeur actuelle que génère le client, soit par rapport à son comportement et ses habitudes, soit par rapport à sa durée de vie avec l'entreprise, soit par rapport à sa valeur future,…..

L'exploitation de la base de données nous permet de réaliser des modèles de Churn grâce à un bon Scoring, d'ici on peut prédire le risque d'attrition et son taux, bien meilleur que cela on peut même déterminer les variables qui influent sur le churn.

- **Optimiser les actions marketing**

Chaque offre de rétention ou bien de fidélisation a pour but final de satisfaire la clientèle au maximum, d'où la nécessité de faire un bon ciblage, grâce aux études data mining on peut atteindre cet objectif et suive la réponse des clients et leur réaction par rapport aux actions faites ainsi on optimisera nos actions futur.

IV. Architecture Datamart

VI.1. Modèle dimensionnelle du datamart

Après le traitement des données au niveau du datawarehouse, elles seront publiées sur le Datamart marketing qui est une vue spécifique des données datawarehouse (dédiée au

Marketing), le modèle dimensionnel est le schéma en constellation des faits.

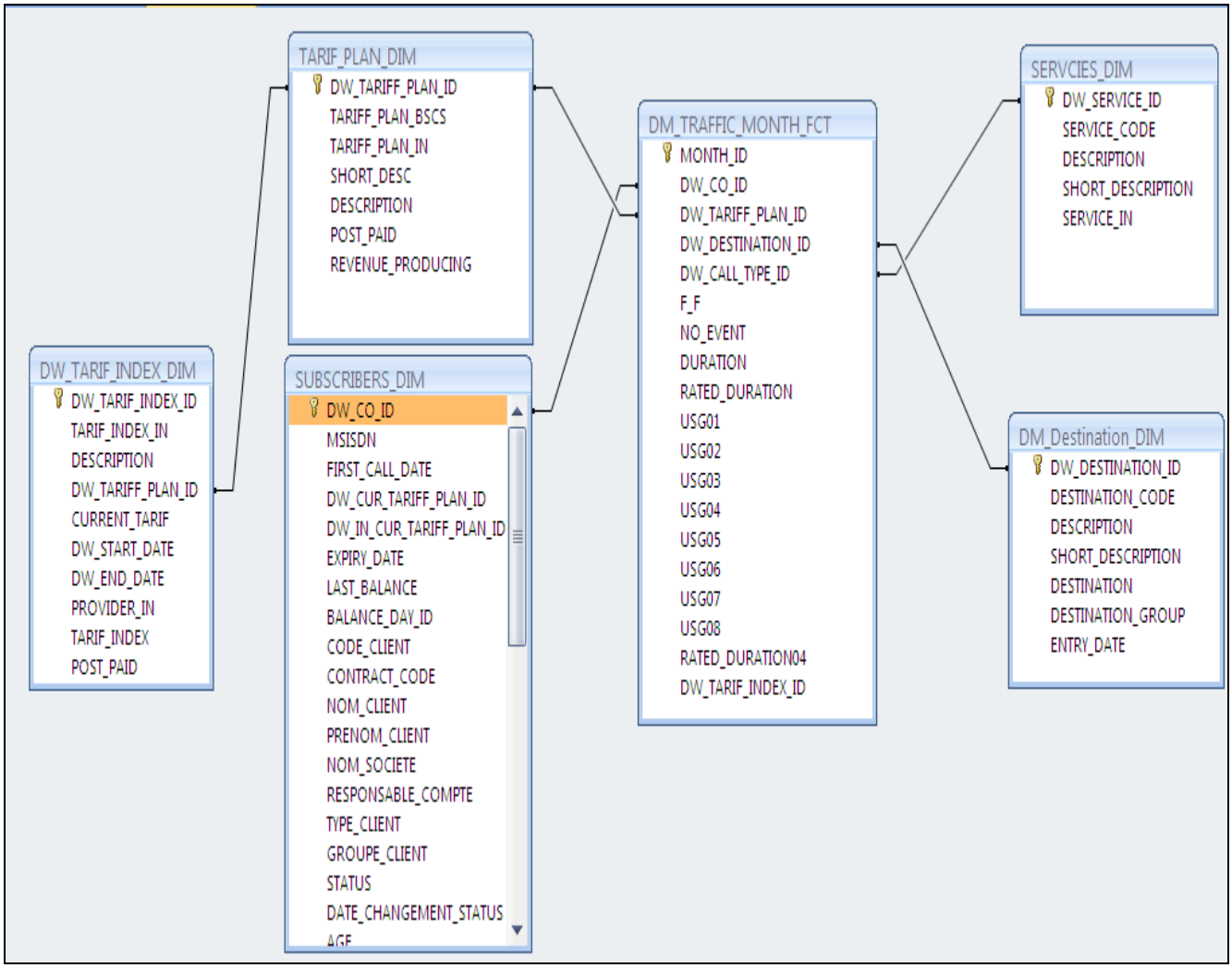


Figure 01 : Architecture datamart marketing

Le schéma ci-dessus nous montre une vision partielle des tables disponibles sur le datamart et les jointures possibles entre les tables.

VI.2. Lecture des tables

Un datamart contient plusieurs tables, chaque table nous donne une description des variables qu'elle renferme, chaque table doit posséder impérativement une clé primaire qui permettra de faire les fusions avec les autres tables.

Une fusion avec d'autres tables s'impose si et seulement si nous avons besoin d'une variable qui ne figure pas sur la table initiale. ***(WALKENBACH, 2003)***

Exemple :

On est sur la table trafic et elle possède des variables liées à la consommation des clients, si on veut obtenir d'autres variables par exemple des données sur le rechargement il faut faire une fusion avec la table rechargement des abonnés, cette jointure se fera grâce à la clé primaire, l'absence de cette dernière dans l'une des tables du datamart signifie que la jointure n'est pas possible.

V. Étude de cas : la modélisation du churn (attrition) chez WTA

Déroulement du projet DM selon le modèle CRISP-DM (Cross industry standard processus for Data Mining)

Une fois les données chargées sur le datamart par l'équipe datawarehouse, on peut commencer le processus de data mining (Figure ci-dessous).

La modélisation du churn passe par les étapes vues précédemment (Processus Data Mining CRISP-DM) :

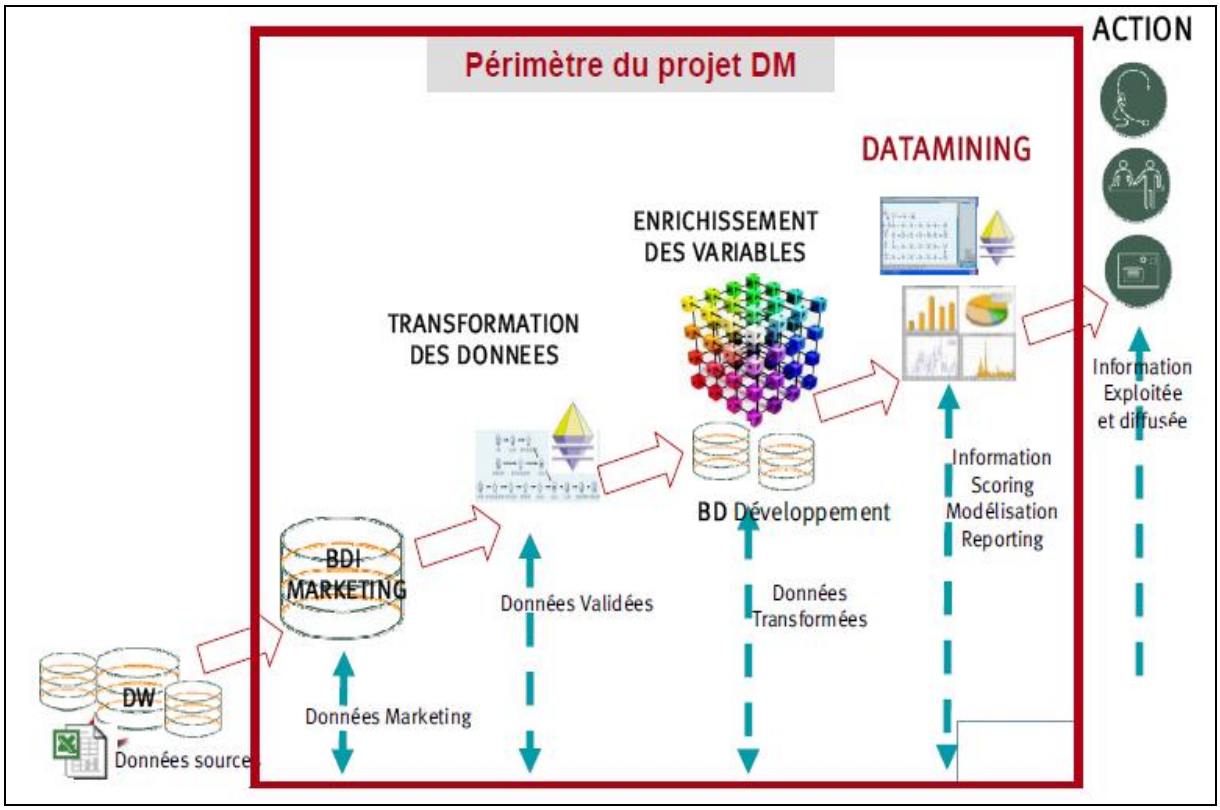


Figure 02 : Processus de data mining après chargement des données

V.1. Étape 1 : la 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. En effet, c'est un domaine dans lequel il y a beaucoup de clients et peu d'opérateurs. Les opérateurs ont donc intérêt de fidéliser leurs clientèles. L'étude statistique des données clients va permettre de mieux cibler les attentes du client. ***(UMMAN TUGBA, 2010)***

Donc la problématique qui se pose est comment minimiser le nombre d'attrition avec un processus de data mining et cela pour un meilleur ciblage qui va servir dans les actions et campagnes de rétention ou de fidélisation.

V.2. Étape 2 : la compréhension des données

Phase de conception :

- Définir les données dont on a besoin au niveau opérationnel(Marketing)
- Vérifier leur disponibilité au niveau DWH
- Définir l'historique à considérer
- Définir la période d'agrégation des données
- Définir le rythme de rafraîchissement de données

Phase de création :

- Préparer les données et les liens des sources
- Chargement des données

Phase de contrôle :

- Vérifier la concordance des données DWH versus Data mart

Pour le cas WTA toutes les phases citées ci-dessous sont respectées et la synchronisation datawarehouse, datamart marketing se fait d'une manière journalière.

V.3. Étape 3 : la préparation des données

Nous devons au préalable établir les principaux indicateurs de performance:

- **Identification client :** MSISDN (numéro client), date d'activation, offre, statut, migration…

- **Usages :** nombre et durée des appels par client, par destination (local, international, fixe…), trafic minutes (incoming, outgoing, onnet(inter réseaux), offnet(intra réseaux)…), nombre et durée d'usage data (SMS, MMS, Wap, téléchargements…), usage de F&F (Friends & Family), roaming, nombre d'appel par wilaya, par site, appels au Call Center, nombre de N° appelés et appelants, date du dernier appel in et out, IMEI (numéro de série du téléphone), lieu d'achat, nombre de jours actifs/non actifs.

- **Revenus :** consommation par type d'usage (appels, data), par origine et destination, solde crédit, nombre, dénominations, et montant rechargé, dates de rechargement.

- **Couverture :** taux de couverture par wilaya

- **Marché :** évènements de la concurrence (baisse tarifs, bonus recharge…)

Ces paramètres ou bien indicateurs vont nous permettre de construire des bases données nécessaires à l'étude et qui vont être à leur tour des inputs pour construire la base finale pour la modélisation.

V.3.1. Extraction des données à partir du datamart

L'étape d'extraction est le point sensible d'une étude data mining ou aucune erreur n'est tolérable, dans cette phase il faut :

- Bien choisir les tables disponibles sur le datamart (souvent on trouve les mêmes tables avec les mêmes noms mais avec des significations différentes).
- Choisir juste les variables nécessaires à l'étude pour minimiser le temps d'exécution des requêtes.
- Vérifier toutes les fusions, agrégations, calculs, transformations, codages, tris, échantillonnages,…...
- Vérifier les valeurs manquantes et si nécessaire de faire un passage par des méthodes d'interpolation ou d'estimation.
- Vérifier les valeurs aberrantes.
- Étudier les statistiques préliminaires pour une meilleure compréhension des données.
- Établir l'audit des données.

- **Extraction des données (cas WTA)**

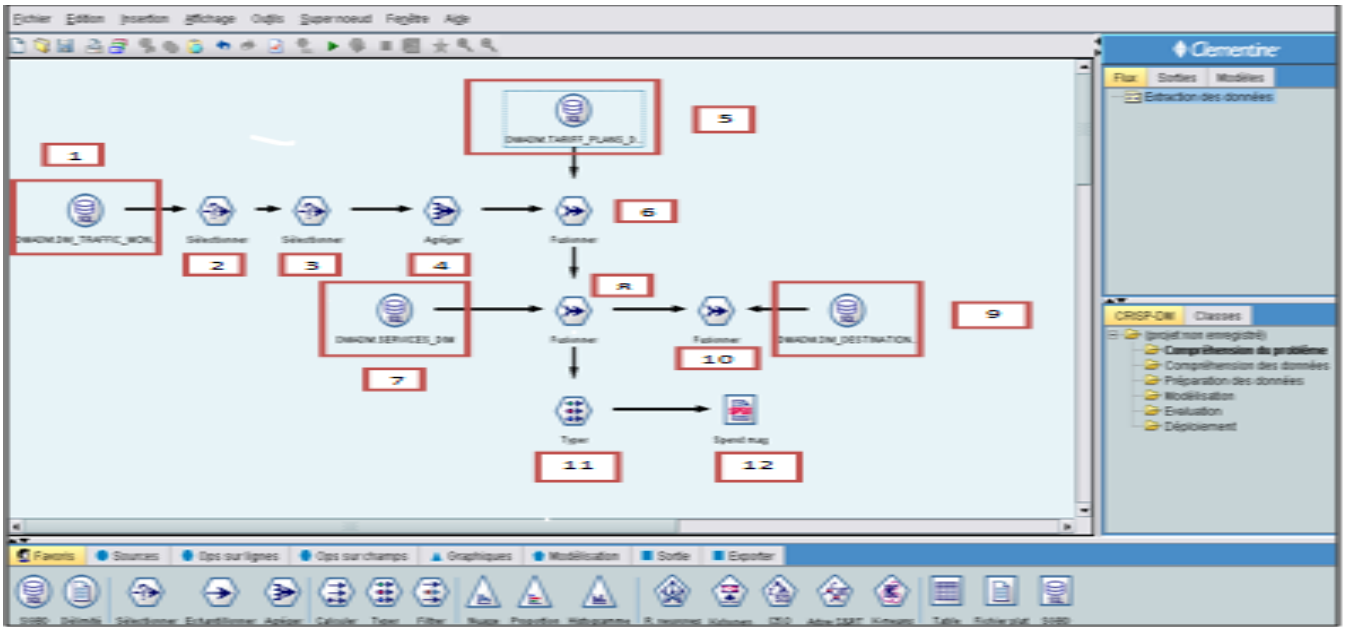
L'extraction se fait à l'aide d'SPSS Modeler (logiciel data mining dédié aux experts appelé aussi Clémentine).

- **Extraction de la base revenu client (Spend)**

Le Data mining avec Clémentine (SPSS Modeler 11.1) est basé sur l'exécution de données via une série de nœuds appelée **flux.** Cette série de nœuds représente les opérations à réaliser sur les données et les liens entre les nœuds indiquent la direction du flux de données. Les flux de données permettent de réaliser les opérations suivantes: envoi de données en mémoire, manipulations diverses des données et envoi de ces dernières vers une destination, telle qu'un fichier SPSS ou le programme Clémentine Solution Publisher.

L'interface unique de Clémentine nous permet d'explorer les données visuellement en travaillant avec des graphiques de flux de données. Au niveau le plus élémentaire, on peut créer un flux de données en utilisant la procédure suivante :

- Ajoutez des nœuds dans l'espace de travail de flux.
- Connectez les nœuds de façon à former un flux.
- Définissez toutes les options de nœud ou de flux.
- Exécutez le flux.



- **Flux 1 :** Extraction du Spend–

- **Étapes d'extraction du Spend**

On commence tout d'abord par faire une synchronisation datamart et SPSS Modeler pour récupérer les tables encadrées en rouge à savoir :

- Table Trafic (contient les informations sur le revenu : heure d'appel, nombre de minutes, montant d'appel,….)
- Table offre (contient toutes les offres)
- Table destination (contient les destinations du trafic : nationales, Internationales,….)
- Table service (contient le type du revenu : appel, SMS, GPRS, ……)

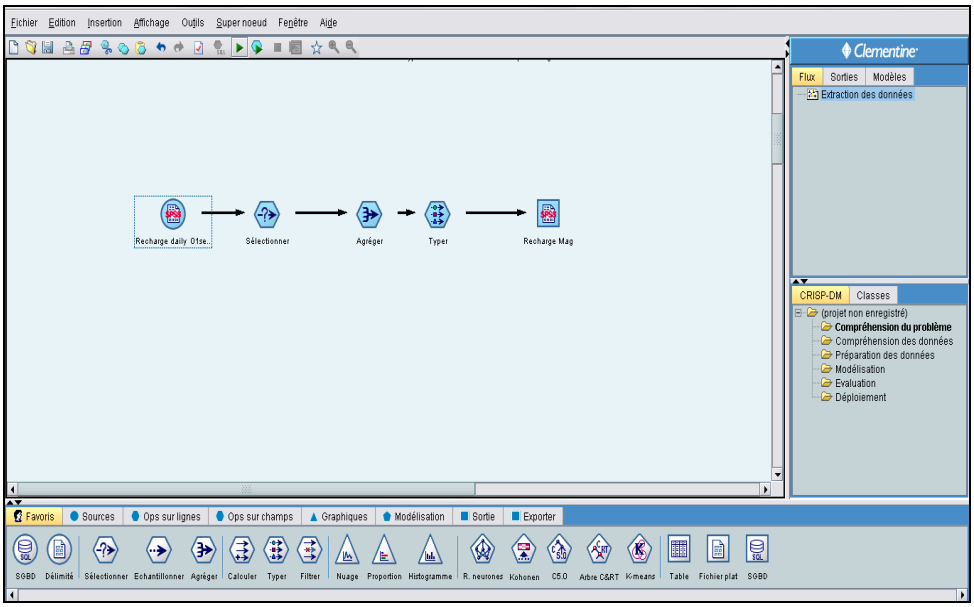
On peut visualiser le contenu de chaque table grâce à l'architecture datamart marketing présentée dans la partie « Architecture datamart **figure 01** ».

Construction du flux :

1- On commence par la récupération de la table trafic.
2- On utilise un nœud « sélectionner » pour définir la période du trafic voulue (EX : de janvier à décembre).
3- On utilise un autre nœud « sélectionner » pour choisir les offres.
4- On utilise un nœud « agréger » pour définir les champs et leurs correspondances.
5- On récupère la table des offres (pour nommer les offres de la table de trafic car ils sont sous forme de codes)
6- On fusionne les deux table trafic et offres.
7- On récupère la table des services.
8- On fusionne le résultat de la fusion trafic, offre avec les services.
9- On récupère la table des destinations.
10- On fusionne le résultat de la fusion trafic, offre et services avec les destinations.
11- On utilise le nœud « Typer » pour donner un type pour chaque variable (Date, ensemble, discrète, sans type…..).
12- On exporte la base trafic.

Extraction de la base rechargement client :

C'est le même principe que le revenu mais avec moins d'étape car nous avons moins de variables sur la table rechargement.

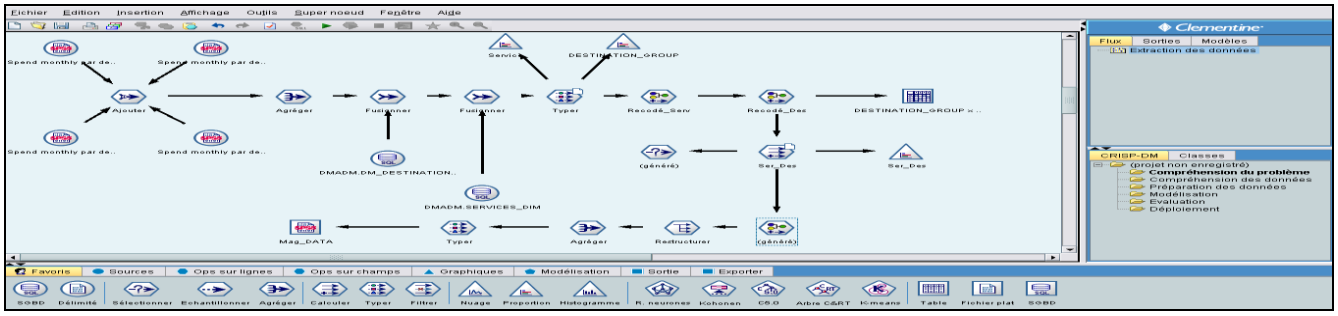


- **Flux 2 :** Extraction du rechargement –

Extraction de la base client avec traitement des données spend (préparations des variables pour la modélisation) :

Cette étape consiste à extraire plusieurs mois de Spend (4 mois de septembre 2010 à décembre 2010) et à la conception des variables qui seront utilisées lors de l'étape modélisation, le **Flux 3** traitera :

- Toutes les données aberrantes
- Toutes les variables d'entrées (le modèle retiendra les variables les plus significatives non corrélées entre elles)
- Le calcul des ratios intéressants (% out, % in, % On net, international, data… sur le total trafic, revenu…)
- L'assemblage de tous les mois de Spend (les bases sont extraites par mois et pour obtenir toute la période il faudra les additionner)



- **Flux 3 :** préparation des données Spend pour la modélisation –

V.4. Étape 4 : la modélisation

Avant de commencer la modélisation il faut définir c'est quoi le Churn.

V.4.1. Définition du Churn

Le mot « churn », est né de la contraction en anglais des mots « change » et « turn », décrit le phénomène de perte d'un client. Il est mesuré par le taux de churn et qui constitue un indicateur important pour les organisations. Ce taux de churn représente le pourcentage de clients perdus sur une période donnée par rapport au nombre total de clients au début de cette période.

Il existe plusieurs types de churn. Nous pouvons distinguer entre le churn interne, externe, volontaire ou involontaire en faisant l'analyse selon la destination du client. Le churn est qualifié d'interne lorsque le client change de produit ou d'offre recouvrant aussi ses besoins tout en restant au sein de la même organisation. On parle de churn externe ou switch lorsqu'il la quitte pour partir chez le concurrent. Le churn est dit volontaire lorsque l'individu quitte délibérément l'organisation, soit pour aller chez le concurrent, soit parce qu'il n'utilise plus le produit ou le service. Le churn involontaire est employé lorsque le client quitte le produit ou le service involontairement, par exemple en cas de décès ou de résiliation du contrat pour impayés. D'autres adjectifs peuvent qualifier le churn. Chez les fournisseurs de services par exemple, les clients qui quittent l'organisation peuvent être répertoriés en deux catégories : ceux qui décident de ne pas renouveler leur contrat à la fin de celui-ci, ce qui crée un churn commercial, et ceux qui arrêtent de payer leur contrat parce qu'ils ne peuvent plus supporter les dépenses, ce qui cause un churn financier. D'autres termes sont employés pour désigner le churn, tels que l'attrition ou la défection.

Pour le modèle du churn, nous avons défini comme **« churneur »** les clients qui n'ont pas fait de transactions payantes durant les trois derniers mois (appel, SMS… out ou in).

L'objectif est de détecter les churneurs assez tôt avant qu'ils churnent, et pas trop tôt pour ne pas avoir des clients à faible risque de churn.

V.4.2. Principe de la modélisation

L'objectif de cette étude est de détecter les clients qui vont churner bien avant qu'ils churnent, pour cela on procède comme suite :

On extrait les clients qui ont churné au mois de **février (le choix du mois de churn est arbitraire)**et on va suivre leurs comportements en termes de Spend (Revenu), rechargement

et d'informations relatives aux clients (date du First call (jour et année d'activation), date du last call (date du dernier appel), passivité (durée de vie d'un client)……..).

La période observée pour le Spend et le rechargement est du **01 septembre 2010** au **31 décembre 2010**, cette période choisie nous donne la possibilité de prévoir les clients qui partent deux mois avant que cela arrive.

Cette durée sera utilisée pour mener des compagnes et actions marketing souvent appelées compagnes de rétention et fidélisation.

Une fois l'extraction des churneurs du mois de février faite, on rajoute leurs comportements en termes de Spend et rechargement et d'autres indicateurs, on passe à l'étape de la modélisation ou on appliquera « un réseau de neurones »,on aura au final une base scorée ou chaque client aura une probabilité de churn.

Grace à cette modélisation on peut prédire nos churneurs à partir de n'importe quel mois souhaité.

Schéma décrivant la démarche :

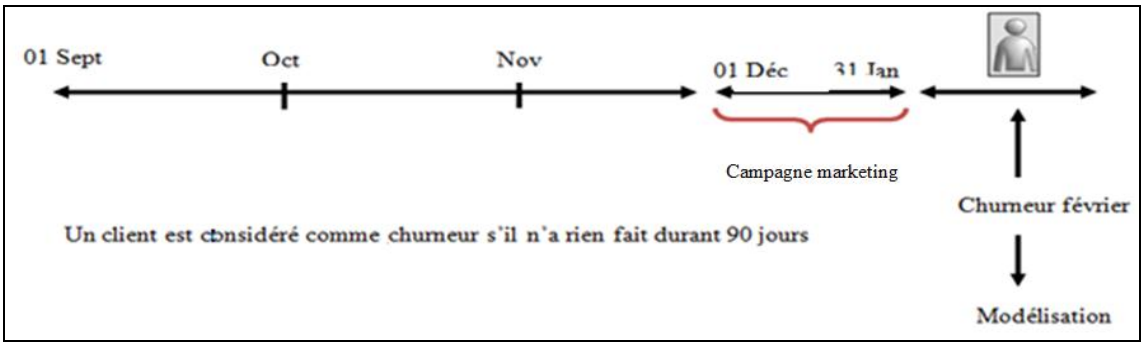


Figure 03 : Modélisation des churneur du mois de février

Schéma décrivant la démarche une fois le modèle obtenu :

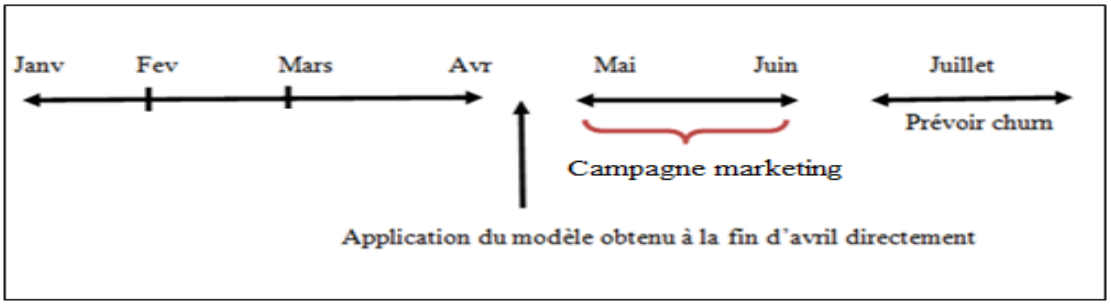


Figure 04 : Application du modèle obtenu pour prédire les churneurs d'avril

V.4.3. 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 chaque trois mois 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 du churn 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 un autre modélisation.

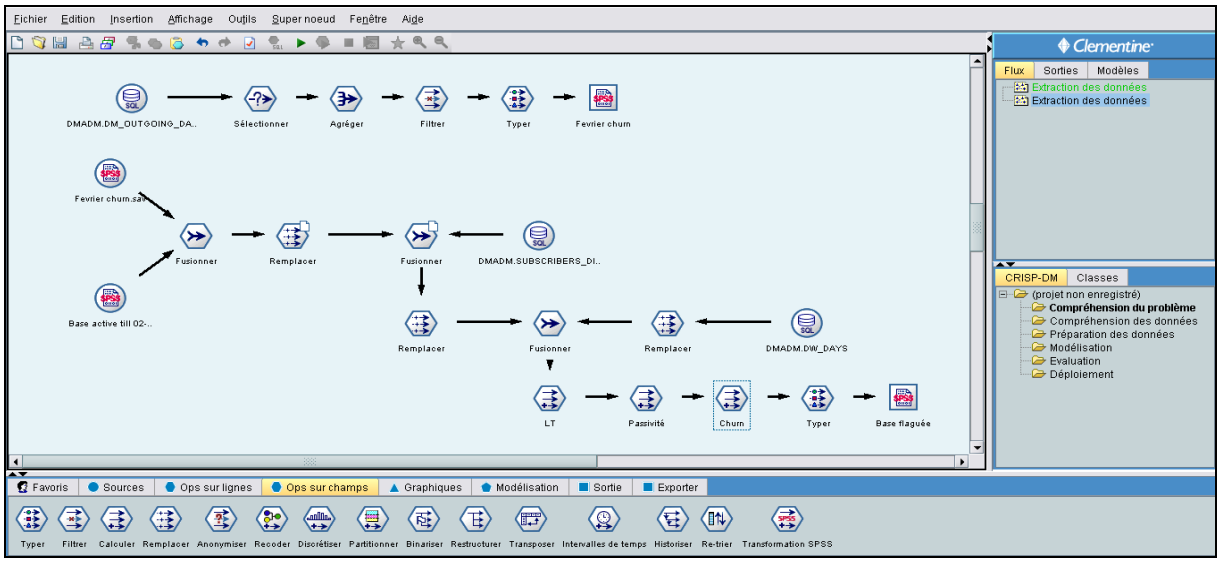
V.4.4. Modélisation de la base WTA

Après avoir fait l'extraction du Spend (Flux 1), l'extraction du rechargement (Flux 2), et la préparation des données spend (Flux 3), on passe à présent à la détection des churneurs du mois de février.

- **Churneurs du mois de février :**

Cette étape consiste à extraire les clients churneurs au mois de février pour cela il nous faut :

- Des clients ayant fait leur dernier appel au mois de février.
- Calculer la durée vie des clients (date du dernier appel- la date du premier appel).
- Calculer la passivité des clients, c'est-à-dire le nombre de jour où ils n'ont pas fait de trafic ou bien de transaction payante.
- Sélectionner ceux qui ont plus de 90 jours (nos churneurs).

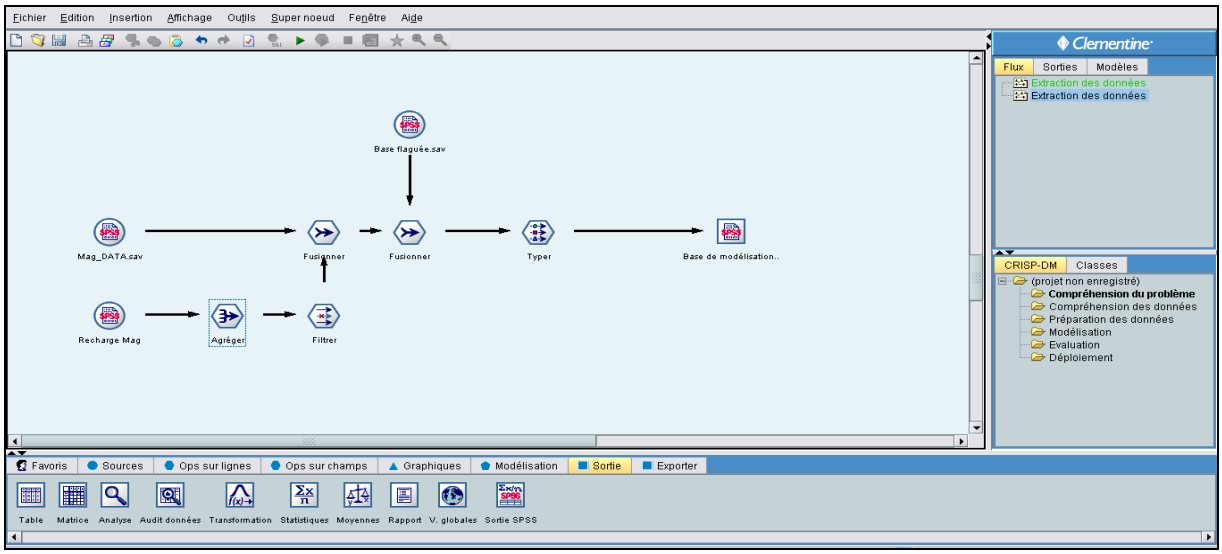


Flux 4 : Détection des churneurs février

- **Construction de la base finale**

Une fois nos extractions faites, il convient à présent de construire la base finale qui sera déployée pour la modélisation.

Dans cette phase on va joindre toutes les extractions réalisées jusqu'à maintenant pour former une seule et unique base.



- **Flux 5 :** Construction de la base de modélisation -

- **Modélisation de la base WTA par les réseaux de neurones**

La modélisation par les réseaux de neurones est une approche très puissante lorsqu'il s'agit d'un volume important de données, elle permet de mieux modéliser les phénomènes et de gérer l'interaction entre un très grand nombre de variables en un temps appréciable (tout dépend du volume des données).

Cette étape commence par mentionner la variable cible à prédire dans notre cas c'est le **« churn** », après il faut générer les échantillons (Apprentissage, test).

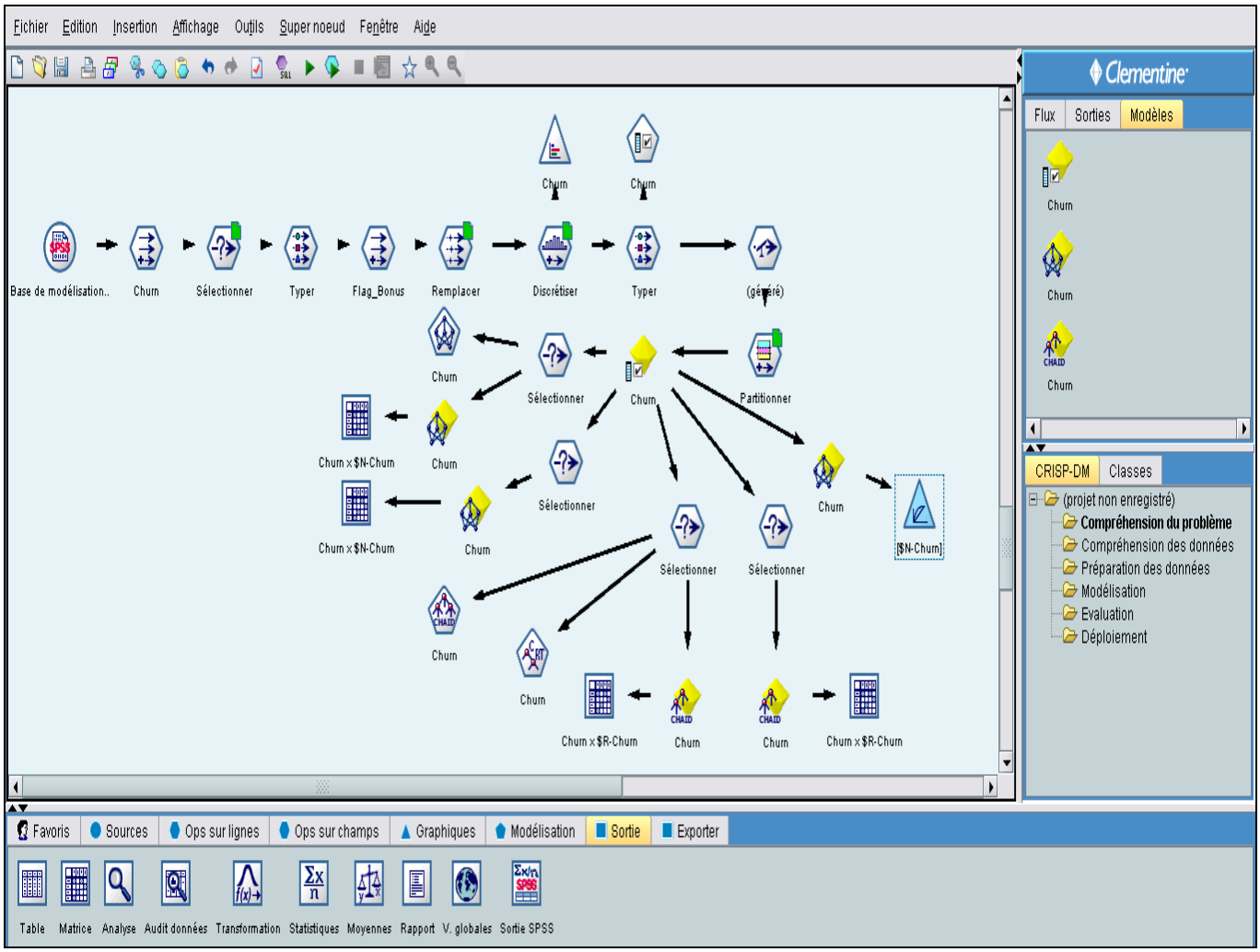
- **Choix des échantillons (Apprentissage, test) :**

- Échantillon d'apprentissage : la modélisation est réalisée sur un échantillon d'apprentissage, ensuite les modèles concluants sont testés sur un 2ème échantillon test représentatif pour vérifier la stabilité des modèles.

Le choix de l'échantillon est fait par le logiciel on doit juste lui indiquer les pourcentages pour chaque échantillon et il va construire un échantillon représentatif et équilibré qui reflète la base globale c'est-à-dire si 20% churnent sur l'échantillon on aura le même taux si on applique le modèle sur toute de la base. Dans notre cas l'échantillon d'apprentissage est à 70% et l'échantillon test est à 30% (standard international).

Après cela on commence la modélisation sur l'échantillon d'apprentissage et on valide par l'échantillon test.

Étape de la modélisation :



Flux 6 : Modélisation par les réseaux de neurones

En premier lieu on récupère la base de modélisation déjà construite par le **Flux 5** et on calcule un Flag (variable booléenne) qui nous donnera les churneurs versus les Actifs qu'on nommera « **Churn »,** cette variable nous permet de partitionner la base de données et obtenir les échantillons test et apprentissage.

En deuxième lieu on choisira la variable à prédire parmi les variables qui participent à la modélisation et pour cela nous avons besoin d'un nœud **« TYPER »** figure 5 ci-dessous.

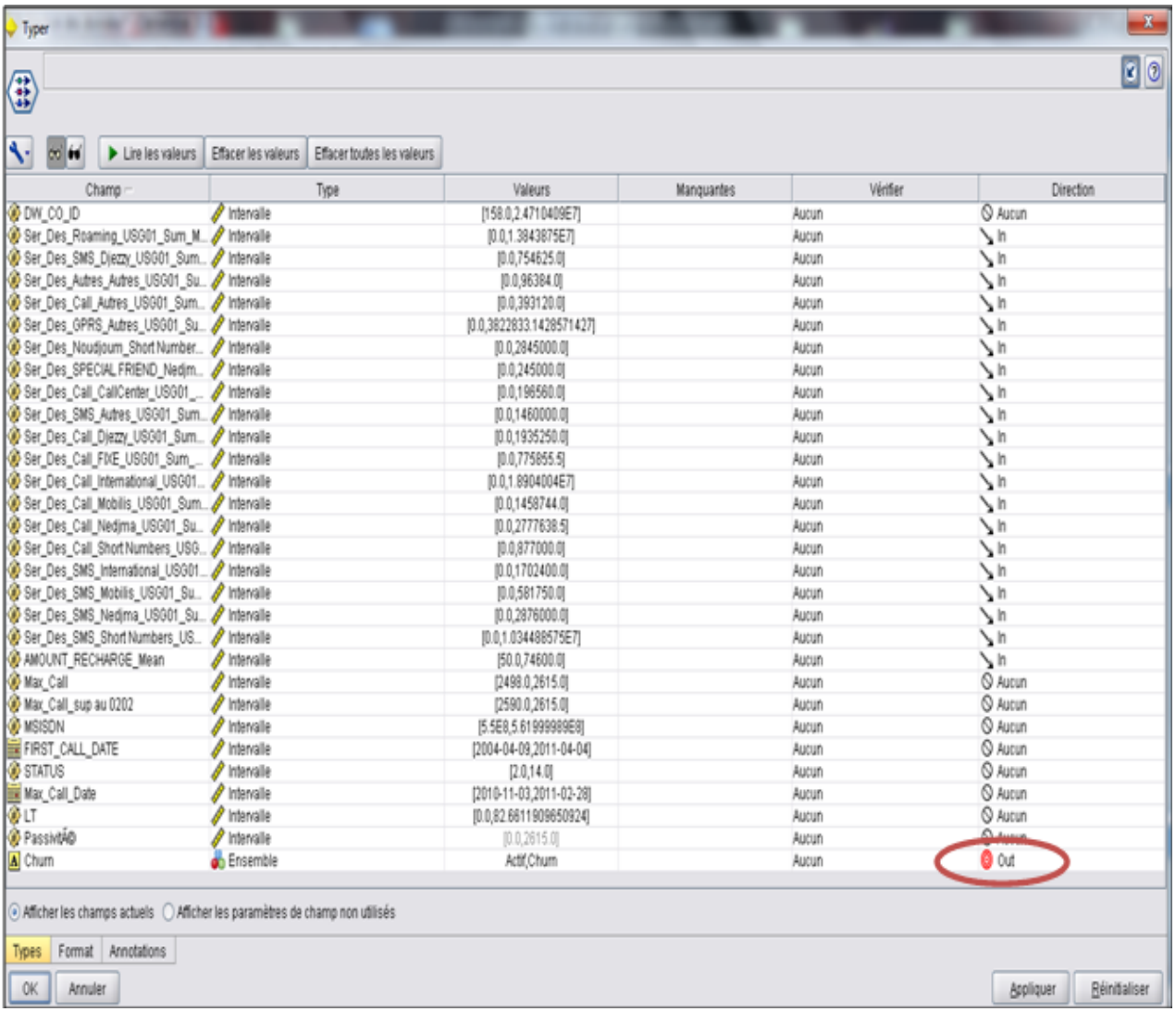


Figure 05 : Ciblage de la variable churn

Sur l'ensemble des variables à modéliser, on doit affecter une direction à chacune d'entre elles, **la figure 05** nous donne une idée sur les directions possibles.

IN : variables qui participent à la modélisation.

AUCUN : variables qui ne participent pas à la modélisation.

OUT : variables cible ou à prédir.

Après avoir calcuer la variable Churn et la mettre comme cible à prédir il convient à présent de créer les échantillons Test et Apprentissage qui seront utilisés pour le choix du modèle à retenir.

Cette étape sera gérée par l'ensemble des nœuds suivants :

- **Proportion :** pour sélectionner un échantillon représentatif et équilibré de base de données globale.
- **Sélection de fonction :** pour donner plus de robustesse à l'échantillon choisi, ce noeud permet d'évaluer les sorties du noeud proportion et de faire un audit complet sur le poids des variables utilisées.
- **Partitionner :** comme son nom l'indique ce noeud va servir à obtenir les deux échantillons Test et Apprentissage avec les proportions souhaitées (l'échantillon d'apprentissage est à 70% et l'échantillon test est à 30%).

Après avoir fait le choix de nos échantillon on passe à une autre étape très importante à savoir la selection de la meilleure méthode de modélisation pour cela nous avons testé un ensemble très varié : les réseaux de neurones, réseaux de kohonen, arbre de décisions chaid et C5, K-means).

Au final nous avons opté pour les réseaux de neurones qui donnent les meilleurs résultats pour la variable Churn.

NB : cette étape de la modélisation est nécessaire pour justifier que le choix des réseaux de neurones, qui n'était pas arbitraire mais après avoir testé plusieurs méthodes de data mining et statistique.

V.4.5. Résultat de la modélisation par les réseaux de neurones

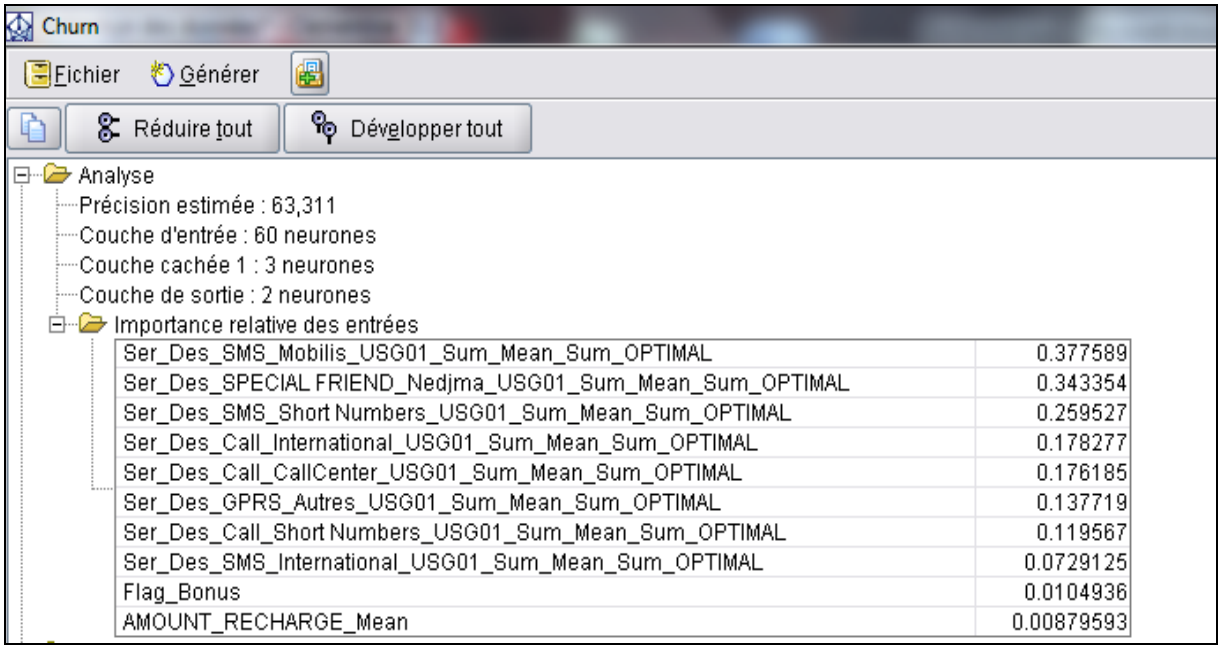
Le nombre de couches d'entrée est : 60.

Le nombre de couches cachées est : 3.

Le nombre de couches de sortie est : 2.

La précision globale du modèle est de **63.311 %** (entre actif et les churneurs)

On peut voir sur la « **figure 06**», l'importance des variables d'entrée qui ont influencé le modèle.



Ser_Des_SMS_Mobilis_USG01_Sum_Mean_Sum_OPTIMAL	0.377589
Ser_Des_SPECIAL FRIEND_Nedjma_USG01_Sum_Mean_Sum_OPTIMAL	0.343354
Ser_Des_SMS_Short Numbers_USG01_Sum_Mean_Sum_OPTIMAL	0.259527
Ser_Des_Call_International_USG01_Sum_Mean_Sum_OPTIMAL	0.178277
Ser_Des_Call_CallCenter_USG01_Sum_Mean_Sum_OPTIMAL	0.176185
Ser_Des_GPRS_Autres_USG01_Sum_Mean_Sum_OPTIMAL	0.137719
Ser_Des_Call_Short Numbers_USG01_Sum_Mean_Sum_OPTIMAL	0.119567
Ser_Des_SMS_International_USG01_Sum_Mean_Sum_OPTIMAL	0.0729125
Flag_Bonus	0.0104936
AMOUNT_RECHARGE_Mean	0.00879593

Figure 06 : Résultat du réseau

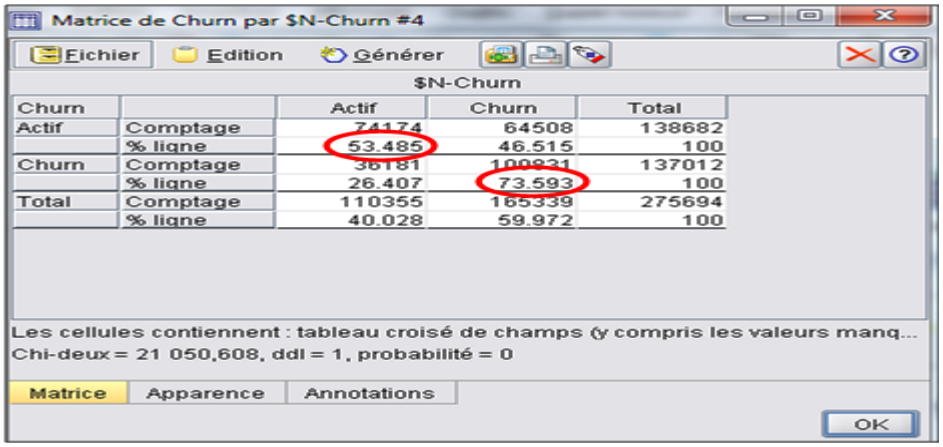
V.4.6. Évaluation du modèle

Matrice d'affectation :

La matrice d'affectation nous permet d'évaluer notre modèle est-il bon ou mauvais.

Matrice d'affectation« échantillon d'apprentissage »:

D'après la **figure 7**, on peut constater que notre réseau prédit les churneurs avec un taux d'environ **74%** contre un taux de **53%** pour les actifs, ce qui nous permis d'obtenir un modèle significatif.

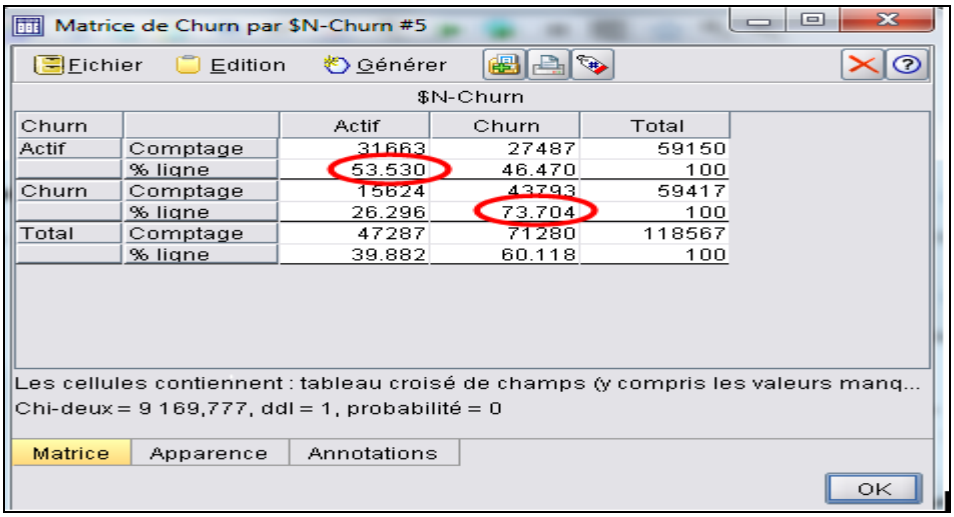


Churn		Actif	Churn	Total
Actif	Comptage	74174	64508	138682
	% ligne	53.485	46.515	100
Churn	Comptage	36181	100831	137012
	% ligne	26.407	73.593	100
Total	Comptage	110355	165339	275694
	% ligne	40.028	59.972	100

Figure 07: matrice d'affectation (Apprentissage)

Matrice d'affectation « échantillon test »:

La **figure 08**, permet de constater que le réseau prédit les churneurs avec un taux d'environ **74%** contre un taux de **53%** pour les actifs, notre modèle répond aux objectifs et les résultats obtenus en comparant avec l'échantillon test, de là on passe à l'évaluation globale de la modélisation.



Churn		Actif	Churn	Total
Actif	Comptage	31663	27487	59150
	% ligne	53.530	46.470	100
Churn	Comptage	15624	43793	59417
	% ligne	26.296	73.704	100
Total	Comptage	47287	71280	118567
	% ligne	39.882	60.118	100

Figure 08 : matrice d'affectation (Test)

Diagramme de gain :

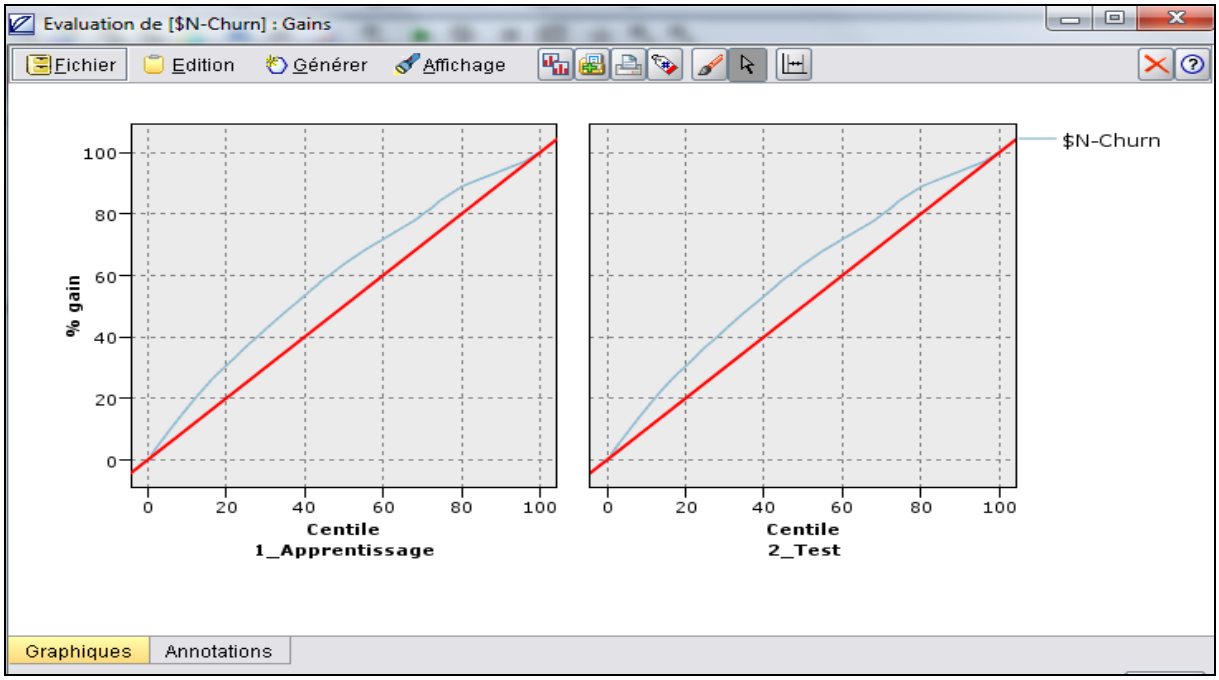


Figure 09 : Diagrammes de gain

Ce graphe nous montre qu'il est possible de prédire par exemple **55%** des "Churneurs" en examinant seulement **40%** de cas.
La courbe en bleu montre notre base modélisée et scorée par le réseau de neurones et la courbe en rouge montre notre base sans score (non modélisée).

Ce digramme de gain nous montre l'efficacité du modèle car il nous permet de prédire au mieux nos clients churneurs (**15%** de plus qu'une base non modéliser ou bien non scorée).

V.5. Étape 5 : l'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 :

V.5.1. En termes de data mining

- La réduction du nombre des clients churneurs.
- Augmenter la durée vie d'un client dans le système.
- Plus de visibilité sur les clients à risque (cela permet de mener à bien les actions et campagnes marketing).
- Réduire les risques d'erreur en travaillant sur une partie de la base (mini population).

V.5.2. En termes de business

- Maintenir stable le nombre d'abonnés.
- Augmenter le revenu des clients.
- Réduire le churn.
- Réduction des coûts de compagne d'up selling et cross selling.
- Booster le rechargement des clients.
- Augmenter l'ARPU (average revenu per user, revenu moyen par client).
- Se servir des résultats obtenus pour mieux gérer les actions avenir.

NB : en télécommunication, si un opérateur peut maintenir la stabilité du nombre d'abonnés toute en augmentant le revenu avec une réduction du churn, il est considéré comme efficient et très compétitif.

V.6. Étape 6 : le déploiement final

V.6.1. Objectifs

- Déployer le modèle sur la base.
- Mener une opération de rétention auprès de nos clients afin de :

- Apprécier la prédiction du modèle de churn sur les clients

  - ✓ Comparer les résultats entre population totale et population ciblée : profil, satisfaction, intention de churn…
  - ✓ Vérifier la pertinence des variables définies dans les modèles auprès des clients nouveaux

- Mesurer l'efficacité des actions de rétention suivant le scoring.

  - ✓ à quel moment doit-on agir?
  - ✓ Retour des actions.

V.6.2. Déploiement des règles du modèle choisi (cas WTA)

Pour le déploiement des réglés nous avons deux façon de le faire :

- **Déploiement via le datawarehouse**

Les scores et règles générées par le modèle issu du processus data mining seront fournies par l'équipe de spécialistes et consultants en marketing pour une implémentation côté technique plus exactement au datawarehouse, des scripts et requêtes SQL vont être développés et intégrés et ensuite ils seront publiés sur le datamart et nous aurons des score et des règles pour nos clients.

Avantage du passage par le datawarehouse

- Plus de rapidité lors du scoring de nouveaux clients.
- Données scorées exploitables directement sur le datamart.
- Mobiliser moins d'équipe au niveau marketing.
- ……

Inconvénient du passage par le datawarehouse

- L'implémentation de la base n'est pas toujours possibles (certaines règles ne peuvent pas être crées).
- Alourdir le système (tache supplémentaire à réaliser).
- La modification des variables reçues n'est pas toujours possibles si elles coïncident avec d'autres requêtes datawarehouse).
- Possibilité de perte d'information ou mal compréhension entre les deux départements marketing et technologique (risque moins présent mais potentiel).
- ……

- **Déploiement via le marketing (équipe data mining)**

Avantage du passage par le Marketing :

- Possibilité de modification.
- Monitoring des scores générés.
- Mener des actions plus rapidement.
- Faire face aux changements de comportement des clients rapidement.
- …….

Inconvénient du passage par le marketing :

- Travail coûteux en termes de temps.
- Absence d'infrastructure de déploiement (serveur, Switch, outils de transformation des données, logiciels, …).
- Immobiliser toute une équipe pour la gestion des bases scorées.
- ……

V.6.3. Exemple de règles à implémenter

Les règles affichées dans la figure ci-dessous sont issues de notre modèle vue auparavant dans la partie modélisation.

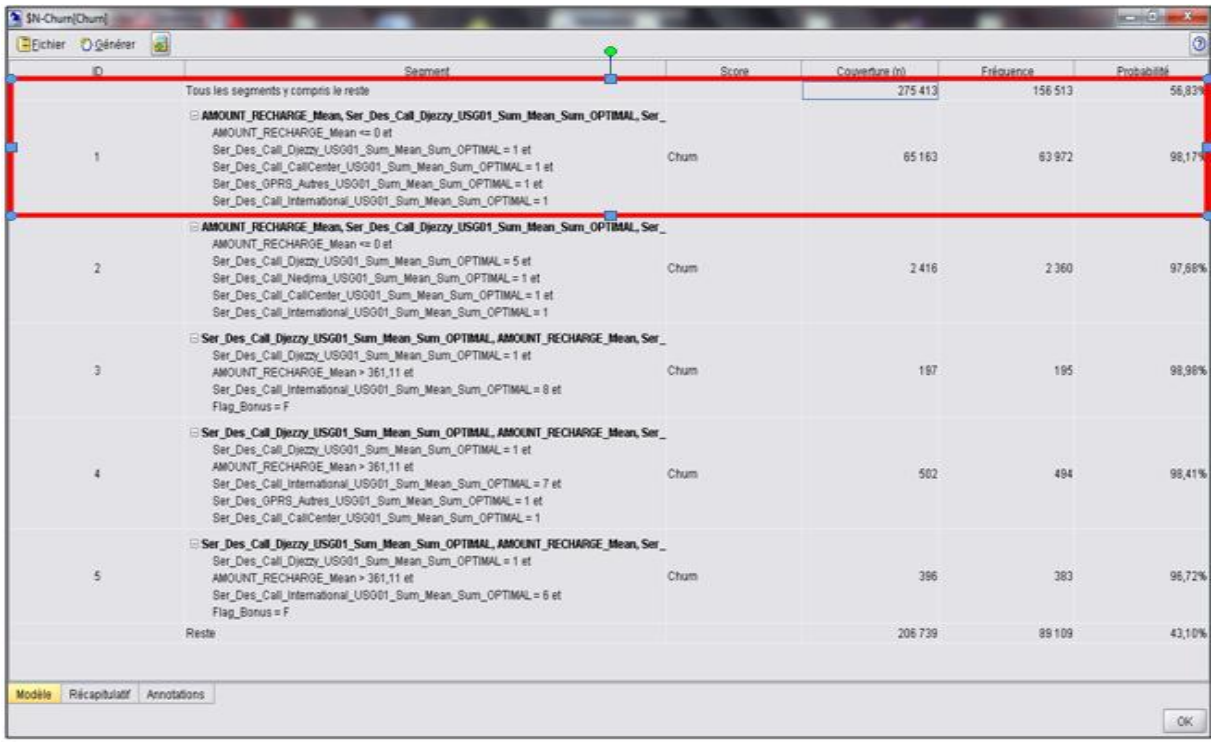


Figure 10 : règles et score du modèle de réseau neurones –

Zoom sur la partie en rouge de la figure ci-dessous :

Cette partie nous montre l'une des règles à adapter avec un score sur les clients churneurs.

ID	Segment	Score	Couverture (n)	Fréquence	Probabilité
1	Ser_Des_Call_Djezzy_USG01_Sum_Mean_Sum_OPTIMAL, A Ser_Des_Call_Djezzy_USG01_Sum_Mean_Sum_OPTIMAL = AMOUNT_RECHARGE_Mean <= 0 et Ser_Des_Call_CallCenter_USG01_Sum_Mean_Sum_OPTI Ser_Des_GPRS_Autres_USG01_Sum_Mean_Sum_OPTIMA Ser_Des_Call_International_USG01_Sum_Mean_Sum_OPT	Churn	65 163	63 949	98,14%

Dans cette règle on peut tirer les informations suivantes :

- Combien de client sont touchées par cette règle.
- Le score du churn (98%).
- Le comportement des clients de cette catégorie durant les 6 dernier mois :

D'après l'exemple ci-dessus on peut dire que ces clients ont effectué un appel vers la concurrence, ils n'ont pas effectué de rechargement, ils ont appelé le call center une fois, aucun appel vers l'international, et enfin aucun trafic GPRS ainsi le score de cette catégorie est donné comme client à haut risque de partir avec une probabilité de 98%.

Exemple de campagne de rétention client

Action CRM Opérationnel : Call Center

Une fois le modèle choisi, la base scorée et les règles implémentées, on peut lancer des campagnes ou actions de rétention, cette phase consiste à :

- Sélectionner tous les clients churneurs.
- Établir des scripts pour chaque catégorie de churneur (faible, moyen, haut risque).

Le script contient souvent :

- La manière d'agir avec un churneur (priorité de passage au call center, meilleurs conseillés clients,..).
- Obtenir le maximum d'information sur le client (est-ce qu'il connait bien l'offre sur laquelle il est, est-ce qu'il souhaite changer, est-ce qu'il connait ses avantages,….).
- …..

- Fournir les scripts à l'équipe CRM.
- Implémentation les scripts sur l'ERP.
- Formation des conseillers clients.
- Retour d'information (réaction d'un client scoré, son comportement après l'entretien avec le conseiller client, …).

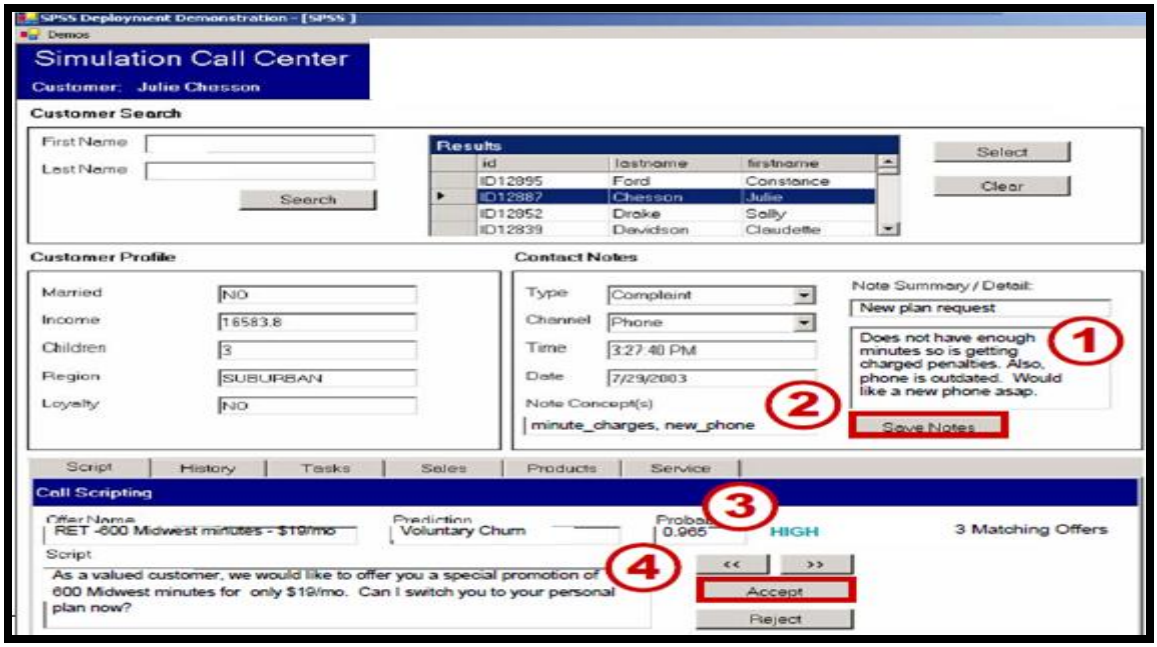


Figure 11: exemple de rétention client via le call center (***SPSS Deploment demonstration)***

Ce schéma nous montre l'affichage qu'aura un conseiller client après le scoring de la base de données, on peut apercevoir qu'il y a un script spécifique à chaque client selon son taux de Churn, s'il est très élevé le conseiller doit le guider dans le choix du meilleur service ou offre avec beaucoup de délicatesse et clarté. Ainsi le client recevra un traitement qui lui convient et qui correspond à ses besoins.

Déroulement de l'opération (voir figure 12):

1. L'opérateur saisie la réclamation du client.
2. Voit les informations disponibles sur le client et essaye de compléter et d'ajouter le maximum sans qu'il se rende compte.
3. La probabilité de churn (il est à un risque 98% de quitter la compagnie).
4. L'opérateur propose le script généré dynamiquement par le système.

Conclusion générale et perspectives

Développer une stratégie CRM est devenu un objectif majeur des entreprises actuelles. Or, la réalité nous a montré que les projets CRM sont des projets risqués et très couteux à mettre en œuvre. Une des conditions nécessaires pour réussir l'implémentation d'une stratégie CRM est la disponibilité de données « *Clients* » fiables, pérennes, précises et répondant aux besoins des décideurs permettant ainsi une gestion efficace de la relation clients.

Pour cette condition, les applications CRM doivent être supportées par des datawarehouses et processus de data mining conçus autour d'objectifs CRM qui ne se limitent pas à recueillir que les données comportementales. Cette vison datawarehouse, 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 la gestion de relation client son évolution et son importance vis-à-vis des entreprises, nous avons montré aussi le cheminement d'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 portera sur la modélisation du churn (attrition), dans ce dernier on peut apprécier la résolution d'un problème réel qui préoccupe la majorité des décideurs dans une entreprise (particulièrement en Télécom), après une bonne étude de la problématique et la mise en œuvre des moyens pour la résoudre (matériels et techniques : le datawarehouse, 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 churneurs potentiels pour les mois avenir et il va contribuer à renforcer les actions et campagnes de fidélisation avec un bon ciblage (choisir les vrai clients et non aléatoirement).

Ainsi et avec le déploiement de ce modèle les 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, on envisage d'intégrer la partie texte mining et web mining pour enrichir la partie data mining déjà réalisée, le principe est d'établir une nouvelle modélisation et de scorer à nouveaux la base client mais en utilisant cette fois les modules texte et web mining.

Le textmining :

Le textmining est l'ensemble des techniques et méthodes destinées au traitement automatique des données textuelles en langage naturel, disponible sous forme informatique, il tient beaucoup de lexicométrie ou « statistique lexicale», dont il est une extension par des outils avancés de statistique multidimensionnelle.

Schématiquement on peut énoncer :

« textmining» = lexicométrie + data mining.

Le web mining:

Le web mining est l'application du data mining au données issues des serveurs internet, des utilisateurs des sites web, des entreprises et organisations. Il permet de fournir des analyses sur le comportement des internautes et peut être éventuellement relié à des analyses portant sur d'autres sources de données.

Bibliographie

[1] BERRY. M. **"Data mining techniques: for marketing, sales, and customer relationship management",** Wiley Publishing, Inc., Indianapolis, Indiana. 2004.

[2] BERADA. I, "**SPSS clémentine"**, SPSS Maghreb. 2010.

[3] BACCINI. A, BESSE. P, **"Data mining et l'exploration Statistique "**, Université Paul Sabatier Toulouse, publication du laboratoire de statistique et probabilité, version septembre 2005.

[4] GRISLIN. E, **"Systèmes d'information décisionnels (Datawarehouse / Data Mining)"**, Université de Valenciennes. 2006.

[5] GUILLERON. R, **"Découverte de connaissances à partir des données"**, Université de Lille 3, novembre 2000.

[6] GOUARNE. JM, **"Le Projet Décisionnel Enjeux, Modèles, Architectures du Data Warehouse"**, publié en 1997 par les Éditions Eyrolles.

[7] IAN. H, "Data **mining: practical machine learning tools and techniques"**, Morgan Kaufmann Publishers. 2005.

[8] KUSIAK. A, AGARD. B, **"exploration des bases de données industrielles à l'aide du data mining– perspectives"**, Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal. 2005.

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

[10] LAROSE. D, **"Discovering knowledge in data: an introduction to data mining"**, Published by John Wiley & Sons, Inc., Hoboken, New Jersey. 2005.

[11] LEFEBURE. R, VENTURI. G, **"La gestion de la relation client".** Eyrolles. 2005.

[12] LAROCQUE. D, **"Techniques quantitatives en marketing",** HEC Montréal. 2008.

[13] LEBART. L, **"Statistiques exploratoire multidimensionnelle : Visualisations et inférences en fouille de données"**, éditions DUNOD. 2006.

[14] MATTISON. R, **"Data Warehousing and Data Mining for Telecommunications"**, Library of Congress Cataloging-in-Publication Data. 2007.

[15] MEIER. A, **"Introduction pratique aux bases de données relationnelles"**, Springer-Verlag France. 2006.

[16] MEYLAN. E, **"Base de données et informatique décisionnelle"**, university of applied sciences western switzerland, school of business administration Neuchâtel business data processing. 12/11/2003

[17] MEIER. A, **"Le CRM analytique : Les outils d'analyse OLAP et le Data Mining"**, Faculté des Sciences économiques et sociales Université de Fribourg, le 26 avril 2008.

[18] NISBET, R. ELDER, J. MINER, G, **"Handbook of statistical analysis and data mining applications"**, Elsevier Inc. 2009.

[19] PREUX. P, **"Fouille de données"**, Université de Lille 3, 9 octobre 2008.

[20] PARIZEAU. M, **"Réseaux de neurones"**, Université de Laval. 2004.

[21] PEPPERS. D, ROGERS. M, **"Le one to one en pratique",** éditions d'organisation. 1999.

[22] RAPHALEN. M, **"Systèmes d'information décisionnels"**, Université de Bretagne Sud. octobre 2002.

[23] PINE. J**, "The New Frontier in Business Competition".** 1999

[24] SMITH. S, WHEELER. J, "**Managing the customer experience"**, published in Great Britain in 2002. www.business-minds.com

[25] SCHARFF. C, **"Entrepôts de données"**, 2004.

[26] SPSS groupe**, " Clémentine 11.1 "**, Crisphelp.2007.

[27] TUFFERY. S, "**Data Mining et statistique décisionnelle"**, éditions TECH NIP. 2010.

[28] TUFFERY. S, **"Data Mining et scoring : Bases de données et gestion de la relation client"**, éditions DUNOD. 2003.

[29] TASLIMANKA. M, **"Initiation au décisionnel (Business Intelligence, DataWarehouse, OLAP)"**, Date de publication : 20/10/2007.

[30] UMMAN TUGBA. SG, **"Customer churn analysis in telecommunication sector"**, Department of Quantitative Methods, School of Business Administration Istanbul University, Istanbul, Turkey. 2010.

[31] VANGENOT. C, **"Le Datawarehouse"**. École polytechnique Fédérale de Lausanne. 2005.

[32] WALKENBACH. J, **"Excel Charts"**, Wiley Publishing, Inc., Indianapolis, Indiana. 2003.

[33] WISRA. P, **"Réseaux de neurones artificiels : architecture et application"**, Université de Haute Alsace, laboratoire MPIS (modélisation, intelligence, processus, systèmes), Avril 2009.