

Le DataMining, qu'est-ce que c'est et comment l'appréhender ?

Extrait d'une conversation téléphonique (réelle) :

« Allô, monsieur Decourt ? Je travaille pour X¹ et nous désirons rencontrer tous les habitants de votre rue (sic). Quand pouvons-nous vous rencontrer pour vous parler de nos offres en matière d'assurances ? »

Cette conversation date de juillet 2000. Elle foule aux pieds tous les concepts véhiculés par le DataMining, ce qui montre combien cette discipline encore jeune a du mal à se faire une place dans les services marketing actuels.

Le plan suivi ici est de prospecter de nouveaux clients, choisis presque aléatoirement (une rue dans Paris !); quel taux de réponse espérer de ce type de campagne ? Il risque d'être dramatiquement bas, ce qui met en doute la rentabilité de cette action. Le DataMining a pour but d'éviter ce genre de prospections dont on a l'impression qu'elles se font au hasard. Rencontrer tous les habitants d'une rue donnée semble moins prometteur que de rencontrer toutes les personnes dont le score d'appétence à un certain produit dépasse les 70 %. La « fouille des données » met au point des typologies descriptives et des modèles afin de faciliter la prise de décision. Les choix sont alors faits en fonction des résultats du score et de la composition de certaines « niches » typologiques, critères statistiques (donc objectifs) et non plus, comme ce fut longtemps le cas, sur le « flair » et l'habitude d'un vieux routier du marketing.

Le but de notre propos est de broser par touches successives le portrait d'une discipline nouvelle. Il s'agit ainsi les avis couramment exprimés par des utilisateurs et des « fouilleurs de données », ainsi que les avis de la littérature.

2.1. Origine et émergence du concept de DataMining

Historiquement, le DataMining est très jeune. Le concept apparaît en 1989 sous un premier nom de KDD (Knowledge Discovery in Databases, en français ECD pour Extraction de Connaissances à partir des Données), avant qu'en 1991 apparaisse pour la première fois le terme de DataMining ou « minage / fouille des données ».

Comme l'expliquent fort bien Michael Berry et Gordon Linoff, ce concept – tel qu'on l'entend aujourd'hui, et surtout tel qu'on l'applique dans les services marketing – est étroitement lié au concept du « *one-to-one relationship* ». C'est à dire la personnalisation des rapports entre l'entreprise et sa clientèle.

La « vision client »

En effet, pendant longtemps, les sociétés se sont focalisées sur leurs produits, se souciant d'offrir une large gamme – plus vaste que celle de son concurrent –, de s'engouffrer dans des niches prometteuses, sans toujours savoir de façon quantifiable si tel ou tel produit était

¹ Lire : une grande société d'assurances

réellement attractif. La découverte de niches tenait plus du miracle – organisé autour d'une connaissance du marché issue de l'expérience, et difficilement reproductible à la demande – que de la recherche organisée et rationnelle.

L'exemple cité ci-dessus reflète parfaitement la politique « orientée produit ». On sélectionne des personnes *au hasard*, et on essaye de leur vendre telle ou telle marchandise de sa gamme. On proposera sans doute toute la gamme au prospect, dans l'espoir qu'il trouvera une solution qui lui convient.

L'approche client découle de la démarche inverse. On choisit soigneusement les prospects en fonction de leur probabilité de vouloir tel ou tel produit. On cherche quel item de la gamme lui convient le mieux, au lieu que ce soit à lui de s'adapter à l'offre de l'entreprise.

Il est évident que la seconde démarche sera la plus appréciée des clients. A notre époque de consommation de masse, se sentir considéré en tant qu'individu particulier et non comme un élément interchangeable d'une population (la société d'assurances se serait-elle souciée de moi si j'avais habité 300 mètres plus loin ?) est valorisant, donc un bon point, marqué d'emblée, dans un souci de vente.

La « vision client » impose un bon nombre de changements dans l'entreprise. En premier lieu, il faut réorganiser les données (traditionnellement, la vision produit entraînait – schématiquement et de façon caricaturale – une base de données, ou au moins une table par produit possédée par le client). Mais il fallait surtout s'intéresser à des outils de description des clients (dont on ne savait rien ou presque, si ce n'est un amas de données que l'on n'avait jamais utilisées) et d'aide à la décision.

Le marketing décisionnel est né ainsi, trouvant un allié dans le DataMining, concept et outil regroupant un certain nombre de techniques statistiques éprouvées (bien que diversement répandues : s'y côtoient arbres de segmentation et régressions, réseaux neuronaux et statistiques descriptives, algorithmes génétiques et tests du khi-2). Le DataMining soutient ce souci de la personnalisation des rapports vendeur/client, en lui donnant les moyens statistiques et chiffrés de se justifier.

Les problématiques décisionnelles dont il est question ici sont prises le plus souvent par des décideurs qui ne sont pas en relation directement avec le service de marketing central. Il s'agit d'un responsable d'agence locale, du chef du démarchage téléphonique, du responsable d'une campagne de démarchage postal. Il fallait donc que les résultats produits soient lisibles et simples à appréhender pour un non-statisticien.

L'amalgame qui se fait alors conduit à une confusion malheureuse : pour utiliser ces techniques, il n'y aurait pas besoin d'un statisticien. Cet amalgame a été récupéré dans un premier temps par les constructeurs d'outils de DataMining avant qu'ils démentent. C'est le pas entre des résultats pouvant être *compris* par un « profane » (il faut quand même quelques notions de base pour lire les résultats d'une régression, fussent-ils présentés sous forme graphique) et des résultats pour être *obtenus* par un profane qui a été allègrement franchi. C'est aller, évidemment, un peu vite en besogne. On se rend compte qu'il n'en est rien, et que la meilleure solution pour mettre en œuvre le DataMining est d'associer un homme de terrain (qui apporte sa « connaissance métier ») et un statisticien (pour régler la mécanique de modélisation, valider les hypothèses, etc...).

Une méthode pour la mise en pratique

Aujourd'hui, le DataMining se présente donc comme un outil incontournable dans un service marketing, et au sein des processus décisionnels d'une entreprise. Il rassemble un faisceau de techniques statistiques qu'il convient d'utiliser au gré des problématiques – descriptives ou décisionnelles. Il s'assortit le plus souvent d'une méthode de travail pour ordonner au mieux hypothèses, modélisations et actions.

L'une d'elles, en particulier, mérite que l'on s'y attarde. Au milieu des méthodes spécifiques proposées par chaque constructeur pour utiliser au mieux son progiciel de DataMining (SEMMA pour SAS Enterprise Miner), CRISP-DM se démarque.

La méthode CRISP-DM² a été développée par un regroupement d'industriels européens, issus de divers marchés. On y compte SPSS (constructeur de logiciels statistiques), NCR (spécialiste en DataWarehouses), OHRA (société d'assurances néerlandaise) ou Daimler-Benz (groupe automobile et aéronautique allemand). Il s'agit de définir une approche rationnelle (et indépendante de l'outil informatique utilisé) du DataMining.

Cette approche se présente comme un cercle vertueux, qu'il convient de suivre. Elle prolonge en cela les réflexions que proposent à ce sujet tous ceux qui ont écrit sur le DataMining (Jambu, Berry et Linoff, Venturi et Lefébure, etc...).

Le cercle vertueux proposé se compose de 6 grandes étapes.

Il commence par la connaissance du contexte (compréhension métier) : connaître la signification de tel ou tel comportement, intérioriser la problématique posée pour être certain de bien la résoudre.

Vient ensuite la connaissance des données elle-même, toujours dans la logique de profiter des connaissances de l'entreprise. Que signifie telle ou telle grandeur, quelle échelle s'y rattache (ordre de grandeurs, évolution) ? La double flèche qui lie ces deux phases rappelle que les grandeurs proposées et les problèmes à résoudre sont indissociables, et que la connaissance parcellaire de l'une ou de l'autre ne peut permettre de réaliser de bons travaux de DataMining.

La mise en forme des données est l'étape suivante. Il s'agit de créer des indicateurs synthétiques sur la foi des deux étapes précédentes. Ce que l'on a appris sur les données et le problème à résoudre, il faut le traduire dans l'approche pratique des chiffres à exploiter.

Il est temps alors de « faire parler » les chiffres. Modèles et typologies sont alors mis en œuvre, afin de produire des réponses au(x) problème(s) posé(s). Cette phase est souvent décrite comme le cœur de la démarche de DataMining. C'est elle, en tout cas, qui a bénéficié d'une grande partie de l'intérêt porté à la discipline.

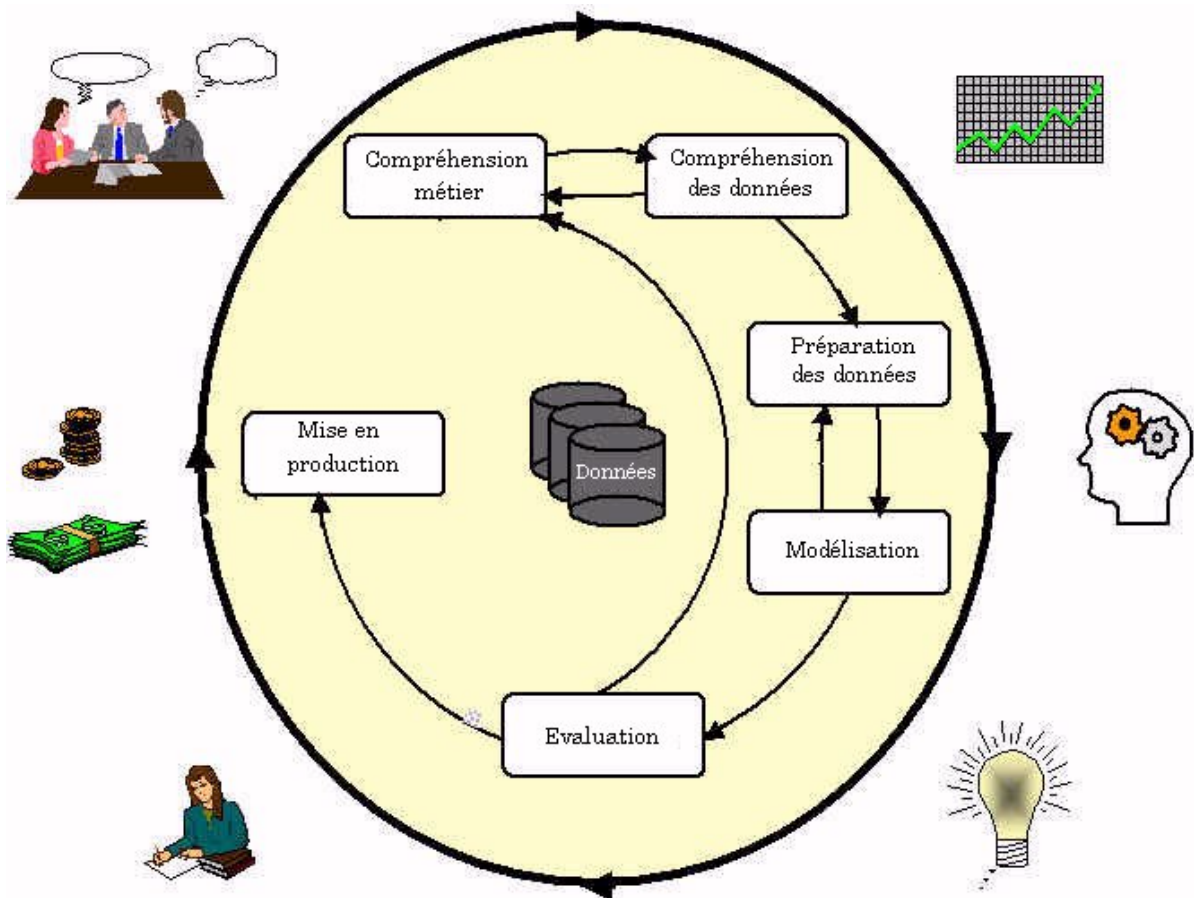
D'un point de vue médiatique (surtout la publicité des constructeurs de progiciels) et universitaire (développement de méthodes nouvelles), c'est ce secteur du DataMining qui a été choyé. Cependant, il est vain de vouloir modéliser ou décrire des données mal connues ou « sales » ; ce qui doit nous amener à toujours insister sur le caractère indispensable de suivre *l'intégralité* de la démarche CRISP-DM. Amputée, elle est bancale.

Rien n'y est superflu, même si le temps à consacrer aux diverses phases peut varier dans d'importantes proportions.

² pour *Cross-Industry Standard Process for DataMining* : processus standard de DataMining pour tous les secteurs de l'industrie

Si plusieurs réponses à la question posée se présentent à l'issue de l'étape précédente, il convient de trancher. C'est la phase d'évaluation, qui permet de choisir la meilleure des solutions.

La mise en production, enfin, est le prolongement concret de l'étude qui vient d'être menée. Elle met en application les résultats proposés, mais ne doit jamais être oubliée une fois effectuée. Il faut encore prendre du recul sur l'action engagée, et la décortiquer une fois qu'elle a produit des effets. Ces résultats, positifs et négatifs, permettront d'améliorer les futurs modèles, et à ce titre devront être réinjectés dans le dispositif. Ce qui boucle le cercle vertueux du DataMining, selon la méthode CRISP-DM.



Quand on connaît le contexte d'émergence du DataMining et le domaine de problématiques que couvre cette approche, que l'on a de surcroît un *modus operandi* à y ajouter, on est à pied d'œuvre. Certes, mais pour faire quoi exactement ? En d'autres termes, que cherche à faire, *concrètement*, une personne qui recourt au DataMining, avec les spécificités de son métier ?

2.2. Les besoins et les attentes des utilisateurs

Nous avons préféré partir des situations rencontrées dans des entreprises réels. Que ce soit par des entretiens avec des chargés d'études et des responsables de « services clients », ou par les cas concrets proposés à l'appui d'une démonstration dans la littérature (comme c'est souvent le cas dans Berry et Linoff), la description du quotidien du DataMining sera faite surtout par l'exemple. Nous allons commencer par une situation qui résume la quasi-généralité des préoccupations rencontrées. Ensuite, nous les détaillerons en voyant d'autres exemples avant de chercher à montrer la diversité des approches rencontrées.

Un exemple paradigmatique

Une responsable des études commerciales de La Redoute définit ainsi les trois attentes de son service marketing...

En premier lieu, structurer et agréger les données : la base clients du vériciste comporte quelque 16 millions d'adresses, assorties de divers renseignements sur chaque client. Trier, recouper et organiser cette masse d'information est effectivement la priorité et le préalable à toute étude sérieuse.

Deuxième point, transformer cette somme de renseignements en connaissances. La distinction entre ces deux termes, « renseignements » et « connaissances » est la clé du DataMining et sa raison d'exister.

Les renseignements sont des données qui existent dans les bases de l'entreprise. Elles y « dorment » et ne servent à personne. Ce fut longtemps le cas des données clientèle, puisque, focalisée sur les produits, l'attention du service clients ne se tournait pas vers les acheteurs.

Les connaissances, en revanche, sont utiles à l'entreprise. Elles aident à la prise de décision. Elles entrent en ligne de compte dans les modèles. Elles permettent de faire des choix et d'orienter la politique commerciale de la firme.

Depuis longtemps, les entreprises achètent à l'INSEE les données du recensement qui ne servaient qu'à juger du taux de pénétration d'un produit sur une aire géographique donnée. On ne corrigeait guère ces taux des profils d'acheteurs spécifiques à chaque région (rurale, urbanisée, frontalière, intérieure, riche ou pauvre, ...) et on appliquait peu ou prou le même calcul à Paris et à Mende. Ne sortant pas de ce domaine qui touche plus les tableaux de bords que les problèmes décisionnels, ces données restaient à l'état de renseignements. Ce n'est que récemment qu'on les a prises en compte. Elles servent aujourd'hui à mieux connaître les acheteurs que l'on est susceptible de rencontrer dans telle ou telle zone de France. Elles sont devenues des connaissances.

Au sein de cette démarche en particulier, la transmutation de l'information en connaissance se traduit par un élagage des données redondantes autant que par la création d'indicateurs synthétiques pour résumer les chiffres à interpréter.

Troisième et dernière préoccupation du service clients de La Redoute, hiérarchiser la clientèle selon des critères de rentabilité. Cette dernière étape pourra sembler pour un statisticien « classique » comme la seule digne d'intérêt. En effet, les Data Warehouses, par les volumes mis en jeu, ont décuplé l'acuité du nettoyage des données avant emploi ; cela n'a pas toujours été le cas. De la même façon, auparavant on avait si peu de données à faire ingérer au modèle qu'on ne distinguait pas connaissances et renseignements. La troisième étape du DataMining

chez le vécipéciste est donc la plus classique de toute. Elle consiste à utiliser sur les données l'arsenal des techniques statistiques à notre disposition pour les faire parler.

On distingue, dans ces techniques, un DataMining descriptif, différent du DataMining explicatif. Le premier cherche à cerner les comportements, et aide à légitimer l'étape précédente où l'on souhaitait décrire avec le moins de chiffres possibles la réalité. Le second vise à expliquer lesdits comportements, et à en inférer une conduite à tenir.

On arrive ainsi à une prise de décision, qui sera suivie d'effets. Ces conséquences, conformément à la méthodologie proposée plus haut, seront quantifiées (logiciels de suivi des actions marketing) en deux étapes : d'abord un test sur un échantillon représentatif avant de lancer, si les résultats sont concluants, la même campagne en grandeur réelle.

Les enseignements tirés de cette démarche seront utilisés pour la définition de la suivante, ce qui nous rappelle que la marche à suivre est cyclique.

La prédiction, premier axe de recherche

L'utilisation massive du DataMining est centrée sur des problèmes décisionnels. C'est à dire sur la prédiction (avec un niveau de probabilité) d'une variable polytomique, ordonnée ou non (elle est souvent dichotomique : accorder un crédit / le refuser, le client répondra au mailing / ne répondra pas, etc...). Une fois ces valeurs prédites avec leur probabilité, on peut construire un score, c'est à dire une échelle continue de valeurs qui note au mieux les individus les plus « intéressants » (il faut alors construire un modèle de coûts ou de profits).

On trouve des exemples de cette utilisation dans tous les secteurs où le marketing décisionnel est développé, ainsi la banque, la finance, l'assurance, la vente par correspondance, la grande distribution, ...

Modéliser sert aussi au cercle vertueux du DataMining, celui qui est explicité dans le processus CRISP-DM. La modélisation permet en effet de tirer les enseignements d'une campagne précédente, en étudiant quels ont été les clients sensibles à une promotion. Quand c'est réalisable, il est bon de comparer le score *a priori*, sur la foi duquel on a ciblé cette campagne promotionnelle, et le score *a posteriori* qui est représentatif des résultats obtenus. On peut ainsi se servir du second pour corriger le premier, déterminer ses forces et ses faiblesses. Pour cela, il suffit de modéliser l'écart entre les deux, avec une matrice de coûts idoine. Les explicatives seront les mêmes que dans la construction des deux scores : leur contribution au modèle rend ainsi compte de leur part dans la réussite ou l'échec du score *a priori*. Elle permet aussi d'éliminer ou de se concentrer sur certaines caractéristiques en vue du score à produire pour la prochaine campagne.

Autre exemple de « minage des données » à but décisionnel : l'appétence à posséder un type particulier de carte bleue, pour les clients d'un groupe bancaire. Quels facteurs font que les clients vont se rendre acquéreurs de ses produits ou, au contraire, vont les abandonner ? Il y a ici une valeur descriptive au processus de modélisation. Mais, par ailleurs, un modèle plus abscons, une « boîte noire » qui rendrait mieux compte du comportement de la clientèle (en mettant en jeu des effets non linéaires), est également le bienvenu. Il permet d'affubler d'un score les futurs acquéreurs, et de détecter ceux qui vont résilier leur contrat parmi les clients déjà possesseurs du produit.

Second axe, la création de typologies

Une demande à peine moins courante est de construire des typologies. Il s'agit d'exploiter les méthodes dites « non supervisées » que regroupe le DataMining, comme les cartes de Kohonen ou les méthodes des nuées dynamiques. Une typologie permet, sans idée d'expliquer a priori UNE variable et une seule, de distinguer des classes d'individus aux mœurs différentes au sein de sa clientèle.

Il s'agit par exemple de la demande du service marketing d'EDF, souhaitant mieux connaître sa clientèle ; on peut aussi citer la Redoute, dont une des préoccupations est de créer des classes de clients dont il faut s'occuper différemment.

On peut avoir sa clientèle composée de médecins, comme c'est le cas dans le cadre de l'industrie pharmaceutique. Dans ce cas, la typologie s'effectue sur le type de prescriptions qu'ils ont coutume de faire, pour leur proposer la gamme de médicaments *ad hoc*.

Cette analyse peut aussi servir aux administrations (Ministère de la Santé, Caisse Nationale d'Assurance Maladie) et aux organismes parapubliques qui font des études dans le domaine de santé et des médications (CREDES³ par exemple).

Les mêmes peuvent encore se pencher sur une typologie des hôpitaux et cliniques afin de légitimer (ou d'infirmer) les conclusions des enquêtes-choc d'associations de défense des consommateurs, lesquelles, tous les ans, nous proposent la liste des établissements de santé à éviter. Une typologie permettrait de cerner les classes où, effectivement, des services sont déficients, et de donner des arguments valables car mathématiques et statistiques pour alimenter les débats.

La grande distribution doit résoudre des problèmes descriptifs d'un type particulier : c'est ce que l'on appelle « l'analyse du panier de la ménagère ». L'exemple le plus parlant consiste à étudier une masse de tickets de caisse. D'un client à l'autre, quelles sont les combinaisons qui reviennent le plus souvent – il s'agit du problème des « règles d'associations » ? Lors d'achats successifs, quels sont les enchaînements d'achats que l'on peut retrouver chez plusieurs clients – ce sont ce qu'on appelle des « séquences fréquentes » ?

Cette étude nécessite des aménagements particuliers dans les données, et n'est pas forcément la plus simple à mettre en place. En effet, il faut posséder, en sus de la « base client » nécessaire au DataMining, un fichier qui compte autant de lignes par client qu'il a acheté d'items. Il faut donc pouvoir accéder également à un fichier non agrégé.

Mais une étude brute, particulièrement dans le cas des tickets de caisse, offre rarement des résultats exploitables : la faute en incombe à la représentation des données. Chaque produit qui passe en caisse est en effet repéré par son code-barre, qui représente une clé à 14 chiffres appelée Gencode. On a ainsi, dans la base, une grande précision sur l'item acheté : marque, produit exact, conditionnement... Il existe une aussi grande différence entre deux bouteilles de vin blanc de mêmes cépage et millésime et de prix voisins, mais de producteurs différents, qu'entre un kilo de pommes et une paire de chaussettes. Tous ont, en effet, des Gencodes différents. De surcroît, pour ne rien arranger, le « radical » d'un Gencode représente le code du fournisseur du produit : il peut donc s'agir d'une grande variété d'items différents. Par exemple, un Gencode commençant par le code de Coca-Cola France peut être celui d'une boisson gazeuse (Coca-Cola, Sprite) ou d'un jus de fruits (Minute Maid)... Sans parler des grands groupes de l'agro-alimentaire comme Nestlé !

Il faut donc définir ce que l'on appelle des taxinomies, c'est à dire des tables de regroupement des produits par familles. On agrège ainsi les Gencodes par items de même type (tous les vins de table, toutes les pièces de porc, toutes les confitures), pour obtenir divers niveaux de

³ Centre de Recherches, d'Etudes et de Documentation en Economie de la Santé

finesse dans l'analyse. On peut ainsi vérifier très rapidement certains lieux communs du marketing, avec des regroupements très grossiers : viande, poisson, fruits, légumes, etc... Il faut ensuite trouver un niveau d'analyse qui apporte des enseignements intéressants, sans tomber dans la finesse excessive qui propose comme phénomène le plus courant l'association de deux Gencodes qui ne se sont trouvés ensemble qu'une dizaine de fois sur un million !

Par ailleurs, il peut être intéressant d'ajouter dans la base d'étude un certain nombre de paramètres pouvant influencer sur le type d'achats : région d'achat, taille du magasin (super, hypermarché, moyenne surface, ...), produit soldé, etc... On appelle ces paramètres des items virtuels, et ils sont mis au même niveau que des items « normaux ».

Le recours aux items virtuels permet d'étendre ce principe d'analyse au-delà de la grande distribution. Une banque pourra par exemple définir les types d'utilisation de ses cartes bleues par cette méthode. Pour un client, on énumérera ses retraits, demandes de RIB, éditions de positions de comptes, transferts de fonds, tous réalisés à des automates avec leur carte. L'ajout d'items virtuels comme le quantième du mois ou l'éloignement géographique de l'agence où est ouvert le compte du client permet alors d'analyser plus finement les comportements. On peut enfin compter au nombre de ces items virtuels les renseignements habituels des analyses descriptives sur un client : CSP, ancienneté dans la banque, et autres variables socio-démographiques.

Le DataMining comme usine à modèles

A toutes ces problématiques de typologie s'oppose une toute autre vision, bien particulière, du DataMining. Elle ne met plus en jeu le côté décisionnel, mais se « contente » de l'utiliser comme outil d'étude. On peut en effet voir cette discipline comme une machine à produire des modèles en grosses quantités, pour ensuite les comparer facilement et efficacement.

On peut par exemple mettre en concurrence cinq régressions logistiques – en jouant sur les fonctions de lien et les méthodes de sélection de facteurs – avec trois arbres de décision – par exemple les méthodes CART, Chaid et C4.5 – et une poignée de réseaux de neurones aux configurations différentes – perceptrons et réseaux à base radiale, en jouant sur la taille de la ou des couche(s) cachée(s).

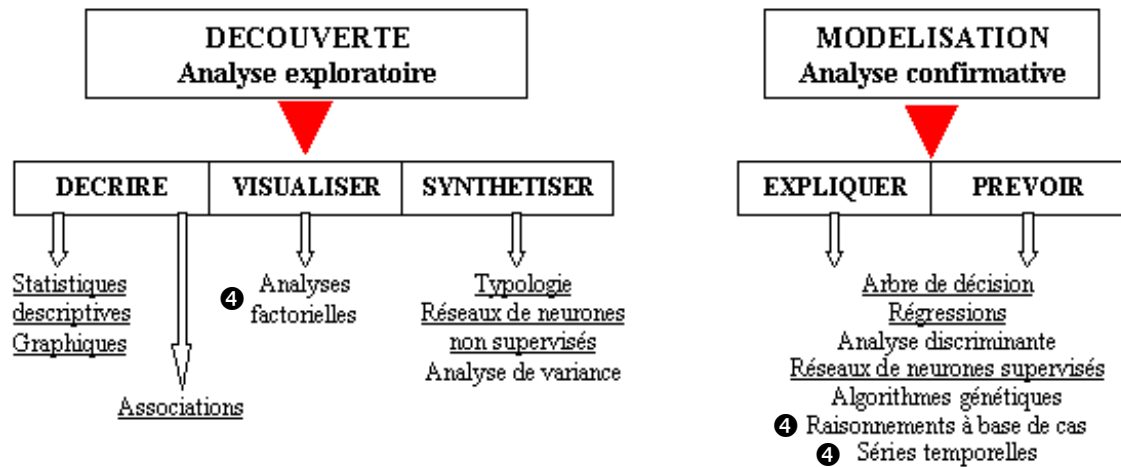
Cette usine à modèles sert au chargé d'étude, le plus souvent, à trouver une explication à une de ses variables importantes (expliquer le salaire en fonction des caractéristiques socio-démographiques de l'employé), voire de façon plus abstraite à lui permettre de transformer un ensemble de variables en un autre. Cette utilisation du DataMining est exploitée au service Qualité de Renault : il s'agit de trouver un lien entre deux ensembles de mesures effectuées selon deux méthodes différentes.

2.3. Synthèse des besoins

Au final, que fait le DataMining ? On peut lister cinq grandes fonctionnalités qui sont la base de tout ce que l'on peut essayer de faire quand on veut « miner » ses données :

- description
- classification
- regroupement par similitudes
- estimation
- prédiction.

Ces 5 fonctions se regroupent en deux familles : exploration et confirmation. On peut illustrer cette « filiation » sur le schéma ci-dessous⁴.



Sur ce schéma, les fonctionnalités directement accessibles depuis Sas Enterprise Miner (version 3) sont soulignées. Un signe ④ accompagne les fonctionnalités apparues avec la version 4.1. Certaines autres sont implémentables manuellement au prix d'une bonne connaissance de la programmation en SAS (on perd alors une bonne partie de l'avantage conféré par l'interface graphique). Il s'agit des techniques factorielles (l'analyse des données dite « à la française »), de l'analyse de la variance, de l'analyse discriminante et des séries temporelles (modèles et prévision). Les algorithmes génétiques, qui souffrent de ne revêtir que des formes trop spécialisées, et le raisonnement à base de cas (ou de mémoire) sont absents du logiciel et ne se mettent pas en œuvre dans l'univers SAS sans une bonne dose de patience, d'efforts et de réflexion algorithmique.

Sur l'ensemble du schéma, nombre des tâches, à cause des techniques qu'elles requièrent pour être menées à bien, sont rattachées à des domaines plus ou moins pointus de la statistique. On peut donc également lister les points de théorie statistique qu'il faudra que les utilisateurs connaissent pour faire fonctionner efficacement SAS Enterprise Miner.

- statistiques descriptives (khi-2, boxplot, adéquation à une loi – QQ plots en particulier –, tests)
- classification (principe de la classification ascendante hiérarchique)
- cartes auto-organisées (type Kohonen)
- régressions (logistique et moindres carrés ordinaires)
- réseaux de neurones (principes généraux et différences entre perceptron et RBF)
- arbres de segmentation (principe général et différences entre Chaid, CART et C4.5)
- techniques de score

Synthèse des problématiques associées au DataMining

Connaître les techniques que l'on peut mettre en place ne permet pas toujours de savoir comment et dans quel cadre les utiliser. En particulier, dans la masse d'exemples proposés ci-dessus, il convient de distinguer quelles sont les grandes problématiques. Celles que l'on peut

⁴ reproduit d'après A-E Sammartino

à bon droit associer à une approche de DataMining, peuvent être regroupées en quelques catégories synthétiques :

- Action commerciale ou marketing sur des **populations très ciblées** (ce que l'on nomme « niches marketing ») : on fait alors appel aux techniques de segmentation pour chercher et caractériser ces microcosmes.
- **Etude du « churn »** ou perte de clients : on cherche à identifier le profil de clients risquant de partir. Pour cela, on use de typologies. On peut aussi construire un score après une modélisation sur la variable parti/resté (sur des données historisées). Une fois identifiés, les clients « à risques » seront relancés afin de mieux cerner leurs griefs et y proposer des solutions.
- **Tarififications ou services adaptés** à des profils de clients particuliers (sans pour autant se fixer sur des micro-populations particulièrement rentables) : une typologie des clients permet, en ne l'affinant pas outre mesure comme dans la recherche de niches, de dégager des grandes « familles » dans la population cliente, et de proposer à son catalogue des offres adaptées sans être trop spécifiques d'une frange très étroite de la clientèle.
- Enfin, reste à part les **problèmes d'associations** : aussi bien l'analyse du panier de la ménagère (étude des tickets de caisse) que les groupements de services sont des problématiques spécifiques, qui ne sont que rarement intégrées aux autres traitements du DataMining.

Qu'est-ce que le DataMining ? Pour en faire la synthèse, on est obligé de répondre à cette question de façon lapidaire et partielle, tant elle fait l'objet de nombreux ouvrages et débats. Il consiste à découvrir des enseignements utilisables et rentables pour l'entreprise dans une grosse masse de données. Les différentes techniques mises en œuvre se retrouvent dans SEM, elles n'ont pas toutes la même complexité. Enfin, les domaines d'applications sont très vastes.