CRM Analytique - L'apport du Data Mining

Françoise Fogelman Soulié

KXEN, 25 Quai Gallieni, 92 158 Suresnes cedex
Francoise.SoulieFogelman @kxen.com
http://www.kxen.com

Résumé. Nous présentons le système d'information CRM et décrivons les composants du CRM Analytique : alors que la Business Intelligence exploite les données du passé pour produire des tableaux de bord synthétisant les évolutions des indicateurs de performance de l'entreprise, le data mining produit, à partir de ces mêmes données, des modèles exploratoires permettant de comprendre les indicateurs de performance (facteurs influençants, structure des variables explicatives, segments) et des modèles prédictifs permettant d'anticiper et donc de planifier des actions CRM plus efficaces. Nous introduisons ces différents éléments et les illustrons sur quelques exemples.
Alors que le volume des données croît de façon exponentielle, peu d'entreprises sont aujourd'hui capables de mettre en œuvre le nombre de modèles nécessaires pour les exploiter toutes, nombre qui devrait lui aussi se trouver en croissance exponentielle. Il faut pour cela mettre en place des «usines à modèles » industrialisant le processus de production de modèles. Nous explicitons ces concepts, indiquons ce que serait le cahier des charges pour une usine à modèles. Nous indiquons comment l'outil de data mining KXEN permet de réaliser des usines à modèles et présentons ensuite quelques exemples d'utilisation de KXEN pour la réalisation de telles usines.

1 Introduction

Les outils de CRM (Customer Relationship Management ou Gestion de la Relation Client) sont aujourd'hui largement déployés dans les entreprises (Lefébure, 2004). Les activités de CRM opérationnel (marketing, vente, service clients) pilotées à travers ces outils permettent de tracer de façon systématique l'ensemble des contacts avec les clients. Toutes les données de contact, consolidées avec les données clients du back-office, peuvent ensuite être exploitées par des outils de CRM analytique pour fournir des tableaux de bord de pilotage de l'activité : suivi des campagnes (nombre de messages envoyés, nombre de réponses, mesure des résultats obtenus …), suivi des processus de vente (nombre de visites clients, de propositions commerciales envoyées, chiffre d'affaire généré …), suivi de l'activité du service clients (nombre de courriers, appels téléphoniques, emails reçus et de réponses envoyés, …) Cependant, les outils de data mining, deuxième composante du CRM analytique, ne sont toujours pas utilisés systématiquement. Bien qu'ils fournissent les moyens d'optimiser l'activité CRM opérationnelle (ciblage des campagnes, recommandations personnalisées …), les entreprises lancent encore de très nombreuses actions (par exemple des campagnes marketing) sans réaliser de modèle prédictif préalable (par exemple modèle de ciblage des campagnes). Toutefois, on voit de plus en plus d'entreprises tenter d'exploiter systématiquement

le data mining en l'intégrant dans toutes les activités de CRM opérationnel : il faut pour cela être capable de produire et industrialiser des centaines – voire des milliers – de modèles par an, sur des volumétries de données très importantes. Ce qui pose des contraintes particulières aux outils qui seront utilisés pour la modélisation prédictive.

Dans cet article, nous présentons le cadre général du CRM analytique ; nous montrons comment les données disponibles dans l'entreprise doivent – et peuvent – être considérées comme une ressource importante que l'entreprise peut exploiter pour améliorer sa performance ; nous présentons ensuite les apports du data mining et les contraintes quand on veut intégrer complètement le data mining dans le CRM opérationnel ; enfin, nous donnons quelques exemples concrets illustrant ces apports et montrons comment certaines entreprises ont pu utiliser l'outil de data mining KXEN [1] pour mettre en œuvre les centaines de modèles dont elles avaient besoin pour leur CRM opérationnel.

2 Qu'est ce que le CRM analytique ?

Le Système de Gestion de la Relation Client comprend plusieurs composants (*FIG. 1*) : le système CRM vise à piloter et animer la relation avec le client à travers des actions lancées sur les différents canaux de contact avec le client. La complexité du système vient de l'hétérogénéité des différents composants, du nombre important de clients et d'offres qu'on peut leur faire, et enfin de la difficulté d'animer une relation multi-canal qui doit apporter à chaque client la bonne offre au bon moment sur le bon canal.

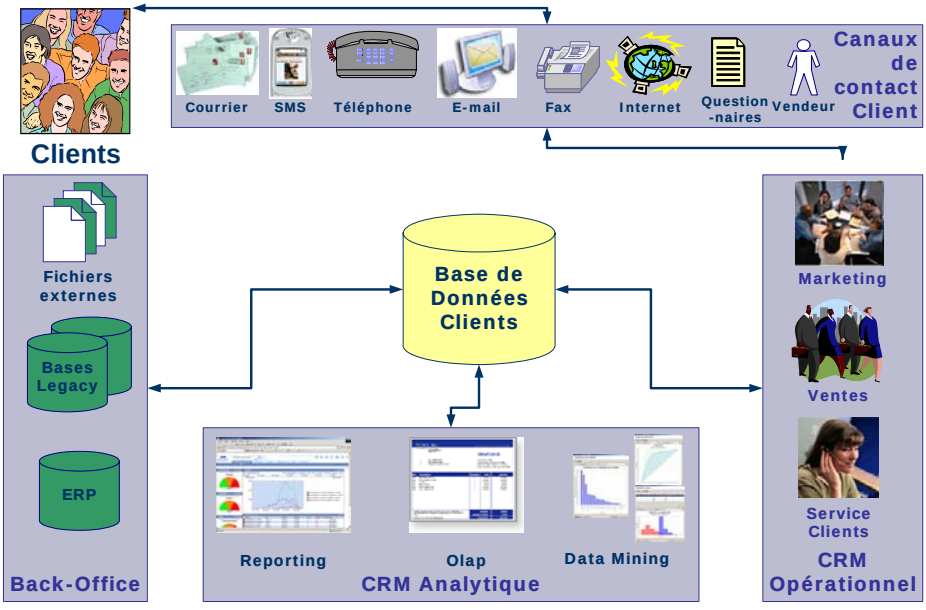

FIG. 1 – *Système de gestion de la Relation Client*

[1] KXEN est un éditeur data mining dont le logiciel KXEN Analytical Framework est considéré par Gartner comme l'un des produits leaders du marché (Herschel, 2007). Voir http://www.kxen.com

2.1 La base de données clients

La base de données clients contient l'ensemble des données disponibles sur les clients (et quelquefois aussi les prospects). Elle a souvent la structure d'un datawarehouse, dont le modèle de données est optimisé pour produire les rapports (§2.4). Les concepts du datawarehouse ont été introduits dans les années 90 par Inmon (1996), Kimball (1998) et ont été très largement développés depuis pour donner naissance à la Business Intelligence (BI ou informatique décisionnelle). La base de données clients est alimentée par des sources multiples.

Les données du back-office de l'entreprise proviennent des bases alimentées par les diverses applications de production (facturation, contrats, litiges et contentieux, …). Ces bases sont souvent des bases transactionnelles non historisées, rarement organisées autour du client (par exemple les bases contrats sont organisées autour du contrat et un client apparaîtra donc autant de fois qu'il a de contrats différents). L'entreprise peut également acheter ou obtenir, selon ses besoins, des fichiers externes (par exemple des résultats d'enquêtes ponctuelles, des fichiers comportementaux …). Ces données sont ainsi hétérogènes, voire contradictoires et de nombreux traitements sont nécessaire pour intégrer ces données dans la base.

Les données contacts : les activités de CRM opérationnel génèrent une grande quantité de données, souvent très riches d'informations sur le client (ce qu'il aime, ce qu'il veut, la façon dont il réagit …) Le CRM opérationnel produit ainsi des données qu'il est important de stocker dans la base de données client. Ces données sont souvent contenues dans un *data mart* géré par l'outil de CRM opérationnel, qui en rationalise ainsi le format et permet d'effectuer des contrôles, parfois assez sophistiqués, sur la qualité des données fournies.

Les données calculées : les analyses data mining exploitent les données de la base et produisent des données (segments, scores, …) qui peuvent être ré-intégrées dans la base. Ces enrichissements peuvent être ensuite exploités pour améliorer les analyses ultérieures (CRM analytique) ou pour optimiser les actions du CRM opérationnel.

2.2 Les canaux de contact client

Les clients ont de nombreux canaux de contact avec l'entreprise (téléphone, courrier, email, SMS, fax, Internet, vendeur, magasin …). L'ensemble de ces contacts apporte des sources d'informations clients très riches … mais hétérogènes : alors que certaines informations sont fiables (par exemple le nom et l'adresse saisis par un client qui commande la livraison d'un achat à son domicile), d'autres peuvent être de qualité moindre, voire carrément fausses (par exemple, un champ date de naissance saisi « par défaut » par un téléopérateur dans un centre d'appel, s'il n'est pas intéressé à la qualité de la donnée saisie, mais seulement au nombre d'appels pris par heure). Les outils de CRM opérationnel, qui permettent la gestion multi-canal, intègrent très souvent des fonctionnalités de contrôle des données au moment de la saisie.

2.3 Le CRM opérationnel

La gestion de la relation client est animée dans l'entreprise à travers trois fonctions.

Le Marketing est responsable de la conception, de l'exécution et du suivi des campagnes marketing. Il utilise pour cela les données disponibles pour générer les listes des clients/ prospects qui recevront les messages, les offres promotionnelles, les invitations … constitutifs des campagnes. Il s'appuie sur les analyses de connaissance clients et les modèles prédic-

tifs développés par le data mining (CRM analytique) pour déterminer la composition optimale des cibles de ses campagnes (i.e. les listes des clients qui recevront les offres). Il réalise des tableaux de bord (CRM analytique) pour suivre le résultat de ses campagnes et évaluer le retour sur investissement.

Les Ventes : le réseau commercial est responsable de vendre les produits de l'entreprise. A travers un réseau de commerciaux terrain, de magasins, de distributeurs (selon les entreprises), il définit les prospects à rencontrer, évalue leurs besoins, établit des propositions commerciales, réalise les ventes. Il utilise pour cela les données disponibles, et recherche éventuellement des sources de données additionnelles pour constituer des listes de prospects potentiels (en achetant des fichiers à des fournisseurs de fichiers, en animant des réunions, des séminaires, des salons …). Il exécute ensuite son processus de vente et en pilote les résultats à travers des tableaux de bord.

Le Service Clients est responsable de toutes les tâches d'assistance au client : renseignement, dépannage, réponse aux réclamations … Il intervient aussi bien pendant les phases d'avant-vente (en support du réseau commercial) que d'après-vente (pour assister le client). Le service clients comprend souvent un centre d'appels que le client peut appeler pour obtenir l'information ou le service désiré, voire un site Web (que le client pourra utiliser en self-service) : le centre d'appels assure généralement aussi la fonction de réception et traitement des courriers et emails des clients. Un ensemble de techniciens de maintenance et dépannage peut compléter le Service Clients.

2.4 Le CRM Analytique

Le CRM Analytique exploite les données stockées dans la base de données clients. Il a trois grandes fonctions mises en œuvre par des techniques de Business Intelligence (reporting & OLAP) ou de data mining (Fig. 2). Le lecteur pourra se reporter aux ouvrages de référence (Imhoff, 2003), (Inmon, 1996), (Kimball, 1998, 2002, 2004) pour la BI, (Hand, 2001) pour le data mining.

Décrire : les données historisées dans la base clients sont exploitées pour produire, pour les utilisateurs métier, des *représentations* des données aussi faciles à comprendre que possible.

1. Les tableaux de bord

- Pour les fonctions de *reporting* standards, des rapports pré-définis sont élaborés par des analystes et diffusés auprès d'un grand nombre d'utilisateurs métier, qui vont les consulter de façon régulière (par exemple : reporting des ventes, chaque semaine). Des rapports ad hoc peuvent aussi être produits pour couvrir un besoin particulier, n'ayant pas vocation à se reproduire.
- Pour les fonctions de *pilotage*, des documents multi-rapports (« dashboards ») intègrent plusieurs rapports pré-définis et peuvent être personnalisés, à partir de rapports communs, selon les besoins des différents utilisateurs. Par exemple, on peut constituer un dash-board pour un utilisateur qui souhaite visualiser le chiffre d'affaire et la marge, les coûts directs et indirects, l'état des stocks et leur évolution, les ventes pour les différents produits (FIG. *3*). Un autre utilisateur aura seulement dans son dash-board les ventes par produits, avec en plus des alertes sur certains produits (de la gamme que pilote cet utilisateur). On peut aussi mettre en place des modes de souscription de rapports en mode push, d'alerte sur événement, voire des services de diffusion multi-canal (PDA, téléphone mobile).

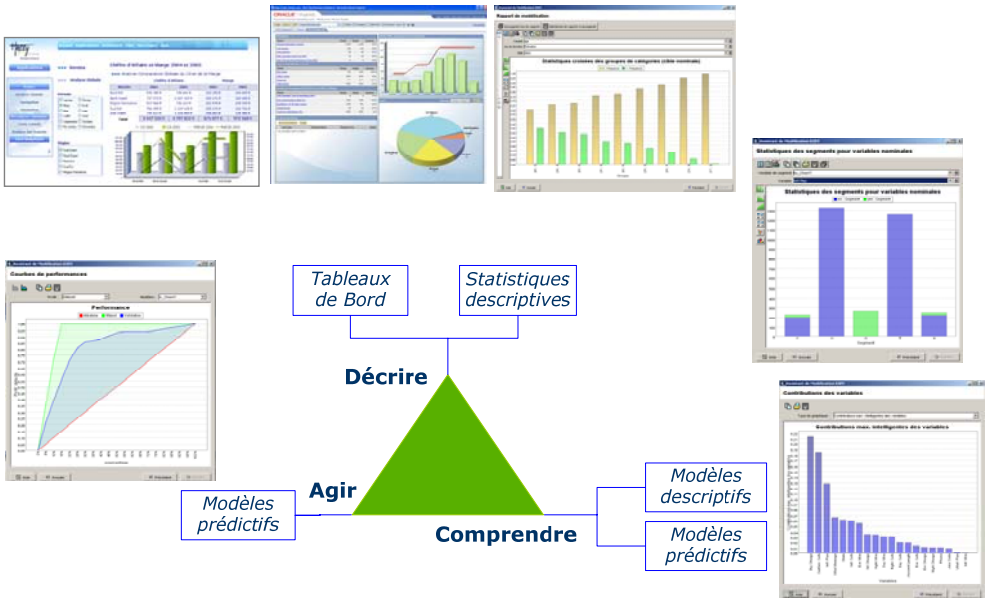

FIG. *2 – Le CRM Analytique*

FIG. 3 – *Dash board (Business Objects)*

2. <u>Les statistiques descriptives</u>

- Elles regroupent un grand nombre de techniques permettant de décrire les données sous forme de graphiques, de tableaux, de valeurs caractéristiques (moyenne, écart-type …) On peut ainsi représenter les distributions des différentes variables, la fréquence des catégories d'une variable, les statistiques croisées de variables 2 à 2 ... La Fig. 4 présente quelques exemples de représentations des données de la base « Adult », introduite dans (Kohavi, 1996).

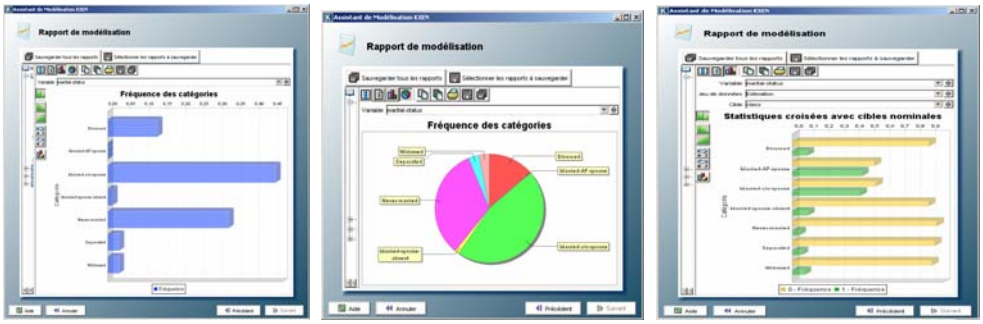

FIG. *4 – Statistiques descriptives (données « Adult ») variable « Marital Status » : de gauche à droite fréquence des catégories et statistiques croisées avec « Class »* [2]

Comprendre : l'exploration des données a pour but de « comprendre » les données, c'est-à-dire d'évaluer leur importance, leurs défauts éventuels (données manquantes ou erronées, outliers), les regroupements de variables ou d'exemples …

1. Modèles exploratoires

- Les tableaux de bord Olap (On Line Analytical processing) fournissent une possibilité d'exploration des données multi-dimensionnelle : on peut obtenir des visions plus détaillées (drill down) ou plus agrégées. Les tableaux de bord Olap sont donc des techniques qu'on utilise pour explorer les données et comprendre ce qui s'est passé (*FIG.* 5)

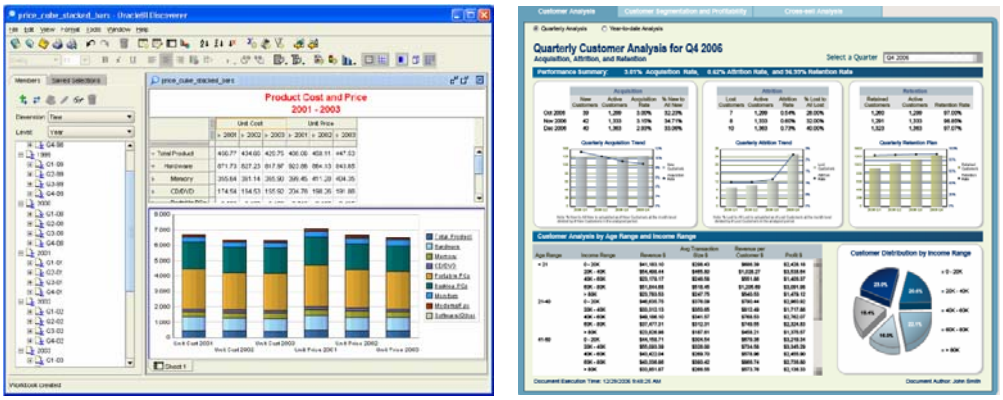

FIG. *5 – Tableaux de bord Olap(à gauche OracleDiscoverer et à droite MicroStrategy)*

- Certaines techniques de data mining permettent de comprendre les données :
  Les techniques de segmentation permettent de construire les segments de clients regroupant des clients ayant un comportement similaire
  Les règles d'association permettent également d'identifier, dans des données de transaction, les achats qui ont une forte probabilité de se produire ensemble (*FIG. 6*)

[2] Les copies d'écran qui figurent dans cet article ont été générées avec le logiciel KXEN, v3.4.1, en utilisant les modules de classification / régression K2R et de segmentation K2S.

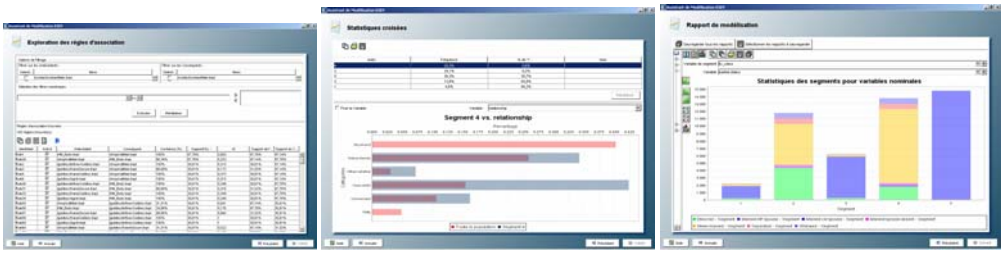

FIG. 6 *Règles d'association produites sur une application de navigation sur le Web (à gauche). Segmentation de la base « Adult » (milieu) : segment 4 par rapport à la population globale. A droite variable « Marital Status » sur les différents segments (KXEN)*

2. Modèles prédictifs

- Les modèles prédictifs (classification ou régression par exemple) peuvent être utilisés pour comprendre l'importance relative des variables, la structure d'une variable (regroupements significatifs) (Fig. 7), les corrélations entre les variables, les déviations d'une période (où on a construit le modèle) sur l'autre (où on applique le modèle) …

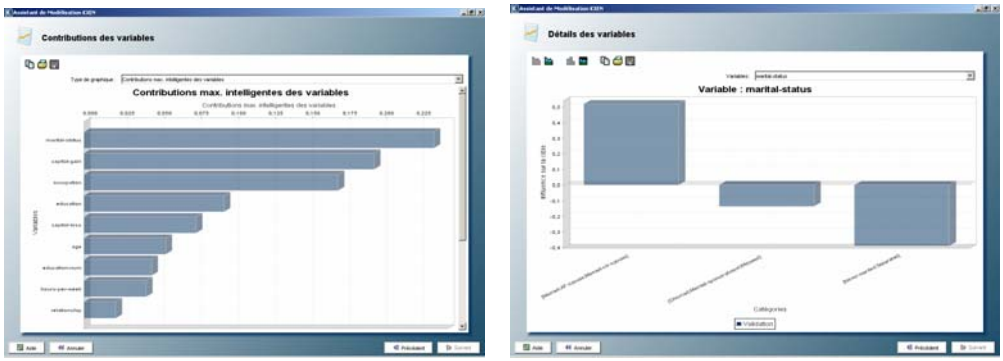

FIG. *7 – Importance des variables dans un modèle de classification de la base « Adult » et importance et regroupement des catégories de la variable « Marital Status » (KXEN)*

Agir

On construit un modèle prédictif pour l'utiliser dans une action CRM. Par exemple, on construit un score de rétention [3], puis à partir de sa courbe de lift (FIG.8 : en abscisse, les clients sont rangés par scores décroissants, en ordonnées, le taux de clients churners), on décide de la fraction des clients à qui on va adresser le message marketing et on lit sur la courbe de lift le taux de churners que la campagne devrait permettre de toucher. On applique ensuite le modèle pour générer la liste de ces clients, et on transmet cette liste à l'équipe qui va exécuter la campagne (CRM opérationnel), c'est-à-dire envoyer le mailing et suivre les retours.

[3] Il s'agit de retenir les clients qui vont passer à la concurrence. On parle aussi de « churn » ou « attrition » pour l'action de passer à la concurrence.

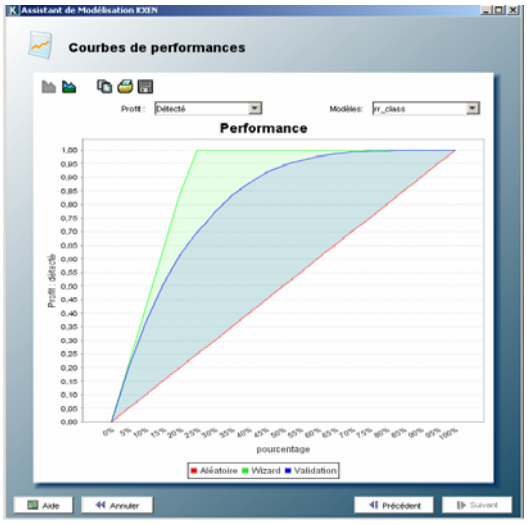

FIG.8 – *Courbe de lift d'un modèle de classification de la base « Adult » (KXEN)*

Le CRM Analytique permet donc d'exploiter les données stockées dans la base de données clients, soit pour expliquer ces données (comprendre ce qui s'est passé), soit pour prévoir ce qui va se passer. Seuls les modèles prédictifs permettent d'anticiper ce qui va se passer et ainsi de lancer des actions de façon *pro-active* ou préventive. Alors que les modèles descriptifs ne permettent que de lancer des actions en *réaction* à ce qui s'est passé. Longtemps considéré comme très complexe et réservé aux problèmes difficiles où un modèle prédictif est absolument nécessaire, le data mining doit en fait être utilisé aussi – et d'abord – en mode exploratoire [4]. Les techniques dites de « data mining » sont donc utilisées pour produire des modèles exploratoires et des modèles prédictifs ; ils sont à la base de l'optimisation des actions CRM.

3 L'apport du data mining

Les analyses data mining sont utilisées soit pour enrichir des rapports (Business Intelligence prédictive) soit pour développer la connaissance client et construire des modèles pour les actions de CRM opérationnel.

3.1 Business Intelligence prédictive

Intégrées dans des rapports de Business Intelligence (on parle alors de BI Prédictive) les analyses data mining permettent de :

[4] « I had always thought of data mining as a tool of last resort when the data is too large or complicated, and nothing else seems to work,… data mining is the first thing that you do when presented with a new business question, or a new data set. You use data mining for the initial analysis of the data to find out which factors in the data really affect the outcome that you are interested in. Once these factors are identified, you can build reports or OLAP cubes using these factors as dimensions to explore in depth what is going on » Richard Taylor (http://bandb.blogspot.com/2005/08/data-mining-insight.html)

- Déterminer les axes d'analyse les plus importants pour l'analyse d'un Key Performance Indicator (KPI) : ceci est souvent fait par les experts métier qui choisissent les axes en fonction de leur connaissance métier. Un rapport ne peut guère inclure plus de 4 axes d'analyse (la dimension de l'espace-temps semble être la limite au-delà de laquelle l'utilisateur ne sait plus lire un rapport). Un modèle prédictif peut automatiquement identifier les variables les plus significatives, ainsi que leur structure (*FIG*. 7). Le tableau de bord obtenu est évidemment beaucoup plus lisible (FIG. 9).

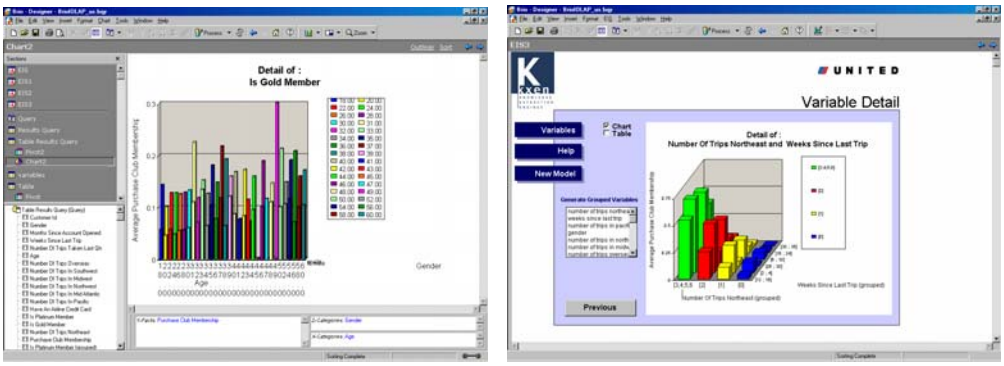

FIG. 9 – KXEN est utilisé pour produire un modèle prédictif de « Gold Member » (données fictives de la compagnie aérienne United). Les 3 variables les plus importantes sont identifiées, ainsi que leur structure : les rapports produits (ici dans le logiciel Brio) qui sont difficiles à interpréter quand on a mal choisi les axes d'analyse (à gauche) sont au contraire très clairs (à droite) quand on utilise l'information fournie par le modèle prédictif.

- Intégrer des prévisions sur un indicateur dans une colonne supplémentaire du rapport ;
- Intégrer des colonnes supplémentaires dans un rapport contenant des indicateurs calculés par un modèle exploratoire (n° de segment par exemple) ou prédictif (score) ;
- Intégrer des indicateurs calculés par un modèle dans des rapports permettant visuellement de naviguer dans les données. Par exemple, Advizor (FIG. 10) intègre KXEN pour produire des modèles don't les résultats sont intégrés dans les rapports et ainsi permettre de sélectionner des axes d'analyse, des sous-populations, ou des affichages plus clairs.

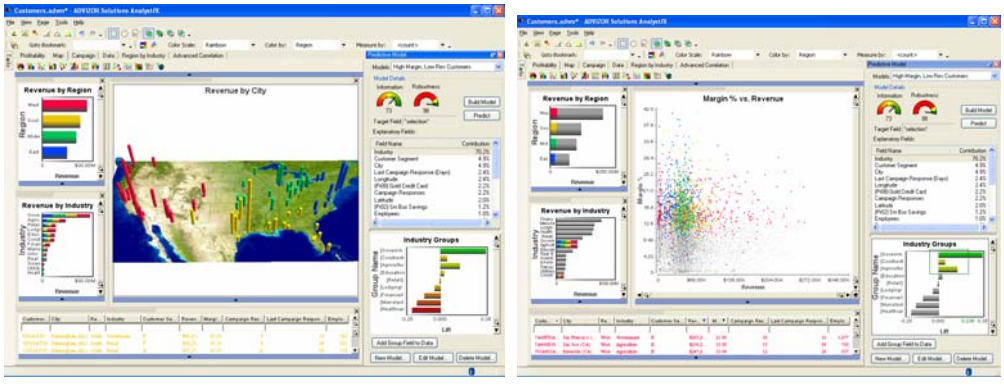

FIG. 10 – Advizor intègre KXEN pour produire un modèle (colonne de droite des écrans) et produire des tableaux de bord très visuels facilitant la compréhension des phénomènes.

- Intégrer des alertes quand la distribution des variables a changé (détection de déviations) ou quand l'information disponible est non significative pour expliquer un indicateur : le modèle prédictif pour cet indicateur n'est pas robuste [5] .

Le domaine de la BI prédictive est considéré comme l'un des domaines les plus prometteurs de la Business Intelligence. On trouvera donc de plus en plus de modèles prédictifs incorporés dans les tableaux de bord du CRM Analytique.

3.2 Analyses pour le CRM opérationnel

Les analyses de data mining permettent de répondre à des questions comme :

- Pourquoi dois-je faire cette offre personnalisée à ce client ?
- Pourquoi ce client répond-il bien à cette campagne ?
- A qui dois-je envoyer ce mailing pour maximiser le taux de retour ?
- Puis-je prévoir si ce client va passer à la concurrence et quand ?
- Quelle est la « valeur » de ce client ?
- Puis-je déterminer qui va frauder et quels sont les facteurs caractéristiques de la fraude ?
- Puis-je déterminer les facteurs qui influencent la signature d'une opportunité commerciale ?
- Puis-je prévoir le montant des signatures de mes opportunités commerciales pour le prochain trimestre ?
- Quels sont les produits que je dois recommander à ce client quand il appelle le centre d'appel ?
- Puis-je prévoir le nombre d'appels et d'emails reçus au service client par heure et par type de sujet abordé ?
- Quels produits puis-je recommander à cet internaute en complément de celui qu'il veut acheter ?
- Quelle évaluation dois-je proposer à cet internaute en fonction de son profil et du produit qui l'intéresse ?

Comme on le voit, les sujets possibles sont très nombreux. Les connaissances ainsi développées sont ensuite utilisées dans tous les processus du CRM opérationnel pour lancer des actions marketing mieux ciblées, pour améliorer le processus commercial, pour optimiser le fonctionnement du Service Clients et ainsi :

- Cibler et conquérir les clients les plus profitables ;
- Augmenter la valeur et la durée de vie des clients existants ;
- Diminuer les départs vers la concurrence ;
- Concentrer les efforts sur les segments de clientèle les plus rentables ;
- Surveiller les clients à risque ;
- Comprendre les comportements des clients ;
- Augmenter l'efficacité des campagnes marketing en améliorant le ciblage ;
- Augmenter les taux de conversion ;
- Se focaliser sur les opportunités commerciales les plus prometteuses ;
- Aider les opérateurs du Service Clients à fournir les meilleures recommandations aux clients ;

[5] Le logiciel KXEN produit un indicateur de robustesse du modèle : KR est la confiance qu'on peut avoir que le modèle, produit sur les données du passé, donnera les mêmes performances à l'avenir.

- Optimiser le planning des agents du Service Clients …

Prenons l'exemple du churn pour illustrer les apports des analyses data mining et la démarche qui est suivie. Supposons qu'un opérateur de télécommunications ait une base de 5 millions de clients, avec un ARPU moyen de 45 € par mois (Average Revenue Per User ou revenu moyen par client). Si le taux de churn est de 2 % par mois (c'est-à-dire que 2% des clients de cet opérateur résilient leur contrat chaque mois), la perte en chiffre d'affaire sur un an est de € 278 Millions environ ; ce qui est bien sûr un montant très important (*FIG. 11*).

	Opérateur Telco
Base Clients	5,000,000
ARPU / mois	45 €
CA sur 1 an	2,700,000,000 €
Taux churn / mois	2.00%
CA après churn sur un an	2,421,936,858 €
Perte de CA sur 1 an	278,063,142 €

FIG. 11 – Churn chez un opérateur télécom (chiffres fictifs)

Les opérateurs telecom ont donc tous mis en place des processus de réduction du churn. Les questions qu'ils se posent sont les suivantes :

- Comprendre qui sont les différents profils des clients passant à la concurrence (churners)
  - On peut réaliser un modèle de classification, qui montre (FIG. 12sur un très petit ensemble de données) que les churners sont les clients ayant une facture journalière de plus de 37 $, ayant appelé le service clients au moins 4 fois, ayant l'option « monde » et n'ayant aucun message dans leur boîte vocale.
  - On peut aussi réaliser une segmentation, par exemple en cherchant 5 segments, on trouve 2 gros segments (représentant près de 72 % de la population) majoritairement non churners (3,3 et 7,8% de churners respectivement) caractérisés par l'état de résidence du client et par la non possession de l'option « monde » ; et 3 plus petits segments majoritairement churners (de 40 à 55 % de churners respectivement) caractérisés par un nombre d'appels au service clients supérieur à 4 et la détention de l'option « monde » (FIG. 13)

FIG. 12 – Score de churn (KXEN)

- Lancer des actions ciblées. On peut ensuite faire un modèle de churn (classification) et, en utilisant la courbe de lift (*FIG.8*) et selon le budget dont on dispose, sélectionner la liste des clients qu'on va contacter pour leur faire une offre susceptible de les retenir. Pour être plus performant, on peut faire un modèle pour chacun des segments identifiés précédemment.

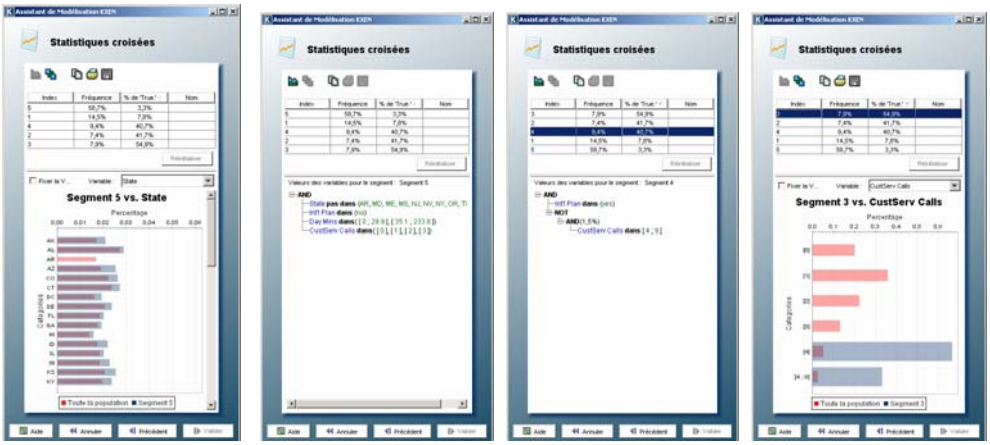

FIG. 13 – Segments de non churners & churners (KXEN)

La performance d'une campagne ciblée est évidemment meilleure que quand on tire au hasard les clients à contacter. Ainsi, une campagne non ciblée peut tout à fait ne générer que des pertes, alors qu'une campagne ciblée, travaillant sur une cible de la même taille mais choisie sur la base du modèle de classification, obtiendra des retours de campagne nettement supérieurs et pourra ainsi générer un résultat positif. Le ciblage permet également de réduire la taille de la cible – pour minimiser le coût de la campagne ou la pression commerciale sur les clients – tout en obtenant des taux de réponse et des bénéfices significatifs (FIG. 14).

	Campagne non ciblée	**Campagne ciblée**	**Campagne ciblée & réduite**
Base Clients	5,000,000	5,000,000	5,000,000
Nb clients ciblés / an	2,500,000	2,500,000	1,250,000
Taux de réponse	15%	30%	26%
Coût du contact	10 €	10 €	10 €
Revenu généré / réponse	40 €	40 €	40 €
Nb répondants	375,000	750,000	325,000
Retour des campagnes	-10,000,000 €	5,000,000 €	500,000 €
Apport du ciblage		15,000,000 €	10,500,000 €

FIG. 14 – Les bénéfices du ciblage

Les analyses data mining sont aussi très souvent utilisées sur le Web. Ainsi, sur les sites de e-commerce, comme par exemple Amazon, quand un internaute lance une requête pour rechercher un produit, le site lui renvoie (FIG. 15), outre les informations concernant ce pro-

duit, des recommandations (produits que d'autres internautes ont acheté en même temps que le produit demandé), des ratings (évaluations produites par les autres internautes) ou même des promotions sur d'autres produits. Tous ces éléments sont produits à partir d'analyses data mining (filtrage collaboratif, règles d'association ou scorings). On comprend aisément que le nombre de modèles nécessaires, à Amazon par exemple qui commercialise plus de 2 millions de livres, est très important (certainement plusieurs milliers). Ces modèles doivent être de plus recalibrés régulèrement, le comportement des internautes changeant très rapidement.

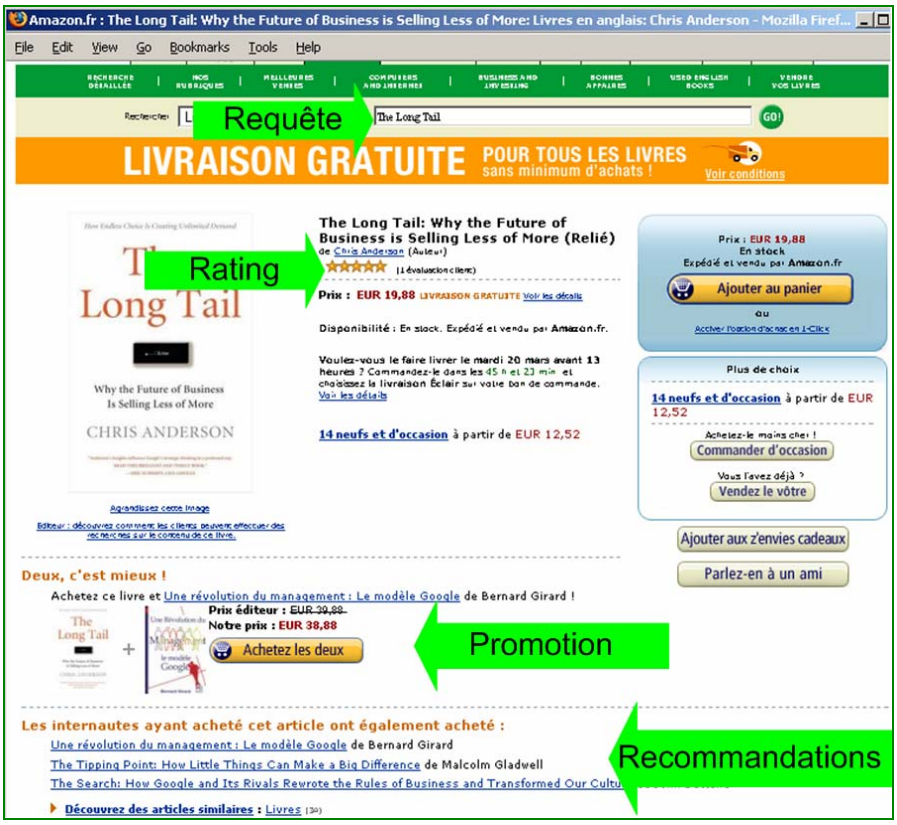

FIG. 15 – Les recommandations sur le site Amazon

L'un des problèmes les plus difficiles quand on veut réaliser des analyses sur un très grand nombre de produits tient à la robustesse des modèles produits : en effet, si le catalogue contient des centaines de milliers de produits, les ventes sur un produit donné peuvent être si faibles qu'on ne disposera pas de suffisamment de données pour construire un modèle fiable.

Enfin, signalons que les analyses data mining utiliseront de plus en plus des données non structurées : pour l'instant seulement le texte, mais demain sans doute aussi les images voire les signaux (parole, video …).

Par exemple, nous avons utilisé les données d'enchères eBay (http://www.data-mining-cup.com/2006/News/1162545659/) pour prévoir si une enchère se termine à un prix supérieur au prix moyen de la catégorie du produit. On peut exploiter les champs descriptifs du

produit en catégorisant les mots du descriptif produit. On obtient ainsi un modèle qui montre que les variables textuelles portent beaucoup d'information prédictive : ainsi après les variables prix de début de l'enchère, la présence des mots « video » et « neuf » dans le descriptif produit est un indicateur fort que le prix de vente final sera supérieur à la moyenne (FIG. 16).

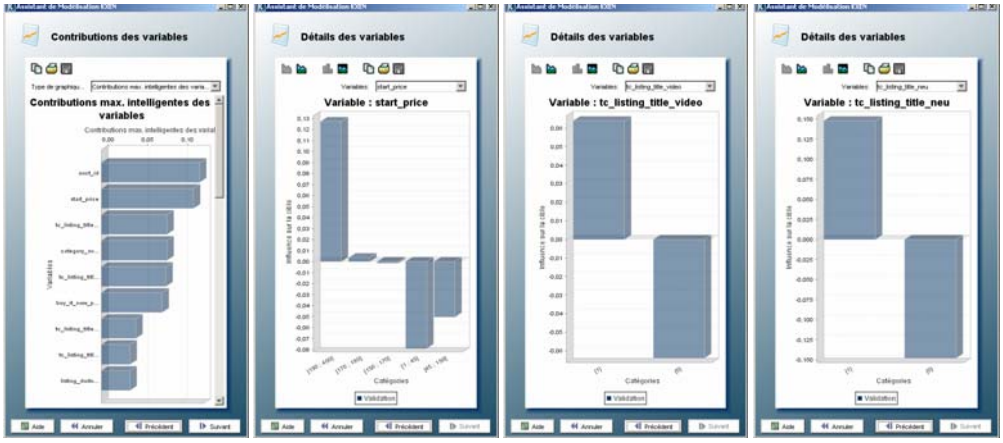

FIG. 16 – Les données textuelles apportent beaucoup d'information discriminante (module KTC : KXEN Text Coder)

Un rapport récent (Aberdeen, 2007) montre ainsi que les entreprises ayant mis en œuvre, mieux que leurs concurrents (« Top performers ») des analyses pour leurs différentes actions marketing obtiennent des performances très nettement supérieures à celles de leurs concurrents (*FIG.* 17 d'après (Aberdeeen, 2007)).

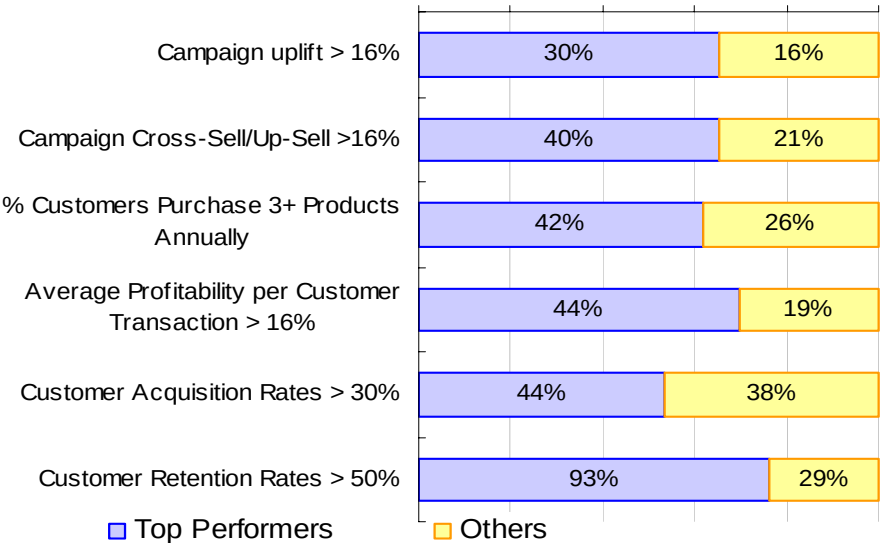

FIG. 17 – Les bénéfices des modèles

Pourtant, malgré ces possibilités d'amélioration de la performance, les entreprises hésitent encore souvent à mettre en œuvre massivement les analyses dont elles auraient besoin. Les raisons en sont nombreuses (Asthana, 2006). Tout d'abord, l'exploitation massive du CRM analytique est souvent perçue comme une révolution culturelle, mal comprise des dirigeants : ces derniers craignent d'exploiter les résultats d'analyses qui leur semblent provenir d'une boîte noire technique mal comprise et préfèrent donc souvent s'en tenir à leurs pratiques courantes. Il peut arriver également que l'exploitation des résultats des analyses soit impossible pour des raisons purement organisationnelles : les processus opérationnels ne sont pas formattés pour cela et la modification de ces processus serait une tâche considérée comme trop lourde au vu des bénéfices attendus.

Par ailleurs, beaucoup d'entreprises, qui ont bien compris que les analyses data mining pourraient leur permettre de développer une connaissance client directement exploitable dans des actions de marketing opérationnel, restent tout simplement incapables de les réaliser. Le cabinet d'analyses Gartner considère ainsi que « Analysis often remains a well-kept secret» (Herschel, 2006) : bien que le volume de données croisse exponentiellement, les entreprises ne sont pas capables de les exploiter toutes pour produire des modèles et ont donc un « gap de connaissances », qui est encore agravé par leur incapacité à exploiter opérationnellement les résultats de ces modèles (« gap d'exécution » : *FIG.* 18 d'après Herschel, 2006a)

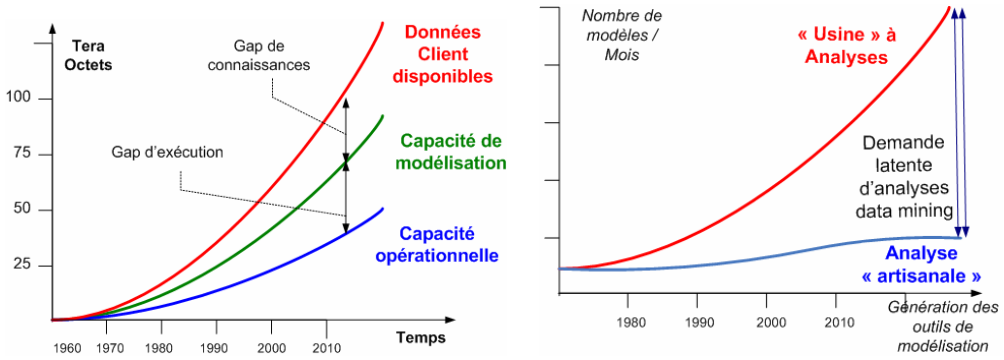

FIG. 18 – Il faut exploiter toutes les données clients (à gauche), mais le nombre de modèles produits de façon « artisanale » ne croit pas assez (à droite) (d'après Herschel, 2006 a et b)

Si on veut vraiment obtenir tous les bénéfices de l'utilisation d'analyses prédictives, il faut que chaque action (par exemple, chaque campagne marketing) exploite toutes les données disponibles, pour produire un modèle adapté et cibler l'action avec ce modèle. On peut alors avoir besoin de produire un très grand nombre de modèles.

Prenons l'exemple d'une entreprise opérant sur un territoire géographique comprenant 20 régions, vendant 30 produits, ayant défini 7 segments clients et voulant réaliser 10 scores (appétence, cross-sell & up-sell, attrition, fraude …). On aura besoin alors de définir un nombre total de modèles de 20x30x7x10 = 42 000 !

Par exemple, Vodafone D2 a analysé ses besoins en analyse et a trouvé que le nombre de modèles nécessaires était de 716 (*FIG.* 19) A partir d'une base Teradata où a été construit un Analytical Data Set contenant toutes les variables disponibles (plusieurs milliers), Vodafone peut maintenant produire les centaines de modèles nécessaires (West, 2005).

		Nb analyses / an
Segmentations	2*2*10	40
Churn (général)	2*3*2*3	36
Churn par produit	2*3*2*4*10	480
Cross-Sell (segments / offres)	2*4*10	80
Acquisition	2*4*10	80
Total		**716**

FIG. 19 – Nombre de modèles nécessaires à Vodafone D2

En pratique cependant, la plupart des entreprises se contentent aujourd'hui de quelques scores « globaux » par famille de produits sur l'ensemble du territoire. Cependant, on voit aisément que les performances d'un modèle « fin » sont meilleures que celles d'un modèle global :

- La performance sur un segment homogène est meilleur : par exemple, un score d'appétence pour une assurance MRH (Multi-Risque Habitation), différent par région (Paris et la Creuse) aura un meilleur taux de retour qu'un score global ;
- Un ciblage plus fin permet de plus d'être plus pertinent dans le message et de mieux le personnaliser ; il évite également la sur-pression commerciale : seuls les clients vraiment intéressés reçoivent le message marketing ;
- Enfin, les volumes touchés par chaque ciblage étant plus petits, on réduit les coûts et on optimise la logistique des opérations.

La plupart des outils actuels de data mining ne permettent pas de produire de façon industrielle les centaines / milliers de modèles nécessaires. Ainsi, Gartner montre que le nombre de modèles nécessaires est en croissance exponentielle au cours du temps (ce qu'il appelle la « factory » analysis), alors qu'il est impossible de les produire avec les outils standards (FIG. 18 d'après (Herschel, 2006b)).

Cette « demande latente » de modèles non réalisés représente ainsi un potentiel significatif de revenu additionnel pour l'entreprise si elle réussit à mettre en œuvre un outil de data mining capable de fournir tous les modèles voulus.

4 L'usine à modèles

L'usine à modèles, c'est la capacité de

- *Traiter des masses de données* : 10-100 Millions de clients caractérisés chacun par des milliers de variables. Ce qui impose des contraintes particulières : les algorithmes data mining utilisés doivent être très simples (par exemple – presque – linéaire) ; avec une manipulation des données minimum, sans duplication et mouvement des données. Les algorithmes de type stream mining sont particulièrement intéressants dans ce contexte (au plus quelques passes pour lire les données, construire le modèle et prendre les statistiques). En particulier, l'arrivée des applications sur le Web amène à manipuler des volumétries très importantes de données structurées ou non (texte, images, videos) et on ne peut donc espérer réaliser des applications pour le Web si on ne satisfait pas à ces contraintes.

- *Produire des masses de projets* : 100-1000 projets par an, par semaine ou même par jour ; ce qui demande la capacité à automatiser la réalisation du modèle (on ne peut pas espérer faire autant de modèles « à la main ».).
- *Produire des modèles « automatiquement »* : pour cela il faut pouvoir industrialiser la production du modèle, l'export du modèle et l'exécution du modèle. Ce qui demande de pouvoir exporter les modèles vers tous formats et d'intégrer ces modèles dans des scripts exécutables automatiquement dans des schedulers.
- *Produire les modèles très rapidement* : en quelques jours ou même heures, notamment pour les applications sur le Web, où la réactivité aux changements de comportement doit être très forte. Ce qui demande un outil convivial, avec automatisation des tâches lourdes (codage des données, sélection des algorithmes, exécution du modèle)
- *Être utilisable par des utilisateurs métier* : l'expertise statistique est rare. Il faut pouvoir donner aux utilisateurs métier le moyen de réaliser eux-mêmes les modèles dont ils ont besoin, pourvu qu'ils aient une connaissance suffisante de leur métier et des données. Ceci demande un outil orienté « utilisateurs » et pas seulement « statisticien ». Le rôle du statisticien devient alors un rôle d'expertise pour assister les utilisateurs métier sur les problématiques complexes.

L'usine à modèles permet ainsi d'augmenter la productivité (plus de modèles, produits plus vite par moins de personnes, avec des personnes moins qualifiées), d'augmenter les bénéfices (en réalisant des modèles pour chaque problème), d'augmenter la vitesse de production des modèles (le « time-to-market » est réduit) et les modèles sont de meilleure qualité parce qu'ils travaillent sur des données plus récentes et des populations plus homogènes.

Le déploiement très large du CRM analytique passe donc par la capacité à mettre en œuvre de véritables « usines à modèles » sur toutes les données disponibles dans l'entreprise. « Competing on Analytics » permet ainsi à l'entreprise de devenir leader sur son marché (Davenport & Harris, 2007).

5 L'apport de KXEN

KXEN commercialise une solution de data mining permettant de mettre en œuvre des « usines à modèles ». Exploitant certains résultats des théories de l'apprentissage statistique développés par V. Vapnik (Vapnik, 1995), KXEN a intégré les principes de la minimisation structurelle du risque (SRM : Structural Risk Minimization) pour automatiser le codage des données et la production de modèles robustes (http://www.kxen.com/).

Cette solution permet de mettre en œuvre des « usines à modèles » (Fogelman Soulié, 2006) en automatisant le codage des variables et la production du modèle tout en contrôlant la robustesse de la solution. Le logiciel accepte les gros volumes de données et obéit de fait au cahier des charges exposé au § 4. Les modèles produits peuvent être exportés dans de nombreux formats (C, SQL, Java, XML, SAS …) ce qui permet de constituer facilement des scripts d'exécution automatisés. De nombreux clients de KXEN ont mis en œuvre le logiciel pour réaliser des « usines à modèles » : nous donnons ici quelques exemples pour des applications de CRM analytique exploitant le logiciel KXEN.

1. Cox Communications (Douglas, 2003) produit aujourd'hui avec KXEN des centaines de modèles pour ses campagnes marketing. Le Département marketing réalise des scores pour ses 26 marchés régionaux en exploitant une base de 10 millions de clients, caractérisés par 800 variables. Les modèles sont réalisés en une semaine par une équipe de 4

analystes et Cox a obtenu son retour sur investissement dans les deux premiers mois d'utilisation de l'outil.

2. Sears (Bibler, 2005) a mis en oeuvre un datawarehouse Teradata contenant 75 millions de clients, caractérisés par 900 attributs. Sears utilise aujourd'hui KXEN pour ses opérations de mailings de catalogues : les modèles sont produits en 1 à 2 jours et le scoring des 75 millions de clients se fait en 30 minutes directement dans la base (« in data base scoring ») .
3. Le Crédit Lyonnais distribue, à travers son offre bancaire, plus de 400 produits. LCL réalise environ 130 actions de marketing direct par an générant 10 millions de contacts annuels clients et prospects. Alors que LCL ne disposait que d'une dizaine de modèles analytiques globaux, il produit aujourd'hui, avec KXEN, plus de 160 modèles par an au niveau deds produits et a plus que doublé le taux de retour de ses opérations de marketing direct dès la première campagne.
4. Proficient – Live Person a réalisé un système d'assistance à la vente en ligne par utilisation du chat. Le système suit le comportement des visiteurs d'un site en temps réel, calcule les scores d'appétence aux produits vendus sur le site, ainsi qu'un score d'acceptation d'une offre de chat avec un opérateur qui va assister l'internaute à identifier le produit qui lui convient. Le système fonctionne de la façon suivante :
    1. Un internaute arrive sur le site web et, éventuellement, s'identifie;
    2. L'historique des données et des préférences du client est récupéré dans la base en temps réel ;
    3. Une page web dynamique est renvoyée à l'internaute;
    4. L'historique des données, des préférences du client, et des pages web visitées sont transmises au service clients de LivePerson
    5. Toutes les six secondes le sytème classe les clients en ligne, en fonction de la vraisemblance qu'ils vont accepter le service de chat et renvoie cette information aux agents du Service clients ;
    6. Les agents du Service clients choisissent les « meilleurs » clients et leur proposent une assistance via chat online ;
    7. Le serveur LivePerson facilite la session de chat entre l'internaute et l'agent ;
    8. Quand la session de chat est refusée ou se termine, le serveur et les agents mettent à jour la base de données de contacts.

    Grâce à l'utilisation des modèles de score développés avec KXEN, LivePerson a augmenté le taux d'acceptation de son offre d'assistance de plus de 75%, et a également amélioré les taux de conversion (internaute achetant le produit) et le taux d'achat des produits proposés. KXEN a permis l'automatisation complète du processus et l'application en temps réel des modèles de score qui sont rafraîchis toutes les nuits
5. Loan Performance utilise KXEN pour déterminer les risques pour les sociétés de prêt immobilier (remboursement anticipé, risque associé au collatéral notamment). LoanPerformance offre à ses clients (la plupart des banques, brokers & investisseurs américains) des solutions de gestion du risque, de rétention, de sécurisation. Pour cela, LoanPerformance alimente une base de données de plus de 40 millions de remboursements d'emprunts par mois, correspondant à 1 milliard de $ de biens. Les portefeuilles d'hypothèques gérés représentent des millions d'emprunts et LoanPerformance doit maintenir de très nombreux modèles spécifiques par produit. Notamment, le système de scoring de remboursement anticipé, PreTell, comprend 75 modèles différents, pour les différents produits hypothéquaires et les différents modes de remboursement anticipé.

En utilisant KXEN, LoanPerformance est capable de produire, déployer et maintenir autant de modèles que nécessaire et les performances obtenues sont de très bonne qualité (Weldon, 2006).

6 Conclusion

Nous avons décrit le système d'information CRM ainsi que les composants du CRM Analytique : alors que la Business Intelligence exploite les données du passé pour produire des tableaux de bord synthétisant les évolutions des indicateurs de performance de l'entreprise, le data mining produit, à partir de ces mêmes données, des modèles exploratoires permettant de comprendre les indicateurs de performance (facteurs influençants, structure des variables explicatives, segments) et des modèles prédictifs permettant d'anticiper et donc de planifier des actions CRM plus efficaces.

Alors que le volume des données croît de façon exponentielle, peu d'entreprises sont aujourd'hui capables de mettre en œuvre le nombre de modèles capables de les exploiter, qui devrait lui aussi se trouver en croissance exponentielle. Il faut pour cela mettre en place des « usines à modèles » industrialisant le processus de production de modèles. Nous avons présenté quelques exemples d'utilisation de KXEN pour la réalisation de telles usines. Les développements en cours dans le domaine du data mining et du CRM analytique renforceront certainement cette approche à l'avenir.

Références

Aberdeen Group (2007) Business Intelligence. A Customer Analysis Solution Selection Guide.

Asthana, Rahul, Crossing the Analytics Chasm. (2006) Business Intelligence Journal, TDWI, march 28, 2006. http://www.tdwi.org/Publications/TenMistake/display.aspx?ID=7892

Bibler, Paul and Bryan, Doug (sept. 2005) Sears: A Lesson in Doing More With Less. TM Tipline. http://ga1.org/tmgroup/notice-description.tcl?newsletter_id=1960075&r=#6

Davenport, Thomas H., Harris, Jeanne G. (2007) Competing on Analytics: The New Science of Winning. Harvard Business School Press

Douglas, Seymour (feb 2003) Product Review – KXEN Analytic framework. DMReview.

Fogelman Soulié, F. (2006) Data Mining in the real world. What do we need and what do we have ? KDD'06, Philadelphia, August 20, 2006. Workshop on Data Mining for Business Applications. 49-53, 2006.
http://labs.accenture.com/kdd2006_workshop/dmba_proceedings.pdf

Hand, D., Mannila, H., Smyth, P. (2001) Principles of Data Mining. MIT Press.

Herschel, G. (2006)a, CRM Analytics Scenario : The Emergence of Integrated Insight. Gartner Customer Relationship Management Summit 2006.

Herschel, G. (2006)b, Customer Data Mining: Golden Nuggets, Not Silver Bullets. Gartner Customer Relationship Management Summit 2006.

Herschel, G. (2007) Magic Quadrant for Customer Data Mining, 2Q07. Gartner.

Imhoff, C., Galemmo, N., GeigerInmon, J.G. (2003) Mastering Data Warehouse Design: Relational and Dimensional Techniques . John Wiley & Sons.

Inmon, W.H. (1996) Building the data Warehouse, 2nd Edition, New York. John Wiley & Sons.

Kimball, R. Reeves, L., Thornthwaite, W. (1998) The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses. John Wiley & Sons.

Kimball, R., Ross, M., Merz, R. (2002) The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. John Wiley & Sons.

Kimball, R. (2004) The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. John Wiley & Sons.

Kohavi, Ron (1996). Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid, In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996. ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult/

Lefebure, R., Venturi, G. (2004) Gestion de la relation client. Eyrolles.

West, Andreas & Bayer, Judy (2005) "Creating a Modeling Factory at Vodafone D2: Using Teradata and KXEN for Rapid Modeling". Teradata Conference, Orlando. http://www.teradata.com/teradata-partners/conf2005/

Vapnik, V.N. (1995). The Nature of Statistical Learning Theory. *Springer Verlag*.

Weldon, D., Wike, E. (2006) LoanPerformance Chooses KXEN to Assess Mortgage Risk, Deploy Prepayment and Collateral Risk. DM Review Magazine. http://www.dmreview.com/article_sub.cfm?articleId=1058064

Summary

We describe the CRM Information System and the components of Analytical CRM. Business Intelligence exploits data from the past to produce reports and dashboards showing the evolution of Key Performance Indicators, whereas data mining produces, from the very same data, both exploratory models – to understand these KPI : key drivers, variables structure, segments – and predictive models – to anticipate and plan more efficient CRM actions.

Even though the volume of data increases exponentially, few companies have today the ability to produce models to exploit them, in numbers which should match this exponential growth. For that, one needs to put in place "model factories" to industrialize the process of models production. We present some examples where KXEN has been used to implement such factories. Present developments in data mining and analytical CRM will allow in the future such model factories to get more and more implemented by companies.