Incertitude du Ciblage Marketing

Stéphane Chauvin

www.R2C-system, Niort, France
stephane.chauvin@R2C-system.com

Résumé. Cet article présente une logique de recherche d'information pour le ciblage des clients dans le cadre de la fidélisation et de la prospection d'un marché pour une marque. Il est mis en exergue l'approche du dataminer qui a pour objectif d'améliorer la connaissance et la pertinence de la cible client. Ceci s'illustre par le besoin de diminuer l'incertitude à prendre une décision sur l'affectation d'un client à une campagne. La méthode présentée consiste à opérer une succession de segmentations pour décider de l'appartenance d'un client à un segment, celle-ci pilotée par le minimum de l'indicateur d'incertitude : la mesure de l'entropie des probabilités a posteriori. L'incertitude se réduit au fur et à mesure que sont ajoutées des informations de sources différentes, en l'occurrence, l'expérience faite dans cet article, des données contextuelles de vie du client. On propose une méthode simple de regroupement des zones géographiques à fort potentiel et un algorithme d'optimisation des micros zones candidates pour une action marketing. Cette méthode a l'avantage de consolider la définition des segments de la population cible pour une interprétation robuste du marqueteur

1 Introduction

Les besoins du marqueteur sont au nombre de deux : d'une part, celui de connaître et de comprendre les évolutions des clients qui constituent le portefeuille d'une enseigne ou d'une marque, et d'autre part, d'adapter une stratégie de communication et d'offre à chaque individu. La complexité de la gestion de la relation client et le coût opérationnel de la mise en place d'une stratégie de communication pour la fidélisation et la prospection, emploient le marqueteur à rechercher des informations sur les clients, à en dégager des connaissances sur les individus. L'enjeu du métier du marqueteur n'est plus de trouver les meilleurs clients pour une campagne mais la meilleure campagne pour chaque client.

Le niveau de complexité que cela engendre impose aux entreprises de disposer d'outils de datamining et de méthodes de recherche de connaissances dans une quantité d'information toujours plus conséquente. Ces outils permettent de réduire la complexité des choix de communication, en calculant des regroupements de l'information par les techniques de segmentation (ou classification sous contraintes). C'est lorsque l'on peut résumer la connaissance que l'on a des clients par le regroupement en catégories et segments que le marqueteur peut faire intervenir sa réflexion quant à la stratégie de communication.

La base de données marketing est une sphère précieuse d'information sur le client. Les données agrégées et orientées client permettent de donner un sens à ce que le segment est, par rapport à la marque : le score d'appétence, le score d'attrition, le nombre d'achats, la mesure de l'activité récente moyenne, la fréquence des réactions aux stimuli commerciaux, la mesure des probabilités de transition entre les différents états socio-démographiques, les dates des événements explicatifs de la vie du client, etc. Ces informations contribuent à calibrer le client en fonction de plusieurs critères de valeur pour la marque. Elles donnent du sens aux différents segments détectés et permettent de mesurer l'adéquation du message avec la stratégie de l'enseigne. Le sens s'accroît si on rend l'ensemble des segments d'une part indépendants et discriminants deux à deux - on recherche une pureté statistique en maximisant le nombre de variables explicatives d'un segment - et d'autre part si on adopte un message précis et exclusif dédié à chacun de ces segments - on favorise le choix des données qui apportent des informations explicitent et opérationnelles d'un segment -, ces deux objectifs pouvant être antagonistes.

On veut, ici, pouvoir mesurer le degré d'incertitude qu'a le marqueteur à prendre une décision sur l'adéquation d'une offre avec un message dédié à une population. Le problème est donc d'étudier le niveau de qualité de l'affectation des individus aux classes prédéfinies. Nous utilisons ici une mesure de qualité de l'information pour la décision basée sur la théorie de l'information de Shannon [Shannon, 1948], [Jaynes, 1957] (section 2). Cette mesure est utilisée pour déterminer le meilleur tableau de bord, au sens du minimum de l'incertitude, efficace pour la prise de décision. Nous illustrons les propos par un exemple pris d'un marché de produits haut de gamme. Ce marché se caractérise par un nombre d'achats faible pour un coût d'achat unitaire élevé. La base client contient le score de la valeur client. L'échantillon est de 1163 individus.

On compare dans la section 3 le résultat du calcul de l'incertitude à partir du taux d'erreur d'affectation (vision statisticienne) à celle de l'ajout d'un nouveau découpage de la population (vision informative sur la valeur client). Dans la section 4, on présente une manière d'ajouter des informations segmentées extérieures à la base de données client. Les résultats montrent que le niveau d'incertitude diminue par le croisement des données de sources très différentes. On détaille les calculs qui déterminent le gain d'information à combiner deux domaines de mesures, plus compréhensible pour le décideur en marketing.

2 Indicateur d'incertitude

Le marqueteur veut mesurer les erreurs d'affectation de l'individu x à une classe et veut savoir si l'individu x répondra favorablement à un stimulus dédié au segment $c_i \in C$ (une offre, un message à une date), $c_i \;\; i = 1, \ldots, K, C(c_1, \ldots, c_k)$, étant une collection de classes prédéterminées. Soit P_i la probabilité a posteriori que l'individu x appartienne à c_i . La réalité est que cette probabilité est entachée d'incertitude i.e. $P_i = 1 - \varepsilon$ avec $\varepsilon > 0$ et $\sum_{i=1}^{K} P_i = 1$. erreur d'affectation en ajoutant une nouvelle variable). Dans cet article, on applique la formule de Bayes pour laquelle aucune ambiguïté d'estimation des probabilités a priori et conditionnelles sont à noter.

Dans le cas où l'incertitude est faible, l'égalité revient à $P_i = e^{-\varepsilon}$, $1-\varepsilon$ étant le développement $e^{-\varepsilon}$ pour ε petit. La quantité ε , devient égale à $-\ln P_i$. Elle est alors considérée comme la mesure de l'incertitude associée à l'affectation de l'individu$x$ à c_i ($c_i$valeur de la variable d'affectation C). Dans le cas où pour tout x la probabilité de répondre favorablement à un stimulus est 1, alors le marqueteur est un homme chanceux et prendra une décision en toute connaissance de cause. On caractérise le niveau de connaissance que l'on a d'un système par la capacité à être déterministe ou pas. L'espérance de la connaissance globale de la décision à prendre sur l'individu x est une fonction monotone croissante et positive donnée par :

$$E(x) = -\sum_{i=1}^{K} P_i \ln P_i.$$

Elle sera nulle si toute décision d'affectation de l'individu x à une classe a une probabilité 1 à l'un des K segments. La situation d'incertitude totale s'obtient quand les probabilités d'appartenir aux K classes sont toutes égales à $1/K$, l'espérance étant bornée par $\ln K$: $0 < E(x) \leq \ln K$. Ces résultats bien connus, sont détaillés dans [Kay, 1995].

La mesure de l'entropie est utilisée ici comme un indicateur de validation d'une segmentation. Dans l'expérience présentée dans cet article, on utilise l'entropie marginale à une classe i permettant ainsi de se focaliser sur la mesure d'incertitude d'une classe. La robustesse d'une segmentation peut aussi s'affirmer par la mesure de l'entropie en combinant plusieurs échantillons [Bercher, 1994]. Si la valeur de l'entropie des différents échantillons, mesurée par l'espérance $E = \sum_x E(x)$, reste stable, alors l'échantillonnage est de bonne qualité et la répartition des individus de l'échantillon sur les segments est la même que celle de la population.

3 Incertitude sur la connaissance du client

Les méthodes statistiques de segmentation permettent d'optimiser le regroupement des individus en catégories homogènes pour lesquelles il se dégage un sens et une interprétation facile à utiliser par le marqueteur. Les différentes mesures de qualité des regroupements consistent en des indicateurs de niveau de discrimination d'une classe. Ce sont les variables qui expliquent le sens caché d'un segment et qui déterminent la qualité du regroupement des individus (faible variance intra classe et variance inter classe élevée, [Saporta, 1978]). L'optimisation sous contraintes des données numériques ne permet pas toujours de dégager un sens logique de marché. Dans le domaine du marketing, on préférera souvent avoir des segments pouvant être interprétés par la constitution sociologique et homogène du groupe que par la qualité discriminante des variables mises en jeu. C'est le dilemme entre les dataminer et les statisticiens qui doivent trouver le juste milieu entre le potentiel de discrimination des segments et le sens des segments [Chauvin, 1997]. Ces enjeux ne sont pas forcément compatibles. Une mauvaise gestion de cet antagonisme augmente l'incertitude qu'a le marqueteur à élaborer une campagne marketing sur une cible : ce sont les cas extrêmes de cibles peu explicites ou de cibles peu discriminantes des clients.

Les techniques de segmentation peuvent être simples à mettre en place. La méthode qui consiste à répartir les clients à fort potentiel contre ceux ayant un potentiel moyen ou faible, opère un tri puis un calcul du cumul d'un score (chiffre d'affaires par exemple). Ce score qui représente

une valeur client est découpé en terciles. Il en est déduit les regroupements, en deux ou trois classes qui favorisent la visibilité du client considéré comme à fort potentiel par rapport aux autres vérifiant un potentiel plus faible. Cette technique génère souvent une variable potentiel s'apparentant à la loi de Pareto i.e . 20% de la population apportant les 80% de l'activité. Le sens des segments est clair mais les paramètres de séparation des populations sont stricts et parfois arbitraires.

Les techniques plus évoluées de segmentation comme celle des arbres de décision, permettent de travailler le sens des segments tout en agrégeant des informations pas à pas. Au final, les règles d'affectation d'un individu à un segment sont clairement établies et conduites statistiquement par des indices éprouvés de séparation (la mesure d'entropie est une des mesures possibles pour constituer un facteur de discrimination inter-classes). D'autres techniques bien connues comme l'analyse factorielle sont très utilisées. Elles favorisent les regroupements homogènes à l'instar des classificateurs comme le K-means. Un bon résumé de l'ensemble de ces techniques peut se consulter dans [Govaert, 2003].

Ces techniques ont toutes leurs particularités et avantages que seul l'utilisateur, le dataminer ou le statisticien, peut juger de l'opportunité à les utiliser. Dans cet article, on oppose de manière caricaturale, le dataminer qui se préoccupera du résultat d'une fonctionnalité d'un logiciel au statisticien qui, quant à lui, sera préoccupé par le calcul qui a permis de constituer le résultat.

L'exemple qui suit porte sur la segmentation d'une population cible issue d'une base client. L'objectif est d'estimer la sous population cible à fort potentiel d'achat sur laquelle il est important économiquement de mieux rentabiliser la relation en incitant le client à fréquenter la marque. Trois catégories de clients ont été détectées, tous jugés en condition d'achat dans un futur proche. On distingue d'une part la population à forte propension à passer à un acte d'achat supérieur en chiffre d'affaires au précédent, d'autre part celle dont la propension portera sur l'achat d'un produit de même standing avec, par voie de conséquence, la stagnation en chiffre d'affaires ou le report d'achat et enfin, la population tentée soit par le changement de marque soit par l'achat d'un produit de qualité inférieure avec pour effet, la diminution du chiffre d'affaires.

La méthode d'affectation se base sur la prédiction du potentiel du chiffre d'affaires de chaque client. Les segments étant fixés, les transitions sont modélisées par les chaînes de Markov. Cela suppose que tout état de l'univers des possibles est atteignable pour n'importe quelle configuration de départ et que l'état considéré n'est conditionné que par l'état précédent. La mise en oeuvre de ce modèle est tout à fait possible pour cette étude puisque, l'échantillon de clients est suffisamment grand pour générer tous les événements possibles. Par ailleurs, entre trois achats, il peut s'écouler quelques années rendant le modèle stationnaire dans le temps. Cette technique sert à déterminer la probabilité d'achat d'un futur produit et le type de produit acheté (pour cette expérience, et de manière à diminuer la complexité calculatoire, les états des matrices de transition ont été au préalable agrégés).

Il est, ainsi, déduit la marge moyenne potentielle de chaque client. Les trois classes sont déterminées par les clients ayant un " Potentiel Fort " (saut de la valeur du panier significatif)

représentant 18%, les clients disposant d'une marge négative ou positive peu significative, la classe " Potentiel Stagnant " représentant 36%, et les clients de " Potentiel Faible " qui représentent 46%. Les variables Style de Vie et les variables de description du dernier produit acheté expliquent à 79% les classes estimées.

Deux types d'incertitude peuvent être calculées : celle qui sera mesurée à partir des erreurs d'affectation, une deuxième résultant de l'ajout d'informations.

3.1 L'incertitude du statisticien

Cette incertitude est calculée à partir de l'estimation des erreurs d'affectation des individus aux trois segments. Sur une base de clients tests pour lesquels l'avant dernier et le dernier achat sont connus ainsi que le score associé, les erreurs d'affectation s'appréhendent en estimant a posteriori les probabilités d'affecter à une classe conditionnellement à l'appartenance à une autre classe. L'incertitude à prendre une décision sur le système d'appartenance aux trois segments est illustrée dans les tableaux 1 et 2

Taux d'erreurs Vérité / Décision	Potentiel Fort	Potentiel Stagnant	Potentiel Faible	Total
Potentiel Fort	76,1%	16,0%	7,9%	100%
Potentiel Stagnant	34,6%	63,1%	2,3%	100%
Potentiel Faible	11,0%	22,5%	66,5%	100%

TAB. 1 – *Estimation des probabilités a posteriori des erreurs d'affectation.* $76,1\%$ *des clients appartenant au segment Potentiel Fort sont bien affectés contre* $23,9\%$ *affectés aux autres*

Mesure d'incertitude E(x) Vérité / Décision	Potentiel Fort	Potentiel Stagnant	Potentiel Faible	Entropie Marginale
Potentiel Fort	0,21	0,29	0,20	0,70
Potentiel Stagnant	0,37	0,29	0,09	0,74
Potentiel Faible	0,24	0,34	0,27	0,85

TAB. 2 – *Mesure de l'incertitude par le calcul de l'entropie. L'incertitude marginale à affecter les clients au segment Potentiel Fort est égale à 0,70. Plus* $E(x)$ *est faible, plus l'incertitude diminue.*

L'incertitude calculée est propre au statisticien qui recherchera à avoir un système d'affectation maîtrisé et contrôlé. A l'inverse, le dataminer acceptera les aléas du calcul et prendra le résultat comme une valeur en tant que telle et jugera plutôt de la pertinence syntaxique de la segmentation utile à la compréhension du phénomène marketing.

3.2 L'incertitude du dataminer

Cette deuxième incertitude consiste à s'affranchir de l'estimation des mesures d'erreurs, et porte sur l'ajout d'une information construite et connue pour compléter le sens du segment

cible. Nous prenons en exemple, ici, le croisement des trois segments précédents avec trois catégories de clients notées GMP (Grand, Moyen et Petit), segmentation très utile pour tout type de marché, appelé souvent Valeur Client : les clients considérés à fort potentiel économique et fidèle à la marque, les clients considérés à potentiel moyen et les petits clients. On sait intuitivement que ce nouveau découpage de la population cible contribue à mesurer et à discriminer plus finement les segments. Le tableau 3 ci-dessous donne la répartition (en nombre et en pourcentage) des individus dans chaque catégorie. Le tableau 4 donne le calcul intermédiaire des probabilités pour en déduire la mesure de l'entropie (tableau 5).

Répartition des clients	Grand		Petit		Moyen	
	Nombre de clients	%	Nombre de clients	%	Nombre de clients	%
Potentiel Fort	155	44,9%	21	4,7%	32	8,7%
Potentiel Stagnant	68	19,7	228	51,0%	127	34,1%
Potentiel Faible	122	35,4%	198	44,3%	212	57,2%
Total	345	29,7%	447	38,5%	371	31,8%

Répartition des clients	Total	
	Nombre de clients	%
Potentiel Fort	208	17,9%
Potentiel Stagnant	423	36,3%
Potentiel Faible	532	45,7%
Total	1163	100,0%

TAB. 3 – *Affectation des clients aux segments " Potentiel " et " GMP ".*

Probabilités Potentiel / GMP	Grand	Moyen	Petit	Probabilités Marginales
Potentiel Fort	74.5%	10.1%	15.4%	17.9%
Potentiel Stagnant	16.1%	54.0%	29.9%	36.3%
Potentiel Faible	23.0%	37.3%	39.7%	45.7%
Probabilités Marginales	29.7%	38.5%	31.8%	100.0%

TAB. 4 – *Probabilités a posteriori P (Potentiel / PMG).*

Entropie Potentiel / GMP	Grand	Moyen	Petit	Entropie Marginale
Potentiel Fort	0.22	0.23	0.29	0.74
Potentiel Stagnant	0.13	0.33	0.36	0.82
Potentiel Faible	0.15	0.37	0.37	0.88

TAB. 5 – *Gain d'information, à l'ajout de la segmentation Valeur Client, mesuré par l'entropie associée.*

L'incertitude calculée est, ici, basée sur la capacité à reconnaître et à mesurer la valeur du client, Grand, Moyen et Petit. Elle est égale à $0,74$ pour la décision d'affecter un individu au segment " Potentiel Fort ".

3.3 comparaison

En comparant les mesures faites, le statisticien et le dataminer auront approximativement les mêmes niveaux d'incertitude à affecter un client au segment Potentiel Fort, respectivement égales à $0,70$ et $0,74$ (comparaison du tableau 2 avec le tableau 5). Cette expérience illustre que dans le cas particulier du marketing, il est parfois souhaitable de préférer porter l'attention à croiser entre elles des connaissances accumulées, en l'occurrence, le potentiel économique réel (GMP), pour mesurer l'incertitude. Cette dernière permet avantageusement de manipuler des concepts du marqueteur (profil du client), au lieu de manipuler des concepts techniques (paramétrage d'outil de calcul d'information et taux d'erreurs d'affectation). L'incertitude du dataminer semble mieux adaptée au métier du marketing.

Le marqueteur a toujours la possibilité d'affiner les mesures et les connaissances sur ses clients. Tout est fait pour agréger des informations contenues dans la base de données " client ". Cependant, il limite sa capacité d'évoluer dans la compréhension de ses clients et se heurte au seuil d'incertitude de plus en plus difficile et coûteux à réduire. Il est nécessaire d'intégrer dans la connaissance du client, des données exogènes et externes aux données endogènes pour enrichir les moyens de discriminer. Le traitement de l'information est la partie importante de la recherche de connaissances [Pyle, 1999].

De nouvelles frontières de mesure de la société s'ouvrent régulièrement. L'apport des nouvelles technologies d'information est fondamental, permettant à distance de sonder, de mesurer les évolutions des marchés. La frontière entre les données de la base client et les données géo-démographiques est une des frontières qui doit être intégrée dans les bases, faisant la jonction entre la connaissance du client et son milieu culturel.

4 Ajout de données externes

Le marché d'achat de données est florissant (les six premières sociétés de sondages absorbent 16 milliards d'euros de chiffre d'affaires en études pour les marchés européens, japonais et des Etats Unis) et offre des mesures riches en enseignement sur les évolutions des comportements, tels les panels de home scanning, les marchés tests contrôlés, les mégas bases, les " access panels ". Les données à disposition sont, soit des données génériques (information standard des mesures socio-démographiques, mesures de comportement d'achat, ...), soit des données spécifiques à la problématique de l'enseigne (recherche de données sur son propre marché, non disponibles a priori).

A supposer que les adresses soient calibrées et géoréférencées, toutes les informations portant sur le contexte de vie du client peuvent être intégrées au niveau du client. Les instituts proposent des informations au niveau de l'IRIS (Ilots Regroupés selon l'Information Statistique), contenant en moyenne 2000 foyers pour IRIS 2000 en France. A chaque zone, sont attribuées les informations génériques : densité de population, nombre de foyers, de personnes adultes de sexe féminins et masculins, le nombre d'enfants par foyer, les pourcentages des typologies d'activité, etc. Le Cycle De Vie (CDV) d'une zone IRIS est une information qui détermine le pourcentage de la population appartenant à une des catégories socio-démographiques définies au préalable. L'exemple est pris ici pour les 7 typologies suivantes, couvrant les cycles éco-

nomiques de transition : " Jeune ou adulte Indépendant " (2 typologies), " Jeune Couple avec deux salaires sans enfants ", " Foyer avec enfants en bas âges ou adolescents " (2 typologies), "Troisième âge de classe aisée ou modeste " (2 typologies). Le CDV peut être agrégé par régions administratives du pays, zones urbaines, suburbaines et rurales, par regroupement des niveaux de densité de population des IRIS, etc.

En prenant l'hypothèse que la réduction de l'incertitude à affecter un client à un segment passe par la connaissance de la cible marketing, on cherche à regrouper les clients ayant des caractéristiques socioculturelles proches. Le tableau 6 donne le détail des répartitions du CDV moyen des trois segments identifiés précédemment. Ce CDV moyen des segments est ensuite comparé aux CDV de tout le territoire, et par conséquent de tous les IRIS etc.

	Profil " Potentiel Fort" " " Moyen Profil	"Potentiel Stagnant " Moyen	Profil " Potentiel Faible "Moyen	Moyenne d'une région	Indice de similitude " Potentiel" Fort / Moyenne Région
Jeune indépendant	18,0%	13,2%	12,6%	4,1%	439
Adulte indépendant	14,0%	7,1%	9,2%	16,4%	85
Jeune couple avec deux salaires sans enfants	11,9%	11,5%	9,4%	11,2%	106
Foyer avec enfants en bas âges	23,3%	25,0%	29,4%	37,0%	63
Foyer avec enfants adolescent	17,5%	16,4%	16,0%	12,7%	138
Troisième âge de classe aisée	12,7%	17,1%	7,8%	6,6%	192
Troisième âge de classe modeste	2,5%	9,7%	15,6%	12,0%	21
	100,0%	100,0%	100,0%	100,0%	100,0%

TAB. 6 – *Répartition moyenne des typologies de cycle de vie des clients de l'enseigne confrontée à la répartition des typologies moyennes d'une région. L'indice de similitude met en exergue les fortes particularités comportementales de la cible, le segment " Potentiel Fort".*

On utilise la mesure du χ^2 pour rejeter ou pas l'hypothèse d'identité entre deux distributions, celle du CDV de l'IRIS et celles moyennes des segments. Le degré de liberté du χ^2 est égal à $n-1=6$. On peut étendre ce principe à l'extrapolation à d'autres variables [1] On utilise la mesure du χ^2 pour rejeter ou pas l'hypothèse d'identité entre deux distributions, celle du CDV de l'IRIS et celles moyennes des segments. Le degré de liberté du χ^2 est égal à $n-1=6$. On peut étendre ce principe à l'extrapolation à d'autres variables La table χ^2_6 du indique un seuil

[1]Les techniques d'extrapolation des données peuvent être mises à profit pour évaluer de nouvelles données. En combinant des données spécifiques sur l'IRIS issues de sondage et des informations agrégées au niveau du territoire, on peut estimer pour le reste du territoire une nouvelle mesure. La technique la plus appropriée est celle des réseaux de neurones qui utilisent aisément les données numériques et symboliques. Le test statistique devient un χ^2 à $\Pi_j(N_j-1)$ degrés de liberté, N_j étant le nombre d'attributs de la variable j.

de rejet égal à 12,59 pour un risque d'erreur de 5%.

Les trois segments initiaux sont découpés chacun en deux sous populations : celle ayant un CDV de l'IRIS proche du profil moyen du potentiel et celle correspondant à une position et un statut social divergeant du CDV du profil moyen de l'enseigne. Le tableau 7 donne les répartitions du nombre de clients appartenant aux trois segments en fonction de l'éloignement au CDV moyen du segment calculé par le χ^2. Une plus grande dispersion des typologies s'avère sur la cible des clients à moyen et faible potentiel. En revanche, pour le segment " Potentiel Fort ", un ensemble homogène se dégage.

En nombre	Clients acceptés		Clients rejetés		Total	
	Nombre de clients	Nombre d'IRIS d'origine	Nombre de clients	Nombre d'IRIS d'origine	Nombre de clients	Nombre d'IRIS d'origine
Potentiel Fort	138	37	70	30	208	67
Potentiel Stagnant	211	87	212	96	422	183
Potentiel Faible	252	50	280	99	531	149
Total	601	174	562	225	1163	399

TAB. 7 – *Répartition de la population sur l'axe des "Potentiel Fort", "Potentiel Stagnant" et "Potentiel Faible" et sur l'axe des Cycles de Vie Moyen. Le tableau en haut donne les comptages et en dessous les pourcentages.* $51,7\%$ *de la population répartie sur* $0,3\%$ *du territoire contribue à donner un profil moyen des trois segments Potentiel.*

Techniquement, il est conseillé de répéter l'exercice itérativement en écartant les clients qui s'écartent de la moyenne de CDV et en relançant le processus sur la population ainsi filtrée. Dans cet exemple, trois itérations suffisent pour disposer de sous-populations stables. Le tableau 8 donne le résultat final après épuration des IRIS divergeant. A noter que les clients écartés des segments peuvent être réinjectés dans les autres segments.

En conclusion, on construit le tableau 9 qui représente le tableau de bord de ciblage des clients. Les trois axes d'analyse sont, d'une part les trois segments des potentiels d'achat à court et moyen terme (" Potentiel Fort ", " Potentiel Stagnant ", " Potentiel Faible "), d'autre part l'axe d'analyse de la valeur du client (trois catégories : Grand, Moyen, Petit) et enfin l'axe d'appartenance au CDV type de l'enseigne.

	Clients acceptés		Clients rejetés	
	% Client	%IRIS	% Client	% IRIS
Potentiel Fort	92,1%	17,5%	7,9%	1,6%
Potentiel Stagnant	97,5%	14,4%	2,5%	1,5
Potentiel Faible	97,2%	19,6%	2,8%	1,1%
% Affection VS TOTAL	41,8%	21,6%	1,7%	1,8%

TAB. 8 – *Filtre sur les IRIS clients ne contribuant pas au Cycles de Vie Moyen des « Potentiel Fort », « Potentiel Stagnant »et « Potentiel Faible », après* 3 *itérations.* $41,8\%$ *de la population de l'échantillon contribue à donner un profil moyen des segments. Les clients sont répartis sur* $21,6\%$ *du territoire de la marque.*

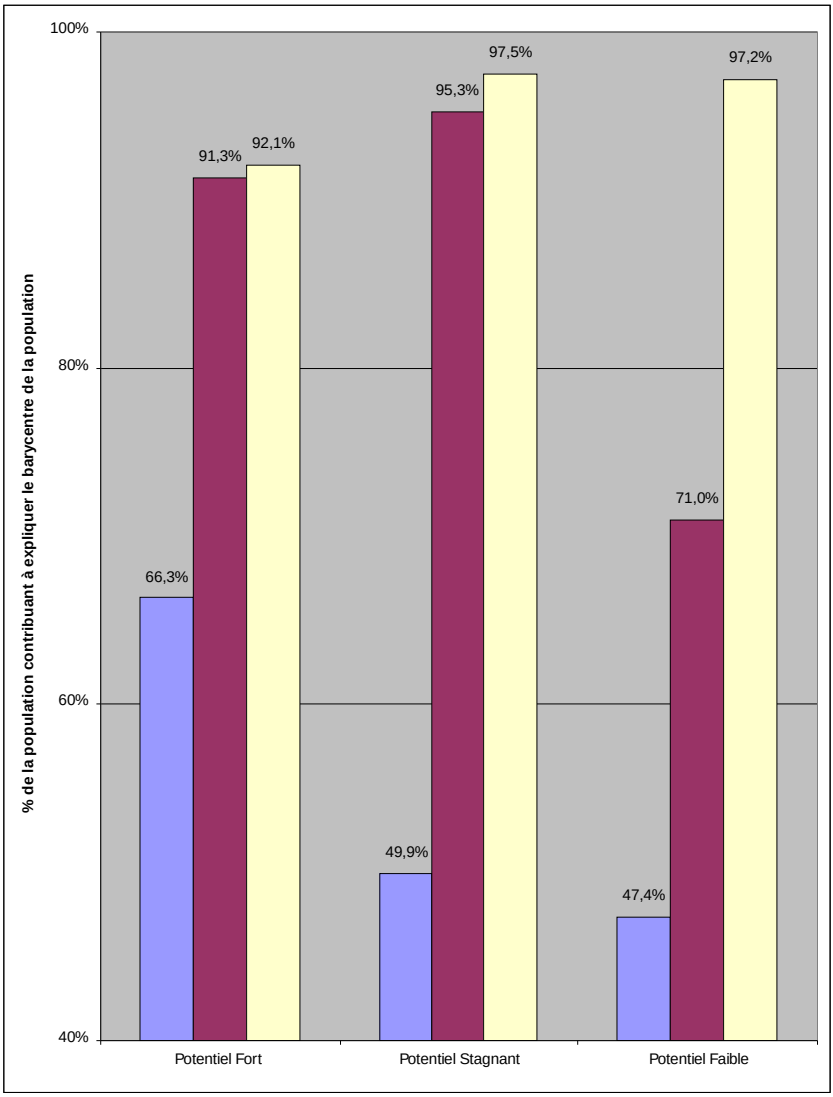

FIG. 1 – *Filtre sur les IRIS clients ne contribuant pas au Cycles de Vie Moyen des « Potentiel Fort », « Potentiel Stagnant »et « Potentiel Faible », après 3 itérations. Représentation graphique associée au tableau 8.*

Les répartitions respectivement en nombre, des probabilités a posteriori et de l'entropie sont données sur ces trois axes d'analyse. La probabilité a posteriori est identique à celle calculée

Répartition en nombre		Grand	Moyen	Petit	Total
Potentiel fort	CDV Moyen	105	7	4	116
	CDV Divergent	2	5	3	10
Potentiel Stagnant	CDV Moyen	48	119	29	196
	CDV Divergeant	6	5	1	12
Potentiel Faible	CDV Moyen	21	59	93	173
	CDV Divergeant	4	6	5	15
		186	201	135	522
Probabilités a posteriori		Grand	Moyen	Petit	Total
Potentiel Fort	CDV Moyen	91%	6%	3%	100%
	CDV Divergeant	20%	50%	30%	100%
Potentiel Stagnant	CDV Moyen	24%	61%	15%	100%
	CDV Divergeant	50%	42%	8%	100%
Potentiel Faible	CDV Moyen	12%	34%	54%	100%
	CDV Divergeant	27%	40%	33%	100%
Incertitude		Grand	Moyen	Petit	Total
Potentiel Fort	CDV Moyen	0,09	0,17	0,12	0,38
	CDV Divergeant	0,32	0,35	0,36	1,03
Potentiel Stagnant	CDV Moyen	0,34	0,30	0,28	0,93
	CDV Divergeant	0,35	0,36	0,21	0,92
Potentiel Faible	CDV Moyen	0,26	0,37	0,33	0,96
	CDV Divergeant	0,35	0,37	0,37	1,09

TAB. 9 – *Répartition de la population sur les axes des " Potentiel Fort ", " Potentiel Stagnant " et " Potentiel Faible ", sur l'axe Grand, Moyen, Petit et sur l'axe des Cycles de Vie Moyen (CDV Moyen ou CDV Divergeant).*

dans les tableaux 2 et 5 :

$$P\left(\frac{Potentiel\ Fort}{CDV Moyen\ et\ Grand}\right) = P\left(\frac{CDV Moyen\ et\ Grand}{PotentielFort}\right) \times \frac{P\left(PotentielFort\right))}{P\left(CDV Moyen\ et\ Grand\right)}$$

Avec l'entropie égale à 0,38 de la probabilité d'affecter un client dans le segment " Potentiel Fort " sachant son appartenance au profil moyen du marché et son appartenance à la catégorie Grand Moyen Petit, soit la probabilité P(Potentiel Fort$/CDV$Moyen, GMP), comparée à l'entropie des tableaux 1 et 2, respectivement égales à $0,70$ et $0,74$, l'ajout de l'information CDV augmente fortement la certitude de prendre une bonne décision, c'est-à-dire à avoir une population fiable et robuste pour la désignation d'une cible marketing. La probabilité de décision sur le segment " Potentiel Fort " augmente au fur et à mesure que des données sont agrégées, passant de $0,74$ à $0,91$.

Le tableau 9 constitue le tableau de bord pour décider si le niveau de connaissance que l'on a de la cible est suffisamment important pour prendre des décisions au lancement d'un stimulus marketing [Becker, 1999]. Les notions d'événements et de sélection du moment du contact n'ont pas été abordées. Mais elles supposent que des sources d'information à terme donneront des connaissances supplémentaires qui pourront éventuellement améliorer une action marketing adaptée au moment. C'est pourquoi, il est nécessaire de rendre dynamique ce type de tableau de bord en industrialisant les calculs de scoring et de segmentation. L'exploitation industrielle exige que la mécanique de segmentation puisse être reconduite régulièrement pour

détecter de nouveaux segments pertinents au cours du temps. L'actualisation au rythme hebdomadaire du tableau de bord, permet de mesurer et de détecter tout événement de la société et de l'environnement : les marchés évoluent, les individus changent, les mouvements socio-démographiques s'accentuent. Le calcul du niveau d'incertitude sur les segments permet de détecter les phénomènes rares et surgissant.

5 Conclusion

Les systèmes d'information dédiés à la gestion de la relation avec les clients offrent des mesures qui délivrent toutes les données nécessaires à l'évaluation d'un événement marketing et la mise en place de stratégies commerciales de la relation client. Si l'information et le caractère granulaire des données recueillies en " front office " de l'enseigne ne permettent pas de créer des axes d'analyses pertinents, alors l'incertitude que l'on a sur les décisions marketing à prendre est réelle et doit être mesurée.

Dans cet article, nous opposons deux types d'approche pour mesurer la qualité d'une population cible : la première approche, proche du statisticien, qui cherchera à mesurer les erreurs des techniques mathématiques, ceci limitant la confiance que l'on a dans l'affectation des personnes dans la cible à celle des outils de calcul, et la deuxième approche, celle du dataminer, qui exploitera de nouvelles informations afin de réduire les incertitudes qui naturellement naissent d'une décision de ciblage [Merckt, 1995].

Les paragraphes 2 et 3 sont consacrés à démontrer qu'il est nécessaire d'enrichir les données " client " par des données extérieures. L'expérience illustre que plus les données sont de sources différentes, plus la discrimination entre sous population est prononcée. Le domaine des données géographiques est exploité ici, très riche en information. On détaille la technique pour mettre au niveau du client les informations socio-démographiques. Cette technique est simple à mettre en place et peut être avantageusement utilisée pour la prospection des clients se trouvant dans des zones de forte similitude avec les typologies moyennes du CDV : pour un ensemble de prospects dont la position géographique est connue, il sera jugé " affine " à la marque s'il habite dans un IRIS compatible avec le profil moyen. Le test du χ^2 valide l'adéquation entre le profil du client et la cible marketing lui correspondant. Cette approche permet d'obtenir une signification renforcée et précisée des segments par l'ajout de nouvelles informations et de réduire l'incertitude sur l'affectation des clients dans les segments. Elle contribue à l'approche marketing : avoir une vision claire et explicite de la population ciblée sans perdre la vision client. Les recherches vont vers l'industrialisation de ces techniques. Les approches sont diverses et celles statistiques ([Vapnik, 1998]) ou celles de combinaison de connaissances ([Desarachy, 1994]) sont les bases des futurs développements.

Cette approche poussée à l'extrême, en ajoutant et combinant de nouveaux critères (la multi-segmentation) doit comporter un volet d'optimisation qui utilise les multiples segmentations disponibles pour retrouver une segmentation optimale qui favorise la lecture des données composant les segments plus homogènes. Un exemple d'outil de multi-segmentation se trouve dans [Dunis and al, 2001]. Ce type d'outil sert à l'optimisation des axes d'analyse dans un cube OLAP - Online Analytical Processing.

Références

Agrabal, R., Imielinski, T. and Swami, A., 1994. Database mining : A performance perspective. IEEE Transactions on Knowledge and Data Engineering, Special issue on Learning a Discovery in Knowledge-Based Databases, 1993.

Anand, S.S., Bell, D.A., Hughes, J.G., 1996. Evidence-Based Discovery of Association Rules. Internal Report, Faculty of Informatics, University of Ulster.

Becker, A., Naïm, P., 1999. Les Réseaux Bayésiens : Modèles Graphiques de Connaissance, Éditions Eyrolles, ISBN 2-212-09065-X, Paris.

Bercher, J-F., Le Besnerais, G. and Demoment, G., 1994. The Maximum Entropy on the mean method, noise and sensibility, In S. Sibisi and J. Skilling, editors, Maximum Entropy and Bayesian Methods, Dordrecht, Kluwer.

Chauvin, S., Jañez Escalada, L., 1997. Tracking Knowledge Data Bases : Fusion of data Software, International conference of Model Recognising Shape, C.I.R.M. institution France, Marseille.

Desarachy, B., 1994. Decision Fusion, IEEE Computer Society Press.

Dunis, C. L. ,Laws, J. et Chauvin, S., 2001. FX Volatility Forecasts and the Informational Content of Market Data for Volatility, Financial Review of Forecasting Model.

Govaert, G., 2003. Analyse des données, Traitement du Signal et de l'image, Hermes Science.

Jaynes, E.T., 1957. Information Theory and Statistical Mechanics, Physics Review, Vol 106, No 4, 620-630.

MacKay, D.J.C.,1995. A Short Course in Information Theory, Cavendish Laboratory.

Van De Merckt, T., Decaestecker,C., 1995. Multiple-Knowledge Representation in Concept Learning, in Lavrac N. and Wrobel S. (Eds), Machine Learning : ECML-95, Lecture Notes in Artificial Intelligence 914, Springer Verlag, Berlin, Heidelberg, New York, 200 - 217.

Pyle, D., 1999. Data Preparation For Data Mining, Morgan Kaufmann Publishers, Inc. San Francisco, California.

Saporta, G., 1978. Théories et méthodes statistiques, Princeton University Press.

Shannon, C.E., 1948. A mathematical theory of communication, Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, July and Octobe.

Vapnick, V.N., 1998. Statistical learning theory, John-Wiley and Sons, New-York, USA.

Summary

This article presents a logical market research to target efficiently clients in terms of fidelisation and for the prospection to get deeper information about the brand market. We underline the approach of the dataminer whose purpose is to improve the knowledge and the relevance of the customer's target. This illustrates the need to dicrease the uncertainty when taking decisions regarding the target market for a campagne. We use here the entropy of posterior probabilities to decide the real property to the segment of client. Applicated on a data base, a succession of segmentation is made with an objective to reduce the indicator of uncertainty. The uncertainty decrease while the information is added from different sources indeed, the experience is made in this article, on contexctual datas of the client's life. We propose a simple method that is to regroup geographic zones which have a strong affinity to the brand potential and an algorythm of optimisation of the micro-zones nominated for received a marketing stimulus. The advantage of this method is to make the population segmentation stronger for a better both interpretation and integration of the activity of the marketing.