Modélisation prédictive de l'incidence et des montants d'achat en marketing direct. Une comparaison a partir de variables RFM.

Michel CALCIU, Francis SALERNO

Introduction

Marketing direct et Datamining

En marketing direct et vente à distance, le ciblage des clientèles est un domaine d'application privilégié des techniques de datamining. Les réseaux de neurones artificiels (RNA), les arbres de classification (CHAID, CART) et la régression logit sont les méthodes souvent utilisées dans ce but (Linder et al.2004) et quelques recherches ont été consacrées à la comparaison des performances des réseaux de neurones et des méthodes statistiques plus traditionnelles de ciblage. Malgré certaines convergences, les résultats de ces travaux incitent cependant à développer les études comparatives de cette nature.

Revue de littérature

Ainsi, les résultats de Kumar et al. (1995) ne permettent pas de départager les RNA et la régression log-linéaire. Lix et Berger (1995) montrent qu'un réseau de neurones particulier est plus performant que les méthodes fondées sur la régression tandis qu'un autre réseau l'est moins. Desmet (1996) confronte les

réseaux de neurones aux modèles prédictifs plus traditionnels dans le domaine du marketing direct de collecte de fonds. Zahavi et Levin (1997) obtiennent des résultats sensiblement identiques avec des modèles de régression logistique et les réseaux de neurones et concluent que « ces résultats ne sont pas encourageants pour l'approche par réseau de neurones car le processus de configuration et de préparation n'est pas simple ». Plus récemment, Potharts et al (2001) comparent les RNA à la méthode CHAID et à la régression logistique et trouvent que les RNA ont des performances prédictives au moins égales à celles des autres modèles. Madeira S. et Sousa J. M. (2002) obtiennent de meilleures performances pour les RNA que pour régression logistique ou pour les arbres de classification (modèle CART). Enfin, Linder et al. (2004) comparent les performances prédictives des RNA, des arbres de classification et de la régression logistique dans un cadre expérimental composé de situations caractérisées par des tailles d'échantillon différentes et par des complexités de données variables.

Une autre direction de recherche

Une autre direction de recherche étend l'analyse au-delà des modèles de choix binaires à des modèles de choix discrets et continus. Levin et Zahavi (1998) comparent les performances prédictives en termes de rentabilité, du modèle logit avec celles de la régression ordinale et de plusieurs modèles de choix continu. Otter, Van der Scheer et Wansbeek (1997) développent un modèle qui prédit simultanément l'incidence et le montant des dons pour optimiser le ciblage dans des opérations de collecte de fonds. On retrouve des tentatives de combiner en deux étapes des modèles de l'incidence de l'achat avec des modèles qui expliquent les montants d'achat chez Courtheoux (1987), Van der Scheer (1998), Levin et Zahavi (1998) et Spring (2001).

Objet de recherche

Ce papier s'inscrit dans cet ensemble de travaux développés en marketing direct. L'objectif est ici de comparer les performances prédictives de plusieurs méthodes de ciblage discrètes (régression logit, probit, ordinale, réseaux de neurones,CART) et continues (régression linéaire, tobit et en deux étapes), en utilisant des variables de la famille des variables comportementales d'achat RFM (Récence, Fréquence, Montant) dans un catalogue à caractère saisonnier. Les données, les variables et l'échantillon d'étude sont présentés dans un premier temps. Les performances des méthodes de ciblage sont ensuite comparées sur la base de critères statistiques et sur celle des critères de rentabilité. Les résultats sont aussi comparés à ceux de ces travaux similaires réalisés dans des contextes différents.

Variables RFM et système R

Dans le champ du marketing direct et de la vente à distance, les travaux sur les

variables RFM et sur leurs contributions dans les modèles prédictifs constituent un domaine de recherche à part entière (tableau 1) et cette étude y apporte les résultats d'une perspective comparative[1]. Enfin, pour le traitement des données, cette étude utilise le système R, c'est-à-dire la version « open source » du système S, qui a comme équivalent commercial le logiciel S-plus. L'exemple d'utilisation de cet outil complet et puissant, mis à la disposition de tous, devrait aussi favoriser le développement d'autres comparaisons[2].

Tableau 1 – Etudes sur la modélisation du réachat en vente par correspondance publiées dans des revues académiques (Source Van den Poel, 2003, p. 5)

Présentation des données, définition des variables, échantillon

Les données

Les données couvrent les comportements d'achat et de réachat d'une cohorte de 315 000 clients et prospects d'un catalogue bi-annuel (printemps, automne) observée durant sept saisons. L'évolution du nombre et de la valeur des commandes est illustrée en Figure 1

Figure 1 –Evolution des commandes durant sept saisons

Focaliser sur le comportement de réachat

Grâce à des taux de transformation importants (prospects qui en passant leur première commande deviennent « clients »), le nombre de commandes se maintien à un niveau relativement constant pour chaque type de saison (printemps/automne). Afin de suivre le comportement de ré-achat, on sélectionne uniquement dans la base des clients et prospects les personnes ayant commandé la première saison, en éliminant ainsi les nouveaux clients qui se sont manifestés par la suite. On élimine ainsi 186190 individus qui n'ont jamais commandé et les individus dont le premier achat est ultérieur à la première saison analysée. Cela conduit à un liste exploitables de 39890 clients. On constate une certaine régularité dans les comportements de réachat de membres de cette cohorte (Figure 2).

Figure 2 – Comportement de réachat des clients ayant commandé la première saison

On peut constater l'épuisement progressif du potentiel de la cohorte, qui justifie le renouvellement de la base de clientèle par des actions de prospection. On constate aussi le caractère saisonnier du catalogue ainsi qu'une fréquence plus

importante des commandes au printemps qu'en automne. Ces régularités dans le nombre des commandes se retrouvent aussi dans les montants des commandes.

Définition de variables RFM

Sept variables de type RFM ont été construites à partir de l'historique d'achat : deux sont des mesures de récence (R1 et R2), deux autres mesurent la fréquence (F1 et F2), et trois variables expriment le montant (M1, M2, M3). Elles sont décrites dans le tableau 2.

Tableau 2 – Définition des variables de la famille RFM

Les valeurs des variables RFM ainsi définies ont été calculées pour les sept saisons. Elles serviront pour expliquer l'incidence d'achat.

Echantillons d'estimation et de validation

Avant toute comparaison des performances, chacun des modèles ou méthodes et utilisés exigent l'application de techniques de validation.

Après élimination des cas exceptionnels les 36630 clients ayant commandé pendant la première saison sont organisés en deux échantillons, les premiers deux tiers forment l'échantillon d'estimation sur lequel les modèles seront calibres et le dernier tiers est mis à part comme l'échantillon test pour vérifier la performance prédictive des modèles

Tableau 3 - Descriptif de la BD clients et des échantillons d'estimation et validation

Modèles de l'incidence de l'achat

Présentation

Un premier groupe de méthodes est constitué par des modèles de prévision de l'incidence de l'achat. Ils ont comme variable expliquée ou dépendante une variable binaire, qui est égale à un quand il y a achat et zéro autrement. Les performances prédictives de plusieurs modèles qui font partie de cette catégorie sont analysées à partir de critères autorisant la comparaison de méthodes aussi variées. Il s'agit des modèles régression logit, probit, des RNA et de CART.

Application logit et probit.

En plus de leur proximité avec les RNA, les modèles logistiques de type probit

ou logit ont l'avantage d'être bien adaptés aux problèmes de décision binaire (achat/non-achat) et d'éviter les désavantages du modèle linéaire de probabilité. Ils supposent l'existence d'une variable latente qui mesure la propension de répondre. Elle dépend d'une série de caractéristiques individuelles représentées dans ce contexte par les variables RFM.

La seule différence entre les modèles de réponse logit et probit réside dans la distribution du terme d'erreur qui suit une loi normale pour le modèle probit et une loi logistique pour le modèle probit.

Par rapport aux RNA, ces modèles ont l'avantage d'être faciles à interpréter et ils sont connus pour être robustes et donner de bons résultats dans les études comparatives. En R, la régression linéaire généralisée (GLM) a été utilisée pour calibrer des modèles logistiques en utilisant les options LOGIT et PROBIT.

Tableau 4 – Estimation de deux types de modèles qui offrent des coefficient interprétables.

L'analyse du tableau 4 montre que la fréquence moyenne des commandes et la récence du dernier achat sont dans l'ordre les catégories de variables RFM qui ont le plus grand effet sur la probabilité d'achat. Les coefficients des variables qui expriment le montant et l'ancienneté de la relation avec le client ne semblent pas significatifs statistiquement. La plupart de ces coefficients deviennent significatifs quand on considère un décalage de deux semestres entre la variable qui exprime l'incidence de l'achat et les variables RFM. Ce décalage est justifié par le caractère saisonnier du catalogue analysé. L'alternance de signe qu'on observe dans les coefficients de la fréquence des dernières commandes quand on applique un décalage de un et deux semestres, montre que les clients ont tendance de ne pas commander deux saisons de suite et d'alterner le réachat d'une saison à l'autre.

Application des réseaux de neurones.

Le RNA utilisé ici est de type feedforward entraîné avec l'algorithme de rétro-propagation classique (Rumalhart et McCleelland, 1986). « On peut considérer les réseaux de neurones feedforward comme des modèles de régression non-linéaires dont la complexité peut être changée » (Bentz et Merunka, 2000, p.183). Dans leur forme la plus simple ils se composent d'une couche d'entrées et d'une seule sortie. Si la fonction de transformation des outputs est sigmoïde, ce modèle de réseaux de neurones devient strictement équivalent au modèle binomial logit.

Figure 3. Le modèle logit binomial comme neurone artificiel (ou un réseau de neurones sans couche cachée et entrées connectées à la sortie)

La structure représentée dans la figure 3 est la forme neuronale d'un modèle

logit calibré à partir des données réelles utilisées dans cette étude. Elle peut être vue comme un neurone artificiel. Un neurone artificiel est caractérisé par des connexions d'entrée (qui représentent les synapses de la cellule et sa dendrite), une valeur de biais (le niveau d'inertie du neurone), un niveau d'activité (qui représente l'état de polarisation du neurone), une valeur et des connexions de sortie (les projections axonales du neurone). A chaque connexion on associe un poids (intensité synaptique) qui détermine l'effet des entrées sur le niveau d'activation de l'unité. Les poids peuvent être positifs (excitants) ou négatifs (inhibants). Le signal de sortie du neurone est donné par une expression qui représente la fonction d'activation (ici la fonction logistique).

Les couches des RNA

Les neurones sont organisés en couches. Il y a trois types de couches dans un réseau de neurones : une couche d'entrée représentant les variables d'entrée (ici les variables RFM), une ou plusieurs couches cachées et une couche de sortie avec une ou plusieurs unités représentant les variables de sortie. Les unités sont connectées avec des intensités de connexion ou poids variables comme le montre la figure 4.

Figure 4 - RNA avec deux neurones dans la couche cachée et un seul neurone en sortie appliqué à un problème d'incidence d'achat

Le modèle RNA utilisé

Le modèle final de RNA comportait sept variables RFM dans la couche d'entrée, un neurone dans la couche de sortie et quatre neurones dans la couche cachée. Une couche cachée est suffisante pour capter les non-linéarités qui peuvent se trouver dans des problèmes de ciblage de ce type. Pour rendre cette étude comparable avec d'autres études utilisant le même type de variables RFM, le nombre de neurones dans la couche cachée a été fixé à quatre. Cela semble suffire et permet d'éviter des risques de sur-calibrage (overfitting) qui ont un effet négatif sur le pouvoir de généralisation et implicitement sur le pouvoir prédictif du RNA. Plusieurs essais avec un plus grand nombre de neurones ont été réalisés mais les résultats sur la qualité de l'ajustement aux données ne se sont pas améliorés de manière significative.

Le nombre de poids à estimer dans une telle architecture est égale à : [(1 + Nombre d'entrées)*Nombre de neurones cachés] + [(1 + Nombre de neurones cachés)*Nombre de sorties], c'est-à-dire (1+7)*4 + (1+4)*1=37.

L'estimation du modèle

Estimation du modèle. Ces poids sont initialisés aléatoirement dans l'intervalle (-0.1,0.1). Le processus d'entraînement s'arrête après un nombre d'époques

fixées ici à 500, ou avant si le niveau de précision est atteint.

En parallèle un autre modèle RNA, qualifié de multiple est obtenu en entraînant successivement et sur les mêmes données dix modèles RNA pour retenir le meilleur. Le nombre d'époques est fixé dans ce cas à 200. Les deux modèles sont complémentaires car le deuxième en reprenant l'entraînement plusieurs fois avec des poids qui s'initialisent avec de valeurs différentes permet d'éviter certains optimums locaux.

La standardisation des variables en entrée améliore la stabilité du processus d'entraînement, car le réseau n'est pas obligé d'opérer avec des poids qui ont des ordres de grandeur différents.

Tableau 5 – Le poids des entrées sur les couches cachée du RNA

Tableau 6 – Les poids de la couche cachée du RNA vers les outputs

On peut observer la répartition des poids positifs (excitants) marqués en gras et des poids négatifs (inhibants) entre les variables RFM et les quatre couches cachées H1 .. H4 (tableau 5) et des couches cachée vers la sortie (tableau 6).

Application de CART.

Une dernière catégorie de méthodes utilisée dans cette étude comparative est représentée par la méthode des arbres de classification et de régression CART. Développée par Breiman et al (1984), elle remplace souvent les autres méthodes de classification explicative telles que CHAID (Kass, 1980) ou AID (Sonquist, 1970) car elle n'est pas limitée à des variables indépendantes catégorielles (nominales) mais accepte aussi des variables continues[3]. La fonction rpart, c'est-à-dire l'implémentation de la méthode CART en R a été utilisée pour obtenir un arbre de classification et de régression à partir des données.

Il s'agit d'un arbre de classification car la variable dépendante est catégorielle (achat/non-achat) et que l'on traite les données individuelles. Le modèle est obtenu en utilisant toutes les variables RFM ; le critère de séparation est la réduction de l'entropie, un minimum de 20 observations dans un nœud terminal (feuille), un paramètre de complexité de 0,001 et une profondeur maximum de l'arbre de 30.

L'arbre à 26 noeuds de la saison 7 utilise dans l'ordre les variables R1, F1, et R2 pour séparer les segments ; les variables F2 et R2 apparaissent à des niveaux inférieurs dans la structure de l'arbre. La variable F1 est le plus souvent en première position, suivi par R1 dans la majorité des saisons analysées. Cela est en accord avec le poids de ces variables dans les modèles logit et probit ajustés.

Dans CART tous les clients qui appartiennent au même segment sont traités de la même manière contrairement aux autres méthodes qui attribuent un score de vraisemblance de réponse différent à chaque client, ce qui leur donne une meilleure granularité par rapport à CART.

Critères statistiques d'évaluation des performances prédictives

Présentation

La grande variété des méthodes de ciblage utilisées exige un choix de mesures de performances suffisamment génériques pour permettre des comparaisons. Toutes les méthodes de ciblage évoquées parviennent à classer les clients dans l'ordre de leur probabilité d'achat ou de réachat estimée. En fait elles calculent pour chaque individu un score qui peut être une probabilité « a posteriori » dans le cas du modèle logit ou probit, ou une autre mesure comme par exemple le logit, ou l'écart entre la probabilité moyenne et la probabilité qui caractérise une cellule comme dans le cas d'une méthode RFM classique.

Lorsqu'une valeur de séparation absolue est choisie (cutoff), tous les clients ayant un score inférieur sont classés non-acheteur ; ceux ayant un score supérieur sont classés acheteur. Le résultat de la classification peut être résumé par une matrice des confusions (Morrison, 1969).

Tableau 7 - La matrice de confusion

Dans de tels problèmes de classification binaire deux types d'erreurs peuvent être distingués (tableau 7). Les statisticiens appelleraient erreur de type 1, la catégorie faux positif, qui correspond ici au fait de classer les non-acheteurs comme acheteurs et erreur de type 2, la catégorie faux négatif le faite de classer des acheteurs comme non-acheteurs. Plusieurs mesures peuvent être tirées de la matrice de confusion (Bradley, 1997):

Précision prédictive, sensibilité et spécificité

La précision prédictive est exprimée par le pourcentage correctement classé (PCC) c'est-à-dire la proportion des résultats positifs et négatifs correctement classés dans l'ensemble des prédictions.

$$PCC = (VP+VN)/(VP+FN+VN+FN)$$

La sensibilité est la probabilité de prédire un résultat positif à condition qu'il soit positif.

Sensibilité = VP/(VP+FN)

La spécificité est probabilité de prédire un résultat négatif à condition qu'il soit vraiment négatif.

Spécificité = VN/(VN+FP)

La courbe ROC

Pour rendre le critère de performance indépendant de la valeur de séparation sélectionnée (cutoff) on calcule et trace la courbe qui représente la sensibilité (sur l'axe vertical) par rapport à un moins la spécificité (sur l'axe horizontal) pour toutes les valeurs de séparation (cutoff) possibles. Le numérateur est aussi appelé pourcentage de hit «positif » et le dénominateur est aussi dénommé probabilité de « fausse alarme » (Green et Swets, 1966). Cette courbe[4] est la courbe des « caractéristiques d'opération du récepteur" (ROC - receiver operating characteristics).

La précision prédictive de procédures de classification telles que celles comparées ici peut être mesuré par l'aire sous la courbe ROC connue sous les initiales AUC ou AUROC. AUC est une mesure de performance qui ne dépend pas de la valeur de séparation (cutoff). Ses valeurs s'étalent de 0,5, la performance d'une procédure de classification non discriminante, aléatoire à 1.0, la performance parfaite.

Les critères choisis pour évaluer les performances prédictives des modèles doivent être de nature prédictive et se différentier ainsi des critères traditionnels de mesure de la qualité d'une estimation comme ceux basés sur les tests d'hypothèses (qui vérifient le caractère significatif des paramètres estimés) ou même du taux de re-substitution (donné par le PCC appliqué à l'échantillon d'estimation).

Intérêt des critères prédictifs

Le choix de tels critères prédictifs est justifié par l'objectif commun à toutes ces méthodes qui modélisent la réponse pour prédire l'incidence d'achat et par la présence de prédicteurs corrélés (comme c'est souvent les cas pour les variables RFM). Quand les prédicteurs sont corrélés, les tests d'hypothèses classiques sont dénaturés par la multicolinéarité qui augmente artificiellement les variances des estimateurs, ce qui peut conduire à des paramètres non-significatifs. Mason et Perreault (1991) indiquent que l'approche prédictive n'est pas affectée par ce phénomène. De plus l'approche prédictive a gagné du terrain par rapport à l'approche fondé sur les tests d'hypothèses en raison de la disponibilité croissante d'ensembles de données de grandes tailles car, sur des grands échantillons, presque tous les coefficients deviennent statistiquement significatifs (Granger, 1998).

L’utilisation de tels critères prédictifs d’évaluation exige l’utilisation d’un échantillon test séparé de l’échantillon d’estimation pour corriger le biais positif qui, selon Morrison (1969) est induit par la classification des mêmes individus utilisés pour estimer le modèle de classification (estimation du taux de re-substitution).

Comparaison des performances selon des critères statistiques

Dans cette étude les données indiquant le comportement de réachat ont été scindés en deux. On a obtenu ainsi un ensemble d’entraînement ou d’estimation sur lequel les modèles ont été calibrés et un ensemble de validation des performances prédictives.

Tableau 8 - L’aire sous la courbe ROC (AUC) et les performances prédictives de plusieurs méthodes de ciblage

On observe que les modèles RNA s’ajustent légèrement mieux que les autres modèles. Les modèles logit et probit ont des performances comparables. La méthode CART enregistre les moins bons résultats. Les performances des RNA sur l’ensemble de validation ne sont pas significativement supérieures aux performances des modèles logit. Des analyses répétées effectuées sur de plus petits échantillons (4000 individus) montrent que les performances des RNA sont souvent légèrement inférieures à celles des modèles logit ou probit sur l’échantillon de validation, tout en restant supérieures sur l’échantillon d’estimation. On constate aussi que la qualité des modèles décroît avec les saisons car moins la campagne est récente moins il y a de données disponibles dans l’historique d’achat.

Comparaison des performances selon des critères de rentabilité

Gainchart des modèles d'incidence de l'achat

Les critères de performance économique utilisés en complément aux critères purement statistiques nous permettront d’étendre la comparaison des performances prédictives au-delà des modèles d’incidence de l’achat analysés jusqu’ici vers des modèles de choix polytomique et continu.

Pour vérifier les performances économiques des modèles nous allons suivre la démarche utilisée par Levin et Zahavi (1998). Elle suppose le calcul de tableaux des gains (gainchart) pour chaque modèle. Ces tableaux présentent le résultat des estimations, appliquées à l'échantillon validation, dans l'ordre décroissant des réponses par décile.

A titre d’illustration le tableau des gains du modèle logit est présenté.

Tableau 9 – Tableau des gains d’un modèle de choix binaire (le modèle logit)

La probabilité dans la colonne 1 représente la limite inférieure de chaque décile ou autrement dit l'équivalent de la probabilité d'achat du dernier individu dans chaque décile. Le tableau donne aussi le profit évalué au niveau de chaque décile. Il est calculé en multipliant le nombre de répondants dans le décile avec le revenu moyen par commande, qui lui est égale à la marge fixée à 10% et multipliée au prix moyen de 405,7F et en déduisant les coûts de mailing (ici 10F). Le taux de réponse d'équilibre est égale ici a 10/(0,1*405.7)= 24,6%. Dans ces conditions il convient d'envoyer les mailings aux premiers 5 déciles ou à 6105 clients qui récupèrent 2565 réponses ou 75,2% des achats (3411) et un profit cumulé de 43019F.

La qualité de la prédiction du modèle logistique est exprimée par le pourcentage des acheteurs qui décline ici fortement avec les déciles, de 816 dans le premier décile à 144 dans le dernier décile. Comme dans Levin et Zahavi (1998, p.10) on peut montrer que le nombre prédit de répondants dans quasi chaque décile se trouve dans les limites de l'intervalle de confiance de 95%[5]. Ce constat est valable pour tous les modèles comparés dans ce papier.

Comparaison des performances de modèle de l'incidence de l'achat

Figure 5 – Comparaison de la rentabilité cumulée de plusieurs modèles de prévision de l’incidence de l'achat

Si on prend comme repère les modèles logit et probit qui offrent des résultats très proches on peut dire que les réseaux de neurones arrivent à mieux isoler les meilleurs clients et d’obtenir une baisse plus accentué du nombre de répondants d’un décile à un autre. Les arbres de classification et régression arrivent sur l’ensemble à de moins bons résultats même si pour les premiers deux déciles leurs performances dépassent même celles des réseaux de neurones.

Si les modèles logit et probit atteignent un profit cumulé maximum d’approximativement 43 milles francs au niveau du 5ème décile et arrivent à regrouper à ce niveau trois quarts des répondants, les réseaux de neurones arrivent après seulement quatre déciles à un profit cumulé maximum légèrement supérieur. Les arbres de régression et classification arrivent aussi à leur profit cumulé maximum après quatre déciles, seulement ce profit est de seulement 40 milles francs.

Un modèle de choix polytomique

La régression ordinale

La régression ordinale correspond à un modèle de choix discret ou les valeurs qui représentent les multiples choix expriment un ordre ou une préférence. Elle prend une position intermédiaire entre les modèles de choix binaire ou d'incidence de l'achat et les modèles de choix continu qui seront utilisés par la suite.

Si on choisit le montant des commandes comme variable à expliquer, on peut transformer cette variable continue en variable discrète en ordonnant ses valeurs dans des intervalles de choix mutuellement exclusifs et en attribuant à chaque intervalle une catégorie ordinale. Le modèle de régression ordinale peut être utilisé ensuite comme un « proxy » pour la régression linéaire (Levin et Zahavi, 1998, p.10). Si on se résume à seulement deux catégories ordinales: 0 pour les non-acheteurs et 1 pour les acheteurs, alors les prédicteurs qui résultent de la régression ordinale devraient être identiques à ceux de la régression logistique. En pratique on préférera d'utiliser plusieurs catégories de choix (0, 1, 2 3) affin d'obtenir des résultats plus nuancés.

Dans cet exemple les catégories de choix sont définies selon les critères évoqués dans le tableau 10.

Tableau 10 – La composition des intervalles de choix ordinal

Gainchart

Le modèle ne se résume pas uniquement à prédire le taux de réponse des clients mais il calcule aussi la probabilité de réponse pour chaque valeur de choix comme le montre le tableau de gains suivant.

Tableau 11 – Tableau des gains d'un modèle de choix polytomique (le modèle de régression ordinale)

A part d'estimer une probabilité de réponse générale par client, la régression ordinale calcule aussi une probabilité de réponse pour chaque modalité de choix. Les coefficients des variables explicatives sont les mêmes pour toutes les modalités de choix. Le seul élément qui diffère est le terme constant dans l'équation de régression.

Comme pour les meilleurs modèles de prévision de l'incidence de l'achat le taux de réponse d'équilibre (24,6%) est atteint déjà dans le quatrième décile. On peut sélectionner comme cible les quatre premiers déciles qui regroupent 2254 répondants (66,1% des acheteurs) repartis entre les trois catégories de choix de la manière suivante: 671, 807, 776 . En utilisant les montants moyens qui correspondent à chaque catégorie de choix et les coûts de mailing aux 4 premiers déciles (4*1221) le profit espéré devient :

10%*(671*171,5F + 807*336,3F + 776*750,1F) - 4*1221*10F = 48016 F

C'est un profit supérieur aux profits calculés pour les modèles d'incidence de l'achat.

On doit rester prudent quand on entame de telles comparaisons des performances en terme de profit entre des modèles de catégories différentes. Même si dans l'ensemble les probabilités de réponse qui séparent les déciles sont les mêmes, la manière dans laquelle on calcule le profit par décile diffère entre les deux catégories de modèles. Si pour les modèles de choix binaire le profit espéré est calculé à partir du revenu moyen, pour la régression ordinale le profit total espéré dépend de la distribution des revenus pour chaque alternative de choix (Levin et Zahavi, 1998, p. 11).

Modèles de choix continu

Présentation

A la différence des modèles de choix discrets présentés, qui utilisent la probabilité d'achat comme critère de sélection des clients cible, les modèles de choix continu utilisent le revenu espéré par client dans ce même but. Le profit total prédit dépend de la distribution des montants individuels par rapport aux quantiles (déciles).

La régression linéaire.

La régression linéaire est la méthode la plus facile à mettre en place pour estimer des modèles de choix continu. Elle est largement disponible sur une gamme large de machines (des calculettes aux ordinateurs) et de logiciels accessibles, comme par exemple les tableurs.

Les estimations obtenues par un modèle linéaire ne sont pas bornées par des seuils inférieurs et supérieurs et produisent des profits négatifs pour les individus qui se trouvent en bas de la liste et des profit supérieurs au seuil maximum pour ceux qui ce trouvent en haut de la liste. La somme des réponses (profit) réelles étant égale à la somme de réponses prévues on arrive à une surestimation des performances des meilleurs clients qui va compenser la sous-estimation des performances des "mauvais" clients qui pour certains se voient attribuer des valeurs négatives.

Gainchart

Tableau 12 – Tableau des gains d'un modèle de choix continu (le modèle de régression linéaire)

Le profit cumulé maximum est atteint au niveau du quatrième décile.

10%*(822*1221)- 4*1221*10F = 51550F

C'est un profit supérieur aux profits calculés pour les modèles de choix discrets. Le bon ajustement du modèle linéaire et la nature des données analysés place les performances prédictives de ce modèle au même niveau que celui des autres modèles de choix continu qui sont présentés par la suite. Ce résultat dans le cas particulier du modèle linéaire est différent de celui obtenu par Levin et Zahavi (1998) qui observent des faibles performances prédictives du modèle et les attribuent aux nonlinéarités qui caractérisaient leurs données. Il est évident que d'un point de vue théorique le modèle linéaire est moins adapté pour représenter les situations de choix continu pour de nombreuses raisons dont quelques-unes ont été évoquées auparavant.

Le modèle Tobit.

La régression linéaire considère que la variable dépendante est observée dans tous les cas. En réalité en marketing direct on observe la variable dépendante le montant des commandes uniquement pour les gens qui ont commandé suite à une campagne. Il n'y a pas d'observations du montant d'achat pour ceux qui n'ont pas commandé. On dit que la variable de réponse est "censuré" à gauche. La variable prend donc une valeur positive pour les répondants et la valeur zéro pour les non répondants et ne satisfait donc pas la condition d'être normalement distribué, comme l'exige la régression linéaire.

Le modèle Tobit est un modèle de régression qui prend en compte de manière explicite le faite que la valeur de la réponse est observée uniquement pour les répondants, qui constituent une minorité dans la population qui caractérise les applications de database marketing.

Même si légèrement inférieures au modèle linéaire les performances prédictives du modèle Tobit sont bien supérieures à celles du meilleur modèle discret (le modèle ordinal).

Régression en deux étapes.

Comme dans la régression linéaire le modèle Tobit suppose que la variable de réponse peut prendre toute valeur, même des valeurs négatives mais comme ces dernières ne sont pas disponibles car il n'y a pas des observations pour les non-répondants elles sont censurées et substitués par la valeur zéro. C'est une supposition qui n'est pas très réaliste car en marketing direct les réponses égales à zéro surviennent parce que certains clients choisissent de décliner l'offre et non pas parce que la réponse était négative et par conséquence censurée à zéro. Pour subvenir à ce problème que présente le modèle Tobit une solution alternative est d'estimer une réponse continue en deux étapes en

utilisant le modèle en deux étapes de Heckman (1979).

Dans une première étape on applique un modèle de choix binaire (par exemple le modèle logistique) à l'échantillon d'estimation. Dans l'étape suivante un modèle linéaire est estimé uniquement au niveau de répondants pour estimer la réponse (conditionnelle) espéré par client à condition que celui-ci soit un répondant

Ensuite un modèle la réponse (inconditionnelle) espérée par client est obtenu en multipliant la réponse conditionnelle par la probabilité d'achat. Le modèle à deux étapes souffre du biais de sélection car il est basé uniquement sur les répondants et ne constitue donc pas un échantillon aléatoire de la population. Pour subvenir au biais de sélection Heckman estime un facteur de correction qu'il utilise comme variable supplémentaire dans le deuxième modèle. Dans le cas d'une distribution logistique ce facteur est la probabilité de réponse estimé par le premier modèle.

Comparaison

Figure 6 – Comparaison de la rentabilité cumulée des modèles de prévision du montant d'achat discrets et continus

Comme chez Levin et Zahavi, le modèle de régression en deux étapes semble obtenir les meilleures performances prédictives selon des critères de rentabilité et confirme les résultats supérieurs enregistrés par les modèles de choix continus par rapport aux modèles de choix discrets.

CONCLUSIONS

Géneralités

L'objectif poursuivi dans cette recherche est d'analyser les performances de plusieurs modèles de prévision de l'incidence et du montant d'achat en marketing direct et de les comparer aux résultats d'autres recherches qui s'inscrivent dans la même direction. L'intérêt d'une telle comparaison est justifié aussi par la nature des données dont on dispose, qui place cette recherche dans un contexte de réachat caractérisé par des taux de réponse relativement élevés.

La majorité des modèles étudiés se classent en deux catégories, des modèles de choix binaire qui expliquent l'incidence de l'achat et des modèles de choix continu qui expliquent les montants d'achat. Le modèle de régression ordinale exprime une catégorie de modèles intermédiaire. Le positionnement de ce modèle de choix polytomique est confirmé par ses performances prédictives

selon des critères de rentabilité.

Choix binaire et critères statistiques

Les résultats obtenus au niveau des modèles de choix binaire (incidence de l'achat) sont en accord avec deux autres études récentes (Potharts et al 2001 ; Madeira et Sousa, 2002) réalisées à partir des mêmes catégories de variables dans des contextes de collecte de fonds. Potharts et al (2001) comparaient les RNA à la méthode CHAID et à la régression logistique et trouvaient que les RNA avaient des performances prédictives au moins égales à celles des autres modèles. Madeira S. et Sousa J. M. (2002) confrontaient quatre méthodes de sélection de cibles en marketing direct : la régression logistique, les RNA, les arbres de classification et la modélisation fuzzy et montraient que les RNA étaient plus performants que la régression logistique ou que le modèle CART. Les méthodes comparées se valent dans le contexte étudié ici. Les RNA s'ajustent en général mieux sur l'échantillon d'estimation (d'apprentissage) que les autres méthodes mais ils perdent cet avantage sur l'échantillon de validation. Il y a sur-apprentissage, les RNA arrivent à capter des phénomènes spécifiques à chaque échantillon et donc non généralisables aux autres. Des historiques d'achat plus longs ont un effet positif sur la qualité des prédictions.

Comparaison des performances selon des critères de rentabilité

Dans l'évaluation des performances prédictives selon des critères de rentabilité des modèles de choix discret et continu, cette recherche reprend la démarche et les modèles utilisés par Levin et Zahavi (1998), et arrive en gros à des résultats semblables. Les différences enregistrées sont en grande partie expliquées par la nature différente des données utilisées. Dans Levin et Zahavi, le nombre de réponses parfois trop faible contribue à pénaliser les performances du modèle intermédiaire polytomique par rapport au modèle binaire (logit), qui lui semble mieux adapté à de telles situations. Dans la situation de réachat qu'on présente, les trois catégories de modèles de choix : binaire, polytomique et continu arrivent à se distinguer très bien du point de vue de leur performances prédictives en terme de rentabilité.

Elargir la gamme des modèles

Si les modèles de choix binaire occupent une place plus importante dans la littérature, en réalité ils ne représentent qu'une catégorie minoritaire parmi les réponses aux actions de marketing direct. Il s'agit d'actions qui engendrent des revenus fixes suite à une réponse favorable (les abonnements, les mailings «single-shot» qui offrent un seul produit etc.). Les autres situations où la réponse a un caractère discret (nombre de produits commandés, durée des abonnements exprimée en mois, trimestres ou années) ou continu (les montants des commandes ou des dons) sont moins fréquemment étudiés dans la littérature. Dans cette étude on essaye de couvrir les trois catégories de

modèles et de montrer l'intérêt des modèles de réponse à caractère discret et continu. Les résultats obtenus incitent à élargir l'analyse effectuée ici à d'autres modèles de réponse à caractère discret, binaires ou polytomique comme les modèles de classification (analyse discriminante) ou de choix multiple (logit multinomial).

Problématiques pour les modèles en plusieurs étaptes

Les deux derniers modèles de réponse continus utilisés, présentent en germe deux facettes d'une problématique de recherche qui vise à distinguer deux ou plusieurs étapes dans la réponse aux actions de marketing direct (van der Scheer, 1998; Spring, 2001). Le plus souvent il s'agit comme dans le modèle à deux étapes évoqué de distinguer une étape primaire où s'exprime l'intérêt ou désintérêt pour une offre et une étape secondaire exprimée en générale en montants et qui détermine la profitabilité de la transaction. Le modèle en deux étapes permet d'estimer chaque étape séparément, même si les deux réponses ne sont pas indépendantes. Le modèle tobit et ses variantes offre une alternative fréquemment utilisée dans la littérature économétrique pour représenter de problèmes liés qui supposent la prise de deux décisions. A la différence du modèle en deux étapes ici la quantité ou le montant de réponse n'est pas conditionnée par la probabilité de réponse. La modélisation jointe d'un enchaînement de réponses primaire, secondaire, tertiaire etc. soit elles le résultat d'une seule action ou d'une succession d'actions de marketing direct constitue un champ de recherche qui conduira à des meilleures méthodes de ciblage et à une gestion plus profitable des bases de données clients.

BIBILOGRAPHIE

Bibliographie

Bauer C. L. (1988) A direct mail customer purchase model. Journal of Direct Marketing, 2,16–24.

Bentz Y. et Merunka. D (2000), Neural networks and the multinomial logit for brand choice modellnig: A hybrid approach , Journal of Forecasting, 19, 3, 177-200

Bradley, A.P. (1997), The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition, 7, 1145-1159.

Breiman, L., J. Friedman, R. Olshen, et C. Stone (1984), Classification and Regression

Trees, Wadsworth, Monterey,

Courtheoux, R.J. (1987), Database modeling: Maximizing the benefits, Direct Marketing, 3, 44–51.

Cullinan, G.J. (1977), Picking them by their batting averages' recency – frequency – monetary method of controlling circulation, manual release 2103, Direct Mail/Marketing Association, N.Y..

Desmet P (1996), Comparaison de la prédictivité d'un réseau de neurones à rétropropagation avec celle des méthodes de régression linéaire, logistique et AID pour le calcul des scores en marketing direct, Recherche et applications en Marketing,11, 2, 17-27

Granger, C.W.J. (1998), Extracting information from mega-panels and high-frequency data, Statistica Neerlandica, 52, 3, 258-272.

Green, D. and Swets, J.A. (1966), Signal detection theory and psychophysics, John Wiley & Sons, NY.

Haughton, D. et S. Oulabi (1993), Direct marketing modeling with CART and CHAID, Journal of Direct Marketing, 7 (3), 16–26.

Kumar, A., Rao V.R., et H. Soni (1995), An empirical comparison of neural network and logistic regression models, Marketing Letters, 6, 251–263.

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data, Appl. Statist., 29 (2), 119-127.

Levin, N. and Zahavi, J. (1998), Continuous predictive modeling: a comparative analysis, Journal of Interactive Marketing, 12, 2, 5-22.

Linder R., Geier J. et Kölliker M. (2004), Artificial neural networks, classification trees and regression: Which method for which customer base? Journal of Database Marketing & Customer Strategy Management,11, 4; 344-357

Lix, T.S. et Berger P.D. (1995), Analytic methodologies for database marketing in the US, Journal of Targeting, Measurement and Analysis for Marketing, 4, 237–248.

Madeira S. et J. M. Sousa (2002) Comparison of target selection methods in direct marketing, European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems, Algarve, Portugal, 19 - 21 September.

Magidson, J. (1994), The CHAID approach to segmentation modeling: CHi-squared Automatic Interaction Detection, in R.P. Bagozzi, editor, Advanced Methods of Marketing Research, Basil Blackwell, Cambridge, MA, 118–159.

Mason, Ch.H., Perreault, W.D. (1991), Collinearity, power and interpretation of

multiple regression analysis, Journal of Marketing Research, 28, 268-280.

Morrison, D.G., 1966, Interpurchase time and brand loyalty, Journal of Marketing Research, 3, 281-291.

Otter, P.W., H.R. van der Scheer, and T.J.Wansbeek (1997), Direct mail selection by joint modeling of the probability and quantity of response, Som Research Report 97B25, University of Groningen.

Potharst R., Kaymak U. et Pijls W. (2001) Neural Networks for Target Selection In Direct Marketing, ERIM Report Series Research In Management, ERS-2001-14-LIS, 18 pgs.

Rumalhart, D.E. et McCleelland J.L. (1986), Parallel Distributed Processing, MA : MIT Press

Shepard, D. (1995), The new direct marketing: How to implement a profit-driven database marketing strategy, second edition, Business One Irwin, Homewood, IL.

Sonquist, J.N. (1970), Multivariate model building, Institute for Social Research, Ann Arbor, MI.

Spring, . (2001), Quantitative approaches for profit maximizationn direct marketing, PhD. Dissertation, Rijksuniversiteit Groningen.

Swets, J.A. (1979), ROC analysis applied to the evaluation of medical imaging techniques, Investigative Radiology, 14, 109-121.

Thrasher, R.P. (1991) CART: a recent advance in tree-structured list segmentation methodology, Journal of Direct Marketing, 5, 1, 35-47.

Van.den Poel D. (2003) Predicting Mail-Order Repeat Buying: Which Variables Matter?, Working Paper 191, Universiteit Gent, Fakulteit vor Bedrijfskunde, August

Van der Scheer, H.R. (1998), Quantitative approaches for profit maximization in direct marketing, PhD. Dissertation, Rijksuniversiteit Groningen.

Zahavi, J. and Levin, N. (1997) Applying neural computing to target marketing, Journal of Direct Marketing, 11, 1, 5-22.

Notes

Notes

[1] Selon Bauer (1988), Cullinan (1977), est le premier à avoir signalé que la récence, la fréquence et la valeur monétaire des achats (RFM) étaient les variables les plus souvent utilisées dans les modèles de database marketing. Depuis, de nombreuses études (tableau 1) ont montré que ces variables constituent le groupe de prédicteurs le plus important pour modéliser l'achat par correspondance. Van den Poel (2003) trouve ainsi que les variables RFM (242 opérationnalisations différentes sont analysées) sont, de loin, les plus importants déterminants du réachat. A elles seules, elles arrivent a couvrir plus de 50% de la « place maximum d'amélioration » (entre les modèle aléatoire et le modèle idéal) et les autres variables n'arrivent qu'à apporter 3,8% d'amélioration additionnelle.

[2] Le système S est un langage de très haut niveau et un environnement d'analyse des données et graphiques. C'est le seul système statistique à avoir reçu le prestigieux Software System Award (1998) de la Association for Computing Machinery (ACM) qui lui reconnaît le mérite d'avoir «définitivement changé la manière dans laquelle les gens analysent, visualisent et manipulent les données». La flexibilité de ce système est telle que même pour le traitement de petits ensembles de données le tableur devient vite un outil encombrant et son rôle ce réduit à celui d'une simple interface pour présenter les résultats des traitements effectués avec R dans un environnement plus familier pour des néophytes. Toutes les analyses de cette recherche ont été réalisées avec le système R. Pour la présentation de certains résultats on a fait appel à l'interface R-(D) COM pour Windows qui permet de connecter des applications client comme Excel avec R.

[3] Une discussion détaillée sur CART peut être trouvée dans Haughton et Oulabi (1993) et dans Thrasher (1991).

[4] Green et Swets (1966) et Swets (1979), pour plus de détails sur cette courbe.