Tirer profit des sources externes pour l'enrichissement des bases clients
Application du Data Mining prédictif aux bases EDF

Christian Derquenne*, Sabine Goutier*, Sylvia Lembo** et Véronique Stéphan*

* Electricité de France, Recherche et Développement
Département ICAME
1, av. du Général de Gaulle, 92 141 Clamart Cedex, France
{christian.derquenne, sabine.goutier, veronique.stephan}@edf.fr
** Ardans SAS
Le Cristal – 2, rue Hélène Boucher, 78286 Guyancourt Cedex, France
slembo@ardans.fr

Résumé. L'objet de l'article est de montrer comment l'utilisation de techniques prédictives de Data Mining améliorent la qualité de l'information utilisée en marketing opérationnel. Cette étude s'inscrit dans la continuité des méthodologies utilisées à EDF pour enrichir les bases clientèles à partir de modèles prédictifs construits sur des variables internes. Ces modèles sont généralement construits à partir de variables explicatives issues des bases de données clients. Afin d'améliorer la qualité des modèles, nous prenons en compte des données externes agrégées. Nous présentons dans cet article comment utiliser conjointement des variables explicatives issues des bases de données clients et l'information supplémentaire issue du recensement de la population de 1999. La méthodologie a été expérimentée sur des bases clients EDF dans le cadre de campagnes de marketing opérationnel.

1. Contexte

Notre travail s'inscrit dans le cadre des techniques prédictives de Data Mining appliquées à EDF pour le segment de clientèle Particuliers. L'utilisation de ces techniques visent à améliorer la connaissance des clients de l'Entreprise permettant ainsi une meilleure efficacité des campagnes de marketing opérationnel. Les variables utilisées pour le ciblage sont en effet très inégalement renseignées dans les bases de données. Pour s'affranchir des valeurs manquantes, une méthodologie d'enrichissement des bases de données clients à l'aide de modèles de prédiction a été définie et expérimentée au sein de l'Entreprise (Derquenne 2000). Les caractéristiques disponibles sur seulement une partie de la clientèle sont ainsi extrapolées à l'ensemble des clients.

Lors de la mise en place d'une campagne de marketing direct, les critères permettant de cibler les clients visés par cette campagne sont ainsi mieux qualifiés, ce qui permet d'augmenter la taille de la cible et d'améliorer le rendement de la campagne.

L'approche proposée dans cet article complète la démarche de Data Mining initiale en intégrant aux modèles de prédiction des nouvelles variables explicatives construites à partir d'informations externes issues du recensement de la population de 1999 (RP99). En effet, depuis ces dernières années, les techniques de ciblage s'affinent prenant également en

compte des informations géo-marketing sur le type d'habitat. L'exploitation des données du RP99 croisées aux données internes répond plus efficacement à ces nouveaux besoins. Les données externes sont exploitées à la maille de l'IRIS qui est un regroupement d'îlots[1] entiers. Les nouvelles variables résument l'information externe sous la forme de profils de logement. Elles sont construites par une classification des IRIS, en fonction de leur distribution sur les valeurs d'observations des variables décrivant le logement. Ainsi les clients EDF d'un même IRIS sont décrits par trois types de variables :

- les caractéristiques propres à EDF telles que la consommation électrique, l'année du contrat, …
- les caractéristiques communes à EDF et aux données du RP99 telles que le type d'habitat,
- les caractéristiques agrégées par le profil d'IRIS auquel le logement appartient et qui décrit les distributions observées à partir du RP99 sur des variables cibles telles que l'année de construction, …

Après avoir décrit le processus de Data Mining mis en oeuvre, nous détaillons l'étape d'élaboration du modèle par régression logistique. Nous montrons alors comment introduire les variables construites à partir des données externes. Plusieurs stratégies de classification sont présentées de manière à inclure dans le modèle des effets qualifiés de simples, interactions ou encore emboîtés. Nous décrivons dans une quatrième partie quelques résultats issus des différentes expérimentations réalisées. Enfin, après avoir replacé notre travail dans le cadre du Data Mining pour le marketing opérationnel, nous concluons sur l'approche proposée.

2. Démarche Data Mining mise en oeuvre

La démarche de Data Mining adoptée est décrite dans la figure ci-dessous. Il est possible de distinguer trois étapes majeures :

- *Préparation des données* : L'extraction et la transformation des données à partir des bases clients s'effectuent avec l'ETL Informatica. Cette étape inclut la normalisation des adresses avec un logiciel dédié : NORMAD5. A l'issue de cette étape, un DataMart est créé sous Oracle pour la phase de modélisation.
- *Data Mining* : La technique de modélisation utilisée est la régression logistique. Celle-ci est précédée d'un sondage post-stratifié aléatoire simple pour respecter les proportions des variables auxiliaires sur l'échantillon des répondants. Le modèle retenu est évalué dans une phase de validation par un taux de bien classés et par la construction d'une courbe de Lift. Ces différents traitements statistiques s'effectuent sous SAS.
- *Enrichissement des bases clients* : Parmi les règles prédictives construites, seules sont retenues pour l'enrichissement celles qui ont un niveau de confiance supérieur à un seuil donné.

[1] Un îlot est la cellule élémentaire de collecte de l'information lors du RP99 et d'agrégation des données statistiques. Il est constitué d'un pâté de maisons, dont les contours sont identifiés par des voies. Leur taille est très variable.

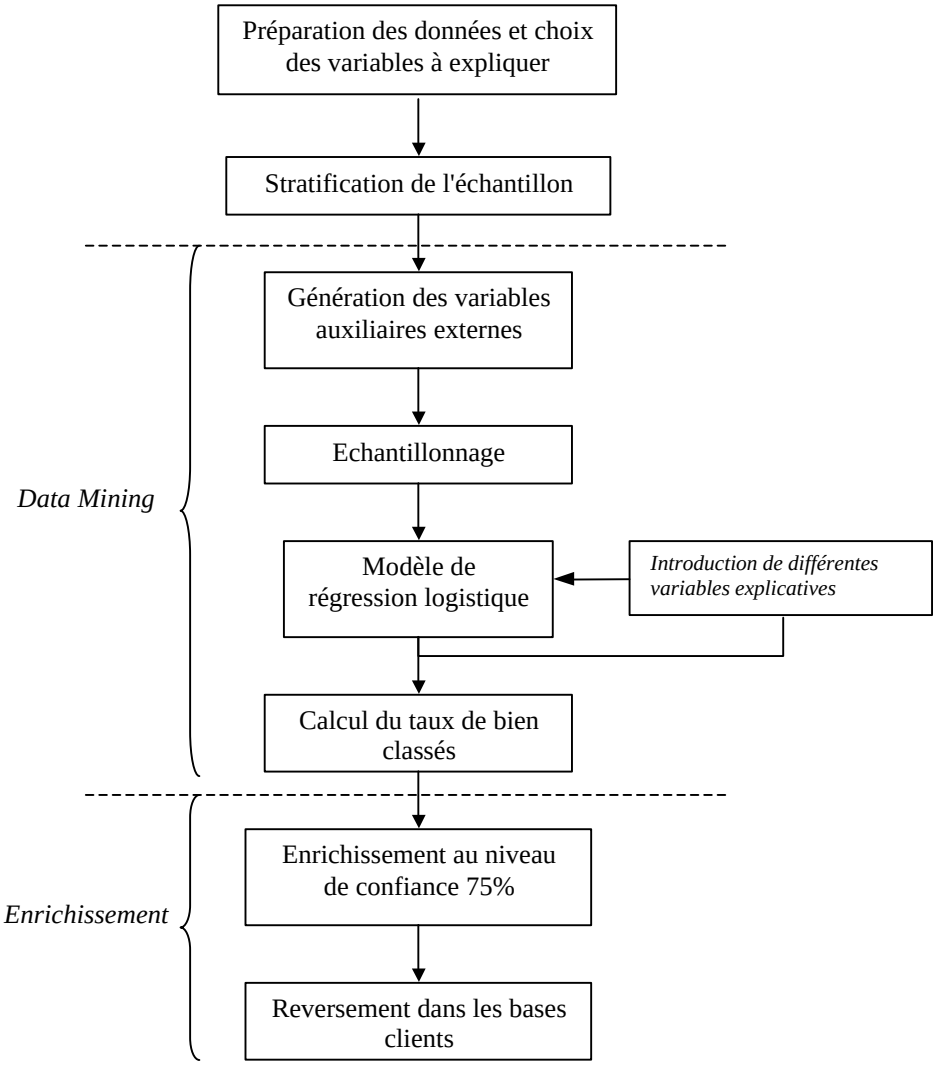

FIG. 1 – *Les principales étapes du processus d'enrichissement*

3. Construction du modèle de prédiction

La nature qualitative des données a motivé le choix de l'utilisation d'une régression logistique réalisée sur un échantillon d'apprentissage et validée sur un échantillon test (constitués lors de l'étape d'échantillonnage). Il aurait été cependant possible d'utiliser des arbres de décision de type CART, mais nous avons préféré nous en tenir à un modèle statistique du type logit.

3.1 Rappel sur la régression logistique

Ce type de modèle nommé LOGIT (LOGarithme of Inverse Transformation) permet d'estimer une probabilité d'appartenance associée à une catégorie de nature booléenne (oui/non), nominale (électricité, gaz, autre) ou ordinale (pas du tout satisfait, plutôt pas satisfait, plutôt satisfait ou très satisfait), en fonction de carctéristiques (variables candidates à l'explication quantitatives ou qualitatives, ou les deux à la fois). Le modèle linéaire gaussien usuel ne peut pas être appliqué, car non seulement la variable à expliquer est qualitative (la probabilité estimée pourrait sortir de [0,1]), mais aussi de nombreuses conditions d'application sont violées (résidus non gaussiens, variance non constante, …). Par conséquent, au lieu de modéliser directement la variable à expliquer par une fonction linéaire, c'est la transformation logistique : $\log(p/(1-p))$ qui l'est (p représente la probabilité à estimer de posséder le caractére associé). Le rapport $p/(1-p)$ est nommé souvent rapport de chances (odds ratio, en anglais). Selon la nature de la variable à expliquer, la transformation sera plus ou moins complexe (modèle logit dichotomique, polytomique ordonné ou non). L'estimation de ce type de modèles sort également du cadre usuel, car les résidus ne sont plus gaussiens, mais de lois discrètes (binomiale, multinomiale). Dans ce cas, la méthode du maximum de vraisemblance remplace la méthode des moindres carrés ordinaires. Enfin, les tests statistiques permettant de valider globalement le modèle, les variables candidates à l'explication, l'analyse des résidus et la construction des intervalles de confiance sont par conséquent aussi différents. Pour plus de détails, on pourra se référer à (Hosmer et Lemeshow 2000).

3.2 Les variables explicatives

Il s'agit de chercher à expliquer au mieux chacune des variables cibles à partir de deux types de variables explicatives :

- Les *variables explicatives internes* : Ces variables sont issues des bases clients. Le tableau de données initial est constitué de p variables notées $X_1, \ldots, X_p$.
- Les *variables explicatives externes* : Ces variables sont construites à partir de données INSEE (fichier Détail…Logements) du RP99 agrégées à l'IRIS. Lorsque l'IRIS a un profil très marqué, la distribution des valeurs observées dans l'IRIS contribue très significativement à la prédiction des clients non renseignés sur la variable à expliquer.

Le problème est donc de faire intervenir des variables explicatives portant sur des niveaux d'unités statistiques différents. En ce sens, ce problème s'inscrit dans le contexte des analyses multiniveaux (Goldstein 1995).

3.2.1 Mode de croisement des deux sources de données

Le niveau élémentaire de croisement choisi, entre les individus des bases clients et la source externe, est l'IRIS. L'affectation à chaque client de son code IRIS a été réalisée à partir du progiciel NORMAD5 avec une étape de normalisation des adresses des clients préalable[2]. Une nouvelle variable correspondant au code IRIS est alors ajoutée au tableau de

[2] Le taux d'irisage observé sur les différentes expérimentations est supérieur à 90%.

données. Elle est notée *IRIS* et possède *I* modalités {1, … , *i*, … ,*I*} où *i* représente le $i^{ème}$ IRIS.

Afin d'améliorer la granularité des informations externes, nous avons également pris en compte le type d'habitation croisé avec l'IRIS comme unité de regroupement. Le type d'habitation[3] a été reconstitué à partir des adresses normalisées. Pour cela, plusieurs règles ont été utilisées. Dans un premier temps, les habitats de type collectif ont été repérés à partir des champs d'adresses contenant des mots-clés tels que "Appartement", "Bâtiment"… Dans un second temps, et à partir des données INSEE, le type d'habitation a été renseigné "individuel" pour les clients résidant dans un IRIS à habitat uniquement individuel. Pour renseigner les types d'habitat non résolus par les deux étapes précédentes, une détermination des occurrences de chaque adresse a permis de considérer comme individuel une adresse unique et comme collectif les autres. Cette variable, notée *Type d'habitation*, intervient de plus comme variable explicative potentielle.

3.2.2 Construction des variables explicatives externes

Les variables explicatives externes sont construites par une classification des IRIS, en fonction de leur distribution sur les valeurs d'observations des variables décrivant le logement. Plusieurs stratégies de classification sont présentées pour l'intégration dans le modèle de différents effets que nous qualifions de simples, interactions ou encore emboîtés. A l'issu de ce premier traitement, plusieurs partitions sont obtenues pour chacune des *k* variables cibles à expliquer. l'élaboration des variables externes est décrit sur la figure 2.

La démarche adoptée est la suivante :

1. Sélection dans le fichier INSEE de l'ensemble des individus appartenant aux *I IRIS* présents dans la base EDF après normalisation des champs d'adresse.
2. Elaboration de différentes partitions par classification des IRIS en fonction des variables INSEE. La méthode de classification est celle d'une classification mixte sur les composantes principales d'une analyse factorielle préalable. Synthétiquement, il s'agit dans un premier temps de résumer par l'analyse factorielle les distributions observées sur les variables au niveau des IRIS. Puis on procède à un regroupement des IRIS par une méthode de classification mixte réalisée à partir de leurs coordonnées sur les premiers axes factoriels. On obtient alors des classes d'IRIS homogènes au sens de leur distribution sur la (les) variable(s) d'intérêt.
   - *Les variables* P_k « *univariées* » : il s'agit de la variable de profils d'IRIS correspondant à la $k^{ème}$ partition. La partition d'indice *k* est obtenue à partir de l'analyse du tableau de fréquences $IRIS \times X'_k$. La variable X'_k correspond dans le fichier INSEE, à la variable cible EDF X_k à expliquer. Dans ce cas, chaque client de l'IRIS *i* est décrit par $P_k(i)$ qui est l'indice de la classe d'appartenance de l'IRIS *i* pour la $k^{ème}$ partition.
   - *Les variables* $P_{k'}$ *« multivariées »* : construites comme précédemment mais cette fois-ci à partir d'un tableau de fréquences multiples $IRIS \times X'_{k_1} \times \ldots \times X'_{k_2}$.
   - *L'interaction entre la variable* P_k *et Type d'habitation.*
   - *Un effet hiérarchique (aussi appelé emboîté)* $P_{k,j}$ (k={1,…,3} et j={1,… 2}) : deux classifications sur la variable INSEE d'intérêt sont réalisées suivant que les

[3] Le type d'habitation est soit individuel (maison individuelle), soit collectif (immeuble).

logements sont individuels ou collectifs. Cette stratégie est intéressante dès lors que l'on observe des IRIS ayant des caractéristiques très différentes selon le type d'habitat. La précaution d'une telle modélisation est cependant de garantir le nombre d'unités pour chaque type d'habitat et pour chacune des strates.

3. Création pour chaque partition retenue d'une variable explicative supplémentaire, notée P_k, $P_{k,i}$ ou $P_{k'}$ par la suite suivant le type de stratégie retenu. Chaque modalité d'une variable $P_.$ correspond à un groupe d'IRIS issu de la partition. Les partitions peuvent être construites à partir de l'analyse d'un tableau de fréquences simples (variables P_k) ou multiples (variables $P_{k'}$). Un effet emboîté avec *Type d'habitation* est créé par la prise en compte de deux partitions distinctes : logements individuels et logements collectifs.
4. Ajout au tableau de données des nouvelles variables P_k, $P_{k,i}$ et $P_{k'}$. Un client est décrit par la modalité correspondant à la classe d'appartenance de son IRIS dans la partition étudiée. Une variable modélisant l'interaction entre *Type d'habitation* et P_k est également ajoutée.

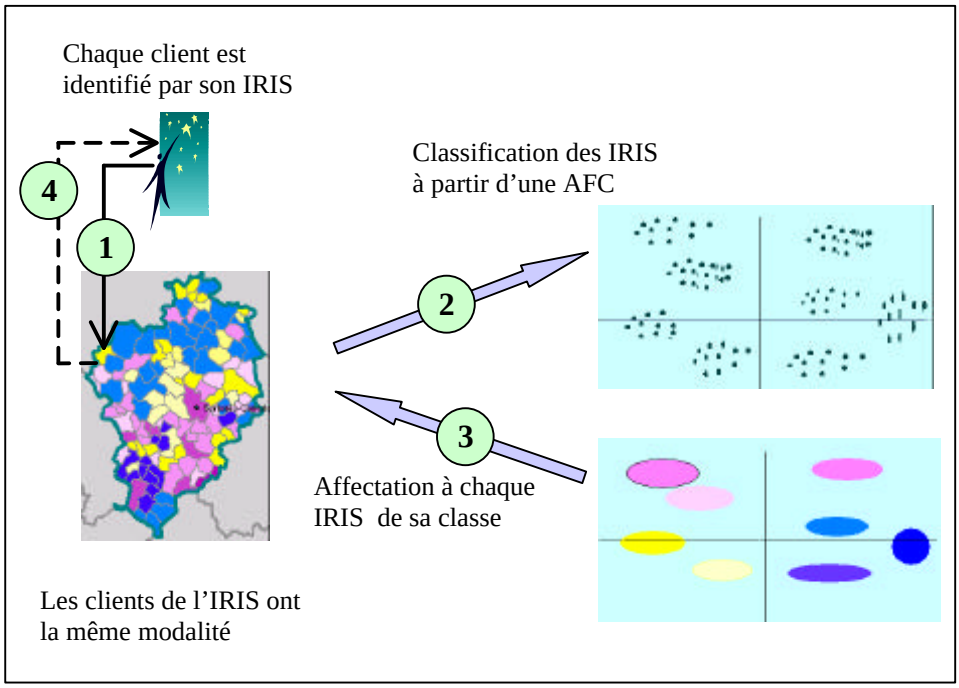

FIG. 2 – *Elaboration des variables sur les données externes*

3.3 Sélection du modèle

La sélection du modèle repose dans un premier temps sur la détermination des variables qui expliquent au mieux la variable à enrichir. Celles-ci sont constituées de variables internes et d'une ou plusieurs variables externes. Ce modèle est ensuite validé statistiquement par

l'application à l'échantillon test des règles construites à partir de l'échantillon d'apprentissage.

Le taux de bien classés permet de sélectionner le modèle explicatif le plus adapté pour chacune des variables. Il est calculé sous forme d'un quotient. Ainsi, pour une variable à expliquer *Y* et une modalité r_1, le numérateur représente le nombre de clients désignés par le modèle comme ayant la modalité r_1 et renseignés comme tels dans la base. Le dénominateur est le nombre total de clients observés r_1 dans l'échantillon test. Pour mieux comprendre ce que représente ce pourcentage, croisons les valeurs observées de la variable *Y* avec celles prédites, on obtient le tableau suivant :

Y observé	*Y* prédit			
		r_1	r_2	Total
	r_1	a	b	a + b
	r_2	c	d	c + d
	Total	a + c	b + d	a + b +c + d

TAB 1 – *Matrice de confusion de la variable Y*

Le pourcentage de bien classés pour la modalité r_1 et le pourcentage de bien classés au total se calculent de la manière suivante :

$$\%\mathrm{BC}(r_1) = \frac{\mathrm{a}}{\mathrm{a}+\mathrm{b}} * 100 \text{ et } \%\mathrm{BC\,Total} = \frac{\mathrm{a}+\mathrm{d}}{\mathrm{a}+\mathrm{b}+\mathrm{c}+\mathrm{d}} * 100$$

De plus, pour chaque groupe de clients ayant les mêmes caractéristiques, en terme de variables explicatives, un niveau de confiance individuel est calculé. Ce niveau de confiance représente un pourcentage de bien classés ; il s'agit de calculer pour chaque profil d'individus (c'est à dire pour chaque combinaison de modalités des variables explicatives) le pourcentage d'individus bien classés dans l'échantillon test.

Une distinction est faite entre les différents types de variables à expliquer, les variables binaires (dichotomiques) et les variables polytomiques (plus de deux modalités ordonnées ou non) :

- *Variables dichotomiques* : Les variables explicatives les plus significatives sont sélectionnées à partir du modèle logit (hiérarchisation des variables par ordre décroissant de contribution à l'ajustement de la variable à expliquer, c'est-à-dire à enrichir). Puis la modélisation est réalisée sur l'ensemble réduit des variables sélectionnées précédemment. De plus, le modèle obtenu est éventuellement simplifié en regroupant les modalités ayant le même apport statistique pour chaque variable explicative. Cette étape permet d'obtenir des règles pour tous les clients ayant les mêmes caractéristiques (profils identiques de variables explicatives).
- *Variables polytomiques* : L'obtention des règles pour tous les clients ayant les mêmes profils se fait également après suppression des variables non significatives, mais aucun regroupement de modalités n'est effectué.

Pour enrichir les données non renseignées de la variable à prédire, nous avons choisi de ne sélectionner que les profils pour lesquels le niveau de confiance est supérieur à 75%. Ainsi, réaliser un enrichissement à un niveau de confiance de 75% signifie que seuls les profils pour lesquels on obtenait au moins 75% de bien classés dans l'échantillon test ont été enrichis. Le niveau de confiance individuel d'un profil i (dans le cas d'une variable à expliquer binaire), noté c_i se calcule avec la formule suivante:

$$c_i = 1_{[\hat{p}_i \geq 0,5]} \frac{n_i^{(r)}}{n_i} + 1_{[\hat{p}_i < 0,5]} \left(1 - \frac{n_i^{(r)}}{n_i}\right)$$

où n_i est le nombre d'individus appartenant au profil i, $\hat{p}_i$ la probabilité estimée sur le profil i et $n_i^{(r)}$ est le nombre d'individus observés pour la modalité cible dans l'échantillon test.

Remarquons que:

$$\%\text{BC Total} = \frac{\sum n_i \ c_i}{\sum n_i}$$

4. Résultats

Une fois sélectionné le modèle, il s'agit de comparer les différents types de résultats que les méthodologies utilisables fournissent sur nos données et ainsi montrer l'apport de l'utilisation des variables explicatives externes.

Le tableau suivant indique:

- en lignes les différentes expérimentations réalisées,
- en colonnes les pourcentages de bien-classés
  - sur le total [**%BC Total**], par modalité [**%BC(r_1)** et **%BC(r_2)**],
    - ces colonnes distinguant le taux de bien-classés selon le modèle sans [***sans***] ou avec variables explicatives externes [***avec***].
    - la colonne niveau [***niveau***] indique le niveau de significativité de l'écart entre les deux taux de bien classés. Lorsque la case est grisée, le taux de bien-classés est significativement supérieur dans le modèle avec variable explicative externe.
  - et selon la règle du maximum [**Règle du max (pour r_1)**].

Le niveau de significativité est calculé grâce à un test statistique (ou un intervalle de confiance) mesurant la contribution d'un modèle statistique par rapport à la règle du maximum et de l'écart entre les taux de bien-classés obtenus sur le modèle avec et sans variable explicative externe.

La statistique de test est calculée par la formule suivante et comparée à la distribution d'une loi normale centrée réduite pour obtenir le niveau de significativité. L'approximation par la loi Normale peut être réalisée car n, la taille de l'échantillon analysé, est élevée.

$$t_{diff} = \frac{f_{\text{BC-Total}}^{avec} - f_{\text{BC-Total}}^{sans}}{\sqrt{\left(1 - \frac{n_{test}}{n}\right) \frac{f_{\text{règle-max}} (1 - f_{\text{règle-max}})}{n_{test}}}}$$

où n et n_{test} sont respectivement les tailles de l'échantillon complet et de l'échantillon test.

%BC Total			%BC(r_1)			%BC(r_2)			Règle du max (pour r_1)
sans	***avec***	***niveau***	***sans***	***avec***	***niveau***	***sans***	***avec***	***niveau***	
82,88	83,23	*0,253*	91,17	91,73	*0,124*	67,57	67,53	*0,397*	74,16
78,93	79,98	*0,001*	76,35	84,24	*0,000*	80,75	76,96	*0,000*	51,96
79,84	82,89	*0,000*	88,66	93,72	*0,000*	61,27	60,07	*0,020*	76,21
81,12	81,66	*0,146*	90,49	92,74	*0,000*	63,02	60,26	*0,000*	65,12
81,81	83,52	*0,000*	91,43	94,19	*0,000*	55,34	54,15	*0,000*	69,85
83,70	84,72	*0,038*	88,80	90,93	*0,000*	76,12	75,49	*0,162*	67,43
81,37	81,62	*0,331*	94,50	93,63	*0,041*	43,01	46,56	*0,000*	74,72
75,48	76,85	*0,004*	78,51	82,94	*0,000*	71,61	69,07	*0,000*	64,59
81,86	83,00	*0,004*	86,85	86,80	*0,395*	76,14	78,65	*0,000*	66,70
76,80	77,33	*0,113*	90,57	90,80	*0,314*	50,86	51,97	*0,002*	66,79

TAB 2 – *Comparaison des modèles avec et sans variable externe*

Par exemple si nous regardons la troisième ligne, nous avons obtenu 79,8% de bien-classés avec le modèle sans variable explicative externe contre 82,9% de bien-classés avec le modèle avec variables explicatives externes. Cette différence est significative par rapport à la règle du maximum (76,2%), le niveau de significativité étant pratiquement nul.

Ces résultats montrent que la prise en compte de données externes améliore grandement (et significativement) les pourcentages de bien classés.

Nous observons également des différences entre le nombre de clients enrichis à partir du modèle construit sans variable externe et celui construit en prenant en compte les variables externes et internes.

Dans le tableau ci-dessous sont ainsi consignées, pour chacune des expérimentations :

- dans le première colonne, le nombre de clients pour lesquels la variable à prédire n'est pas renseignée
- dans la seconde (resp. la troisième) les enrichissements effectivement réalisés en se basant sur le modèle sans variable externe (resp. avec variables externes).

Propositions d'enrichissement	Enrichissements effectués sans variable externe	Enrichissements effectués avec variables externes
1417		171
3436		1120
1239	6	286
3647		3201
2346		375
1140		285
4733	227	2943
1936	5	1632
1341	38	347
1188		439

TAB 3 – *Comparaison du nombre de clients enrichis*

Par exemple si nous regardons la troisième ligne, 1239 clients n'étaient pas qualifiés pour la variable cible (sur un échantillon de certaines communes), 6 ont pu être enrichi avec le modèle sans variable externe contre 286 avec.

Remarquons à ce niveau que n'ont été effectivement chargés dans les bases que les règles ayant un niveau de confiance individuel supérieur à 75%. De plus, dans certaines des expérimentations, aucun des clients n'aurait été enrichi par le modèle construit sans variable externe.

Nous avons également utilisé un autre critère pour juger de la qualité du modèle retenu. Il s'agit de la courbe de Lift.

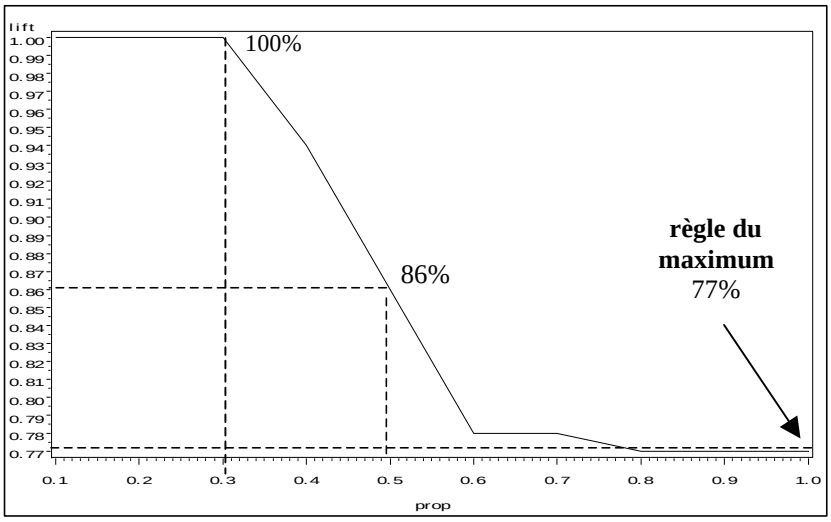

FIG. 2 – *Courbe de lift*

Cette courbe est tracée point par point, pour chaque modalité de la variable à prédire selon le principe décrit ci-dessous.

Soit *Y* la variable que nous cherchons à expliquer. Les individus de l'échantillon test sont classés par ordre décroissant de la probabilité prédite (de ceux qui ont le plus de chances d'être prédit comme ayant la modalité cible de *Y* à ceux qui ont la plus faible probabilité). Un calcul de taux de bien-classés (noté $t_{0,1}$) est réalisé sur les 10 premiers pourcents de ces individus et placé comme premier point de la courbe; ce point ayant pour coordonnées $(0,1 ; t_{0,1})$. Une série de *k* points $(k * 0,1 ; t_{k*0,1})$ est alors obtenue par itération pour *k* variant de 1 à 10.

Ainsi un modèle de bonne qualité est un modèle pour lequel le pourcentage de bien classés reste bon au fur et à mesure que les individus sont cumulés, c'est à dire lorsque la courbe de lift décroît le plus tard possible.

Ainsi, si l'on considère la moitié des individus (celle des mieux prédits), pour une proportion de 50% (lue en abscisses), 86% de ceux-ci sont bien classés; soit un gain opérationnel de 1,12 (=86/77). Ceci s'interprète en disant que pour deux fois moins de clients ciblés, le retour positif est 1,12 fois meilleur. Et pour une proportion de 30%, le gain est de 1,3 (=100/77).

5. Conclusions et perspectives

Dans le cadre du marketing opérationnel, de nombreuses études montrent la pertinence de l'utilisation des techniques de Data Mining. Des exemples d'applications dans le cadre du Marketing et du CRM sont développées dans « Les 5èmes Journées Modulad » consacrées au Data Mining des Données Clientèles (MODULAD 2000) alors qu'un survol des différentes méthodes, dans le cadre du marketing, est fourni dans (Berry et al. 1997). Plus récemment, lors de la conférence KDD 2002, deux exemples d'applications ont été présentés. Le premier (Rosset et al. 2002) s'inscrit dans les problématiques de modélisation du cycle de vie de la valeur client. La méthode proposée est implémentée dans un logiciel commercial : Amdocs' Business Insight. Le second (Storey et al. 2002) traite du problème de l'amélioration de la gestion de la relation client, dans le secteur des banques, par la recherche de combinaison optimale entre le choix du client à cibler, l'offre de produits, le choix du canal de communication et le moment de la campagne.

Comme nous l'avons dit en introduction, dans le contexte d'EDF, la démarche de Data Mining prédictif a été expérimentée et industrialisée à l'ensemble des clients Particuliers avec succès (Derquenne 2000). Cette démarche permet ainsi d'augmenter de façon significative la taille des cibles des campagnes de marketing. En effet les critères permettant le ciblage des individus présentent dorénavant des taux de qualification plus importants.

A l'issue de notre étude, nous pouvons conclure de manière très positive sur l'apport des données externes pour l'élaboration des modèles de prédiction. Dans tous les modèles testés, au moins une des variables *P.* est retenue comme variable explicative. De plus, on note une augmentation significative de l'enrichissement des bases clients lors de l'ajout des variables explicatives externes. Statistiquement, cette observation tendrait à démontrer que l'apport de telles variables joue significativement sur la qualité de la courbe Lift du modèle.

Une piste non exploitée dans l'étude est la prise en compte de l'analyse multiniveaux pour le modèle de prédiction (Goldstein 1995) et (Snijders et al. 1999). Cette approche vise à utiliser dans une modélisation des informations de différents niveaux. Dans notre cas, il s'agit d'utiliser conjointement les variables qui décrivent les profils d'IRIS et des variables

qui décrivent des clients. Dans l'étude, nous nous sommes ainsi limités à une analyse contextuelle qui consiste à prédire la caractéristique du client en fonction de ses caractéristiques individuelles et de ses caractéristiques agrégées. La contribution de tels modèles est d'étendre cette approche en introduisant des effets aléatoires (modèle mixte) décrivant la variance intra-groupes du niveau hiérarchique supérieur.

Bibliographie

Berry M. et Linoff G. (1997), Data Mining Techniques appliquées au marketing, à la vente et aux services clients, Paris, Editions Masson, 1997.

Derquenne C. (2000), Mise en œuvre d'un démarche statistique complète pour la prédiction de variables dans une base de données clientèles, 5èmes Journées MODULAD Data mining des Données Clientèles, Clamart, 16-17 novembre 2000, pp 47-64.

Goldstein H. (1995), Multilevel Statistical Models second edition, London, Kendall's Library of Statistics, 1995.

Hosmer D. et Lemeshow S. (2000), Applied Logistic Regression second edition, John Wiley & Sons, 2000.

MODULAD (2000), 5èmes Journées MODULAD Data mining des Données Clientèles, Clamart, 16-17 novembre 2000.

Snijders T. et Bosker R. (1999), An introduction to basic and advanced multilevel modelling, London, SAGE Publications, 1999.

Rosset S., Neumann E., Eick U., Vatnik N. et Idan Y. (2002), Customer Lifetime Value Modeling and Its Use for Customer Retention Planning, Eighth ACM SIGKDD Int[l] Conference on Knowledge Discovery and Data Mining, Edmonton Canada, Juillet 2002, pp 332-340.

Snijders T. et Bosker R. (1999), An introduction to basic and advanced multilevel modelling, London, SAGE Publications, 1999.

Storey A. et Cohen M.-C. (2002), Exploiting Response Models - Optimizing Cross-Sell and Up-Sell Opportunities in Banking, Eighth ACM SIGKDD Int[l] Conference on Knowledge Discovery and Data Mining, Edmonton Canada, Juillet 2002, pp 325-331.

Summary

The aim of this article is to show how the use of Data Mining forecasting methods can improve the quality of data used for operational marketing. This study takes place in the continuity of classical methods used by EDF to enrich the databases with forecasting models based on internal variables. In general, these models are based on internal data came from customers databases. In order to improve the quality of classical models, we took into account external aggregated data. We present in this article how to use in the same time internal variables and supplementary informations coming from the census of population done in 1999. The methodology was tested on internal databases for operational marketing campaigns.