

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université de Carthage
._._*._*

Ecole Supérieure de la Statistique et de l'Analyse de l'Information



Projet de Fin d'Etudes

En vue de l'obtention du

Diplôme d'Ingénieur National en Statistique et Analyse de l'Information

Elaboration d'un modèle de score permettant
de détecter des personnes susceptibles de
pratiquer un acte de bonnes œuvres.

Réalisé par

Aymen ZAAFOURI

Sous la direction de

Mokhtar KOUKI
Professeur

Hichem ABDELHAK
Responsable Filiale

Kheireddine KAMOUN
Chargé d'études statistiques

Année universitaire : 2013/2014

*A mes chers parents,
qui ont tout sacrifié pour faire de moi ce que je suis aujourd'hui
Que Dieu vous procure une longue vie pleine de santé et de bonheur*

*A mes deux frères,
Que Dieu vous garde et que le sourire éclaire toujours notre vie*

*A toute ma famille et mes meilleurs amis,
Qui n'ont jamais cessé de m'encourager*

Remerciements

Je tiens à exprimer mes vifs remerciements à mon encadrant universitaire M. Mokhtar KOUKI pour avoir dirigé ce travail, pour m'avoir soutenu tout au long du projet ainsi que pour ses précieux conseils.

Je voudrais aussi remercier mes deux encadrants à INBOX Tunisie : M. Hichem ABDELHAK responsable filiale qui a porté un regard critique, ouvert et constructif sur mon travail, et M. Kheiredine KAMOUN chargé d'études statistiques pour son suivi et son aide durant toutes les phases du projet.

Mes remerciements s'adressent aussi à toute l'équipe de INBOX Tunisie : Wajdi, Imen, Housseem, Tarek et Nada pour le soutien et l'aide qu'ils m'ont apporté durant les quatre mois du stage.

Mes remerciements sont adressés aussi aux messieurs les membres de jury, pour avoir l'extrême gentillesse de bien vouloir évaluer mon travail.

Enfin, je tiens à exprimer mon profond respect à tous mes enseignants à qui je dois ma formation.

Aymen ZAAFOURI

Résumé

Le présent projet a été réalisé au sein de « Inbox Tunisie », société de conseil en marketing relationnel. Il s'inscrit dans le cadre d'un projet de fin d'étude en vue de l'obtention d'un diplôme d'ingénieur en statistique et analyse de l'information.

Le but de ce projet est de cibler des personnes susceptibles de pratiquer un acte de bonnes œuvres pour une action marketing de collecte de fonds en utilisant les données d'une base de données mutualisée myLIST. Cette action est menée par l'association Secours Catholique, membre du réseau Caritas Internationalis.

Ce projet est constitué de deux étapes : Au cours de la première étape, nous avons construit une variable « score prénom » qui affecte à chaque individu un score selon son prénom. Dans la seconde étape, nous avons expliqué la probabilité qu'un individu réponde positivement à la campagne de Secours catholique. Cette probabilité a été calculée à l'aide d'une régression logistique, la variable score prénom et d'autres variables sociodémographiques et comportementales.

Mots clés :

Scoring, myLIST, Secours Catholique, score prénom, régression logistique, Cross validation.

Table des matières

Remerciements	ii
Résumé	iii
Liste des tableaux	vii
Liste des figures	viii
Introduction générale.....	1
Chapitre I Cadre d'analyse et méthodologie	3
1 Présentation de l'entreprise d'accueil	4
2 Présentation de myLIST	5
3 Problématique et objectifs	7
4 Méthodologie	8
4.1 Data mining	8
4.1.1 Définition	8
4.1.2 Déroulement d'un projet Data mining	8
4.2 Scoring	9
4.2.1 Définition	9
4.2.2 Méthodes de scoring.....	10
4.3 Régression logistique	10
4.4 Tests d'indépendance et mesures d'association.	11
Chapitre II Etude empirique et modélisation	12
1 Construction de la variable score prénom	13
1.1 Introduction	13
1.2 Création de la base d'étude et définition des variables	13
1.2.1 Présentation des quatre partenaires	14
1.2.2 Les variables TOP	15
1.2.3 Méta_age	16
1.2.4 Création du périmètre d'étude.....	17
1.3 Traitement sur les prénoms	17
1.3.1 Correction des prénoms mal orthographiés.....	17
1.3.2 Suppression des prénoms rares	18

1.4	Choix de la variable qui renseigne l'âge	18
1.4.1	Procédure.....	19
1.4.2	Suppression des prénoms	19
1.4.3	Calcul des âges moyens.....	19
1.5	Restriction par âge.....	20
1.6	Affectation des niveaux d'affinité.....	21
1.6.1	Calcul de l'indice	21
1.6.2	Regroupement des prénoms	22
1.7	Affinité par tranches d'âge	24
1.7.1	Résultats des modèles.....	24
1.8	Conclusion.....	25
2	Score Secours Catholique.....	26
2.1	Introduction	26
2.2	Présentation de Secours Catholique	26
2.3	Normalisation des adresses et déduplication.....	26
2.4	Construction de la base d'étude	27
2.5	Restriction de la base d'étude.....	27
2.6	Regroupement des modalités	29
2.6.1	Typologie myLIST	30
2.6.2	Age	31
2.6.3	Sexe	31
2.6.4	Niveau d'études.....	32
2.6.5	PCS: Professions et catégories socioprofessionnelles.....	32
2.6.6	Canal de commande	33
2.6.7	Activité presse	34
2.6.8	Univers solidarité	34
2.6.9	Score Prénom	35
2.6.10	Comptes en affinité	36
2.6.11	Autres variables.....	37
2.7	Elaboration du modèle	37
2.7.1	Méthode.....	37
2.7.2	Résultats	38

2.8	Validation du modèle	42
2.9	Conclusion.....	46
	Conclusion.....	47
	Bibliographie.....	49
	Annexes	51
	Annexe 1 : Les indices des modalités	52

Liste des tableaux

Tableau 1. Statistiques descriptives de la variable Méta âge	16
Tableau 2. Statistiques descriptives de la variable Âge déclaré.....	16
Tableau 3 : Extrait du tableau des prénoms triés selon le méta âge moyen.	20
Tableau 4. Prénoms avec les indices les plus élevés et les indices les plus faibles	22
Tableau 5. Résultats regroupement	23
Tableau 6. Résultats des modèles généralisés	24
Tableau 7. Indice des différentes classes de typologie.....	28
Tableau 8. Exemple d'une variable à éliminer	29
Tableau 9. Indice des différentes classes de typologie.....	30
Tableau 10. Indice des différentes tranches d'âge	31
Tableau 11. Indice des modalités de la variable sexe	31
Tableau 12. Indice des modalités de la variable niveau d'études	32
Tableau 13. Indice des différentes modalités de la variable PCS	32
Tableau 14.Indice des différentes modalités de la variable Canal de commande.....	33
Tableau 15. Indice des différentes modalités de la variable Activité presse.....	34
Tableau 16. Indice des différents centres d'intérêts de l'univers solidarité	35
Tableau 17. Indice des différentes modalités de la variable Score prénom	35
Tableau 18. Indice des différentes modalités de la variable Comptes	36
Tableau 19. Significativité des paramètres	38
Tableau 20. Tableau des estimations des paramètres.....	39
Tableau 21. Tableau des Odds-ratio.....	40
Tableau 22. Les variables significatives après Cross validation.....	42
Tableau 23. Estimation des paramètres par Cross Validation.....	43
Tableau 24. Odds-ration calculés par Cross Validation.....	44
Tableau 25. Courbe ROC	45

Liste des figures

Figure 1: Clients d'Inbox	4
Figure 2.Schéma de la procédure	6
Figure 3: Déroulement d'un projet data mining	9
Figure 4: Nombre des individus nés portant le nom PASCAL	10
Figure 5: Exemple d'un tableau de contingence	11
Figure 6. Répartition des individus dans les comptes	15
Figure 7. Répartition des BO.....	15
Figure 8: Effectifs et pourcentages du TOP 15 des prénoms de la base d'études	18
Figure 9 : Variation du taux des BO dans la base d'étude.....	21
Figure 10. Fonction de répartition de l'indice	23
Figure 11. Variation de l'indice selon la classe de typologie	29
Figure 12. Courbe LIFT	41

Introduction générale

En France, la baisse des subventions publiques après la crise de 2008 oblige les associations et les ONG¹ à développer l'appel au don privé, alors même que les besoins et l'effectif des populations aidées augmentent chaque année. Cette équation délicate nécessite de recourir à des techniques de collecte venues du monde de l'entreprise à l'instar du Marketing Direct.

Le Marketing Direct est une technique qui regroupe l'ensemble des actions de communication et de vente. Il consiste à diffuser un message personnalisé afin de toucher directement une cible. Par ce concept, la cible reçoit l'offre via les canaux classiques comme le courrier, le téléphone, le fax et récemment le SMS. Mais depuis quelques années, avec l'avènement d'internet, c'est surtout le recours à l'email qui s'est fortement développé. Ceci s'explique par le fait que l'annonceur peut toucher des milliers de personnes en quelques minutes et au moment désiré. Cette démarche de ciblage doit permettre d'obtenir un résultat positif, rapide et mesurable par le calcul du taux de retour. La différence par rapport aux autres types de marketing, est la disposition d'une base de données qui répertorie des informations sur les individus à cibler. Les bases de données les plus exhaustives sont les bases de données mutualisées car elles contiennent des informations collectées à partir d'un ensemble de partenaires.

C'est dans ce cadre que s'inscrit ce projet de fin d'étude qui permet d'extraire des données de la base myLIST afin de détecter des donateurs potentiels suite à une action conduite par l'association *Secours Catholique*.

Ce présent rapport est composé de deux chapitres. Le premier sera consacré au cadre du stage et à la méthodologie. Tandis que le deuxième présentera la partie empirique.

Dans le premier chapitre, Nous présenterons l'entreprise d'accueil, la problématique, l'objectif du stage, la base de données myLIST ainsi que la méthodologie utilisée. Dans le deuxième chapitre, nous nous intéresserons aux aspects pratiques du ciblage à travers la construction d'une variable « score prénom » et l'élaboration d'un modèle de score. Finalement, nous clôturerons notre travail par une conclusion générale résumant les résultats.

¹ Organisation non gouvernementale

Chapitre I

Cadre d'analyse et méthodologie

1 Présentation de l'entreprise d'accueil



Inbox est une société de conseil fondée en 2001 spécialisée en marketing relationnel. Depuis 2011, Inbox a intégré différentes cultures en s'implantant en Russie et en Tunisie, puis sur le continent nord-américain à Washington et Montréal.

A l'aide des consultants marketing, statisticiens, et développeurs web, INBOX aide ses clients à mieux appréhender le comportement des consommateurs.



Figure 1: Clients d'Inbox

Les domaines de compétences d'Inbox sont :

- Conseil marketing : Expertise et réflexion stratégique.
- Datamining : Segmentations, Scoring, profiling.
- Applications opérationnelles : Hébergement mise à jour des bases de données marketing et CRM opérationnel.

2 Présentation de myLIST



MyLIST est une base de données mutualisée², multicanal, multipartenaire et multi-secteurs. Elle est Créée en Juin 2010 par une entreprise de conseil en plan fichier de prospection nommée « Critère Direct ».

Cette base de données est un réservoir de 22 millions d'adresses et de plus de 300 variables de sélection. Ces variables fournissent des informations **sociodémographiques** (âge, CSP³, Revenu,...), **comportementales** (nombre de paiement par, canal de commande,...) et de **type de contact** (adresse postale, adresse mail, numéro de téléphone fixe et mobile).

Ce volume de données est obtenu grâce aux données partagées par les partenaires qui sont essentiellement des groupes de presse (*Le nouvel observateur, Mondadori,...*), sites web de vente à distance (*Rue du commerce*), associations de collecte de fonds (*CFRT, Ordre de Malte*) et des communautés (*Aide à l'Eglise en Détresse*).

Les annonceurs comme « *France Abonnement* », « *Michael KORS* » et « *HSBC* » utilisent myLIST en vue d'augmenter la rentabilité de leurs campagnes marketing, acquérir de nouveaux clients, Aussi, enrichir leurs bases de données. Les partenaires peuvent refuser qu'un annonceur ait des adresses de leurs comptes pour une campagne spécifique.

MyLIST permet aux annonceurs de louer et d'échanger des adresses qualifiées multisources. Il s'agit de solutions de conquête de nouveaux clients conçues à partir de sélections d'adresses optimisées. La solution myLIST se fonde sur une démarche d'enrichissement des données qui a pour but d'accroître la connaissance client pour plus d'efficacité.

² Une base de données mutualisée est le résultat de la mise en commun des bases de données clients ou prospects de plusieurs partenaires.

³ Catégorie SocioProfessionnelle

Inbox est chargé de la mise à jour des données ainsi que des analyses statistiques comme les extractions des adresses, les segmentations et les scores afin d'augmenter la rentabilité des campagnes marketing.

La procédure à suivre afin d'extraire des données de myLIST, commence par l'annonceur qui contacte « Critère Direct » en indiquant la nature de l'analyse (Score, Segmentation, Extraction, ...). De son tour, « Critère Direct » communique la demande aux partenaires pour avoir leurs accords ou refus.

Une fois les partenaires ont répondu, « Critère Direct » transmet la commande à INBOX joignant le fichier des réponses des partenaires.

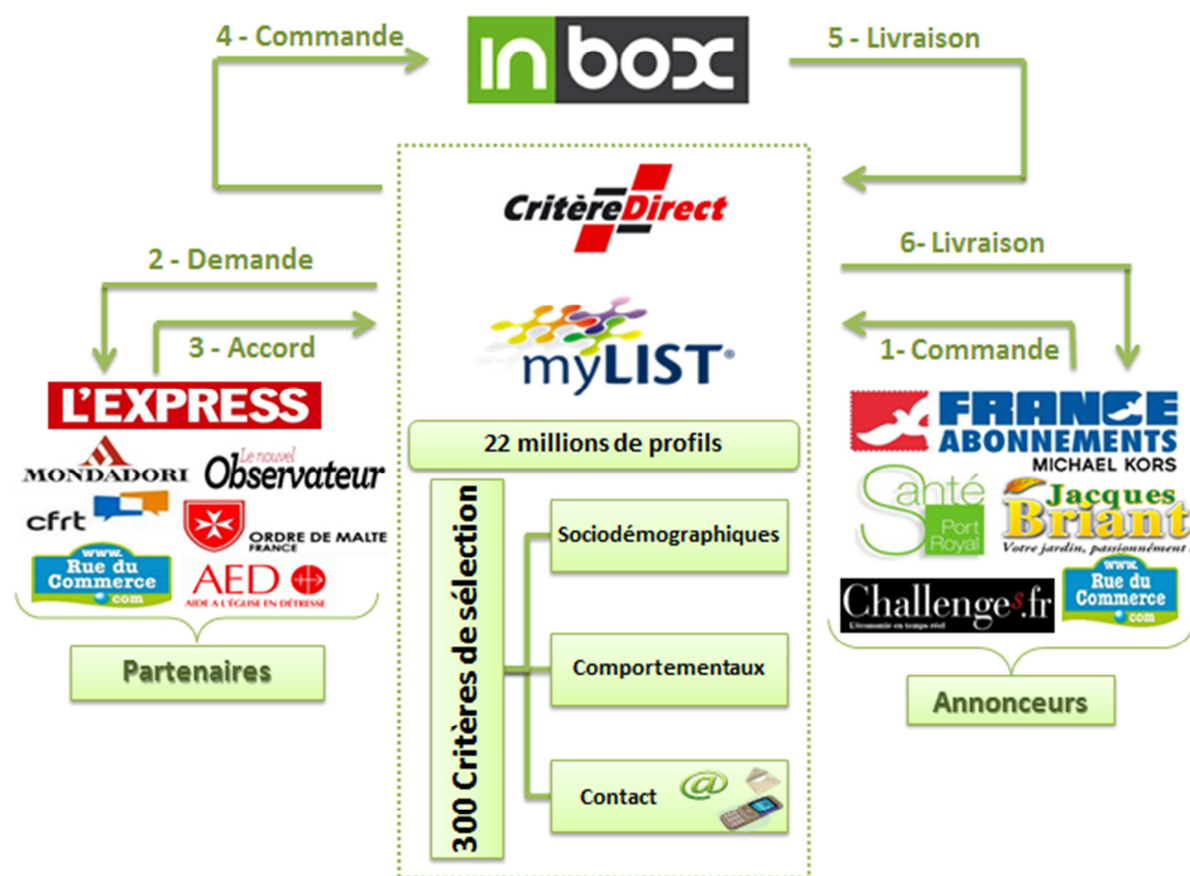


Figure 2.Schéma de la procédure

3 Problématique et objectifs

La mission est de construire un modèle de score afin de cibler des prospects susceptibles de pratiquer un acte de bonnes œuvres. La différence entre le score élaboré à la fin de ce projet et les scores précédents est l'utilisation d'une variable « score prénom ». Cette variable affecte à chaque individu un score selon son prénom qui servira à calculer la probabilité de réponse à une action marketing.

Le but du projet est d'augmenter la rentabilité d'une action marketing par le ciblage des prospects. Un bon ciblage permet d'améliorer la ROI⁴ et d'éviter de tomber dans le piège de la cannibalisation⁵. L'objectif est donc d'élaborer un modèle de score pour cibler des donateurs potentiels lors d'une campagne de collecte de fonds pour l'association *Secours catholique*.

L'idée est de détecter les prénoms des personnes qui ont un taux de présence élevé dans les comptes des partenaires qui contiennent des adresses des individus susceptibles de pratiquer des actes de bonnes œuvres. Ensuite, affecter à chacun des prénoms un score calculé à partir d'un indice. Enfin, expliquer la probabilité de réponse des individus à l'aide du score prénom et d'autres variables sociodémographiques et comportementales.

⁴ Return On Investment

⁵ Phénomène observé lorsque les ventes d'un produit augmentent en contrepartie d'une diminution de celles d'un autre. <http://www.linternaute.com/dictionnaire/fr/definition/cannibalisation-produit>

4 Méthodologie

4.1 Data mining

4.1.1 Définition

Le « Data mining » est un ensemble de techniques et de méthodes du domaine de la statistique, des mathématiques et de l'informatique permettant l'extraction, à partir d'un important volume de données brutes, des connaissances et des informations inconnues auparavant. Il s'agit d'une " fouille intelligente des données " visant à découvrir des informations qui aident dans la procédure de la prise de décision [2].

Dans un contexte marketing, le « data mining » regroupe l'ensemble des technologies susceptibles d'analyser les informations d'une base de données marketing pour y trouver des informations utiles à l'action marketing et d'éventuelles corrélations significatives et utilisables entre les données grâce à deux familles d'analyses qui sont l'analyse descriptive et l'analyse prédictive [1].

4.1.2 Déroulement d'un projet Data mining

Généralement, un projet « Data mining » passe par les phases suivantes :

- 1) Définition des objectifs
- 2) Inventaire des données existantes et collecte des données manquantes
- 3) Exploration et préparation des données.
- 4) Elaboration et validation des modèles.
- 5) Déploiement des modèles.
- 6) Suivi et mise à jour des modèles



Figure 3: Déroulement d'un projet data mining

4.2 Scoring

4.2.1 Définition

Le « scoring » est une méthode qui consiste à affecter une note (un score) à chaque individu afin de cibler et prospecter avec une meilleure efficacité. Ce score peut être déterminé à partir de données externes ou de données calculées à partir des variables comportementales.

Un des types de scoring en marketing, est le « scoring prénom » qui est une technique qui permet d'affecter un score de probabilité de réponse à partir du prénom des individus présents dans une base marketing. Le scoring prénom se base sur les âges moyens correspondants aux différents prénoms.

En France, le scoring prénom se base sur les statistiques émises par l'INSEE⁶ sur la fréquence d'un prénom par rapport à une année de naissance donnée. Par exemple, les « Marie » sont souvent des femmes d'un certain âge assez élevé alors que « Marine » ou « Océane » sont plutôt des prénoms de jeunes filles. D'autres prénoms, comme « Pierre », restent stables dans le temps. Autre exemple, à l'époque du « Général De Gaulle », plusieurs enfants avaient « Charles » comme prénom [5].

⁶ Institut National de la Statistique et des Etudes Economiques

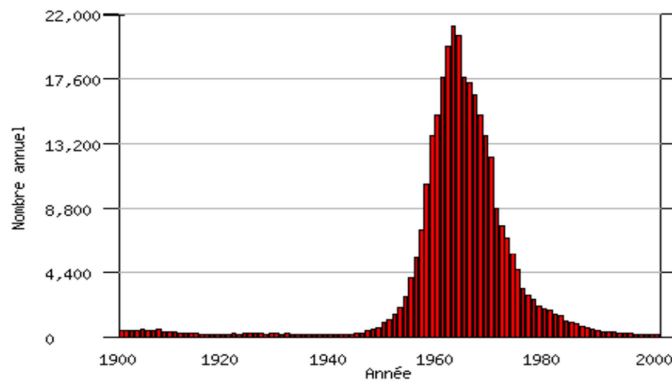


Figure 4: Nombre des individus nés portant le nom PASCAL

4.2.2 Méthodes de scoring

Les méthodes de scoring peuvent être divisées en deux catégories : Les méthodes de classement comme l'arbre de décision et l'analyse discriminante, et les méthodes de prédiction comme la régression [7].

4.3 Régression logistique

La régression logistique dichotomique est un modèle dont la variable à expliquer est binaire $\{0,1\}$ et les variables explicatives sont quantitatives ou qualitatives. [4]

Le modèle généralisé s'écrit :

$$Prob(Y = 1 / x_1 \dots x_n) = \frac{e^{\alpha + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\alpha + \beta_1 x_1 + \dots + \beta_k x_k}}$$

Comme dans toute démarche de modélisation, plusieurs étapes se posent :

- 1) Estimation des paramètres $\{\alpha, \beta_1, \beta_2, \dots, \beta_k\}$ du modèle à partir d'un échantillon
- 2) Évaluer la significativité des estimations
- 3) Mesurer le pouvoir explicatif du modèle
- 4) Vérifier s'il existe une liaison significative entre l'ensemble des descripteurs et la variable dépendante
- 5) Identifier les descripteurs pertinents dans la prédiction de Y, évacuer celles qui ne sont pas significatives et/ou celles qui sont redondantes
- 6) Calculer le score pour les individus à classer, déterminer la probabilité de réponse

4.4 Tests d'indépendance et mesures d'association.

Ces tests sont applicables à partir des tableaux de contingences. Ils sont utilisés principalement pour choisir les variables pertinentes. En fait, dans le cas d'une modélisation avec plusieurs variables. [3]


X \ Y		j		Total
i				n_{ij}
Total		$n_{.j}$		N

Figure 5: Exemple d'un tableau de contingence

- $n_{i,j}$: effectif des observations ayant comme modalités i pour la variable X et j pour la variable Y
- $n_{.j}$: effectif des modalités de la variable colonne J
- $n_{i.}$: effectif des modalités de la variable ligne I
- $e_{i,j} = \frac{n_{i.} * n_{.j}}{N}$: effectif théorique
- N : effectif total
- L : Nombre total de lignes
- C : Nombre total de colonnes

Pour tester s'il existe un lien entre les deux variables (explicative et à expliquer), on calcule le χ^2 (**Khi²**). Il s'agit de la somme sur toutes les cases (i,j) du tableau, des carrés des écarts entre l'effectif observé $n_{i,j}$ et l'effectif théorique $e_{i,j}$ divisé par l'effectif théorique.

$$\sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2((L-1) * (C-1))$$

L'indicateur global d'association χ^2 représente la statistique du test :

$$\begin{cases} H_0: \text{Pas d'association} \\ H_1: \exists \text{ une association} \end{cases}$$

Chapitre II

Etude empirique et modélisation

1 Construction de la variable score prénom

1.1 Introduction

Ce chapitre sera consacré à la construction de la variable « Score prénom ». Généralement, ce type de variables est utilisé pour prédire l'âge, le sexe ou une probabilité de réponse à une action lors des enrichissements géomarketing des bases de données clients. Dans le cas du présent projet, elle servira au calcul de la probabilité de réponse d'un client à une action marketing en vue de participer à un acte de collecte de fonds menée par l'association Secours Catholique, membre du réseau international CARITAS. Durant cette partie, les étapes de la construction seront détaillées : en commençant de la préparation de la base d'étude jusqu'à l'affectation des niveaux d'affinité en passant par le traitement des prénoms et le calcul de l'indice.

Pour la suite du rapport, Toute personne ayant pratiqué un acte de bonnes œuvres sera notée BO.

1.2 Création de la base d'étude et définition des variables

L'étape de la construction de la variable « Score prénom » commence par la détection des prénoms ayant un effectif élevé dans les comptes des partenaires de myLIST susceptibles d'avoir dans la liste de leurs clients des personnes catholiques ayant pratiqué des actes de bonnes œuvres. Ces partenaires sont : LA CROIX, PELERIN, Comité français de radio-télévision (CFRT) et Aide à L'Eglise en Détresse (AED). La base d'étude contiendra donc tous les individus appartenant à ces quatre partenaires ainsi que tous les individus possédant des adresses postales et électroniques sollicitables. Des variables de dummy (indicatrices) renseignant le compte d'où provient le client seront ajoutées à la base d'étude. Ces variables seront suffixées par TOP_ suivi par le compte du partenaire. La base d'étude contiendra donc 4 variables TOP : TOP_CFRT, TOP_AED, TOP_LACROIX et TOP_PELERIN. L'âge aussi est important dans cette partie de l'étude. Les variables qui renseignent l'âge des individus seront ajoutées aussi à la base d'étude. Ces variables sont : Age_declaré et Méta_age.

1.2.1 Présentation des quatre partenaires

Ces partenaires ont été choisis pour cette étude car ils regroupent des catholiques en une grande partie de ses clients. Ils existent ceux qui participent à des actes en faveur de l'église comme CFRT et AED. Les deux autres partenaires sont des magazines possédant un grand nombre d'abonnés de religion catholique.

1.2.1.1 Comité français de radio-télévision

Le Comité français de radio-télévision (CFRT) est une association responsable de la production de l'émission religieuse française *le Jour du Seigneur* diffusée tous les dimanches matin de 10h30 à 12h sur France 2. Cette association qui vient d'intégrer myLIST en Janvier 2014 englobe plus de 300,000 donateurs qui veulent permettre aux personnes âgées, isolées, malades, ou handicapées de bénéficier d'un réconfort spirituel.

1.2.1.2 Aide à l'Eglise en Détresse (AED)

Aide à l'Eglise en détresse (AED) est une fondation internationale catholique qui a pour mission d'aider les chrétiens menacés, persécutés, réfugiés ou dans le besoin. Elle a été fondée en 1947. Chaque année, AED collecte près de 80 millions d'euros pour soutenir l'Eglise dans le monde. Un séminariste⁷ sur six dans le monde est soutenu par AED. Elle a diffusé 46 millions de « Bibles pour enfant ».

1.2.1.3 PELERIN

PELERIN est un hebdomadaire catholique français créé le 12 juillet 1873. Il est édité par le groupe de presse Bayard Presse. Consacré initialement à la question des pèlerinages catholiques (Notre-Dame de la Salette, Lourdes, etc.), il se présentait alors comme une sorte de bulletin de liaison, évoquant également plusieurs aspects liés à ces mouvements. Les fondateurs visaient deux objectifs : contribuer au mouvement de restauration religieuse et sociale, et, affirmer une présence catholique dynamique à travers des manifestations de masse (pèlerinages, enseignements, presse, etc.).

⁷ Dans le catholicisme, les séminaristes sont des hommes qui suivent une formation comprenant de la théologie et de la philosophie suivant l'enseignement de l'Eglise Catholique.

1.2.1.4 LA CROIX

La Croix est un journal quotidien français, propriété du groupe Bayard Presse depuis 1880. Fondé par la congrégation des assomptionnistes, le journal se réclame ouvertement chrétien et catholique.

1.2.2 Les variables TOP

Une variable TOP ne prend que la valeur 1 ou 0. Les quatre variables TOP_CFRT, TOP_AED, TOP_PELERIN et TOP_LACROIX prennent la valeur 1 si l'individu appartient au compte du partenaire et 0 pour le reste. Pour la variable TOP_BO, elle prend la valeur 1 si l'individu appartient à l'un des comptes. Ci-dessous une figure qui décrit la répartition des individus de la base d'étude dans les comptes.

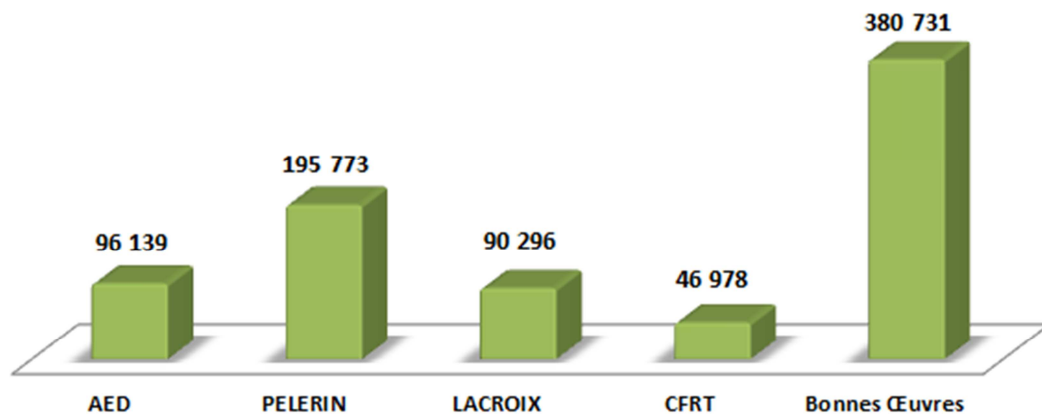


Figure 6. Répartition des individus dans les comptes

La somme des effectifs des quatre comptes n'est pas égale à l'effectif des BO car ils existent des individus appartenant à plusieurs comptes.

CFRT possède l'effectif le plus petit car il vient de rejoindre les partenaires de myLIST récemment et il a pris le choix de partager seulement une partie de ses clients.

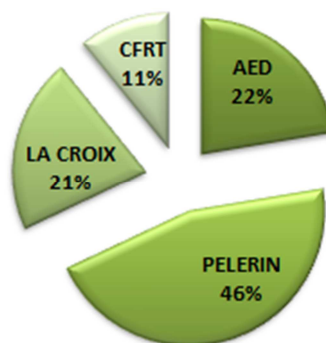


Figure 7. Répartition des BO

La plupart des BO appartiennent au compte de PELERIN. En effet le secteur presse constitué par PELERIN et LA CROIX englobe 67% des individus tandis que les associations ont participé de 33% seulement.

1.2.3 Méta_age

C'est une variable qui synthétise toutes les données en relation avec l'âge de l'individu. C'est l'une des variables enrichies de la base myLIST. Elle représente l'âge estimé des individus à partir des variables comme : L'âge moyen des individus portant le même prénom dans toute la base myLIST, le nombre d'enfants, la date de naissance du premier enfant, etc.

Moyenne	51.17	Std Deviation	18.84
Médiane	50	Variance	355.26
Mode	64	Range	99
75% - Q3	64	Interquartile Range	28
25% - Q1	36		

Tableau 1. Statistiques descriptives de la variable Méta âge

D'après le tableau 1 ci-dessus, les informations suivantes peuvent être tirées :

- Le méta âge moyen de tous les individus de la base d'étude est 51 ans.
- 25% des individus de la population sont âgés plus de 64 ans. C'est l'âge le plus fréquent dans la base.
- La moitié des individus ont des méta âges supérieurs à 50 ans.

1.2.3.1 Age déclaré

Cette variable est calculée à partir de la date de naissance fournie par le partenaire.

Moyenne	45.97	Std Deviation	16.83
Médiane	43.91	Variance	283.5
Mode	66.91	Range	112.91
75% - Q3	57.4	Interquartile Range	24.88
25% - Q1	32.51		

Tableau 2. Statistiques descriptives de la variable Âge déclaré

Ce tableau révèle les informations suivantes :

- 50% des individus ont un âge déclaré supérieur à 43 ans
- L'âge déclaré le plus fréquent est 66 ans

- Contrairement à la variable méta âge, la variable âge déclaré présente une variance inférieure à celle de la variable méta âge. Ce résultat est attendu puisque la variable enrichie résume les informations apportées par d'autres variables.

1.2.4 Création du périmètre d'étude

Afin de définir le périmètre d'étude, il faut prendre en considération deux critères pour la sélection des individus : Les prénoms doivent être renseignés et les individus doivent être sollicitables par voie postale et par e-mail.

Cette étape peut être divisée en trois parties :

- La première est la création d'une première table qui contient tous les individus ayant un prénom renseigné. Cette table qui a deux champs «id_client_inbox » et «prenom», et contient **18,105,225** observations. Ensuite, récupération des **17,793,477** « id_p » qui correspondent aux « id_client_inbox ».
- La deuxième partie est la récupération des « id_p » distincts sollicitables par voie postale et par e-mail. Cette table contient **11,412,811** « id_p ».
- Dernière étape de la construction du périmètre par la jointure des deux tables. La table résultante est de **10,233,532** « id_p ».

1.3 Traitement sur les prénoms

Comme myLIST se présente comme une base de données mutualisée, les prénoms des personnes venant de plusieurs comptes ne sont pas homogènes. Par conséquent, ils vont être traités : Deux types de prénoms se présentent, les prénoms mal orthographiés et les prénoms rares.

1.3.1 Correction des prénoms mal orthographiés

La correction des prénoms fournis par les différents partenaires se fait en deux étapes :

- Remplacement des lettres « J » et « M » par « JEAN » et « MARIE »
- Remplacement de toutes les lettres minuscule en majuscule et sans accents

Exemples :

J Christophe → JEAN CHRISTOPHE

Hélène → HELENE

M JOSEPH → MARIE JOSEPH

1.3.2 Suppression des prénoms rares

Soit \mathcal{P}_r la probabilité de tirer un prénom renseigné au hasard.

$$\mathcal{P}_r = \frac{1}{176102} = 5,68 * 10^{-6}$$

Avec 176,102 le nombre total de prénoms

Le prénom est considéré rare si son taux de présence dans la table myLIST est inférieur à \mathcal{P}_r .

Ainsi,

$$\text{Seuil} = \mathcal{P}_r * \text{Effectif_table_myLIST} = 50$$

Le seuil représente donc le nombre de fois que chaque prénom devrait figurer dans la base si les prénoms étaient équitablement répartis

Après la suppression des prénoms dont l'effectif est inférieur à 50, **171,868** prénoms ont été supprimés, Ce qui correspond à **391,872** observations. Seuls **4,234** Prénoms réparties sur **9,841,660** observations restent dans la base d'étude.

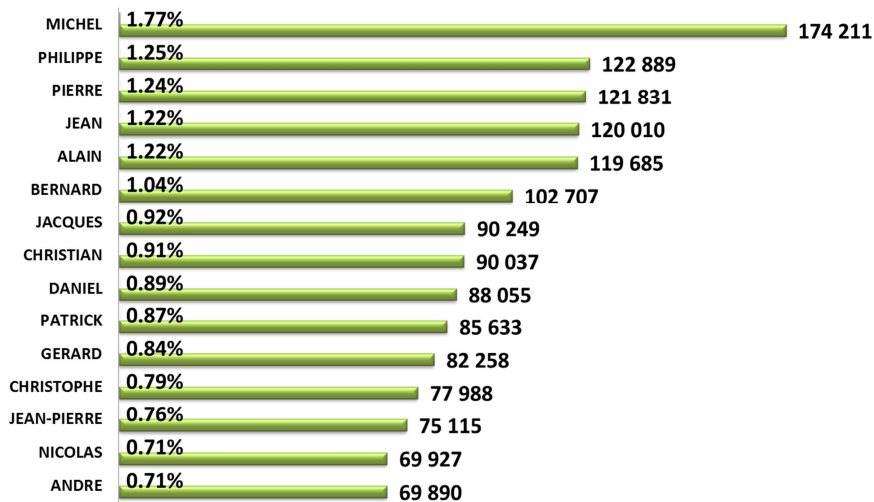


Figure 8: Effectifs et pourcentages du TOP 15 des prénoms de la base d'études

1.4 Choix de la variable qui renseigne l'âge

La base d'étude contient deux variables qui renseignent l'âge de l'individu : la première est l'âge déclaré et la deuxième variable est Méta_âge. Le but de cette étape est de choisir laquelle parmi ces deux variables à utiliser pour la suite de l'étude.

S'il existe une différence entre les différents âges moyens des prénoms, la variable « âge_déclaré » sera plus précise car elle provient du partenaire. Sinon, « Méta_âge » sera sélectionnée pour la suite de l'étude car elle englobe d'autres informations qui proviennent d'autres variables comme l'âge moyen des personnes portant le même nom dans la base myLIST.

1.4.1 Procédure

La procédure consiste à :

- Suppression des prénoms ayant un âge déclaré ou un méta âge non renseigné.
- Calcul de l'âge déclaré moyen et méta âge moyen pour chaque prénom.
- Calcul du coefficient de corrélation.

1.4.2 Suppression des prénoms

Au cours de cette étape, les individus ayant un âge déclaré ou un méta âge non renseigné seront éliminés. Après l'exécution de cette opération **4,735,803** individus ont été éliminés. Ils restent **5,105,857** individus. Il faut rappeler que ces individus seront supprimés seulement pour le calcul du coefficient de corrélation.

Cette étape est très importante car généralement les logiciels remplacent la valeur manquante par zéro.

1.4.3 Calcul des âges moyens

On calcule l'âge déclaré moyen et le méta âge moyen de chaque prénom.

PRENOM	META_AGE_MOYEN	AGE_DECLARE_MOYEN
CLAUDIUS	78,59	78,63
GERMAINE	78,18	78,07
SIMONNE	78,07	77,84
⋮	⋮	⋮
MEGANE	22,22	22,17
MAURANE	22,32	22,36
CASSANDRA	22,52	22,44

Tableau 3 : Extrait du tableau des prénoms triés selon le méta âge moyen.

Ces résultats sont attendus et conformes à l'étude⁸ publiée le 29 Avril 2014 par LE MONDE qui révèle les prénoms les plus populaires depuis 1946.

Le coefficient de corrélation permet de choisir la variable la plus pertinente pour l'étude. Ce coefficient est égal à :

$$r = 0.9998$$

Les deux variables sont fortement corrélés ce qui permet d'affirmer que la variable « Méta âge » sera utilisée pour la suite de l'étude vu qu'elle résume des informations de plusieurs autres variables.

1.5 Restriction par âge

Comme le montre la figure 9 ci-dessous, le taux de présence des BO dans la base d'étude est de l'ordre de 3,9%. Le but de cette partie est de restreindre la base afin d'augmenter le taux. Le choix de restreindre par âge est dû car les études précédentes montrent que la plupart des personnes qui pratiquent des actes de bonnes œuvres sont assez âgées (plus de 50 ans généralement). De plus, les annonceurs exigent souvent dans les demandes d'extractions ou de scores de cibler des personnes âgées de plus de 50 ans.

⁸ http://www.lemonde.fr/les-decodeurs/article/2014/04/29/la-carte-des-prenoms-les-plus-donnees-en-france_4408677_4355770.html

Seuil = 50 ans

	Appartenance au groupe BONNES OEUVRES					
	NON		OUI		Total	
+ 50 ans	4 703 869	93,3%	337 904	6,7%	5 041 773	100,0%
- de 50 ans	4 757 060	99,1%	42 827	0,9%	4 799 887	100,0%
Total	9 460 929	96,1%	380 731	3,9%	9 841 660	100,0%

Seuil = 60 ans

	Appartenance au groupe BONNES OEUVRES					
	NON		OUI		Total	
+ 60 ans	3 166 501	91,2%	304 878	8,8%	3 471 379	100,0%
- de 60 ans	6 294 428	98,8%	75 853	1,2%	6 370 281	100,0%
Total	9460 929	96,1%	380 731	3,9%	9 841 660	100,0%

Seuil = 70 ans

	Appartenance au groupe BONNES OEUVRES					
	NON		OUI		Total	
+ 70 ans	1 363 027	86,2%	217 581	13,8%	1 580 608	100,0%
- de 70 ans	8 097 902	98%	163 150	2%	8 261 052	100,0%
Total	9 460 929	96,1%	980 731	3,9%	9 841 660	100,0%

Figure 9 : Variation du taux des BO dans la base d'étude.

Jusqu'à cette étape, la base d'étude contient **9,841,660** individus dont **380,731** qui ont déjà pratiqué des actes de bonnes œuvres. Les prospects sont les **337,904** personnes qui ont des âges supérieurs au seuil et qui appartiennent au groupe des pratiquants des bonnes œuvres. Pour le seuil de 50 ans, l'étude se limite aux personnes dépassant le seuil, **42,827** personnes qui représentent presque 12% des Bonnes œuvres seront écartées et le taux des bonnes œuvres passera de 3,9% à 6,7% dans la base d'étude qui gardera plus de la moitié des observations. En variant le seuil à 60 et 70 ans : Le taux des bonnes œuvres passe respectivement à 8,8% et 13,8% dans la base d'étude mais 24% et 74 % des personnes appartenant au groupe des bonnes œuvres seront supprimées et la moitié des observations seront supprimées.

Le choix le plus approprié est de restreindre la base d'étude aux personnes âgées de plus de 50 ans. Le taux des bonnes œuvres augmente de 3,9% à 8,8% suite à cette restriction.

1.6 Affectation des niveaux d'affinité

1.6.1 Calcul de l'indice

Au cours de cette étape, chaque groupe de prénoms aura un niveau d'affinité⁹. Ce niveau est affecté à partir d'un indice calculé de la manière suivante :

⁹ Est un ratio exprimé en indice ou pourcentage qui met en évidence la proximité d'une population cible avec un support - <http://www.e-marketing.fr/Definitions-Glossaire-Marketing/Affinite-6965.htm>

$$Indice = \frac{Taux_presence_BO}{Taux_presence_myLIST}$$

Avec,

$$Taux_presence_BO = \frac{Effectif\ du\ prénom\ avec\ TOP_BO = 1\ et\ Meta_Age\ entre\ 50\ et\ 100\ ans}{Total\ des\ effectifs\ des\ prénoms\ ayant\ un\ TOP_BO = 1\ et\ Meta_Age\ entre\ 50\ et\ 100\ ans}$$

Et

$$Taux_presence_myLIST = \frac{Effectif\ du\ prénom\ ayant\ un\ Meta_Age\ entre\ 50\ et\ 100\ ans}{Total\ des\ effectifs\ des\ prénoms\ ayant\ un\ Meta_Age\ entre\ 50\ et\ 100\ ans}$$

Prénom	Indice
KRISTELLE	16.271
MARGUERITE-MARIE	11.541
MARIE-MAGDELEINE	10.983
EVE-MARIE	10.848
MARIE-MARGUERITE	9.973

Prénom	Indice
TONY	0.046
MALIKA	0.027
ALI	0.024
AHMED	0.019
MOHAMED	0.016

Tableau 4. Prénoms avec les indices les plus élevés et les indices les plus faibles

Le taux de présence du prénom « KRISTELLE » dans la base d'étude est 16 fois de plus que le taux de présence dans toute la table myLIST.

Les personnes qui portent des prénoms comme KRISTELLE et MARGUERITE-MARIE sont susceptibles de répondre positivement à une action de collecte de fonds et à participer des actes de bonnes œuvres pour la religion catholique. Par contre, ceux qui portent un prénom comme MOHAMED, AHMED ou ALI ont une tendance à ne pas répondre. C'est tout à fait logique, ces prénoms sont d'origines arabes et généralement, ils correspondent à des personnes de religion musulmane.

1.6.2 Regroupement des prénoms

En examinant la fonction de répartition, trois classes se distinguent : La première englobe les prénoms ayant un indice très élevé, la deuxième contient les prénoms avec un indice moyen et une troisième classe formée par les prénoms avec des indices faibles.

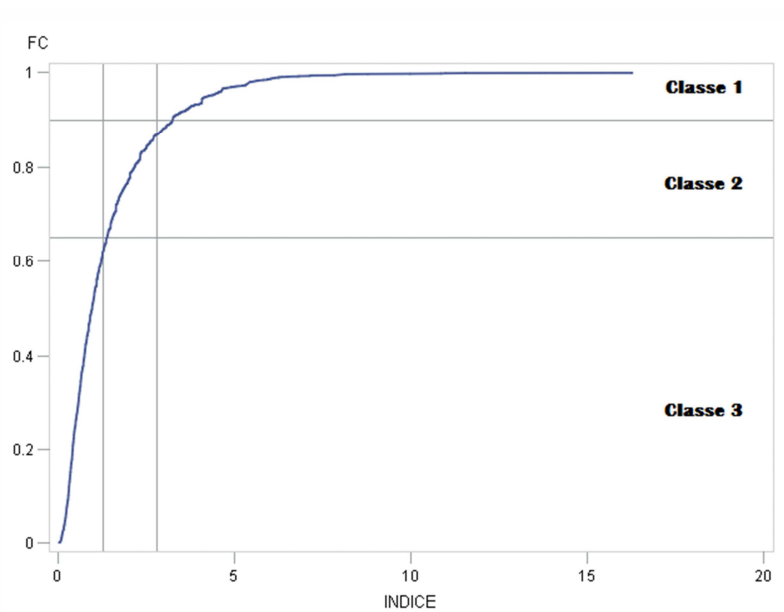


Figure 10. Fonction de répartition de l'indice

Ainsi :

- La première classe regroupera les 10% meilleurs prénoms
- La deuxième contiendra les 25% qui suivent
- La troisième aura les 65% des prénoms restants.

Ci-dessous, un tableau résumant l'effectif de prénoms dans chaque groupe ainsi que les bornes de l'indice et son écart type.

Groupe	Effectif		INDICE_MAX	INDICE_MIN	Ecart_Type_Indice
	N	%			
1	198	10%	16.27	3.27	1.64
2	552	25%	3.25	1.38	0.55
3	1394	65%	1.38	0.02	0.35

Tableau 5. Résultats regroupement

1.7 Affinité par tranches d'âge

Pour vérifier si les niveaux d'affinité des prénoms dépendent de l'âge, il suffit de créer un modèle généralisé qui a pour variable à expliquer l'indice et comme variables explicatives : le niveau d'affinité global (pour les personnes âgées de plus de 50 ans) et niveau dans la tranche d'âge. Cinq modèles seront créés : un pour chaque tranche d'âge.

Si le niveau d'affinité global dépendrait de la tranche d'âge, une variable score prénom sera ajoutée à chaque tranche.

1.7.1 Résultats des modèles

	Source	DF	Khi-2	Pr > Khi-2
Tranche 50 – 60 ans	Niveau	9	948.01	<.0001
	Classe_Age	9	7.33	0.6024
	Niveau*Classe_Age	73	52.71	0.9648
Tranche 60 – 70 ans	Niveau	9	1147.25	<.0001
	Classe_Age	9	6.42	0.6972
	Niveau*Classe_Age	72	127.83	<.0001
Tranche 70 – 80 ans	Niveau	9	948.39	<.0001
	Classe_Age	9	4.90	0.8432
	Niveau*Classe_Age	77	99.07	0.0460
Tranche 80 - 90	Niveau	9	1347.49	<.0001
	Classe_Age	9	6.85	0.6524
	Niveau*Classe_Age	76	89.89	0.1319
Tranche 90 - 100	Niveau	9	1268.39	<.0001
	Classe_Age	9	9.18	0.4205
	Niveau*Classe_Age	77	97.12	0.0605

Tableau 6. Résultats des modèles généralisés

Le modèle généralisé est construit une variable à expliquer quantitative et des variables explicatives qualitatives ce qui permet de transformer le problème à une analyse de la variance ANOVA¹⁰.

D'après le tableau des résultats des modèles :

- le niveau d'affinité global est significatif à l'ordre de 5% pour toutes les tranches. Ce résultat est attendu puisque le niveau d'affinité global est affecté selon l'indice.
- Le niveau au sein d'une tranche d'âge n'a pas d'effet direct. Ceci est valable pour toutes les tranches.
- L'effet croisé des variables : *Niveau* et *Classe_age* peut être considéré comme non significatif.

1.8 Conclusion

Dans cette partie, la variable score prénom a été construite. En commençant par la définition du périmètre d'étude, la création de la base et le traitement sur les prénoms jusqu'au calcul d'indice et l'affectation des niveaux d'affinité. Cette étape a été finalisé par l'étude de la relation entre ces niveaux et l'âge ce qui a affirmé d'une seule variable au lieu d'une pour chaque tranche. La variable score prénom caractérisera tous les individus et s'ajoutera aux autres variables de myLIST.

Dans la prochaine partie, la variable construite sera utilisée dans un score pour l'association *Secours Catholique*.

¹⁰ *ANalysis Of VAriance*

2 Score Secours Catholique

2.1 Introduction

Pour réaliser un score avec les données de myLIST, il faut passer par la procédure suivante :

- 1) L'annonceur envoie un fichier contenant les caractéristiques (Nom, Prénom, Adresse,..) de ses clients
- 2) Identification de ces individus (client, donateur, abonné,...) dans la base myLIST
- 3) Construction de la base d'étude
- 4) Restriction de la base selon les classes de typologie de myLIST
- 5) Regroupement des modalités
- 6) Elaboration du modèle du score

2.2 Présentation de Secours Catholique

Secours catholique est une association à but non lucratif créée le 8 septembre 1946 . Elle est surtout attentive aux problèmes de pauvreté et d'exclusion et cherche à promouvoir la justice sociale.

Reconnue d'utilité publique en 1962, l'association a été déclarée grande cause nationale en 1988. Elle établit aussi des rapports pour l'information du gouvernement, en matière sociale notamment. Elle constitue la branche française du réseau *Caritas Internationalis*.

Elle présente aussi un service de l'Église catholique en France. L'association prend son appui dans la doctrine sociale de l'Église pour venir en aide aux plus démunis "sans distinction de race, de religion ou de nationalité", dans le respect de la Charité chrétienne.

2.3 Normalisation des adresses et déduplication

Après la réception du fichier des adresses envoyé par l'annonceur « *Secours Catholique* » et l'import, la phase de la déduplication commence. Elle consiste à reconnaître les individus envoyés par l'annonceur dans la base myLIST. Cette opération exige la normalisation des adresses postales.

La normalisation est divisée en deux étapes : La première consiste à remettre dans le bon ordre les différentes lignes du pavé adresse, soit complément d'adresse, voie, boîte postale, code postal et localité. Puis, durant la deuxième étape, chaque ligne du pavé adresse doit répondre à certaines règles pour la longueur (32 ou 38 caractères), l'abréviation de certains mots (général, maréchal,...), la casse (minuscules ou majuscules obligatoires pour la localité), la ponctuation (caractères interdits),...

Une fois la normalisation des adresses est réalisée, les donateurs fournis par l'annonceur vont être identifiés à l'aide de la déduplication¹¹.

2.4 Construction de la base d'étude

La base d'étude pour le score *Secours Catholique* comporte tous les individus ayant des adresses exploitables : Elle englobe **10,270,802** individus dont **10,892 communs** soit un taux de **0,11 %**.

Deux Variables ont été ajoutées, la variable « Score prénom » et la variable à expliquer « Y ». Cette dernière prend la valeur 1 pour les individus issus de la déduplication et 2 pour le reste.

2.5 Restriction de la base d'étude

Dans ce paragraphe, la base d'étude sera restreinte dans le but d'augmenter le taux de commun.

La mise à jour de la base myLIST se fait après chaque introduction d'un nouveau partenaire ou 4 fois par an. Après cette opération, Inbox effectue une classification de tous les individus. Le résultat est résumé par la variable classe de typologie qui représente le numéro de la classe de chaque individu. 16 classes d'individus homogènes se distinguent.

Pour chaque classe, un indice est calculé par le rapport entre le taux de présence et le taux de commun.

$$\textbf{Indice de taux de commun} = \frac{\frac{\text{Effectif dans la base cible}}{\text{Effectif dans la base d'étude}} * 100}{\text{Taux de commun}} * 100$$

¹¹ C'est l'identification des individus présents à la fois dans la base myLIST et la base de données clients de Secours catholique.

Si l'indice de taux de commun d'une classe est égal à 100, cela signifie que le taux de présence est égal au taux de commun. Les classes à garder sont celles qui possèdent un indice de taux de commun élevé.

			BASE ETUDE		BASE CIBLE			
			Effectif	Poids	Effectif	Poids	Taux	Indice
Chiffres clés			10 270 802	100%	10 892	100%	0.11%	100
Variable	Modalité							
Typo myLIST	0	Typo myLIST	103 410	1%	-	-	0.00%	0
	1	Typo myLIST	701 963	7%	683	6%	0.10%	92
	2	Typo myLIST	1 302 467	13%	267	2%	0.02%	19
	3	Typo myLIST	212 183	2%	273	3%	0.13%	121
	4	Typo myLIST	251 251	2%	251	2%	0.10%	94
	5	Typo myLIST	270 018	3%	218	2%	0.08%	76
	6	Typo myLIST	683 141	7%	1 811	17%	0.27%	250
	7	Typo myLIST	623 384	6%	1 734	16%	0.28%	262
	8	Typo myLIST	892 954	9%	84	1%	0.01%	9
	9	Typo myLIST	1 039 797	10%	417	4%	0.04%	38
	10	Typo myLIST	846 828	8%	178	2%	0.02%	20
	11	Typo myLIST	458 551	4%	1 031	9%	0.22%	212
	12	Typo myLIST	468 659	5%	1 729	16%	0.37%	348
	13	Typo myLIST	919 147	9%	1 437	13%	0.16%	147
	14	Typo myLIST	289 242	3%	452	4%	0.16%	147
	15	Typo myLIST	801 150	8%	290	3%	0.04%	34
	16	Typo myLIST	406 657	4%	37	0%	0.01%	9

Tableau 7. Indice des différentes classes de typologie

L'objectif de cette partie est de restreindre la base d'étude en utilisant les classes de typologie afin d'augmenter le taux de communs 0,11%. Pour fixer le seuil, il faut trier les classes selon l'indice et tracer l'indice en fonction des classes.

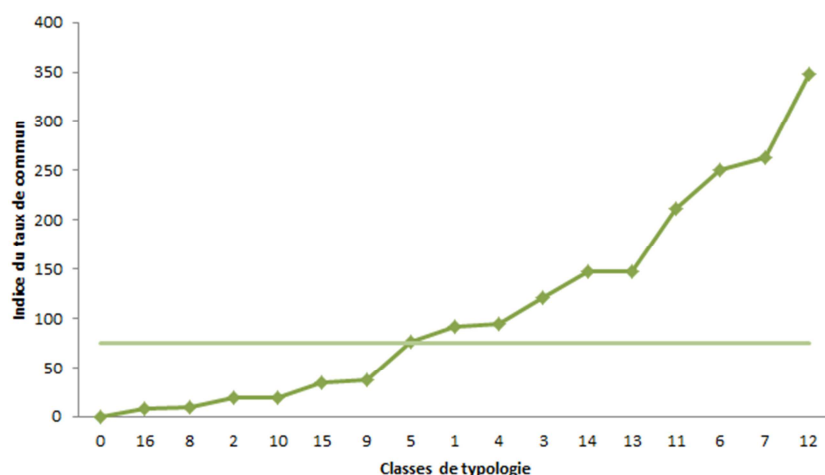


Figure 11. Variation de l'indice selon la classe de typologie

La courbe de l'indice subit un décrochage à partir de la classe 5. En prenant 75 comme seuil minimal d'indice, la base d'étude gardera les individus appartenant aux classes de typologie 5, 4, 1, 3, 13, 11, 14, 6, 7 et 12. **4,877,539** individus dont **9,619** communs restent.

Le seuil augmente suite à cette restriction et atteint **0,2%**.

2.6 Regroupement des modalités

L'indice de taux de commun est aussi calculé pour les autres variables. Cette fois, il est utilisé pour choisir les variables explicatives du modèle tout en regroupant leurs modalités. Les variables à choisir présentent des modalités avec un indice supérieur à 100 et d'autres modalités inférieures à 100. La figure ci-dessous représente une variable à éliminer du choix des variables explicatives. Cette variable donne une information sur la date du premier achat du client à partir d'un site web de vente à distance.

			BASE ETUDE		BASE CIBLE			
			Effectif	Poids	Effectif	Poids	Taux	Indice
Chiffres clés			4 880 059	100%	9 619	100%	0.20%	100
Date premier acte VAD	Jamais	Date premier acte VAD	3 415 702	70%	7 282	76%	0.21%	108
	0 - 6 mois	Date premier acte VAD	15 050	0%	10	0%	0.07%	34
	7 - 12 mois	Date premier acte VAD	90 114	2%	74	1%	0.08%	42
	13 - 18 mois	Date premier acte VAD	137 572	3%	172	2%	0.13%	63
	18 - 24 mois	Date premier acte VAD	182 257	4%	247	3%	0.14%	69
	Plus de 24 mois	Date premier acte VAD	1 039 364	21%	1 834	19%	0.18%	90

Tableau 8. Exemple d'une variable à éliminer

Dans la base cible, 95% des individus appartiennent à seulement deux modalités. De plus la plupart des indices sont faibles et une seule modalité a un indice supérieur à 100.

2.6.1 Typologie myLIST

		BASE ETUDE		BASE CIBLE			
		Effectif	Poids	Effectif	Poids	Taux	Indice
Chiffres clés		4 880 059	100%	9 619	100%	0.20%	100
Variables	Modalité						
Typo myLIST	0	Typo myLIST	-	-	-	-	-
	1	Typo myLIST	700 913	14%	683	7%	0.10%
	2	Typo myLIST	-	-	-	-	-
	3	Typo myLIST	209 833	4%	273	3%	0.13%
	4	Typo myLIST	252 711	5%	251	3%	0.10%
	5	Typo myLIST	272 588	6%	218	2%	0.08%
	6	Typo myLIST	684 091	14%	1 811	19%	0.26%
	7	Typo myLIST	623 194	13%	1 734	18%	0.28%
	8	Typo myLIST	-	-	-	-	-
	9	Typo myLIST	-	-	-	-	-
	10	Typo myLIST	-	-	-	-	-
	11	Typo myLIST	459 851	9%	1 031	11%	0.22%
	12	Typo myLIST	471 789	10%	1 729	18%	0.37%
	13	Typo myLIST	917 047	19%	1 437	15%	0.16%
	14	Typo myLIST	288 042	6%	452	5%	0.16%
	15	Typo myLIST	-	-	-	-	-
	16	Typo myLIST	-	-	-	-	-

Tableau 9. Indice des différentes classes de typologie

Pour cette variable, le regroupement proposé permet de réunir les classes : 1, 3, 4 et 5 en un seul groupe. L'indice pour ce groupe est compris entre 41 et 66. Le groupe 2 réunira les deux classes 13 et 14. En effet, ce groupe est caractérisé par un indice supérieur à celui du groupe précédent mais inférieur à 100. En ce qui concerne les classes ayant un indice supérieur à 100 seront divisées en deux groupes : un groupe pour la classe 2 car elle présente un indice assez élevé et un taux de présence de 18% et 10% respectivement dans la base cible et la base d'étude et un groupe pour les classes : 11, 6 et 7. Il faut rappeler que la base cible contient les clients en communs issus de la déduplication et la base d'étude comprend les individus de myLIST y compris les communs.

Une variable *cl_typo* sera créée de la façon suivante :

$$Cl_typo = \begin{cases} 1 & \text{si typologie myLIST} = 1,3,4,5 \\ 2 & \text{si typologie myLIST} = 6,7,11 \\ 3 & \text{si typologie myLIST} = 13,14 \\ 4 & \text{si typologie myLIST} = 12 \end{cases}$$

2.6.2 Age

				BASE ETUDE		BASE CIBLE		
				Effectif	Poids	Effectif	Poids	Taux
Chiffres clés	min	max		4 880 059	100%	9 619	100%	0.20%
Meta age	0	.	.	167 543	3%	183	2%	0.11%
	1	16	41	483 090	10%	290	12%	0.06%
	2	42	54	540 490	11%	530	22%	0.10%
	3	55	58	485 008	10%	518	21%	0.11%
	4	59	61	439 016	9%	556	23%	0.13%
	5	62	64	694 244	15%	1 294	53%	0.19%
	6	65	66	384 342	8%	882	36%	0.23%
	7	67	68	277 318	6%	658	27%	0.24%
	8	69	78	472 726	10%	1 596	65%	0.34%
	9	79	85	530 192	11%	1 792	73%	0.34%
	10	86	109	406 090	9%	1 320	54%	0.33%

Tableau 10. Indice des différentes tranches d'âge

La variable Méta âge sera divisée en 3 classes : La première classe avec des indices compris entre 30 et 64. Elle regroupera les individus ayant un âge non renseigné ou inférieur à 62 ans. Une deuxième classe pour les individus âgés entre 62 ans et 69 ans et une dernière classe pour ceux âgés de plus de 69 ans.

Une variable Cl_age sera créée de la manière suivante :

$$Cl_Age = \begin{cases} 1 & \text{si } Meta_age < 61 \text{ ans} \\ 2 & \text{si } Meta_age \text{ entre } 62 \text{ ans et } 69 \text{ ans} \\ 3 & \text{si } Meta_age > 69 \text{ ans} \end{cases}$$

2.6.3 Sexe

		BASE ETUDE		BASE CIBLE			
		Effectif	Poids	Effectif	Poids	Taux	Indice
Chiffres clés		4 880 059	100%	9 619	100%	0.20%	100
Sexe	Hommes	2 395 728	49%	5 308	55%	0.22%	112
	Femmes	2 098 567	43%	3 467	36%	0.17%	84
	Couple	382 492	8%	842	9%	0.22%	112

Tableau 11. Indice des modalités de la variable sexe

Les femmes sont présentes dans la base cible avec un taux inférieur que celui de la base d'étude ce qui mène à un indice inférieur à 100. Par contre, pour les deux modalités « Couple » et « Homme » possèdent le même indice 112.

La variable Sexe sera remplacée par une variable *cl_sexe* :

$$Cl_sexe = \begin{cases} 1 & \text{si } sexe = \text{"Homme"} \text{ ou } \text{"Couple"} \\ 2 & \text{si } sexe = \text{"Femme"} \end{cases}$$

2.6.4 Niveau d'études

		BASE ETUDE		BASE CIBLE		
		Effectif	Poids	Effectif	Poids	Taux
Chiffres clés		4 880 059	100%	9 619	100%	0.20%
Niveau d'études	< BAC	3 147 522	65%	5 542	58%	0.18%
	BAC-BAC + 2	985 154	20%	2 384	25%	0.24%
	> BAC + 2	711 217	15%	1 597	17%	0.22%
	NR	36 166	1%	96	1%	0.27%

Tableau 12. Indice des modalités de la variable niveau d'études

Pour cette variable, la modalité « NR¹² » possède un indice supérieur à 100 mais le poids est de l'ordre de 1% dans les deux tables. La modalité « < BAC » présente un indice faible en le comparant avec les deux modalités restantes. La variable *cl_geo_etude* regroupera les individus qui ont un niveau d'étude \geq BAC en une classe et les autres dans une deuxième classe.

$$Cl_geo_etude = \begin{cases} 1 & \text{si } niveau_etude = \text{"NR"} \text{ ou } \text{"<BAC"} \\ 2 & \text{si } niveau_etude \geq \text{BAC} \end{cases}$$

2.6.5 PCS: Professions et catégories socioprofessionnelles

		BASE ETUDE		BASE CIBLE		
		Effectif	Poids	Effectif	Poids	Taux
Chiffres clés		4 880 059	100%	9 619	100%	0.20%
PCS	ETUDIANT	148 193	3%	93	1%	0.06%
	PCS - -	447 053	9%	233	2%	0.05%
	PCS -	289 674	6%	214	2%	0.07%
	PCS +	246 676	5%	276	3%	0.11%
	PCS ++	405 179	8%	559	6%	0.14%
	RETRAITE	3 328 106	68%	8 216	86%	0.25%
	NR	15 178	0%	28	0%	0.18%

Tableau 13. Indice des différentes modalités de la variable PCS

¹² Non renseigné

La modalité « Retraité » est la seule à avoir un indice supérieur à 100. Elle se retrouvera seule dans une classe. Comme le cas pour la variable niveau d'étude, la variable « NR » possède un effectif faible, elle sera donc regroupée avec les modalités da faibles indices. Pour les autres modalités, elles seront regroupées en une troisième classe.

$$CI_PCS = \begin{cases} 1 \text{ si } PCS = "NR" \text{ ou } "PCS - -" \text{ ou } "PCS - " \text{ ou } "PCS - " \\ 2 \text{ si } PCS = "PCS+" \text{ ou } "PCS + +" \\ 3 \text{ si } PCS = "RETRAITE" \end{cases}$$

2.6.6 Canal de commande

			BASE ETUDE		BASE CIBLE			
			Effectif	Poids	Effectif	Poids	Taux	Indice
Chiffres clés			4 880 059	100%	9 619	100%	0.20%	100
Variables	Modalité							
Canal	Courrier	Canal	2 646 197	54%	6 477	67%	0.24%	124
	Domicile	Canal	180	0%	-	0%	0.00%	0
	Web	Canal	953 909	20%	1 079	11%	0.11%	57
	Point de vente	Canal	1 404	0%	4	0%	0.28%	145
	Telemarketing	Canal	212 156	4%	536	6%	0.25%	128
	Mobile	Canal	4 495	0%	5	0%	0.11%	56
	Autres	Canal	145 951	3%	291	3%	0.20%	101

Tableau 14. Indice des différentes modalités de la variable Canal de commande

La variable canal de commande désigne la façon utilisée par le client pour passer ses commandes. Pour cette variable, une seule modalité « Courrier » sera prise en considération car elle possède un poids assez élevé et un indice supérieur à 100.

La variable existe déjà dans la base myLIST sous le nom : top_canal_co

$$top_canal_co = \begin{cases} 1 \text{ si } canal_commande = "Courrier" \\ 0 \text{ pour le reste} \end{cases}$$

2.6.7 Activité presse

			BASE ETUDE		BASE CIBLE		
			Effectif	Poids	Effectif	Poids	Taux
Chiffres clés			4 880 059	100%	9 619	100%	0.20%
Activité presse	0	Activité presse	2 320 844	48%	3 364	35%	0.14%
	1	Activité presse	1 982 284	41%	3 884	40%	0.20%
	2	Activité presse	362 848	7%	1 328	14%	0.37%
	3	Activité presse	112 623	2%	493	5%	0.44%
	4	Activité presse	51 282	1%	252	3%	0.49%
	5	Activité presse	26 378	1%	148	2%	0.56%
	6	Activité presse	13 237	0%	77	1%	0.58%
	7	Activité presse	5 687	0%	37	0%	0.65%
	8	Activité presse	2 617	0%	17	0%	0.65%
	9	Activité presse	1 269	0%	9	0%	0.71%
	10	Activité presse	990	0%	10	0%	1.01%

Tableau 15. Indice des différentes modalités de la variable Activité presse

Pour la variable Activité presse, 2 modalités de faibles indices ont un poids égal à 89%. Ces deux modalités se regrouperont ensemble dans une seule classe et tous les autres modalités dans une deuxième classe.

$$cl_activite_presse = \begin{cases} 1 & \text{si } activite_presse \geq 2 \\ 0 & \text{si } activite_presse < 2 \end{cases}$$

2.6.8 Univers solidarité

Un partenaire peut avoir plusieurs comptes dans myLIST. Chaque compte correspond à un ou plusieurs centres d'intérêts. Par exemple, le groupe de presse Mondadori possède 7 comptes : Mondadori TV, Mondadori Sciences, Mondadori Femme Senior, ...

Un ou plusieurs centres d'intérêts forment un Univers. Parmi les univers qui existent dans myLIST et rentrent dans le cadre de cette étude, l'univers solidarité qui contient les centres d'intérêts suivants : Bonnes œuvres, Spiritualité, Caritatif multi-causes, Recherche médicale,...

Une variable *top_U_solidarite* sera créée : Elle prendra la valeur 1 si l'individu appartient à un compte où la solidarité figure dans ses centres d'intérêts et 0 pour le reste.

			BASE ETUDE		BASE CIBLE			
			Effectif	Poids	Effectif	Poids	Taux	Indice
Chiffres clés			4 880 059		9 619		0.20%	100
Solidarité	Bonnes oeuvres, Spiritualite	Solidarité	43 828	1%	298	3%	0.68%	345
	Caritatif multi-causes	Solidarité	-	0%	-	0%	-	-
	Pauvrete en France	Solidarité	6 377	0%	37	0%	0.58%	294
	Parrainage d'enfants	Solidarité	7 312	0%	22	0%	0.30%	153
	Urgence humanitaire	Solidarité	5 485	0%	25	0%	0.46%	231
	Social	Solidarité	-	0%	-	0%	-	-
	Recherche medicale	Solidarité	14 484	0%	54	1%	0.37%	189
	Tiers-Monde	Solidarité	-	0%	-	0%	-	-
Total			77 486	2%	436	5%	0.56%	285

Tableau 16. Indice des différents centres d'intérêts de l'univers solidarité

Une variable *top_U_solidarite* sera créée : Elle prendra la valeur 1 si l'individu appartient à un compte où la solidarité figure dans ses centres d'intérêts et 0 sinon.

2.6.9 Score Prénom

Modalité	Base d'étude		Base Cible		Résultats	
	Effectif	Poids	Effectif	Poids	Taux	Indice
0	1 692 041	16.5%	867	7.96%	0.07%	64
1	6 858 055	66.8%	7 047	64.70%	0.10%	96
2	1 435 135	14.0%	2 437	22.37%	0.17%	160
3	285 566	2.8%	541	4.97%	0.19%	178
Total	10 270 797	100%	10 892	100%		

Tableau 17. Indice des différentes modalités de la variable Score prénom

Même la variable score prénom subira un regroupement de modalités de la façon suivante :

$$\text{Top_SC} = \begin{cases} 0 & \text{si score prénom} = 0 \\ 1 & \text{si score prénom} = 1 \text{ ou } 2 \text{ ou } 3 \end{cases}$$

2.6.10 Comptes en affinité

	BASE ETUDE		BASE CIBLE			
	Effectif	Poids	Effectif	Poids	Taux	Indice
Chiffres clés	4 880 059	100%	9 619	100%	0.20%	100
<u>Comptes</u>						
AIDE A L'EGLISE EN DETRESSE	-	0%	-	0%	-	-
CHALLENGES	102 519	2%	349	4%	0.34%	173
EXPRESS ROULARTA ACTU-ECO	311 101	6%	1 441	15%	0.46%	235
EXPRESS ROULARTA AUTRE	180 483	4%	583	6%	0.32%	164
EXPRESS ROULARTA MIEUX VIVRE	196 652	4%	1 372	14%	0.70%	354
FRANCE ABONNEMENTS	194 351	4%	521	5%	0.27%	136
FRANCE ABONNEMENTS ENTREPRISE	149 689	3%	479	5%	0.32%	162
FRANCE ABONNEMENTS INTERNET	393 075	8%	335	3%	0.09%	43
MONDADORI AUTO	62 895	1%	145	2%	0.23%	117
MONDADORI CHASSE	262 108	5%	368	4%	0.14%	71
MONDADORI FEMME GRAND PUBLIC	284 339	6%	509	5%	0.18%	91
MONDADORI FEMME SENIOR	336 422	7%	1 032	11%	0.31%	156
MONDADORI HAUT DE GAMME	146 162	3%	332	3%	0.23%	115
MONDADORI SCIENCES	98 509	2%	289	3%	0.29%	149
MONDADORI TV	440 814	9%	604	6%	0.14%	70
NOUVEL OBSERVATEUR	180 205	4%	515	5%	0.29%	145
ORDRE DE MALTE 1	-	0%	-	0%	-	-
ORDRE DE MALTE 2	-	0%	-	0%	-	-
PELERIN	-	0%	-	0%	-	-
PEPINIERES JACQUES BRIANT	435 953	9%	853	9%	0.20%	99
PLAY BAC PRESSE	21 924	0%	34	0%	0.16%	79
SCIENCES ET AVENIR	126 401	3%	441	5%	0.35%	177
LE MONDE SEM	108 060	2%	250	3%	0.23%	117
LE MONDE DIPLOMATIQUE	32 452	1%	42	0%	0.13%	66
LE MONDE COURRIER INTER	103 160	2%	220	2%	0.21%	108
DIRECT PERFORMANCE	129 732	3%	452	5%	0.35%	177
PROFILS SENIORS	881 100	18%	2 210	23%	0.25%	127
SANTE PORT ROYAL 1	70 061	1%	251	3%	0.36%	182
SANTE PORT ROYAL 2	44 578	1%	148	2%	0.33%	168
LA CROIX	-	0%	-	0%	-	-
RUE DU COMMERCE	736 208	15%	728	8%	0.10%	50
CFRT	-	0%	-	0%	-	-
MONDADORI NATURABUY	29 259	1%	19	0%	0.06%	33

Tableau 18. Indice des différentes modalités de la variable Comptes

Pour la variable Compte, vu qu'il existe plusieurs modalités ayant un indice supérieur à 100. Le seuil augmentera pour atteindre 120. Une variable *compte_en_affinité* sera créée qui aura pour valeur 1 si l'individu provient d'un compte dont l'indice est supérieur à 120 et 0 sinon.

La base d'étude ne contient aucun individu appartenant aux comptes comme LA CROIX ou CFRT. En effet, ces partenaires ont refusé que les données de leurs clients soient exploitées pour le score Secours catholique.

2.6.11 Autres variables

De la même manière, les variables suivantes ont été créées : *top_metarevenus_30*, *cl_dernier_act_press*, *Top_U_Gestion* et *Top_U_Formation*.

Les tableaux résumant les indices des modalités sont dans l'annexe.

2.7 Elaboration du modèle

2.7.1 Méthode

L'objectif de tout utilisateur d'une régression logistique est d'arriver, à partir d'un ensemble de variables explicatives, à un modèle final qui retiendrait le plus grand nombre de variables explicatives qui s'avèrent significatives dans l'explication de la variation dépendante Y.

Plusieurs méthodes de sélection automatique de variables sont proposées dans les logiciels statistiques pour effectuer un choix du meilleur ensemble de variables explicatives. Les plus courantes sont :

- La méthode d'élimination progressive (en anglais « backward selection »).
- La méthode d'introduction progressive (en anglais « forward selection »).
- La méthode de régression pas à pas (en anglais « stepwise regression »).

Il est important de noter que ces méthodes peuvent ne pas conduire au même choix de variables explicatives à retenir dans le modèle final. Elles ont l'avantage d'être faciles à utiliser et de traiter le problème de la sélection de variables de façon systématique.

Le modèle de score est créé à l'aide d'une régression logistique en choisissant la méthode de régression pas à pas pour le choix des variables explicatives. Cette méthode propose après

l'introduction d'une nouvelle variable dans le modèle deux actions : La première est un test de corrélation avec toutes les variables anciennement admises et la deuxième qui permet de retirer la variable la moins significative entre elles. Ce processus continue jusqu'à ce que aucune variable ne puisse être introduite ni retirée du modèle.

2.7.2 Résultats

Toutes les variables créées dans la partie regroupement des modalités ont été utilisées. Le tableau suivant montre les résultats du choix des variables significatives :

Variable	DDL	Wald Chi-Square	Pr > Khi-2
Top_SC	1	4.5649	0.0326
Comptes_en_affinite	1	543.5394	<.0001
cl_typo_mylist	3	163.9536	<.0001
cl_sexe	1	73.9310	<.0001
cl_age	2	337.5365	<.0001
cl_pcs	2	75.3763	<.0001
cl_geo_etudes	1	137.5385	<.0001
top_canal_co	1	230.6051	<.0001
cl_activite_presse	1	98.8991	<.0001
Top_U_Solidarite	1	147.8079	<.0001

Tableau 19. Significativité des paramètres

Les résultats de la méthode pas à pas pour le choix des variables explicatives montrent bien que les 10 variables du tableau 20 sont non corrélées entre elles et sont significatives au seuil de 5%. Elles fournissent des informations qui aident à bien discriminer les individus étudiés.

Après l'étude de significativité des paramètres, vient l'étape de l'estimation des paramètres. Dans le cas de ce modèle, le nombre de paramètres à estimer est la somme des modalités de toutes les variables plus l'estimateur de la constante.

Parameter		DDL	Estimate	Standard Error	Wald Chi-Square	Pr > Khi-2
Intercept		1	-6.3319	0.0405	24402.5216	<.0001
Top_SC	0	1	-0.0616	0.0288	4.5649	0.0326
Comptes_en_affinite	0	1	-0.3725	0.0160	543.5394	<.0001
cl_typo_mylist	1	1	-0.3062	0.0314	94.9678	<.0001
cl_typo_mylist	2	1	-0.1282	0.0262	23.9678	<.0001
cl_typo_mylist	3	1	0.0978	0.0207	22.4040	<.0001
cl_sexe	1	1	0.1238	0.0144	73.9310	<.0001
cl_age	1	1	-0.4303	0.0321	179.1754	<.0001
cl_age	2	1	0.0330	0.0229	2.0788	0.1494
cl_pcs	1	1	-0.3182	0.0430	54.7966	<.0001
cl_pcs	2	1	0.0193	0.0383	0.2543	0.6141
cl_geo_etudes	1	1	-0.2126	0.0181	137.5385	<.0001
top_canal_co	0	1	-0.2434	0.0160	230.6051	<.0001
cl_activite_presse	0	1	-0.1701	0.0171	98.8991	<.0001
Top_U_Solidarite	0	1	-0.2441	0.0201	147.8079	<.0001

Tableau 20. Tableau des estimations des paramètres

En terme de quantité d'information apportée, les variables compte_en_affinite, top_canal_co et cl_age sont les plus riches car ils possèdent des variances élevés.

La personne qui n'utilise pas les courriers comme canal de commande, n'appartient pas à un compte en affinité ainsi qu'à l'une des classes de typologie 6, 7 ou 11 a de faibles chances de répondre positivement à l'action de Secours Catholique. Par contre, toute personne appartenant à un compte en affinité, retraitée et âgée de plus de 68 ans se caractérise par une probabilité élevée de réponse positive.

Après les estimateurs, il est intéressant de décrire les odds-ratio¹³ afin de dresser le profil du meilleur donateur.

¹³ Les Odds ratio, également appelés rapport des chances, ou rapport des côtes, est une mesure statistique, exprimant le degré de dépendance entre des variables qualitatives. Il permet de mesurer l'effet d'un facteur.

Effect	Odds-Ratio	95% Wald Confidence Limits	
Top_SC 0 vs 1	0.884	0.790	0.990
Comptes_en_affinite 0 vs 1	0.475	0.446	0.505
cl_typo_mylist 1 vs 4	0.526	0.471	0.587
cl_typo_mylist 2 vs 4	0.628	0.571	0.692
cl_typo_mylist 3 vs 4	0.788	0.725	0.856
cl_sexe 1 vs 2	1.281	1.211	1.355
cl_age 1 vs 3	0.437	0.396	0.482
cl_age 2 vs 3	0.695	0.654	0.738
cl_pcs 1 vs 3	0.540	0.469	0.621
cl_pcs 2 vs 3	0.756	0.668	0.855
cl_geo_etudes 1 vs 2	0.654	0.609	0.702
top_canal_co 0 vs 1	0.615	0.577	0.654
cl_activite_presse 0 vs 1	0.712	0.666	0.761
Top_U_Solidarite 0 vs 1	0.614	0.567	0.664

Tableau 21. Tableau des Odds-ratio

Pour qu'un paramètre β soit significatif, il faut que l'intervalle de confiance de son estimateur ne contienne pas 0. La même chose pour les Odds-Ratio, mais l'intervalle de confiance ne doit pas contenir la valeur 1 car $\text{Odds} = e^{\beta}$. ($\text{Odds-ratio} = 1 \Rightarrow \beta = 0$).

Pour ce modèle tous les rapports de chance sont significatifs.

D'après le tableau 22, toute personne ayant 0 comme valeur pour la variable Top_SC a 12% de chances de moins de répondre positivement à l'action de secours catholique. Aussi, Les hommes ou les couples, ont 28% plus de chance de répondre positivement que les femmes.

L'annonceur Secours Catholique peut prévoir les performances du modèle à partir de la courbe LIFT¹⁴.

¹⁴ Une courbe qui mesure la performance d'un model prédictif. Elle est construite en calculant pour chaque décile du score, le pourcentage de donateurs qui s'y trouvent.

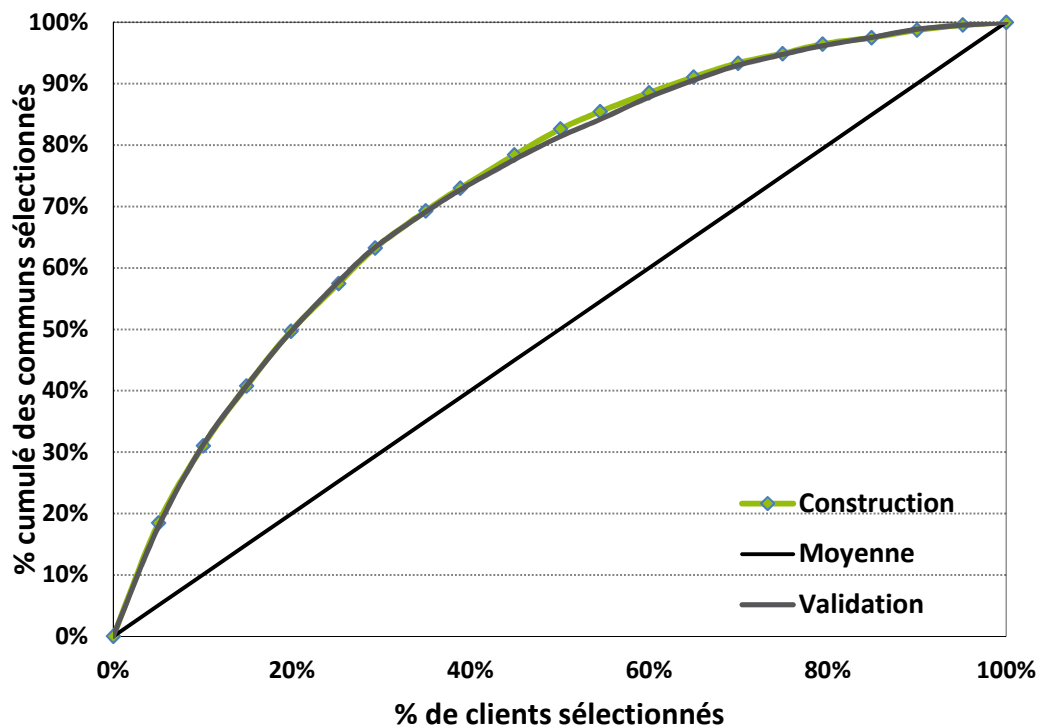


Figure 12. Courbe LIFT

L'annonceur peut interpréter cette courbe de deux manières :

- En fixant sur l'axe des abscisses le seuil des clients qui seront sélectionnées, il peut déduire le taux de retour attendu. Exemple : En ciblant 5% de la base d'étude soit 200,000 individus, 18% des communs répondront positivement.
- En fixant sur l'axe des ordonnées un taux de retour, l'annonceur peut connaître combien de clients doivent être sélectionnés.

Avant de construire le modèle, un tirage stratifié de deux échantillons représentatifs a été effectué sur la base d'étude : Une première table de construction contenant 60% des observations et une deuxième table de validation contenant les 40% des individus restants ont été créées.

La figure 12 montre la superposition des deux courbes LIFT construites à partir des deux échantillons, c'est presque le même modèle qui a été construit. Il peut être considéré robuste.

2.8 Validation du modèle

La validation des modèles de score consiste à montrer leurs capacités à bien discriminer la variable à expliquer. La meilleure méthode de valider un modèle créée à partir d'une régression logistique est la *Cross Validation* : Il s'agit d'une méthode d'estimation de fiabilité d'un modèle fondée sur une technique d'échantillonnage. En effet, l'échantillon sera divisé en 2 sous-échantillons, puis un échantillon sélectionné comme ensemble de validation et l'autre échantillon constituera l'ensemble d'apprentissage. L'opération se répète ainsi 2 fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation.

Effect	DDL	Wald Chi-Square	Pr > Khi-2
Top_SC	1	8.0435	0.0046
Comptes_en_affinite	1	906.4981	<.0001
cl_typo_mylist	3	258.4127	<.0001
cl_sexe	1	110.1970	<.0001
cl_age	2	532.0029	<.0001
cl_pcs	2	132.1477	<.0001
cl_geo_etudes	1	264.3359	<.0001
top_canal_co	1	354.1160	<.0001
cl_activite_presse	1	173.7416	<.0001
Top_U_Solidarite	1	222.0171	<.0001

Tableau 22. Les variables significatives après Cross validation

Les variables explicatives trouvées à partir de la régression en utilisant la méthode Stepwise sont aussi significatives au seuil de 5% en appliquant la méthode de Cross validation.

Parameter		DDL	Estimate	Standard Error	Wald Chi-Square	Pr > Khi-2
Intercept		1	-4.1512	0.0446	8668.1066	<.0001
Top_SC	0	1	-0.1271	0.0448	8.0435	0.0046
Comptes_en_affinite	0	1	-0.7444	0.0247	906.4981	<.0001
cl_typo_mylist	1	1	-0.6110	0.0438	194.2871	<.0001
cl_typo_mylist	2	1	-0.4208	0.0382	121.6494	<.0001
cl_typo_mylist	3	1	-0.1916	0.0331	33.5811	<.0001
cl_sexe	1	1	0.2339	0.0223	110.1970	<.0001
cl_age	1	1	-0.8195	0.0386	449.7635	<.0001
cl_age	2	1	-0.3383	0.0238	202.8227	<.0001
cl_pcs	1	1	-0.6224	0.0552	127.2460	<.0001
cl_pcs	2	1	-0.3353	0.0492	46.4861	<.0001
cl_geo_etudes	1	1	-0.4537	0.0279	264.3359	<.0001
top_canal_co	0	1	-0.4665	0.0248	354.1160	<.0001
cl_activite_presse	0	1	-0.3492	0.0265	173.7416	<.0001
Top_U_Solidarite	0	1	-0.4663	0.0313	222.0171	<.0001

Tableau 23. Estimation des paramètres par Cross Validation

Les modalités de toutes les variables sont significatives au seuil de 5%. Ce résultat est légèrement différent à celui trouvé précédemment car il ne s'agit pas du même échantillon utilisé.

Effect	Point Estimate	95% Wald Confidence Limits	
Top_SC 0 vs 1	0.881	0.807	0.961
Comptes_en_affinite 0 vs 1	0.475	0.453	0.499
cl_typo_mylist 1 vs 4	0.543	0.498	0.592
cl_typo_mylist 2 vs 4	0.657	0.609	0.708
cl_typo_mylist 3 vs 4	0.826	0.774	0.881
cl_sexe 1 vs 2	1.264	1.210	1.320
cl_age 1 vs 3	0.441	0.409	0.475
cl_age 2 vs 3	0.713	0.681	0.747
cl_pcs 1 vs 3	0.537	0.482	0.598
cl_pcs 2 vs 3	0.715	0.649	0.787
cl_geo_etudes 1 vs 2	0.635	0.601	0.671
top_canal_co 0 vs 1	0.627	0.597	0.658
cl_activite_presse 0 vs 1	0.705	0.670	0.743
Top_U_Solidarite 0 vs 1	0.627	0.590	0.667

Tableau 24. Odds-ratios calculés par Cross Validation

Les résultats du tableau 25 sont presque conformes aux résultats du tableau 22. Tous les rapports de chance sont significatifs : 1 n'appartient pas à aucun intervalle de confiance.

Même les résultats des estimations des Odds sont légèrement différents et parfois ce sont les mêmes constatations qui se répètent comme : Les personnes qui ont 0 comme valeur pour la variable Top_SC ont 12% de chances de moins pour répondre positivement à l'action de secours catholique.

Appartenir à un compte en affinité et avoir un âge supérieur à 68 ans sont les deux critères les plus discriminants.

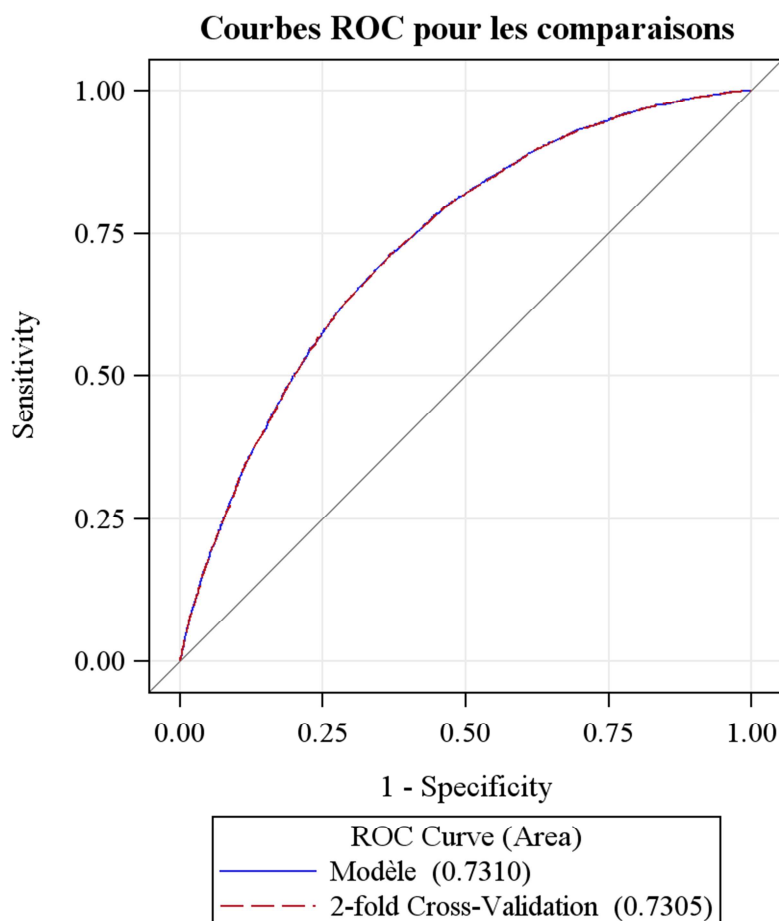


Tableau 25. Courbe ROC

Une autre manière pour tester la robustesse d'un modèle prédictif est de tracer la courbe ROC¹⁵. L'idée de la courbe ROC est de faire varier le « seuil » de 1 à 0 et, pour chaque cas, calculer le *Taux de Vrai Positifs* et le *Taux de Faux Positifs* que l'on reporte dans un graphique. Ces deux taux représentent aussi la probabilité de bien détecter un positif (la sensibilité) et la probabilité de détecter un mauvais positif (1-spécificité).

La courbe ROC passe par le point (0,25 ; 0,55). Ce point correspond à un seuil **S** tel que le modèle détecte 25% de faux positifs (des individus non positifs ayant un score > **S**) et 55% De vrais positifs (des individus positifs qui ont un score > **S**).

Les deux courbes ROC sont superposées ce qui confirme la robustesse du modèle.

¹⁵ Receiving Operating Characteristics

D'autre part un modèle est d'autant meilleur que l'AUC ¹⁶ est plus proche de 1. AUC=0.5 signifie que le modèle est similaire à une prédiction au hasard. L'aire sous la courbe ROC du modèle étudié est de l'ordre de 0,7 ce qui permet d'affirmer la performance du modèle

2.9 Conclusion

Dans ce chapitre, un modèle permettant de détecter des donateurs potentiels susceptibles de pratiquer un acte de bonnes œuvres a été élaboré. Ce modèle qui a prouvé sa robustesse et sa performance permet d'affecter à chaque individu une probabilité de réponse calculée à l'aide des paramètres estimés.

Contrairement aux autres modèles élaborés par Inbox pour divers clients, Ce modèle est le premier à expliquer une probabilité de réponse en fonctions d'une variable caractérisant le prénom.

¹⁶ Il représente l'aire de la surface sous la courbe ROC

Conclusion

Au cours de ce projet, nous avons identifié des donateurs potentiels pour l'association *Secours Catholique*. Ce travail a été réalisé à l'aide d'un modèle de score. La méthode de validation croisée a affirmé la robustesse de ce modèle.

Le scoring a trouvé encore une fois son application dans le domaine du marketing direct en permettant un ciblage efficace et ainsi un taux de retour assez élevé. Ceci justifie les investissements financiers consacrés à ce domaine.

La modélisation des scores continue de nos jours à guider la prise de décisions pertinentes avec des garanties d'optimalité non seulement dans le domaine du marketing mais aussi dans divers domaines comme la médecine et la finance.

Le modèle obtenu a bénéficié en partie de l'exploitation d'une base de données exhaustive regroupant plusieurs critères mais de nos jours avec l'apparition des nouveaux concepts comme le « *big data* » et devant l'afflux énorme des données massives, nous sommes en droit de se demander si le statisticien saura relever le défi et jusqu'où pourrait-il répondre au besoin du marché quant à la prédiction du comportement des consommateurs ?

Bibliographie

1. BATHELOT, B. (2011, Mai 12). Consulté le 02 28, 2014, sur <http://www.definitions-marketing.com/>: <http://www.definitions-marketing.com/Definition-Datamining>
2. BESSE, A. B. (2005). *Data mining : Exploration Statistique*. 31062 – Toulouse cedex 4: Laboratoire de Statistique et Probabilites - Universite Paul Sabatier .
3. CONFAIS, J., GRELET, Y., & LE GUEN, M. (2005). *TESTS D'INDEPENDANCE ET MESURES D'ASSOCIATION DANS UN TABLEAU DE CONTINGENCE*. Université Pierre et Marie Curie (Paris 6) - ISUP, Boîte 157, 4 Place Jussieu, 75252 Paris: Revue MODULAD.
4. RAKOTOMALALA, R. (s.d.). *Université Lumière Lyon 2*. Consulté le 02 27, 2014, sur <http://eric.univ-lyon2.fr/>.
5. SAPORTA, G. (s.d.). *Introduction au Data Mining et à l'apprentissage statistique*. Consulté le 02 28, 2014, sur Chaire de Statistique Appliquée & CEDRIC: <http://cedric.cnam.fr/~saporta/DM.pdf>
6. TENENHAUS, M. (s.d.). *La Régression Logistique*. Consulté le 03 05, 2014, sur studies2.hec.fr:
https://studies2.hec.fr/jahia/webdav/site/hec/shared/sites/tenenhaus/acces_anonyme/home/fichier_ppt/regression_logistique.ppt
7. TUFFERY, S. (2012). *DataMining et statistique décisionnelle*. TECHNIP.

Annexes

Annexe 1 : Les indices des modalités

Top_meta_revenu_30

			BASE ETUDE		BASE CIBLE			
			Effectif	Poids	Effectif	Poids	Taux	Indice
Chiffres clés			4 880 059	100%	9 619	100%	0.20%	100
Meta revenus	> 120 000	Meta revenus	1 021	0%	1	0%	0.10%	50
	> 100 000	Meta revenus	21 603	0%	93	1%	0.43%	218
	> 75 000	Meta revenus	355 836	7%	966	10%	0.27%	138
	> 50 000	Meta revenus	488 829	10%	1 349	14%	0.28%	140
	> 30 000	Meta revenus	917 174	19%	2 584	27%	0.28%	143
	< 30 000	Meta revenus	-	-	-	-	-	-

Cl_dernier_act_presse

			BASE ETUDE		BASE CIBLE			
			Effectif	Poids	Effectif	Poids	Taux	Indice
Chiffres clés			4 880 059	100%	9 619	100%	0.20%	100
Date dernier acte Presse	Jamais	Date dernier acte Presse	2 176 371	45%	3 071	32%	0.14%	72
	0 - 6 mois	Date dernier acte Presse	52 360	1%	190	2%	0.36%	184
	7 - 12 mois	Date dernier acte Presse	406 907	8%	1 387	14%	0.34%	173
	13 - 18 mois	Date dernier acte Presse	393 107	8%	1 127	12%	0.29%	145
	18 - 24 mois	Date dernier acte Presse	270 290	6%	670	7%	0.25%	126
	Plus de 24 mois	Date dernier acte Presse	1 581 024	32%	3 174	33%	0.20%	102

Top_U_Gestion

			BASE ETUDE		BASE CIBLE			
			Effectif	Poids	Effectif	Poids	Taux	Indice
Chiffres clés			4 880 059	100%	9 619	100%	0.20%	100
Gestion	Droit, pratique	Gestion	332 025	7%	1 215	13%	0.37%	186
	Banque, Assurance	Gestion	-	0%	-	0%	-	-
	Immobilier	Gestion	261 958	5%	1 568	16%	0.60%	304
	Placements, Epargne, Bourse	Gestion	717 980	15%	2 640	27%	0.37%	187
	Total	Gestion	1 311 963	27%	5 423	56%	0.41%	210

Top_U_Formation

			BASE ETUDE		BASE CIBLE			
			Effectif	Poids	Effectif	Poids	Taux	Indice
Chiffres clés			4 880 059	100%	9 619	100%	0.20%	100
Formation	Langues	Formation	10 284	0%	44	0%	0.43%	217
	Education, Enseignement	Formation	-	0%	-	0%	-	-
	Developpement personnel	Formation	-	0%	-	0%	-	-
	Formation / Carriere professionnelle	Formation	130 082	3%	452	5%	0.35%	176
Total			140 366	3%	496	5%	0.35%	179

Tous les centres d'intérêts ont été utilisés pour les deux variables Top_U_Formation et Top_U_Gestion.