Communauté Économique et Monétaire de l'Afrique Centrale

(C.E.M.A.C)

Institut Sous régional de Statistique et d'Économie Appliquée

(I.S.S.E.A)

Organisation Internationale

B.P. 294 Yaoundé, Tel. (237) 22 22 01 34, Fax. (237) 22 22 95 21, Site: www.issea-cemac.org

(République du Cameroun)

PROJET DATA MINING

THEME :

ANALYSE DU COMPORTEMENT DES CLIENTS SUSCEPTIBES DE SOUSCRIRE FORTEMENT À UNE PROMOTION

Rédigé par :

ETEME Serge Brice

KAPNANG Herrman Brice

&

KWEDOM Annick Yolande

Elèves Ingénieurs d'Application de la Statistique, 4ème année

Enseignant :

Monsieur BOBDA Jean Christophe

Ingénieur Statisticien Economiste

Janvier 2013

LISTE DES TABLEAUX

LISTE DES GRAPHIQUES

INTRODUCTION

Face à la concurrence accrue dans le domaine de la téléphonie, l'objectif premier des opérateurs est de se différencier. Ils doivent désormais jouer sur l'originalité pour attirer et fidéliser le client car il est plus couteux d'acquérir un nouveau client que de le fidéliser.

Ainsi, il appartient aux opérateurs d'évaluer l'impact de leur stratégie de promotion chez leurs clients. Autrement dit, d'identifier et convertir les clients les moins appétant aux offres promotionnelles. C'est dans cette logique que s'inscrit le présent projet qui est une analyse du comportement des clients afin de dégager parmi les moins appétant ceux susceptibles de devenir « High promophile».

Le présent travail s'organise comme suit : dans la première partie, nous présentons la préparation des données notamment, la description des variables, l'apurement de la base et le traitement des données manquantes. La deuxième partie consiste en une analyse descriptive et explicative qui permettra de mettre en exergue les liens entre les habitudes de consommation du client et son appétence aux promotions. Les méthodes d'intelligence artificielle notamment les réseaux de neurones et les supports vector machines (SVM) seront présentées dans la troisième partie. Enfin, le meilleur modèle sera sélectionné et utilisé dans la dernière partie pour identifier les clients « none promophile » susceptibles de devenir « high promophile ».

N.B. Tous les graphiques, tableaux et figures de ce travail ont pour source nos données.

CHAPITRE I : ANALYSE ET PRÉPARATION DES DONNÉES

I.1 Présentation de la base de données brutes

La base de données soumise à notre étude provient d'une entreprise de téléphonie mobile. Elle contient 11079 observations et chaque observation se réfère à un détenteur de téléphone mobile et comprend les données de participation aux différentes promotions que l'entreprise offre sur le marché. La base de données contient 67 variables; dont 20 variables sont des données de panels étalées sur les mois d'août, septembre et octobre. Nous avons une variable d'identification des clients et 5 variables qui sont recueillies uniquement pour le mois d'octobre. En plus, nous avons la variable cible à modéliser PROMO_TYPE. Cette variable renseigne sur les abonnés qui souscrivent à des promotions appelés « promophiles» et contient 6 catégories qui renseignent sur les non promophiles (Not promo users, verylow promo users et low promo users) et les promophiles. L'idée est de dégager parmi les non promophiles ceux susceptibles de devenir promophiles. Le travail de notre groupe (groupe 2) est d'identifier parmi les non promophiles ceux susceptibles de devenir high promophiles. Les variables de la base peuvent se résumer dans le tableau1 suivant

Tableau 1: description des variables de la base

VARIABLE	DESCRIPTION
NB_CALLS	Nombre total d'appels émis
ACTUAL_DURATION	Total de minutes appelées
BILLABLE_DURATION	Total de minutes facturées
OUT_NB_SMS	Nombre de sms émis
OUT_VOICE_AMT	Total consommation voix
OUT_SMS_AMT	Total consommation sms
GPRS_AMT	Total consommation Internet
OUTGOING_ONNET_DURATION	Total de minutes appelées intra réseau
OUT_OFFNET_DURATION	Total de minutes appelées inter réseau
OUT_INTERNATIONAL_DURATION	Total de minutes appelées à l'international
OUT_PEAK_DURATION	Total de minutes appelées entre 5h et 22h59
OUT_OFFPEAK_DURATION	Total de minutes appelées entre 23h et 4h59
INC_NB_SMS	Nombre de sms reçus
INC_NB_CALLS	Nombre d'appels reçus
INC_ONNET_DURATION	Total de minutes reçus intra réseau
INC_OFFNET_DURATION	Total de minutes reçus inter réseau local
INC_INTERAT_DURATION	Total de minutes reçus de l'international
ONNET_SOI	Nombre de personnes différentes appelées Intra réseau
OFFNET_SOI	Nombre de personnes différentes appelées Inter réseau local
INTERNATIONAL_SOI	Nombre de personnes différentes appelées à l'international
SC_NAME	Option tarifaire
TOWN	Ville de résidence
REGION	région de résidence
CVS	Score mettant en évidence la valeur d'un client
CVS_SEGMENT	Segment de valeur associé à ce client
PROMO_TYPE	Type de client que l'on est face à une promotion

I. 2 Extraction de la base de travail et traitement

I. 2.1 Extraction de la base de données

Le travail de notre groupe consiste à détecter les clients non promophiles susceptibles de devenir high promophiles. Pour ce faire, une variable binaire appelée PROMO a été créée. La base extraite est donc constituée de deux groupes de clients : d'une part les « none promophiles » en nombre de 6616 (85.6%) et d'autre part, les « high promophiles » en nombre de 1116 clients (14.4%) soit un total de 7732 individus.

I. 2.2 Traitement de la base

I.2.2.1 Définition des hypothèses

Après extraction de la base de travail, le problème est de reconnaître un client non promophile susceptible de devenir high promophile au vue des données d'aout à octobre.

Nous avons fait les hypothèses de travail suivantes :

- Les clients non promophiles, ayant connu des changements **« brusques »** d'attitude de consommation sur les trois mois, peuvent laisser apparaitre une attrition forte pour une promotion lancée. En effet, l'arrêt ou le lancement d'une promotion sms par exemple peut amener une personne à diviser d'au moins de moitié ou à doubler sa consommation sms.
- les clients non promophile dont la consommation est élevée et constante sur les trois mois s'apparente à des « gros clients » pour qui les promotions (baisse des prix, sms gratuits) n'ont pas d'effet sur les habitudes de consommation.

I.2.2.2 Choix et création des variables

Après définition des hypothèses de travail voici quelques traitements opérés sur la base extraite.

- Les variables : total de minute appelée, total de consommation sms et total de consommation voix ont été enlevées, car chacun de ces totaux est obtenu par combinaison linéaire des variables existant dans la base. En effet, la régression logistique est très sensible au problème de multi colinéarité.
- Les variables (town) et (région) n'ont pas été retenues à cause de la forte proportion des données manquantes et la pertinence de l'information révélée avec notre problématique.
- Les variables scores (CVS10), segment (CVS_SEGMENT10) ont été utilisées pour imputer les données manquantes.

Certaines variables binaires ont été créées pour capter l'effet de changement « brusque » sur les trois mois. Il s'agit de :

- La variable CHANG_ONNET, qui attribue 1 aux personnes qui ont changé brusquement leurs nombres de personnes appelées intra réseau local pendant les trois mois et 0 sinon.
- La variable CHANG_OUT_OFFPEAK qui attribue 1 si la personne a connu un changement brusque dans son nombre de minutes appelées entre 23h et 5h et 0 sinon.
- La variable CHANG_OUT_PEAK qui attribue 1 si la personne a connu un changement brusque dans son nombre de minutes appelées entre 5h et 23h et 0 sinon.
- La variable CHANG_OUT_GOING qui attribue 1 si la personne a connu un changement brusque dans son nombre de minutes appelées intra réseau et 0 sinon.
- La variable GPRS_USERS, retrace le comportement du client face à l'offre internet. Un client qui a consommé l'internet les trois mois, peut ne pas être influencé par les promotions, contrairement à celui qui, au cours des trois mois a consommé l'internet uniquement au cours d'un mois. Cette variable vaut 1 si la personne a consommé l'internet au cours de 1 ou de 2 mois et 0 si elle a consommé les trois mois ou aucun mois.
- La variable INTERNA_USERS retrace le comportement du client face à l'offre d'appel à l'international. Un client qui a appelé des personnes à l'international au cours des trois mois, peut ne pas être influencé par les promotions, contrairement à celui qui au cours des trois mois a appelé des personnes à l'international uniquement au cours d'un mois. La variable vaut 1 si l'individu a appelé à l'international au cours de 1 ou de 2 mois et 0 s'il a appelé les trois mois ou aucun mois.

Cette liste n'est pas exhaustive.

I.2.3 Traitement des données manquantes

Plusieurs variables dans la base de données n'étaient pas renseignées chez certains individus. Le traitement des données manquantes s'est fait par imputation par la moyenne. La variable (segment) a été utilisée. La valeur d'un individu pour qui une variable n'est pas renseignée mais dont on connait le segment sera remplacée par la moyenne de cette variable dans le segment d'appartenance de l'individu.

CHAPITRE II : ANALYSE DESCRIPTIVE ET EXPLICATIVE

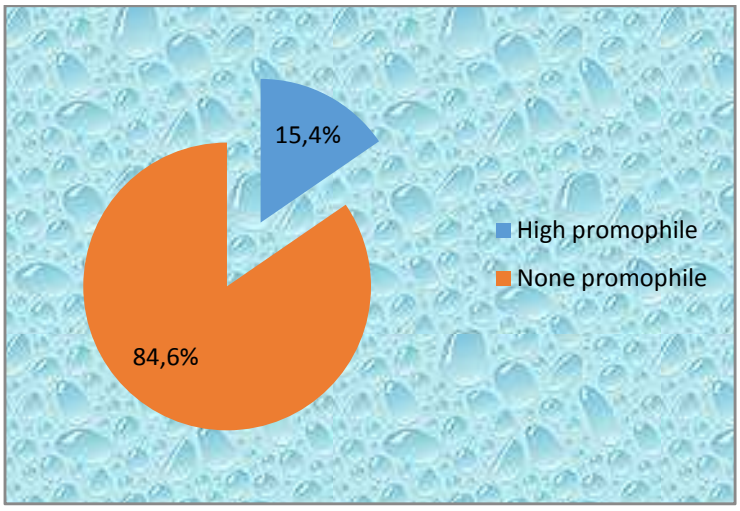
II.1 Analyse exploratoire des données

II.1.1 Analyse uni variée

II.1.1.1 Variable cible

Après apurement du fichier, la base finale retrace les habitudes de consommations sur 7019 individus. La répartition de l'appétence des clients aux offres promotionnelles montre qu'environ 85% de notre échantillon est constitué des clients « non promophiles » contre 15% de clients « high promophile ».

Graphique 1: Distribution par type de client



II.1.1.2 Description des variables quantitatives

Notre base possède 11 variables quantitatives pour le mois d'Octobre. Les caractéristiques de tendance centrale et de dispersion de chacune d'elle sont contenues dans le tableau ci – dessous.

Tableau 2: Tendance et dispersion des variables quantitatives

	Minimum	Maximum	Moyenne	Ecart-type
billable	1	229189	3485,02	6698,3
inc_nb_c	0	888	66,6	76,98
inc_nb_s	1	2232	69,59	129,87
inc_inte	0	54756	225,54	1702,85
inc_onne	1	111917	4009,4	6104,67
onnet	0	667	17,82	27,04
out_inte	0	38253	87,16	791,24
out_nb_s	0	1691	11,54	51,54
out_offn	0	34228	429,03	1211,21
out_offp	0	118204	420,87	2367,7
out_peak	1	224631	3280,17	6309,73
outgoing	1	161675	3158,65	6514,38
offnet	0	315	5,03	8,11

Nous pouvons relever dès lors que le nombre de minutes facturées (*billable*) varie de 1 à 229189 pour un nombre moyen d'environ 3485 minutes par individu. De même, le plus petit nombre de sms reçu *(inc_nb_s)* est de 0 alors que le maximum est de 888 pour un nombre moyen de 66,60 et une dispersion de 76,98. On peut ajouter que le nombre total de minutes appelées entre 5h et 22h59 *(out_peak)* varie entre 1 et 224631 pour une moyenne de 3280,17 et une dispersion de 6309,73. Nous notons aussi de manière générale que la plus petite valeur prise par nos variables est 0 et la plus grande est 229189 (*billable*), les moyennes sont comprises entre 5,03 en plus petite valeur pour *offnet_soi* (Nombre de personnes différentes appelées Inter réseau local) et 4009,40 en plus grande valeur pour *inc_onne* (Total de minutes reçus intra réseau) et enfin les dispersions vont de 8,11 en plus petite valeur pour *offnet_soi* et 6698,30 en plus grande pour *billable.*

II.1.1.3 Variables qualitatives

Au sortir de l'apurement de la base qui nous a été transmise, nous en dénombrons avec 15 variables qualitatives : 14 variables qui ont été créées pour capter les éventuels changements et la dernière, qui existait déjà dans la base d'origine et qui porte sur les options tarifaires.

- **La variable existant**

En parlant de la variable option tarifaire, nous remarquons que l'option la plus prisée par les individus de notre base est MTN ONE, trois personnes sur cinq (60,91%) qui l'utilisent et elle est suivie de l'option MTN BOOSTER avec un peu plus d'une personne sur cinq (23,42%). Les options les moins représentées de notre base sont : BIZ200, Business Prepaid, CALL BOX Providence, MTN Bundle Prepaid qui enregistre chacune 0,01 % de personnes (Voir tableau 3 ci après).

Tableau 3: Distribution de l'option tarifaire

	Fréquence	Pourcentage
BIZ200	1	0,01
Business Prepaid	1	0,01
CALL BOX_Providence	1	0,01
DDS	717	10,22
MTN BOOSTER	1644	23,42
MTN Bundle prepaid	1	0,01
MTN ONE	4275	60,91
MTN99	286	4,07
New SC For Call Box	2	0,03
New Trace Mobile	89	1,27
SAMSUNG_GALAXY	2	0,03
Total	7019	100

Nous allons dès à présent nous intéresser aux différentes variables qui ont été créées.

- **Variables créées**

Dans ce commentaire, nous nous intéressons à la comparaison des proportions (ou pourcentages) des individus chez qui il semble avoir été observé un changement. En effet, cette comparaison nous permet de prime abord d'avoir une idée sur le penchant général des individus qui constituent notre base.

En analysant donc de manière globale le tableau ci –après, il ressort que les variables chan_billa, chan_inc_nb_s, chan_offnet, chan_out_sms, chan_out_off, chan_out_ofp, chan_out_p, chan_outg ont une proportion d'individus chez lesquels l'on a observé un changement largement supérieure à celle chez lesquels on n'observe pas de changement. De même, les variables gprs_use, chan_inc_inter, internat_user, chan_out_int indiquent une

proportion forte plutôt chez les personnes pour lesquelles on n'a pas observé de changement au détriment de l'autre catégorie. Une indécision reste quand à l'appréciation des proportions des deux catégories dans les variables suivantes : chan_inc_nb_c, chan_onnet où l'écart entre les deux proportions n'est pas assez considérable.

Tableau 4: Distribution des variables captant le changement

		Changement observé				Changement observé	
		non	oui			non	oui
Chan_billa	Effectifs	1978	5041	chan_onnet	Effectifs	3419	3600
	Pourcentage	28,18	71,82		Pourcentage	48,71	51,29
gprs_use	Effectifs	5941	1078	chan_out_int	Effectifs	4624	2395
	Pourcentage	84,64	15,36		Pourcentage	65,88	34,12
Chan_inc_inter	Effectifs	5088	1931	chan_out_sms	Effectifs	1510	5509
	Pourcentage	72,49	27,51		Pourcentage	21,51	78,49
Chan_inc_nb_c	Effectifs	3214	3805	chan_out_off	Effectifs	1379	5640
	Pourcentage	45,79	54,21		Pourcentage	19,65	80,35
chan_inc_nb_s	Effectifs	997	6022	chan_out_ofp	Effectifs	2034	4985
	Pourcentage	14,2	85,8		Pourcentage	28,98	71,02
internat_user	Effectifs	5046	1973	chan_out_p	Effectifs	2182	4837
	Pourcentage	71,89	28,11		Pourcentage	31,09	68,91
chan_offnet	Effectifs	2674	4345	chan_outg	Effectifs	1837	5182
	Pourcentage	38,1	61,9		Pourcentage	26,17	73,83

Il ressort de manière globale à l'issue de cette analyse uni variée qu'un peu plus de quatre individus sur cinq de la base de notre projet sont des Non promophiles. De même, on note que l'option tarifaire la plus utilisée est MTN ONE alors que les moins utilisées sont MTN Bundle prepaid, BIZ200…. La statistique uni variée nous a ainsi permis de décrire nos différentes variables mais ne nous permet pas de répondre aux questions comme : Existe –t-il un lien entre le type de promophile et l'option tarifaire? Ou encore, les Non Promophiles ont –ils plus de minutes facturées que les promophiles? Et bien d'autres encore. D'où l'entrée en statistique bi variée pour apporter des éléments de réponse à ces questions.

II.1.2 Analyse bi variée

Comme on peut l'observer sur le tableau qui suit, les high promophiles sont à peu près 5,5 fois moins nombreux que les non promophiles. De plus, il ressort que les high promophiles ont généralement une consommation moyenne supérieure à celles des non promophiles. Par ailleurs la consommation maximale se rencontre presque toujours chez les non promophiles

exception faites pour les variables : total consommation internet, nombre de sms reçu, et nombre de personnes différentes appelées intra-réseau.

On remarque chez les high promophiles, que le total de minutes appelées intra réseau (écart type=10218,46) puis le total de minutes reçues intra réseau (écart type= 9072,24) sont les variables qui varient le plus tandis que le nombre de personnes différentes appelés à l'international (écart type=0,762), puis le nombre de personnes différentes appelées inter réseau local (écart type=4,411) varient le moins. Il s'en suivrait donc que les high promophiles entreprennent beaucoup plus des actions téléphoniques intra réseau qu'inter réseau ou avec l'international.

Pour ce qui concerne les non promophiles, le total de minutes facturées (écart type =6090), total de minutés appelées entre 5h et 22h59 (écart type= 5677,87) sont les variables qui varient le plus et cela conforte le fait qu'ils sont des non promophiles car les minutes facturées élevées et les appels diurnes élevés reflètent le fait qu'ils ne sont pas intéressés ou presque pas par la promotion.

Tableau 5: Statistiques descriptive par type de client

Tableau de bord									
promo_type		**billable**	**gprs**	**inc_inte**	**inc_nb_call**	**inc_nb_sms**	**inc_onne**	**Internat**	**offnet**
Highpromophile	Moyenne	6998,0572	57,6990	126,8763	96,5605	69,6759	7173,9123	,2862	4,4109
	N	1083	1083	1083	1083	1083	1083	1083	1083
	Ecart-type	8540,43978	783,68073	865,98572	86,46088	129,53209	9072,23886	,76261	6,28822
	Minimum	1,00	,00	,00	1,00	1,00	1,00	,00	,00
	Maximum	82125,00	25061,00	20565,00	666,00	2232,00	86933,00	11,00	99,00
None promophile	Moyenne	2844,0775	46,2493	243,5428	61,1351	69,5730	3432,0529	,5056	5,1440
	N	5936	5936	5936	5936	5936	5936	5936	5936
	Ecart-type	6090,49079	487,89228	1813,83620	73,82473	129,94823	5186,67346	1,83414	8,39103
	Minimum	1,00	,00	,00	,00	1,00	1,00	,00	,00
	Maximum	229189,00	15434,00	54756,00	888,00	2139,00	111917,00	36,00	315,00
Total	Moyenne	3485,0178	48,0160	225,5417	66,6011	69,5889	4009,4049	,4717	5,0309
	N	7019	7019	7019	7019	7019	7019	7019	7019
	Ecart-type	6698,29599	544,06786	1702,84879	76,97636	129,87490	6104,66576	1,71490	8,10620
	Minimum	1,00	,00	,00	,00	1,00	1,00	,00	,00
	Maximum	229189,00	25061,00	54756,00	888,00	2232,00	111917,00	36,00	315,00

		onnet	**out_inte**	**out_nb_s**	**out_offn**	**out_offp**	**out_peak**	**outgoing**
Highpromophile	Moyenne	34,4949	21,3426	33,5272	252,1865	1531,7128	6654,3749	7803,0748
	N	1083	1083	1083	1083	1083	1083	1083
	Ecart-type	52,28651	163,08604	89,44526	500,95066	4239,54917	8241,47082	10218,45815
	Minimum	,00	,00	,00	,00	,00	1,00	1,00
	Maximum	667,00	4127,00	1096,00	6278,00	68466,00	75915,00	118585,00
None promophile	Moyenne	14,7807	99,1712	7,5340	461,2985	218,2067	2664,5590	2311,2877
	N	5936	5936	5936	5936	5936	5936	5936
	Ecart-type	17,49962	857,04156	39,72984	1297,00931	1756,68556	5677,87378	5146,90063
	Minimum	,00	,00	,00	,00	,00	1,00	1,00
	Maximum	521,00	38253,00	1691,00	34228,00	118204,00	224631,00	161675,00
Total	Moyenne	17,8225	87,1626	11,5447	429,0335	420,8748	3280,1695	3158,6457
	N	7019	7019	7019	7019	7019	7019	7019
	Ecart-type	27,04064	791,24083	51,54146	1211,21114	2367,69915	6309,72791	6514,38039
	Minimum	,00	,00	,00	,00	,00	1,00	1,00
	Maximum	667,00	38253,00	1691,00	34228,00	118204,00	224631,00	161675,00

La variable d'intérêt PROMO_TYPE est corrélée au seuil de 5% avec toutes les variables quantitatives d'Octobre hormis la variable INC_NB_SMS qui représente le nombre d'sms reçus et la variable GPRS_AMT qui est la consommation totale d'internet. Toutefois, mentionnons que ce lien est le plus fort avec le total de minutes appelées intra réseau (Fisher= 717,364) puis le nombre de personnes différentes appelées intra réseau (Fisher= 523,039). Ainsi, ces deux variables seraient les plus pertinentes pour discriminer un high promophile d'un non promophile. De plus, ce lien est le moins probable avec le nombre total de minutes

reçues de l'international (INC_INTERAT_DURATION), (Fisher=4,301), puis le nombre de personnes différentes appelées à l'international (OFFNET_SOI), (Fisher= 7,499). Voire tableau ci-dessous.

Tableau 6: Table ANOVA

Tableau ANOVA					
	Somme des carrés	**df**	**Moyenne des carrés**	**F**	**Signification**
outgoing * promo_ty	27623231608,307	1	27623231608,307	717,364	,000
onnet * promo_ty	355965,670	1	355965,670	523,039	,000
out_peak * promo_ty	14579842241,088	1	14579842241,088	386,318	,000
billable * promo_ty	15804321571,970	1	15804321571,970	370,808	,000
inc_onne * promo_ty	12823956387,215	1	12823956387,215	361,802	,000
out_offp * promo_ty	1580197295,857	1	1580197295,857	293,630	,000
out_nb_s * promo_ty	618823,675	1	618823,675	240,908	,000
inc_nb_c * promo_ty	1149412,600	1	1149412,600	199,468	,000
out_offn * promo_ty	40050229,782	1	40050229,782	27,403	,000
internat * promo_ty	44,055	1	44,055	15,010	,000
out_inte * promo_ty	5547859,510	1	5547859,510	8,871	,003
offnet * promo_ty	492,291	1	492,291	7,499	,006
inc_inte * promo_ty	12466358,009	1	12466358,009	4,301	,038
gprs * promo_ty	120069,347	1	120069,347	,406	,524
inc_nb_s * promo_ty	9,692	1	9,692	,001	,981

De ce qui précède, on constate que la variable PROMO_TYPE est corrélée à la presque totalité des variables d'Octobre. Toutefois, pour capter l'effet de ces corrélations sur les mois précédents, nous avons utilisé les variables changements au moyen desquels nous avons effectué des Khi-deux tel qu'illustré dans le tableau ci-dessous.

Tableau 7: Table des Khi deux

	Khi-deux de Pearson			
	valeur	ddl	P valeur	v-cramer
chang_billab * promo_ty	3,158	1	0,076	
gprs_user * promo_ty	37,332	1	0,000	0,073
chang_inc_inter * promo_ty	0,554	1	0,459	
chang_inc_nb_call * promo_ty	31,85	1	0,000	0,067
chang_inc_nb_sms * promo_ty	13,744	1	0,000	0,044
chang_offnet * promo_ty	0,425	1	0,518	
chang_onnet * promo_ty	11,575	1	0,001	0,041
chang_outgoing * promo_ty	13,812	1	0,000	0,044
sc_name1 * promo_ty	410,364	1	0,000	0,242
chang_out_inter * promo_ty	1,016	1	0,312	
chang_out_sms * promo_ty	0,234	1	0,63	
chang_out_offnet * promo_ty	4,958	1	0,027	0,027
chang_out_offpeak * promo_ty	127,185	1	0,000	0,135

chang_out_peak * promo_ty	7,622	1	0,006	0,033
interna_ user* promo_ty	13,82	1	0,000	0,044

Il ressort du lien entre la variable cible et les variables qualitatives, illustré dans le tableau ci-dessus que, 10 des 15 variables sont corrélées avec la variable PROMO_TYPE au seuil de 5%. Toutefois, pour chacune des variables corrélées qui captent les changements de consommation sur les trois mois, on note que ce lien est très faible (v-cramer<0,1). Exception faite pour la variable CHANG_OUT_OFFPEAK (v-cramer=0,135) qui capte les changements brusques du nombre de minutes appelées entre 23h et 5h. Ainsi les minutes appelées entre 23h et 5h permettrait de mieux discriminer le fait que l'on soit high promophile du fait qu'on ne le soit pas.

Au sortir de l'analyse bi variée, for est de constater que parmi les 30 variables explicatives, 23 sont corrélées avec la variable type de promophile. Toutefois, nous mentionnons que ce lien est très faible pour plus de la moitié des variables corrélées. La question qui se pose est donc celle de savoir si les analyses multidimensionnelles qui vont suivre pourront confirmer ou infirmer les liens présagés.

II.1.3 Analyse en correspondance multiple (ACM)

L'ACM s'est effectué plusieurs fois (3) en cherchant à avoir une meilleure qualité de projection. Les variables dont la contribution sur les 5 premiers plans factoriels était très faible (inférieure à 20%) ont été enlevées.

- ✓ Choix du nombre d'axes factoriels

Le choix du nombre d'axes est basé sur le critère du coude. Sur le graphique 6 en annexe on observe une cassure au niveau de la deuxième valeur propre. De ce fait, nous allons analyser les deux premiers axes factoriels qui expliquent 31,66% de l'inertie totale.

- ✓ Qualité de la représentation

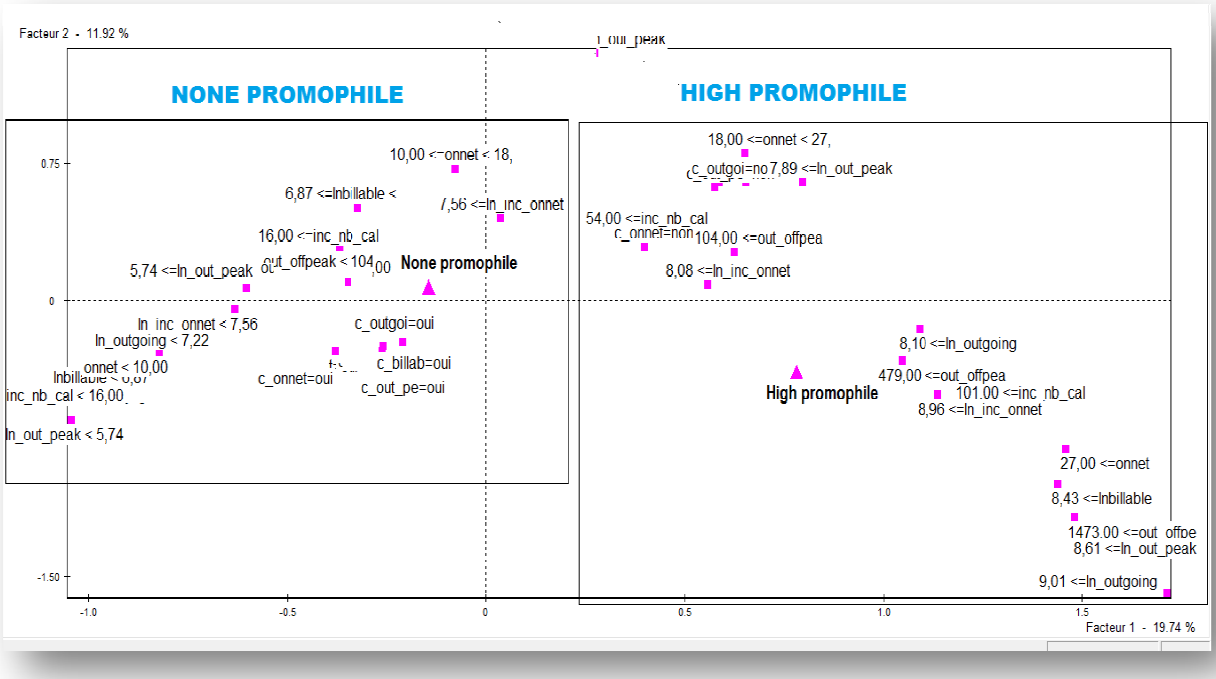
La qualité de représentation d'une variable sur l'axe dépend de la valeur de son cosinus carré. Un cosinus carré élevé traduit une bonne représentation sur l'axe. Nous allons retenir les variables ayant un cosinus carré supérieur à 0,15. (Voire tableau 15 en annexe)

- ✓ Interprétation des résultats

Le premier plan factoriel résume environ 31,66% de l'inertie totale soit 19,74% d'information sur le premier axe factoriel et 11,92 sur le second axe. La variable cible (promo_type) a été mise en illustrative. Elle n'a pas participé à la construction du nuage. Par contre, parmi les variables actives, celles qui ont le plus contribuées sont : le totale de minute fracturé (billable), le nombre total de personnes différentes appelées intra réseau (onnet), le total de minute appelée entre 5h et 22h59 (out_peak) et le total de minute appelé intra réseau. Les variables liées aux changements brusque dans les attitudes de consommation ont très peu contribuées à la construction du premier plan factoriel toutefois, leurs modalités jouissent d'une qualité de projection acceptable (cos^2 >0,15).

Le graphique 2 ci-dessous présente le premier plan factoriel.

Graphique 2: Premier plan factoriel



Le plan ci-dessus montre l'opposition des deux groupes de consommateurs selon leurs habitudes de consommation et leur comportement vis-à-vis des promotions. À gauche, on observe les non promophiles caractérisés par : total de minute facturé inférieur à 962 pour le mois d'octobre, le nombre de personnes différentes appelées est inférieur à 16, leur nombre total de minute appelée aux heures de pointe est inférieur à 310 minutes. A l'opposé, le côté

droit met en exergue les high promophiles. Ces derniers ont un nombre total de minute facturé pour le mois d'octobre supérieur à 4582, le nombre de différentes personnes appelées le même mois est d'environ 30 et le nombre de minute appelée aux heures de pointe est supérieur à 5767.

Il ressort ainsi de l'analyse des correspondances multiples que le premier axe factoriel oppose d'une part les clients ayant une forte valeur d'usage et un répertoire assez fourni aux clients ayant une faible activité sur le réseau et un carnet d'adresse peu fourni.

Nous allons dans la suite effectuer une régression logistique pour confirmer ou infirmer les résultats de l'analyse descriptive, et aussi pour avoir l'effet des variables influentes sur le profil d'un abonné.

II.2 Analyses explicatives

II.2.1 Échantillonnage des données

Notre base de données contient 7019 abonnés avec environ 84,6% de non promophile et 15,4% de high promophiles. Les analyses qui vont suivre ont été faites sur un échantillon d'apprentissage représentant 70% de la base de données soit alors 4913 abonnés. Les résultats que nous avons obtenus ont été évalués sur un échantillon test de 2106 abonnés.

II.2.2 Régression logistique

II.2.2.1 Présentation du modèle

L'étude de l'influence et de l'effet des variables sur la qualité du client se fera par une modélisation économétrique.

La modélisation économétrique se fera au moyen du modèle de régression logistique simple. Il consiste à faire une régression d'une variable qualitative dichotomique par plusieurs autres variables (qualitatives ou quantitatives) en utilisant la méthode du maximum de vraisemblance. Le choix de cette méthode s'est fait sur la base du type de notre variable expliquée. Ainsi cette méthode nous permettra de rechercher les facteurs qui déterminent et

influencent significativement la qualité du client et de quantifier les effets individuels de ceux-ci.

La variable dépendante qui est la qualité du client peut être présentée de la façon suivante : $y = \begin{cases} 1 \text{ si le client est High promophile} \\ 0 \quad \text{si il est None promophile} \end{cases}$

Notons X= (X_1, X_2, X_3…, X_P) le vecteur de variables expliquées, ces variables étant dichotomisées au niveau des modalités. Le modèle de régression logistique à estimer est spécifié comme suit :

$$P(y_i = 1/X) = \frac{e^{X'\beta}}{1 + e^{X'\beta}}$$

Avant toute interprétation des résultats du modèle logistique, il convient de valider les hypothèses sous-jacentes à la qualité du modèle.

II.2.2.2 Evaluation de la qualité du modèle

- ✓ **Test de significativité globale du modèle**

La significativité globale du modèle sera évaluée par le test du rapport de vraisemblance. Ce test permet d'éprouver l'hypothèse nulle :

Ho : Tous les coefficients des variables explicatives sont nuls

Le tableau 16 en annexe nous renseigne que la p-value de ce test est égale à 0,000 qui est inférieure au seuil fixé à 5%. Nous pouvons donc conclure que le modèle est globalement significatif et donc qu'il existe au moins une variable influente sur le comportement du client à être promophile.

- ✓ **Test d'ajustement global du modèle**

Le test de HOMER et LEMESHOW est utilisé pour juger de l'ajustement global du modèle à partir de l'hypothèse nulle :

Ho : Le modèle s'ajuste bien aux données

Nous constatons qu'au seuil de 5% le modèle s'ajuste bien aux données car p-value = 0.23 supérieur à 0,05. On est alors amené à ne pas rejeter l'hypothèse nulle de bon ajustement du modèle aux données.

Par ailleurs, le R^2 de McFadden (0,22) apporte davantage d'évidence sur la qualité du bon ajustement du modèle aux données.

II.2.2.3 Interprétation des résultats

Plusieurs variables incluses dans le modèle sont significatives. Le tableau ci-dessous renseigne sur les coefficients des modalités des différentes variables, leur occurrence (odds ratio) et leur influence sur la probabilité relative d'être un « High promophile » (effets marginaux).

Tableau 8: Récapitulatif sortie de la regression

Variables	Modalités	Coefficient	Odds Ratio	effets marginaux (mfx)	P_valeur
chang_onnet	Oui	REF			
	Non	0,28	1,32	0,028	0,002
Nombre d'appels reçus en octobre	0-16	1,007	2,73	0,131	0
	16-54	1,07	2,94	0,124	0
	54-101	0,54	1,71	0,061	0
	plus de 101	REF			
Nombre total de minutes reçues intra réseau	0-1913	-0,419	0,657	-0,041	0,007
	1913-3215	REF			
	3215-7824	0,509	1,66	0,057	0
	plus de 7824	1,03	2,82	0,139	0
Nombre de personnes différentes appelées Intra réseau	0-10	-0,11	0,892	-0,011	0,469
	10-18	REF			
	18-27	0,17	1,19	0,018	0,241
	plus de 27	0,52	1,68	0,059	0,002
Total de minutes appelées entre 23h et 4h59	0-104	-0,079	0,923	-0,008	0,553
	104-479	REF			
	479-1473	0,59	1,814	0,073	0
	plus de 1473	1,3	3,69	0,199	0
Total de minutes appelées entre 5h et 22h59	0-311	-0,59	0,55	-0,052	0,006
	311-1540	REF			
	1540-2670	-0,443	0,923	-0,04	0,035
	2670-5486	-0,8	0,449	-0,066	0,001
	plus de 5486	-1,05	0,34	-0,083	0
Total de minutes appelées intra réseau	0-1366	-1,12	0,32	-0,119	0
	1366-3294	REF			
	3294-8184	0,789	2,2	0,098	0,006
	plus de 8184	1,187	3,28	0,17	0

Les consommateurs qui ont au cours des trois mois connu un changement brusque dans le nombre de personnes intra réseau qu'ils ont appelé ont environ 1,3 fois plus de chance

d'être des « promophile » que ceux qui n'ont pas connu de changement brusque[1]. On observe également que même si les p_valeur associées au nombre d'appel reçu au cours du mois d'octobre sont significatives, le signe des coefficients et les odds ratio ne permettent pas d'interpréter les modalités en rapport avec notre problématique. En effet, très peu de promotion sont offertes aux consommateurs en matière de réception téléphonique. On n'observe toutefois que les consommateurs qui reçoivent assez d'appels intra réseau ont plus de chance que ceux qui en reçoivent moins d'être des « promophiles ». En matière de flux téléphonique, les opérateurs distinguent deux période celle dite de pointe (5h-22h59) et celle dite de morte (23h- 4h59), ces opérateurs sont ainsi plus prédisposés à offrir des promotions pendant ces périodes mortes. Ce résultat se confirme dans cette étude, où les personnes qui appellent beaucoup pendant la nuit (23h59-05h) ont environ 4 fois plus de chance d'être des « High promohile » et ceux qui appellent beaucoup en journée ont environ 3 fois (1/0,34) plus de chance d'être des non promophiles. De même, les résultats de la régression logistique révèlent que contrairement à ceux qui n'ont pas diversifié leurs contacts, les personnes qui ont au cours du mois d'octobre appelé au moins 27 différentes personnes ont 1,69 fois plus de chance d'être des « high promophile » que non.

Globalement, il ressort des estimations que les consommateurs les plus â même d'être des « high promophiles » sont ceux qui ont connu un changement brusque dans le nombre de personnes différentes qu'ils appellent au cours d'une période (3 mois), ils reçoivent beaucoup d'appels intra réseau, appellent généralement la nuit et très peu en journée et ils ont au cours du mois d'octobre appelé plusieurs personnes différentes.

II.2.2.4 Pouvoir prédictif du modèle

Le tableau ci-dessous résume les informations sur le pouvoir prédictif du modèle sur l'échantillon test exécuté sur le logiciel STATA. Il en ressort que le modèle a un pouvoir prédictif global de 79,41%. Ce pouvoir prédictif est de 72,18% chez les consommateurs « high promophile » et de 80,50% chez les « none promophile ». Le modèle semble mieux classer les non promophiles que les promophiles.

[1] Ils ont doublé ou réduit de moitié le nombre de personnes différentes appelées sur le réseau

Tableau 9: Répartition des clients dans la base de sondage

```
. estat classification in 5000/7019, cutoff(0.15)

Logistic model for promot1

                 True
Classified |         D          ~D |      Total
-----------+----------------------+-----------
         + |       192         342 |        534
         - |        74        1412 |       1486
-----------+----------------------+-----------
     Total |       266        1754 |       2020

Classified + if predicted Pr(D) >= .15
True D defined as promot1 != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   72.18%
Specificity                     Pr( -|~D)   80.50%
Positive predictive value       Pr( D| +)   35.96%
Negative predictive value       Pr(~D| -)   95.02%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   19.50%
False - rate for true D         Pr( -| D)   27.82%
False + rate for classified +   Pr(~D| +)   64.04%
False - rate for classified -   Pr( D| -)    4.98%
--------------------------------------------------
Correctly classified                        79.41%
--------------------------------------------------
```

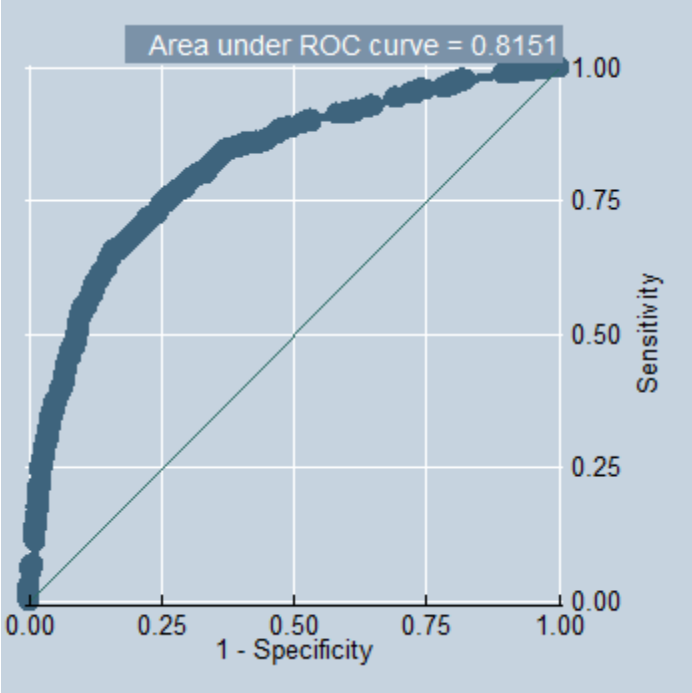
II.2.2.5 Le pouvoir discriminant du modèle

La courbe ROC nous donne l'information en ce qui concerne la capacité à discriminer du modèle. Elle est en fait une représentation de la sensibilité du modèle en fonction de sa spécificité. Lorsque :

- Aire ROC = 0,5 il n'y a pas de discrimination ;
- 0,5 < aire ROC < 0,8 la discrimination est acceptable ;
- 0,8 ≤ aire ROC <0,9 la discrimination est excellente ;
- Aire ROC ≥ 0,9 la discrimination est exceptionnelle.

Le graphique ci-après présente la courbe de ROC de notre modèle et nous permet de dire que la discrimination est excellente étant donnée la valeur de la surface en-dessous de la courbe ROC = 0,8151.

Graphique 3: Courbe ROC



CHAPITRE III : METHODES D'INTELLIGENCE ARTIFICIELLE

Il est difficile de parler de Data Mining aujourd'hui sans parler de réseau de neurones et des Support Vector Machines (SVM). Ces méthodes sont largement répandues grâce à leur puissance de modélisation. Elles peuvent approcher n'importe quelle fonction suffisamment régulière. Cependant, leur utilisation est souvent freinée par les difficultés qu'elles présentent : le côté « boîte noire », la délicatesse des réglages à effectuer, la puissance des ordinateurs requise et le risque de sur- apprentissage.

III.1 RESEAUX DE NEURONES

III.1.1 Description de la méthode

Il s'agit d'un modèle non linéaire constitué de trois couches (entrée, cachée et sortie). Les variables continues et les différentes modalités des variables qualitatives correspondent à un nœud de premier niveau, appelé couche d'entrée. La variable à expliquer (binaire) correspond à un nœud de deuxième niveau (couche de sortie). Entre ces deux couches, sont généralement connectés des nœuds appartenant à un niveau intermédiaire (couche cachée). Cette couche est composée de neurones cachés qui ont chacune pour tâche d'ajuster une partie des points connus tout en tenant compte de l'activation des autres neurones cachés. Les étapes de mise en œuvre sont :

- L'identification des données en entrée et en sortie
- La normalisation des données
- La constitution d'un réseau avec une structure adaptée
- L'apprentissage du réseau
- Le test du réseau
- L'application du modèle généré par l'apprentissage
- La dénormalisation des données en sortie

III.1.2 Mise en œuvre

La mise en œuvre s'est faite sur le logiciel TANAGRA. Les variables quantitatives ont été discrétisées en classes et dichotomisées. De sorte qu'on a autant de nœuds que de modalités. En effet, la normalisation peut faire apparaître une relation d'ordre qui n'existe pas et peut induire le réseau en erreur. La couche cachée est constituée de deux neurones et la fonction de transfert utilisée dans chacune des couches est la fonction logistique.

Les poids synaptiques des différents nœuds (modalités des variables discrétisées) sont présentés sur le tableau 10 ci-dessous.

Tableau 10: Couches cachées du Neurone

-	Neuron "1"	Neuron "2"
c_onnet2	0,41949096	-0,91505987
onnet2	-3,00260615	1,39343818
onnet3	-5,52466854	1,94545330
onnet4	-2,76042997	1,25030356
inc_nb_cal2	-4,18276059	-1,78389089
inc_nb_cal3	-2,22767027	0,58581435
inc_nb_cal4	-1,55979524	-2,91855483
inc_onnet1	4,08450483	1,17288865
inc_onnet2	1,27485136	1,28443791
inc_onnet4	2,93430411	2,78278190
out_offpeak2	-3,83842128	-2,89084946
out_offpeak3	1,02106144	-4,62312374
out_offpeak4	0,99370163	-0,92331341
ln_out_peak2	0,62037424	0,47567305
ln_out_peak3	1,01057807	-0,03074138
ln_out_peak4	4,90862612	-2,30443062
ln_out_peak5	0,60435837	4,28167124
ln_outgoing2	0,94740677	-1,59795556
ln_outgoing3	-1,57982043	-3,05216600
ln_outgoing4	2,15163312	2,23648336
bias	1,55461138	-2,65901834

La matrice de confusion réalisée sur l'échantillon test est présentée ci-dessous.

Tableau 11: Matrice de confusion des Neurones

Error rate	0,1372

Values prediction

Value	Recall	1-Precision
None promophile	0,9709	0,1200
Highpromophile	0,2500	0,3969

Confusion matrix

	None promophile	Highpromophile	Sum
None promophile	1738	52	1790
Highpromophile	237	79	316
Sum	1975	131	2106

Le taux d'erreur est de 13,72% soit un taux de bon classement d'environ 86.28%.

III.2 SUPPORT VECTOR MACHINE (SVM)

III.2.1 Description de la méthode

Ils sont basés sur les travaux de Vladimir Vapnik en théorie de l'apprentissage. Il s'agit en quelque sorte d'une analyse discriminante généralisée effectuée dans un espace de dimension assez grande pour qu'existe une séparation linéaire. En effet, l'idée est de trouver parmi une infinité de séparateurs linéaires entre les classes, celui à vastes marges, c'est-à-dire qui maximise l'écart entre les classes (robustesse**).**

Elle se déroule en deux étapes :

- Dans la première, une transformation fait passer de l'espace d'origine dans un espace de dimension plus grande.
- Dans la seconde, on cherche un séparateur linéaire qui est un hyperplan.

III.2.2 Mise en œuvre

La mise en œuvre s'est faite sur le logiciel TANAGRA. Le filtre utilisé est un polynôme de degré 2. Le tableau 12 ci-dessous présente la matrice de confusion de l'échantillon test

Tableau 12: Matrice de confusion des SVM

Error rate	0,1292

Values prediction

Value	Recall	1-Precision
None promophile	0,9751	0,1141
Highpromophile	0,2408	0,3846

Confusion matrix

	None promophile	Highpromophile	Sum
None promophile	1762	45	1807
Highpromophile	227	72	299
Sum	1989	117	2106

La SVM fournit un taux d'erreur de 0.1292 correspondant à un taux de bon classement d'environ 87%.

Les paramètres d'exécution de cette méthode sont présentés dans le tableau 13 qui suit.

Tableau 13: Paramètres des SVM

SVM Parameters	
Exponent	2
Filter type	NORMALIZE
Use polynom space normalization	1
Use RBF kernel	0
Gamma for RBF kernel	0,0100
Complexity	1,0000
Calculation parameter	
Epsilon for rounding	1,0E-012
Tolerance for accuracy	1,0E-003

CHAPITRE IV : COMPARAISON DES TROIS METHODES

IV.1 POURVOIR PREDICTIF

Le tableau ci-dessous compare le pouvoir prédictif des trois méthodes sur un échantillon test. Le support vecteur machine est la méthode qui a le meilleur taux de bon classement (87,08). En effet, les SVM ont l'avantage de modéliser les phénomènes non linéaires, ils bénéficient par conséquent d'une meilleure précision dans la prédiction. Cependant, cette règle est en proie à deux types d'erreur. La fixation du seuil entraine des effets antagonistes sur la sélection des promophiles et des non promophiles. La courbe ROC vient en quelque sorte pallier à ce problème de choix du seuil dans la comparaison des modèles.

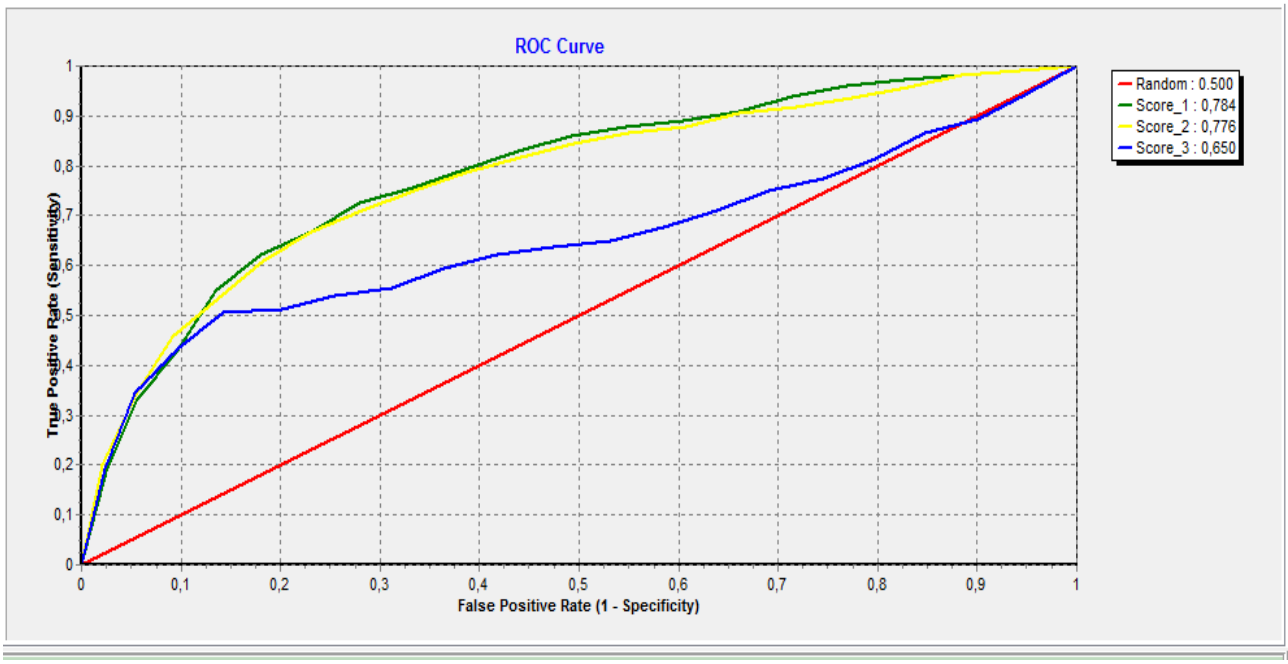
Tableau 14: Comparaison des trois modèles

comparaison des trois méthodes d'apprentissage	ROC (AUC)	TBC
Régression logistique	0,784	85,660%
Réseau de neurones	0,776	86,280%
Support vecteur machine (SVM)	0,65	87,080%

IV.2 COURBE ROC

L'option scoring implémenté sur le logiciel TANAGRA a permis de générer pour chaque méthode d'apprentissage, le score des consommateurs. Ces modèles ont été utilisés sur l'échantillon test pour construire la courbe ROC. L'aire en dessous est un bon critère de comparaison. Plus cette aire est proche de 1, meilleur est le modèle. Voire graphique qui suit

Graphique 4: Comparaison Courbes ROC des modèles



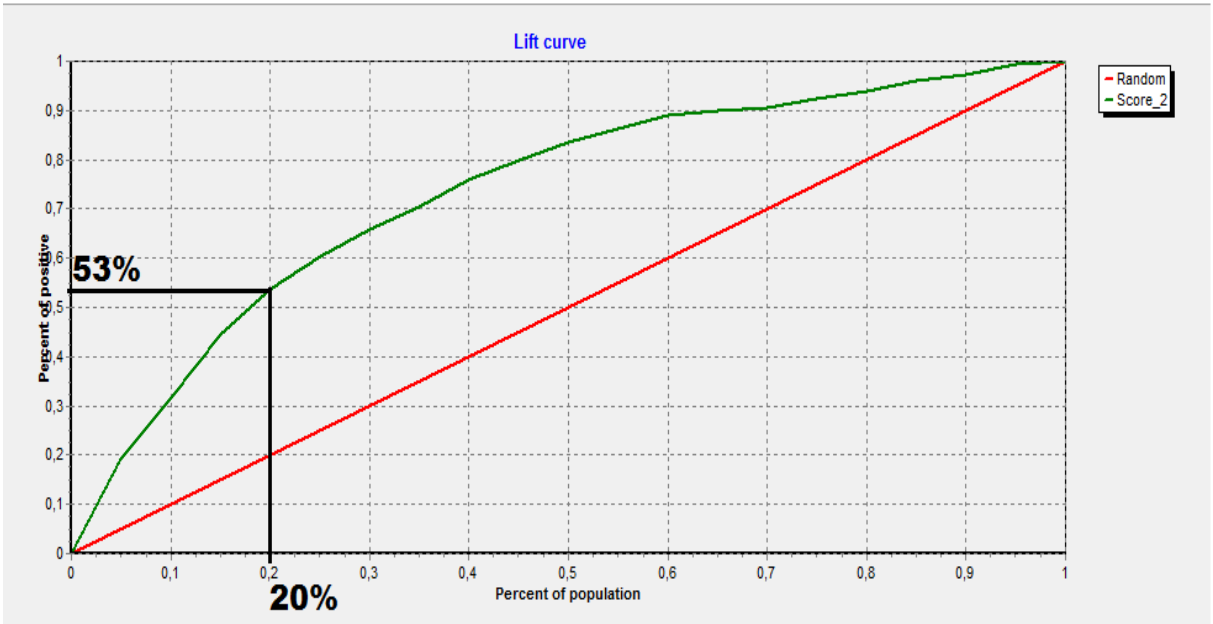
La comparaison des aires en dessous de la courbe ROC (AUC) des trois modèles (vert=logistique, jaune=neurone et bleu=SVM) montre que le support vecteur machine a la plus petite aire sous la courbe ROC (AUC=0.65). La régression logistique a la plus grande aire sous la courbe (AUC=0.784).

Cette dernière sera utilisée pour effectuer un ciblage marketing afin d'identifier les clients les plus susceptibles de devenir « High promophile ».

IV.3 COURBE LIFT ET CIBLAGE

Il est question ici de pouvoir identifier les clients les plus réceptifs des promotions. Car faisant face à un budget limité, il n'est également pas nécessaire d'agacer les clients « hostiles ».

Graphique 5: Courbe LIFT de la régression



Pour identifier parmi les consommateurs de cet opérateur les plus susceptibles de devenir des « high promophiles », ayant trouvé le bon modèle (modèle logistique) et généré pour chaque individu son score, nous avons trié la base de départ par ordre décroissant. Les individus « none promophile » ayant des scores élevés sont ceux qui ont plus de chance d'être « high promophile ».

Le graphique ci-dessus montre que si on cible par exemple 20% des individus de la base triée, 53% d'entre eux sont susceptibles de devenir high promophiles.

CONCLUSION ET RECOMMANDATIONS

Il était question pour nous dans ce travail de prédire si un client qui est *non promophile* peut devenir ou non high promophile en se servant de quelques informations recueillies sur son comportement en téléphonie. La base de données a été traitée afin de corriger les incohérences présentes dans les données. Les méthodes d'analyse descriptive et d'analyse factorielle ont été utilisées pour appréhender le lien entre la variable cible et les variables pertinentes. Le modèle de régression logistique et les méthodes d'intelligence artificielle notamment les réseaux de neurones et SVM ont été utilisées. La régression logistique s'est avérée être le modèle le plus pertinent. Elle nous a permit de construire un score à partir duquel les non promophiles ayant un score élevé (supérieur à 0,6) se sont révélés être les plus susceptibles de devenir high promophiles.

Les recommandations qui se dégagent de ce travail sont les suivantes :

- L'opérateur doit repenser sa stratégie de positionnement des offres sms, des appels internationaux et des services internet (GPRS) parce que les résultats issus de nos analyses montrent que l'appétence des clients aux promotions n'est pas influencée par ces variables d'usage.
- Les variables retraçant le comportement d'appel inter réseau ne discriminent pas assez les deux groupes. Ainsi, l'opérateur doit davantage offrir des promotions d'appel inter réseau.

ANNEXE

Graphique 6: Histogramme des valeurs propres

NUMERO	VALEUR PROPRE	POURCENTAGE	POURCENTAGE CUMULE	
1	0.4666	19.74	19.74	**********
2	0.2816	11.92	31.66	**********
3	0.2220	9.39	41.05	**********
4	0.1500	6.35	47.39	**********
5	0.1336	5.65	53.04	**********
6	0.1106	4.68	57.72	**********
7	0.1001	4.24	61.96	**********
8	0.0930	3.93	65.89	**********
9	0.0911	3.86	69.75	**********
10	0.0852	3.60	73.35	**********
11	0.0796	3.37	76.72	**********
12	0.0776	3.28	80.00	**********
13	0.0629	2.66	82.66	**********
14	0.0616	2.60	85.27	**********
15	0.0565	2.39	87.66	**********
16	0.0546	2.31	89.97	**********
17	0.0519	2.20	92.17	**********
18	0.0481	2.03	94.20	**********
19	0.0317	1.34	95.54	**********
20	0.0286	1.21	96.75	**********
21	0.0202	0.86	97.61	**********
22	0.0169	0.71	98.32	**********
23	0.0162	0.69	99.01	**********
24	0.0128	0.54	99.55	**********
25	0.0074	0.31	99.86	**********
26	0.0033	0.14	100.00	**********

Tableau 15: Informations sur les coordonnées des points

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES MODALITES ACTIVES
AXES 1 A 5

MODALITES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN - LIBELLE	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
3 . c_billable																	
AA_1 - c_billab=non	2.56	2.55	0.65	0.64	1.09	-0.28	-0.03	2.4	3.8	13.8	1.3	0.0	0.17	0.16	0.47	0.03	0.00
AA_2 - c_billab=oui	6.53	0.39	-0.26	-0.25	-0.43	0.11	0.01	0.9	1.5	5.4	0.5	0.0	0.17	0.16	0.47	0.03	0.00
						CONTRIBUTION CUMULEE =		3.3	5.2	19.2	1.8	0.0					
19 . c_onnet																	
AH_1 - c_onnet=non	4.43	1.05	0.40	0.29	0.43	-0.03	-0.01	1.5	1.3	3.7	0.0	0.0	0.15	0.08	0.18	0.00	0.00
AH_2 - c_onnet=oui	4.66	0.95	-0.38	-0.28	-0.41	0.03	0.01	1.4	1.3	3.5	0.0	0.0	0.15	0.08	0.18	0.00	0.00
						CONTRIBUTION CUMULEE =		2.9	2.6	7.2	0.0	0.0					
30 . c_out_peak																	
AM_1 - c_out_pe=non	2.83	2.22	0.58	0.62	1.05	-0.26	-0.04	2.0	3.8	14.0	1.3	0.0	0.15	0.17	0.49	0.03	0.00
AM_2 - c_out_pe=oui	6.26	0.45	-0.26	-0.28	-0.47	0.12	0.02	0.9	1.7	6.3	0.6	0.0	0.15	0.17	0.49	0.03	0.00
						CONTRIBUTION CUMULEE =		2.9	5.5	20.2	1.9	0.0					
33 . c_outgoing																	
AN_1 - c_outgoi=non	2.38	2.82	0.59	0.64	1.11	-0.30	0.01	1.8	3.5	13.2	1.4	0.0	0.12	0.15	0.43	0.03	0.00
AN_2 - c_outgoi=oui	6.71	0.35	-0.21	-0.23	-0.39	0.11	-0.01	0.6	1.2	4.7	0.5	0.0	0.12	0.15	0.43	0.03	0.00
						CONTRIBUTION CUMULEE =		2.4	4.7	17.8	1.9	0.0					
37 . lnbillable_casier																	
AR_1 - 6,87 <=lnbillable <	1.39	5.55	-0.32	0.50	0.37	1.47	0.11	0.3	1.3	0.9	20.1	0.1	0.02	0.05	0.03	0.39	0.00
AR_2 - 7,47 <=lnbillable <	2.08	3.37	0.48	1.18	-0.87	-0.47	-0.40	1.0	10.2	7.0	3.0	2.6	0.07	0.41	0.22	0.07	0.05
AR_3 - 8,43 <=lnbillable	1.86	3.88	1.44	-1.00	0.07	0.00	0.31	8.3	6.6	0.0	0.0	1.3	0.53	0.26	0.00	0.00	0.02
AR_4 - lnbillable < 6,87	3.76	1.42	-0.86	-0.34	0.31	-0.29	0.03	6.0	1.6	1.6	2.0	0.0	0.53	0.08	0.07	0.06	0.00
						CONTRIBUTION CUMULEE =		15.6	19.6	9.6	25.2	4.0					
38 . inc_nb_cal_casier																	
AS_1 - 101,00 <=inc_nb_cal	1.86	3.89	1.23	-0.43	-0.01	-0.01	-0.55	6.0	1.2	0.0	0.0	4.3	0.39	0.05	0.00	0.00	0.08
AS_2 - 16,00 <=inc_nb_cal	3.35	1.71	-0.37	0.27	0.10	0.25	0.14	1.0	0.9	0.1	1.4	0.5	0.08	0.04	0.01	0.04	0.01
AS_3 - 54,00 <=inc_nb_cal	1.96	3.64	0.37	0.37	-0.32	0.30	0.70	0.6	0.9	0.9	1.2	7.2	0.04	0.04	0.03	0.02	0.14
AS_4 - inc_nb_cal < 16,00	1.92	3.73	-0.92	-0.44	0.16	-0.74	-0.43	3.5	1.3	0.2	7.0	2.6	0.23	0.05	0.01	0.15	0.05
						CONTRIBUTION CUMULEE =		11.1	4.4	1.3	9.6	14.6					
40 . ln_inc_onnet_casier																	
AU_1 - 7,56 <=ln_inc_onnet	1.49	5.11	0.04	0.45	-0.11	0.41	0.71	0.0	1.1	0.1	1.7	5.6	0.00	0.04	0.00	0.03	0.10
AU_2 - 8,08 <=ln_inc_onnet	2.14	3.25	0.56	0.09	-0.16	0.18	0.49	1.4	0.1	0.3	0.5	3.9	0.10	0.00	0.01	0.01	0.07
AU_3 - 8,96 <=ln_inc_onnet	1.25	6.30	1.14	-0.51	-0.05	0.12	-0.94	3.5	1.2	0.0	0.1	8.2	0.21	0.04	0.00	0.00	0.14
AU_4 - ln_inc_onnet < 7,56	4.22	1.15	-0.63	-0.05	0.13	-0.27	-0.22	3.6	0.0	0.3	2.1	1.6	0.35	0.00	0.02	0.06	0.04
						CONTRIBUTION CUMULEE =		8.5	2.3	0.7	4.3	19.2					
41 . onnet_casier																	
AV_1 - 10,00 <=onnet < 18,	2.15	3.23	-0.08	0.71	-0.09	0.89	-0.14	0.0	3.9	0.1	11.4	0.3	0.00	0.16	0.00	0.25	0.01
AV_2 - 18,00 <=onnet < 27,	1.26	6.22	0.65	0.80	-0.71	-0.49	0.35	1.1	2.8	2.9	2.0	1.1	0.07	0.10	0.08	0.04	0.02
AV_3 - 27,00 <=onnet	1.76	4.16	1.46	-0.81	0.08	-0.07	0.01	8.0	4.1	0.0	0.1	0.0	0.51	0.16	0.00	0.00	0.00
AV_4 - onnet < 10,00	3.92	1.32	-0.82	-0.28	0.24	-0.30	-0.04	5.7	1.1	1.0	2.4	0.1	0.51	0.06	0.04	0.07	0.00
						CONTRIBUTION CUMULEE =		14.9	11.9	4.0	15.8	1.5					
45 . out_offpeak_casier																	
AZ_1 - 104,00 <=out_offpea	1.07	7.52	0.62	0.26	-0.26	-0.12	0.19	0.9	0.3	0.3	0.1	0.3	0.05	0.01	0.01	0.00	0.00
AZ_2 - 1473,00 <=out_offpe	0.59	14.36	1.48	-1.18	-0.03	0.37	-1.00	2.8	2.9	0.0	0.6	4.4	0.15	0.10	0.00	0.01	0.07
AZ_3 - 479,00 <=out_offpea	0.74	11.34	1.05	-0.33	-0.39	0.02	0.41	1.7	0.3	0.5	0.0	0.9	0.10	0.01	0.01	0.00	0.01
AZ_4 - out_offpeak < 104,00	6.69	0.36	-0.35	0.10	0.09	-0.02	0.01	1.7	0.2	0.2	0.0	0.0	0.33	0.03	0.02	0.00	0.00
						CONTRIBUTION CUMULEE =		7.1	3.7	1.0	0.7	5.7					
46 . ln_out_peak_casier																	
BA_1 - 5,74 <=ln_out_peak	3.38	1.69	-0.60	0.07	0.47	0.71	0.27	2.6	0.1	3.3	11.4	1.8	0.21	0.00	0.13	0.30	0.04
BA_2 - 7,34 <=ln_out_peak	1.29	6.05	0.27	1.34	-0.74	0.15	-1.24	0.2	8.2	3.2	0.2	14.8	0.01	0.30	0.09	0.00	0.25
BA_3 - 7,89 <=ln_out_peak	1.40	5.49	0.80	0.64	-0.75	-0.92	0.85	1.9	2.1	3.6	8.0	7.6	0.12	0.08	0.10	0.16	0.13
BA_4 - 8,61 <=ln_out_peak	1.44	5.32	1.54	-1.27	0.21	0.25	-0.04	7.3	8.2	0.3	0.6	0.0	0.45	0.30	0.01	0.01	0.00
BA_5 - ln_out_peak < 5,74	1.59	4.73	-1.04	-0.65	0.08	-1.05	-0.27	3.7	2.4	0.0	11.6	0.9	0.23	0.09	0.00	0.23	0.02
						CONTRIBUTION CUMULEE =		15.8	20.9	10.4	31.8	25.0					
47 . ln_outgoing_casier																	
BB_1 - 7,22 <=ln_outgoing	1.78	4.12	0.40	1.30	-0.75	-0.05	-0.75	0.6	10.7	4.5	0.0	7.6	0.04	0.41	0.14	0.00	0.14
BB_2 - 8,10 <=ln_outgoing	1.39	5.54	1.09	-0.16	-0.44	-0.69	1.28	3.6	0.1	1.2	4.5	17.0	0.22	0.00	0.03	0.09	0.30
BB_3 - 9,01 <=ln_outgoing	0.87	9.41	1.71	-1.60	0.33	0.59	-0.89	5.5	7.9	0.4	2.0	5.1	0.31	0.27	0.01	0.04	0.08
BB_4 - ln_outgoing < 7,22	5.05	0.80	-0.74	-0.14	0.33	0.11	0.07	5.9	0.5	2.4	0.4	0.2	0.68	0.02	0.13	0.01	0.01
						CONTRIBUTION CUMULEE =		15.5	19.0	8.6	6.9	29.9					

Tableau 16: Récapitulatif du modèle logit

Logistic regression — Number of obs = 5000; LR chi2(20) = 962.30; Prob > chi2 = 0.0000; Log likelihood = −1698.8668; Pseudo R2 = 0.2207

promotion2	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Ic_onnet_2	.7788559	.0716196	−2.72	0.007	.6504067	.9326727
_Iinc_nb_c~2	.8321616	.1451452	−1.05	0.292	.5912096	1.171315
_Iinc_nb_c~3	.4571358	.1005469	−3.56	0.000	.297046	.7035045
_Iinc_nb_c~4	.2850822	.0715503	−5.00	0.000	.1743148	.4662361
_Iinc_onne~2	1.524091	.2389956	2.69	0.007	1.120811	2.072476
_Iinc_onne~3	2.403669	.3810592	5.53	0.000	1.761699	3.279575
_Iinc_onne~4	4.3913	.8416412	7.72	0.000	3.016136	6.39345
_Ionnet_ca~2	1.347492	.2101262	1.91	0.056	.9926384	1.829201
_Ionnet_ca~3	1.366211	.2590744	1.65	0.100	.9421167	1.981213
_Ionnet_ca~4	2.255482	.4624834	3.97	0.000	1.509043	3.371142
_Iout_offp~2	1.071892	.1449224	0.51	0.608	.8223681	1.397125
_Iout_offp~3	1.928191	.2670936	4.74	0.000	1.469743	2.52964
_Iout_offp~4	4.06389	.6479145	8.79	0.000	2.973262	5.554575
_Iln_out_p~2	2.107834	.4873024	3.23	0.001	1.339831	3.316061
_Iln_out_p~3	1.207766	.3642313	0.63	0.531	.6687784	2.181139
_Iln_out_p~4	1.021709	.3289635	0.07	0.947	.5435786	1.9204
_Iln_out_p~5	.7889332	.2799605	−0.67	0.504	.3935349	1.581602
_Iln_outgo~2	3.558758	.7479048	6.04	0.000	2.357278	5.372619
_Iln_outgo~3	6.920597	1.859866	7.20	0.000	4.086849	11.71922
_Iln_outgo~4	11.22271	3.673476	7.39	0.000	5.908515	21.31658