Rapport de projet en programmation et traitement statistique des données

Rendu le 20 décembre 2020

Analyse des facteurs liés à la souscription d'un compte à terme

---

Réalisé par : Diamondra RAKOTONDRAZAKA, étudiant en 4e année du Cursus Master en Ingénierie (CMI) en sciences des données à l'Université Bretagne Sud (UBS) de Vannes

Professeur : Salim LARDJANE, maitre de conférence UBS

Résumé

Dans le système bancaire, la souscription à un compte à terme ou dépot à terme est la création d'un compte épargne bloqué pour une durée qui peut être renouvelée.
Ce placement financier est rémunéré suivant un taux connu par l'épargnant au départ, et ne peut être réapprovisionné durant la durée du blocage.

— L'interêt pour un particulier de souscrire à un compte à terme, est de placer une somme d'argent (idéalement très importante) de manière sécurisée. Pour ensuite en tirer, suivant le taux mis en place durant la souscription et la durée, une rémunération.

— L'intérêt pour la banque est de disposer de fonds placés, et d'en tirer une marge en y applicant un taux inférieur à celui du taux moyen pratiqué sur le marché interbancaire.

Cette décision sera prise au terme d'une campagne de marketing direct de la banque pour ce produit d'épargne. Les clients ciblés par la campagne, ont été sollicités par appel téléphonique.

L'objectif de cette étude est de savoir quels facteurs impactent la décision d'un client, à souscrire à un compte à terme.

Plus précisément, de voir quelle stratégie de marketing améliore le succès d'une campagne. Et quels sont les profils des clients qui sont succeptible d'accepter de souscrire à un dépôt à terme.

Les clients susceptibles de souscrire à un compte à terme, sont ceux qui disposent d'un capital suffisant. Cela en fonction de leur situation professionnelle ou familiale.
La stratégie de marketing direct de la campagne, sera d'autant plus efficace selon les périodes de l'année choisies et si le client a déjà répondu favorablement à une précédente campagne.

Mot clés : compte à terme, dépot à terme, marketing direct

Table des matières

1 Introduction

Les données de cette étude proviennent d'une campagne de marketing direct, d'une institution bancaire portugaise.[1]

Le choix de cette base de données est due au fait que la thématique bancaire, n'a pas beaucoup été abordée dans les cours ou travaux dirigés de la formation de M1 data science et modélisation statistique de l'UBS.

Les campagnes de marketing direct de compte à terme ont ciblés au total 45211 clients. Le fichier csv contenant les données à été pris sur le site de l'université américaine d'Irvine (Californie), Donald Bren School of Information and Computer Sciences. [**?**] Dans la rubrique menant à leur entrepôt de jeux de données dédiés au Machine Learning.

La variable d'intérêt y est catégorielle (binaire) et porte sur la survenue d'une souscription d'un client à un compte à terme. La variable duration est étroitement liée à y. Car il est évident que si la durée d'appel téléphonique d'un client devient longue, le client sera engagé et souscrira à un compte à terme. On connait d'ailleurs à la fin de chaque appel, la réponse des clients.
Il y a 16 autres variables majoritairement de type catégorielle. La description des variables se trouve en annexe, sur la Figure 19. Il existe une version du jeu de données comportant 3 variables de plus, mais celles-ci portent sur des indicateurs économiques. Comme ils ne sont pas intrinsèquement liés aux clients et à la stratégie de marketing, on décide donc prendre la version avec 17 variables.

La problématique de cette étude est de savoir quels sont les facteurs principaux qui amènent un client accepter la souscription à un compte à terme. En les connaissant, une banque pourra mieux ajuster la stratégie de sa campagne de marketing direct. Afin de faire souscrire plus de client et de manière plus efficace. Car des clients ciblés par les appels peuvent être sollicité plus d'une fois, sans pour autant souscrire à un compte à terme.
Ce qui peut entraîner un coût en terme de temps et de productivité pour les conseillers effectuant les appels.

La méthodologie utilisée est dans un premier temps, la préparation des données pour notre étude. Ensuite l'analyse univariée, portant sur une approche exploratoire et descriptive des données donnera une description globale des variables de l'échantillon.
Puis à partir de l'analyse bivariée, l'étude va se diviser en 2 grands axes :

— Les données personnelles des clients.
— La stratégie de marketing direct.

. Cette analyse constitue une étape faite au préalable précédant l'analyse multivariée. L'analyse bivariée aura pour objectif de détecter les liaisons statistiques entre y et chaque variable explicative.
Pour finir, l'analyse multivariée comportera toutes les variables explicatives retenues sur à l'analyse bivariée, dans le but de savoir quels sont les profils des clients qui sont succeptible de souscrire à un compte à terme. Et voir quels sont les facteurs associés à la stratégie de marketing direct, qui augmentent la probalité qu'un client réponde favorablement à la campagne.

2 Traitements des données

Afin de préparer les données pour notre étude statistique, nous avons besoin de nous assurer que ces dernières soient consistantes et cohérentes.

2.1 Données manquantes

Une gestion des valeurs manquantes doit tout d'abord être mise en place. Seules certaines variables catégorielles en contiennent. Les valeurs manquantes sont associées à la classe "unknown". Voici ci-dessous sur la Figure 1, les variables contenant des valeurs manquantes et leur quantification de la modalité "unknown" :

Caractéristique	N = 45 211[1]
job	
admin.	5 171 (11%)
blue-collar	9 732 (22%)
entrepreneur	1 487 (3,3%)
housemaid	1 240 (2,7%)
management	9 458 (21%)
retired	2 264 (5,0%)
self-employed	1 579 (3,5%)
services	4 154 (9,2%)
student	938 (2,1%)
technician	7 597 (17%)
unemployed	1 303 (2,9%)
unknown	288 (0,6%)

[1] Statistique présentée: n (%)

Caractéristique	N = 45 211[1]
education	
primary	6 851 (15%)
secondary	23 202 (51%)
tertiary	13 301 (29%)
unknown	1 857 (4,1%)
contact	
cellular	29 285 (65%)
telephone	2 906 (6,4%)
unknown	13 020 (29%)
poutcome	
failure	4 901 (11%)
other	1 840 (4,1%)
success	1 511 (3,3%)
unknown	36 959 (82%)

[1] Statistique présentée: n (%)

FIGURE 1 – Tables des variables catégorielles possédant des valeurs manquantes

On voit que pour les variables job et éducation, il y a un nombre négligeable de valeurs manquantes (inférieur à 5%). On va donc simplement supprimer les individus associés.

En revanche pour les variables contact et poutcome, il y a des parts très importantes de valeurs manquantes (29% et 82% des données). On pourrait trouver un moyen de donner une valeurs pour ces valeurs manquantes, comme dans le cas de variable quantitatives où on prendrait la médiane. Ici on pourrait prendre la classe la plus représentée de la variable. Ce qui ne semble pas judicieux car cela ne donne aucune idée approchée, sûre et pertinente de la classe réelle manquante.

On conservera dans ce cas la modalité "unknown" comme classe à part entière des variables contact et poutcome.

2.2 Variables non utilisées

Il y a 2 variables qu'on ne prendra pas en compte dans notre étude.
La variable pdays semble inutile, de part la définition de ce qu'elle représente. On va supposer que le nombre de jours passés après le dernier contact pour une précédente campagne, ne présente aucun d'intérêt (aucun apport d'information).
Ensuite, la variable day portant sur le jour du mois (numériquement) pourrait plus précisément être vu comme une variable multi-classe (1 à 30). Ce qui devient fastidieux, au vu du grand nombre de modalité. On va donc partir de l'hypothèse qu'une banque s'intéressera à l'influence du mois où elle a contacté un particulier et non le jour.

2.3 Recodage de variable

Concernant le recodage de variable, il est intéressant de recoder la variable age en classes. On fera de même pour la variable balance, afin d'avoir une description facilement interprétable plutôt que des valeurs numériques pour le solde bancaire.
Plus tard pour la suite (pour faciliter la construction de modèles de regression logistique), on décomposera la variable month en plusieurs variables catégorielles pour chaque mois.
Voici une synthèse des variables ajoutées pour notre étude :

$$\text{cat\_age} = \begin{cases}]17\,;30] & \text{si age} \leq 30 \\ ]30\,;40] & \text{si } 30 < \text{age} \leq 40 \\ ]40\,;60] & \text{si } 40 < \text{age} \leq 60 \\ ]60\,;100] & \text{si } 60 < \text{age} \leq 100 \end{cases}$$

$$\text{c\_balance} = \begin{cases} \text{critic balance} & \text{si balance} < 0 \\ \text{low balance} & \text{si } 0 \leq \text{balance} < 1000 \\ \text{average balance} & \text{si } 1000 \leq \text{balance} < 5000 \\ \text{high balance} & \text{si } 5000 \leq \text{balance} \end{cases}$$

Pour chaque mois :

$$\text{mois} = \begin{cases} \text{yes} & \text{si month} = \text{mois} \\ \text{no} & \text{sinon} \end{cases}$$

3 Analyse univariée

Suite aux traitements des données, une analyse univariée a été effectuée afin de décrire la population de notre échantillon.
Suite à la suppression d'individus possédant des valeurs manquantes, la taille de l'échantillon est désormais de 43193 observations.
Par soucis de simplicité, nous ne prendrons pas en compte dans cette section des variables catégorielles associées aux mois. Elles seront utilisées et décrites dans l'analyse bivariée.
Voici sur la Figure 2 un bref résumé des caractéristiques des variables de l'échantillon :

```
      age                job          marital          education     default
 Min.   :18.00   blue-collar:9278   divorced: 5028   primary  : 6800   no :42411
 1st Qu.:33.00   management :9216   married :25946   secondary:23131   yes:  782
 Median :39.00   technician :7355   single  :12219   tertiary :13262
 Mean   :40.76   admin.     :5000
 3rd Qu.:48.00   services   :4004
 Max.   :95.00   retired    :2145
                 (Other)    :6195
    balance       housing      loan            contact           day             month
 Min.   : -8019   no :18901   no :36086   cellular :28213   Min.   : 1.00   may    :13192
 1st Qu.:    71   yes:24292   yes: 7107   telephone: 2694   1st Qu.: 8.00   jul    : 6601
 Median :   442                          unknown  :12286   Median :16.00   aug    : 6037
 Mean   :  1354                                            Mean   :15.81   jun    : 4980
 3rd Qu.:  1412                                            3rd Qu.:21.00   nov    : 3842
 Max.   :102127                                            Max.   :31.00   apr    : 2820
                                                                           (Other): 5721
    duration        campaign          pdays          previous          poutcome
 Min.   :   0.0   Min.   : 1.000   Min.   : -1.0   Min.   :  0.0000   failure: 4709
 1st Qu.: 103.0   1st Qu.: 1.000   1st Qu.: -1.0   1st Qu.:  0.0000   other  : 1774
 Median : 180.0   Median : 2.000   Median : -1.0   Median :  0.0000   success: 1424
 Mean   : 258.3   Mean   : 2.758   Mean   : 40.4   Mean   :  0.5849   unknown:35286
 3rd Qu.: 318.0   3rd Qu.: 3.000   3rd Qu.: -1.0   3rd Qu.:  0.0000
 Max.   :4918.0   Max.   :58.000   Max.   :871.0   Max.   :275.0000

  y            cat_age              c_balance
 no :38172   (17,30] : 6779   critic balance : 3634
 yes: 5021   (30,40] :17178   average balance:11222
             (40,60] :18167   high balance   : 2698
             (60,100]: 1069   low balance    :25639
```

FIGURE 2 – Résumé des variables du jeu de données après traitements

Nous allons donner par la suite des représentations graphiques pour étailler la description de certaines variables de l'échantillon.

3.1 Variables quantitatives

La grande majorité des personnes ciblées par les campagnes de marketing sont agées entre 30 et 50 ans. Ils n'ont généralement pas été contacté avant la campagne dont ils ont été ciblés (previous proche de 0 en moyenne).
On observe une possible grande disparité entre les valeurs de la variable balance, due à l'écart entre la valeur maximale et minimale.
Il en va de même pour les variables duration et campaign, où on voit que en général les particuliers sont contactés 2 à 3 fois durant une campagne. Il est donc pertinent de récupérer l'écart-type de chaque variable quantitative, une information qui manque au résumé des données sur la Figure 2.
On donne l'écart-type de chaque variable sur la Figure 3 :

```
      age     balance    duration    campaign    previous
10.618762 3044.765829  257.527812    3.098021    2.303441
```

FIGURE 3 – Ecart-type de chaque variable quantitative

On a bien un écart-type assez élevé pour la variable balance, il y a une grande disparité entre les soldes bancaires des clients solicités par la campagne. Il y en a une aussi pour la variable duration car à l'échelle de la seconde pour un temps d'appel, on peut considérer qu'une durée d'appel allongée ou réduite de 4 minutes est importante.

Il faudrait pour certains tests statistiques, vérifier si ces variables quantitatives, suivent une loi normale. Notamment pour des tests de comparaison (en particulier ici pour la variable duration). Dans ce cas, on s'intéressera aux distributions dans chaque groupe de la variable à expliquer y, ce qui sera fait lors de l'analyse bivariée.

A titre d'exemple, pour avoir une idée de la distribution, voyons l'histogramme de la variable duration.

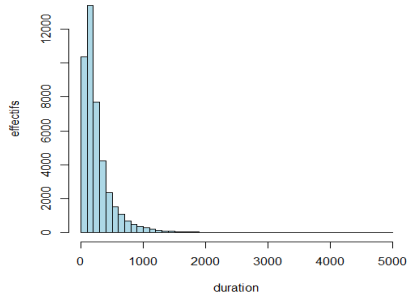


FIGURE 4 – Histogramme de la variable duration

D'après l'histogramme précédant sur la Figure 4, duration ne semble pas suivre une distribution normale. Ce qui sera à confirmer dans une analyse bivariée, où nous nous intéresserons à sa distribution pour chaque classe de y et aussi des variables catégorielles explicatives liées au marketing direct.

3.2 Variables qualitatives

On va donner ici sur la Figure 5 des graphiques présentant la répartition des modalités pour les variables job et month. Car dans le résumé global des variables, on ne voit pas l'effectif de toutes les modalités de ces variables.

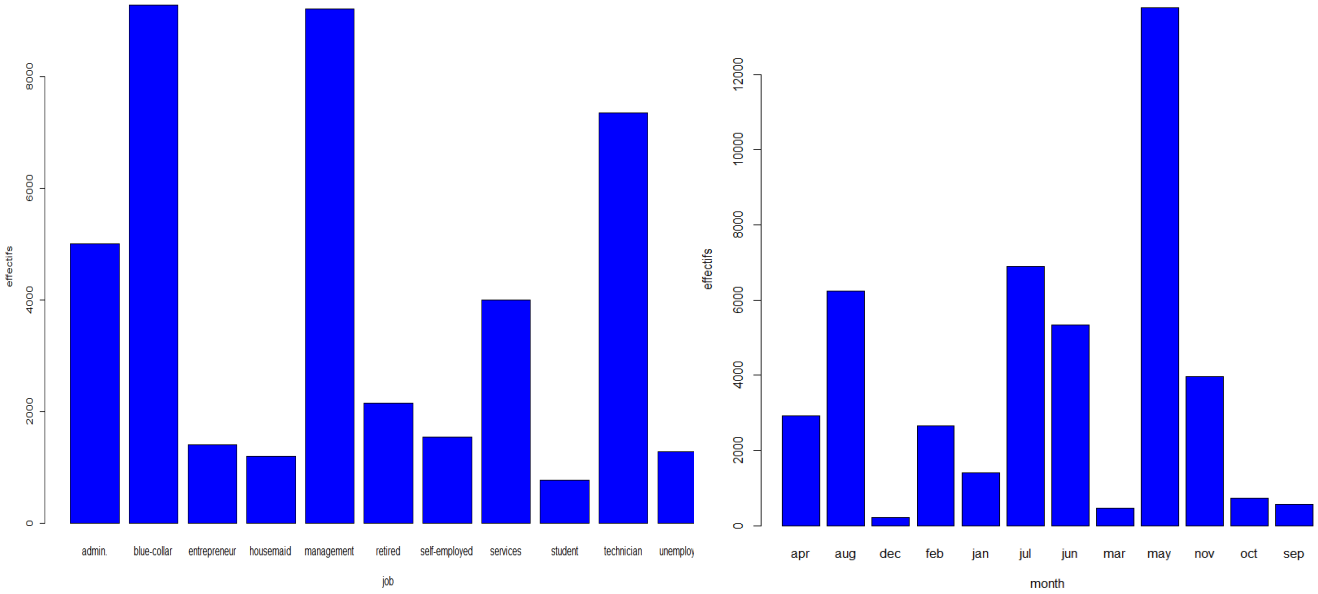


FIGURE 5 – Diagrammes en barres des variables job et month

Les particuliers ciblés par les campagnes sont majoritairement des travailleurs manuel, en management et technicien. La sollicitation des clients se concentre essentiellement sur le mois de mai.

Pour l'aboutissement de l'étude portant sur les facteurs influant la souscription à un compte à terme. Nous allons voir dans l'analyse bivariée, les liens de chaque variable explicative et y.

4 Analyse bivariée

Dans cette partie nous allons voir pour chaque variable explicative, au moyen de tests et modèles statistiques, l'influence qu'elle a sur y. Les variables n'influant pas y de manière significative, ne seront pas sélectionnées dans l'analyse multivariée.
L'étude va se diviser en 2 champs. Tout d'abord les profils des clients et ensuite la stratégie de marketing direct.

4.1 Profils des clients

Pour l'étude des profils des clients favorables à la souscription à un compte à terme, nous allons prendre les variables (qui sont toutes catégorielles) portant sur ce champs, en comptant celle qu'on a ajouté aux données :

cat_age, job, c_balance, marital, education, default, housing, loan, y

4.1.1 Test d'indépendance : Khi-deux de Pearson

Voyons au moyen d'un test de Khi-deux de Pearson, s'il existe un lien statistique entre y et chaque variable explicative. Ce test convient bien pour des grands échantillons (de taille supérieure à 30), ce qui est notre cas.
Sous l'hypothèse H0, les 2 variables sont indépendantes (à un risque de première espèce $\alpha = 5\%$). On rejetera cette hypothèse, lorsque la p-value sera inférieure à α.
Voici sur la Figure 6, les résultats de ce tests avec les effectifs des classes de chaque variable explicative pour chaque modalité de y :

Caractéristique	**no**, N = 38 172[1]	**yes**, N = 5 021[1]	**Total**, N = 43 193[1]	**p-value**[2]
job				<0,001
admin.	4 387 (11%)	613 (12%)	5 000 (12%)	
blue-collar	8 603 (23%)	675 (13%)	9 278 (21%)	
entrepreneur	1 295 (3,4%)	116 (2,3%)	1 411 (3,3%)	
housemaid	1 090 (2,9%)	105 (2,1%)	1 195 (2,8%)	
management	7 963 (21%)	1 253 (25%)	9 216 (21%)	
retired	1 659 (4,3%)	486 (9,7%)	2 145 (5,0%)	
self-employed	1 358 (3,6%)	182 (3,6%)	1 540 (3,6%)	
services	3 654 (9,6%)	350 (7,0%)	4 004 (9,3%)	
student	549 (1,4%)	226 (4,5%)	775 (1,8%)	
technician	6 538 (17%)	817 (16%)	7 355 (17%)	
unemployed	1 076 (2,8%)	198 (3,9%)	1 274 (2,9%)	

[1] Statistique présentée: n (%)
[2] Test statistique réalisé: test du khi-deux d'indépendance

Caractéristique	**no**, N = 38 172[1]	**yes**, N = 5 021[1]	**Total**, N = 43 193[1]	**p-value**[2]
cat_age				<0,001
(17,30]	5 688 (15%)	1 091 (22%)	6 779 (16%)	
(30,40]	15 421 (40%)	1 757 (35%)	17 178 (40%)	
(40,60]	16 452 (43%)	1 715 (34%)	18 167 (42%)	
(60,100]	611 (1,6%)	458 (9,1%)	1 069 (2,5%)	
education				<0,001
primary	6 212 (16%)	588 (12%)	6 800 (16%)	
secondary	20 690 (54%)	2 441 (49%)	23 131 (54%)	
tertiary	11 270 (30%)	1 992 (40%)	13 262 (31%)	

[1] Statistique présentée: n (%)
[2] Test statistique réalisé: test du khi-deux d'indépendance

Caractéristique	**no**, N = 38 172[1]	**yes**, N = 5 021[1]	**Total**, N = 43 193[1]	**p-value**[2]
c_balance				<0,001
critic balance	3 427 (9,0%)	207 (4,1%)	3 634 (8,4%)	
average balance	9 516 (25%)	1 706 (34%)	11 222 (26%)	
high balance	2 281 (6,0%)	417 (8,3%)	2 698 (6,2%)	
low balance	22 948 (60%)	2 691 (54%)	25 639 (59%)	
marital				<0,001
divorced	4 430 (12%)	598 (12%)	5 028 (12%)	
married	23 343 (61%)	2 603 (52%)	25 946 (60%)	
single	10 399 (27%)	1 820 (36%)	12 219 (28%)	
default	734 (1,9%)	48 (1,0%)	782 (1,8%)	<0,001
housing	22 418 (59%)	1 874 (37%)	24 292 (56%)	<0,001
loan	6 634 (17%)	473 (9,4%)	7 107 (16%)	<0,001

[1] Statistique présentée: n (%)
[2] Test statistique réalisé: test du khi-deux d'indépendance

FIGURE 6 – Tableaux d'effectifs et résultats des tests d'indépendance de chi deux de Pearson pour chaque variable explicative en fonction de y

Toutes les variables catégorielles ont pour leur test une p-value inférieure à α, donc on rejète dans tous les cas l'hypothèse d'inpédendance avec y.

Maintenant à partir de ces résultats, voyons quels profils de clients semblent être favorable à une souscription à un compte à terme.

Comme c_balance (solde bancaire) a un lien avec y. La souscription à un compte à terme, nécessite que le client possède des ressources financières suffisante voir importante. Et d'après les effectifs et proprotion de ses modalités en fonction de y. On voit que plus le solde bancaire d'un client est haut, plus il a de chance à faire la souscription. Et plus il est bas, moins le client répondra favorablement à une campagne. Ce qui est logique car le montant minimal pour un dépôt sur un compte à terme, est loin d'être négligeable (quelques milliers d'euros en général). Afin d'éviter une surcharge visuelle, on retrouvera le reste des représentations visuelles en annexe. Voyons quels profils sur la Figure 7 de clients possèdent un haut solde bancaire, en croisant la variable c_balance avec les autres variables explicatives :

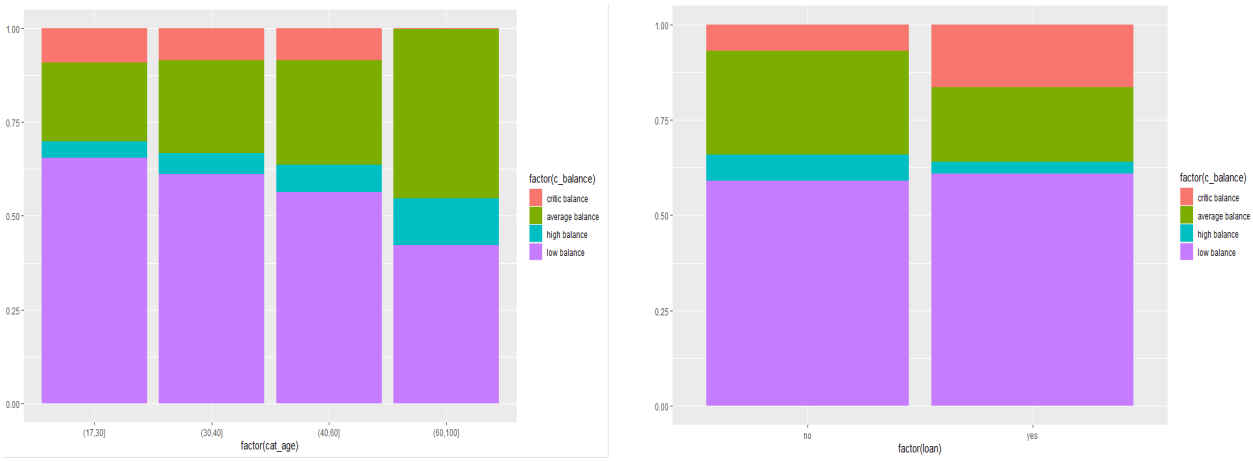


FIGURE 7 – Diagrammes en barres des variables explicatives (données clients) en fonction de c_balance

Synthèse :
✓ Les personnes agées (à partir de 60 ans), sont celles qui ont généralement un solde bancaire élevé. Elles tendent à souscrire à un dépot à terme, possiblement en vu d'un héritage familial.
✓ Ce qui est en lien avec le fait que parmi les catégories socio-professionnel, la catégorie ayant la proportion la plus élevée de compte à haut solde sont les retraités avec à suivre les clients occupant un poste dans le managment.
✓ Pour ce qui est de l'éducation, c'est logiquement avec des études plus longues que les particuliers acquièrent des connaissances qui leurs donnent accès à des postes mieux payés.
✓ Pour la situation marital, il n'y a pas de différences flagrantes entre un client marié et célibataire. On voit que le divorce impacte négativement le solde bancaire.
✓ Un client avec un défaut de crédit aura généralement un solde très bas. Il y aura donc très peu de chance qu'il accepte de soucrire à un compte à terme
✓ De même pour ceux qui possèdent un prêt immobilier ou quelconque, ils ont globalement un plus faible solde bancaire que ceux qui n'ont pas besoin de de prêt.

4.2 Stratégie de marketing direct

Pour trouver quels sont les facteurs influant sur l'efficacité de la stratégie de marketing direct mise en place par la banque, l'approche adoptée sera le calcul d'un risque relatif (Odds-Ratio ou OR) au moyen d'une regression logistique.
Pour cette partie, nous allons prendre les variables en liée à la stratégie de marketing direct de la banque :

contact, duration, campaign, previous, poutcome, chaque variable binaire portant sur les mois, y

Un point essentiel de la stratégie marketing d'une campagne est le choix de la période de déploiement de celle-ci. Regardons sur la Figure suivante, sur quels mois la campagne de la banque s'est le plus concentré selon les classes de y (pour avoir une idée de la réussite de la campagne selon les mois).

Caractéristique	**no**, N = 38 172[1]	**yes**, N = 5 021[1]	**Total**, N = 43 193[1]	**p-value**[2]
jan	1 186 (3,1%)	132 (2,6%)	1 318 (3,1%)	0,071
feb	2 115 (5,5%)	418 (8,3%)	2 533 (5,9%)	<0,001
mars	207 (0,5%)	241 (4,8%)	448 (1,0%)	<0,001
apr	2 277 (6,0%)	543 (11%)	2 820 (6,5%)	<0,001
may	12 304 (32%)	888 (18%)	13 192 (31%)	<0,001
jun	4 457 (12%)	523 (10%)	4 980 (12%)	0,009
jul	6 015 (16%)	586 (12%)	6 601 (15%)	<0,001
aug	5 378 (14%)	659 (13%)	6 037 (14%)	0,067
sep	281 (0,7%)	251 (5,0%)	532 (1,2%)	<0,001
oct	393 (1,0%)	297 (5,9%)	690 (1,6%)	<0,001
nov	3 452 (9,0%)	390 (7,8%)	3 842 (8,9%)	0,003
dec	107 (0,3%)	93 (1,9%)	200 (0,5%)	<0,001

[1] Statistique présentée: n (%)

FIGURE 8 – Effectifs et proportions des mois où la banque a contacté des clients, selon y

On voit que la campagne de souscription à un compte à terme s'est faite principalement au mois de mai. Ensuite au mois de juillet et août. En terme de saisonnalité, on se retrouve ici une campagne plus concentrée sur la saison d'été. A partir des information de la Figure 8, nous pouvons calculer des rendements concernant le nombre de client ayant accepté de souscrire à un compte à terme, par rapport à ceux qui ont refusé.

Voici les résultats de ces rendements, dans la Figure 9 : Le mois de mai étant celui choisi

mois	rendement
jan	0.11
feb	0.20
mars	1.16
apr	0.24
may	0.07
jun	0.12
jul	0.10
aug	0.12
sep	0.90
oct	0.76
nov	0.11
dec	0.87

FIGURE 9 – Rendements du nombre de client ayant accepté de souscrire à un compte à terme, par rapport à ceux qui ont refusé pour chaque mois

principalement par la stratégie de la campagne de la banque pour solliciter le plus ses client, est malgré cela le mois le mois rentable pour la banque. On voit que les clients ont tendance à être plus favorable à la souscription à un compte à terme au printemps et en automne plutôt qu'en été.

Nous savons que la variable duration est fortement liée à y. En effet, un client ayant une conversation très longue avec un conseiller durant la phase de campagne, montre de cette manière qu'il est très engagé. Il y a de fortes chances qu'il ait accepté de souscrire à un compte à terme et donc souhaite avoir plus de détail sur le taux par exemple.
Visuellement sur la Figure 10, on voit bien que les médianes des 2 groupes selon y de la variable duration, sont sensiblement différentes. La durée d'appel est bien globalement plus élevée pour les clients ayant accepté de soucrire à un compte à terme.

Examinons les distributions de la durée de contact d'un lors lors d'une campagne, pour chaque modalités des variables contacts, poutcome et de chaque mois. Cela au moyen de tests de comparaison.

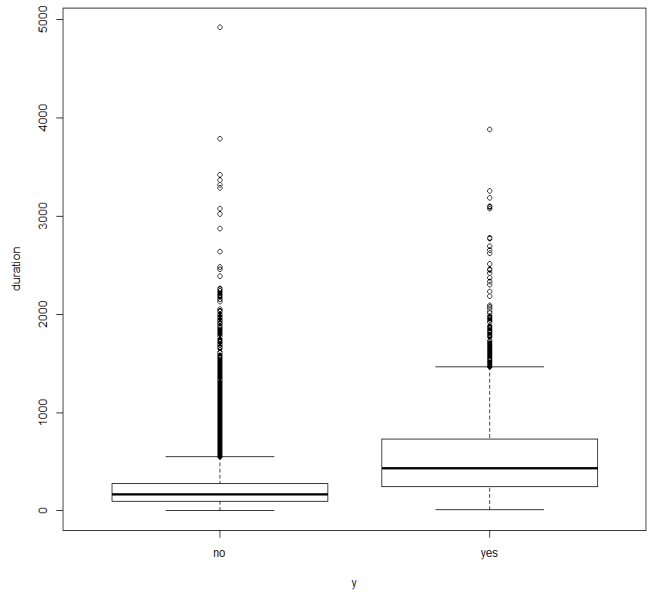


FIGURE 10 – Boxplot de la variable duration en fonction de y

4.2.1 Test de normalité : Kolmogorov Smirnov

Il faut d'abord voir si les distributions de duration dans chaque modalité des variables qualitatives, suivent une loi normale à l'aide du test de Kolmogorov Smirnov. Afin de savoir si nous allons effectuer des tests de comparaison paramétrique, ou non-paramétrique.
Sous l'hypothèse H0, les 2 distributions des 2 échantillons appariés d'une variable quantitative suivent une loi normale (toujours à un risque de première espèce $\alpha = 5\%$). On rejetera cette hypothèse lorsque la p-value sera inférieure à α.
La Figure 11 répertorie les résultats des tests de normalité.

Les p-values des tests sont inférieures à α, donc la durée d'appel d'un client ne suit jamais la loi normale, selon chaque modalité des variables catégorielles liées au marketing direct.

On va donc utiliser un test de comparaison non-paramétrique.

variables	p-value
contact	$<$ 2.2e-16
poutcome	$<$ 2.2e-16
jan	$<$ 2.2e-16
feb	$<$ 2.2e-16
mars	$<$ 2.2e-16
apr	$<$ 2.2e-16
may	$<$ 2.2e-16
jun	$<$ 2.2e-16
jul	$<$ 2.2e-16
aug	$<$ 2.2e-16
sep	$<$ 2.2e-16
oct	$<$ 2.2e-16
nov	$<$ 2.2e-16
dec	$<$ 2.2e-16

FIGURE 11 – Résultats des tests de normalité de Kolmogorov Smirnov de la variable duration, selon chaque modalité des variables catégorielles liées au marketing direct

4.2.2 Test de comparaison : Kruskal-Wallis

Les tests non-paramétriques permettent de comparer la distribution de deux échantillons indépendants. Ils sont utiles quand l'hypothèse de normalité des échantillons n'est pas vérifiée pour l'utilisation des tests paramétriques (comme le test de Student).
Le test de Kruskal-Wallis permet de tester l'hypothèse H0, sous laquelle plusieurs échantillons (nombre supérieur ou égal à 2) ont la même distribution.
La Figure 11 répertorie les résultats des tests de Kruskal-Wallis.

Les distributions de duration ne sont pas significativement différentes, pour les mois de mars et juillet. On va donc admettre que ces mois ne sont pas déterminant, pour améliorer la performance de marketing direct de la banque. Nous n'utiliserons donc pas ces variables pour la suite. D'autant plus que les rendements associés à la souscription à un compte à terme d'un client ne sont pas élevés.

Maintenant, il va falloir déterminer quels sont les facteurs les plus importants à la souscription à un compte à terme. Et voir dans quelles dispositions ou mesures (augmentation, dimunition, choix d'une modalité...), ces facteurs pourraient augmenter les chances qu'un particulier souscrit à un compte à terme.

variables	p-value
contact	$<$ 2.2e-16
poutcome	$<$ 2.2e-16
jan	0.000982
feb	0.001643
mars	0.135
apr	$<$ 2.2e-16
may	1.27e-09
jun	1.491e-13
jul	0.3149
aug	$<$ 2.2e-16
sep	1.334e-06
oct	0.0001679
nov	1.733e-05
dec	6.513e-08

FIGURE 12 – Résultats des tests de comparaison de Kruskal-Wallis de la variable duration, selon chaque modalité des variables catégorielles liées au marketing direct

4.2.3 Régression logistique simple

Une régression logistique simple a été réalisée pour y, en fonction de chaque variable explicative (liées au marketing). Afin de mettre en évidence le lien de chaque variable explicative et la variable d'intérêt, à un seuil de significativité fixé à $\alpha = 0,05$. Si la p-value est inférieure à α (test de Wald), alors l'OR est significativement différent de 1.
La référence prise pour Y est le refus de souscription du client à un compte à terme (modalité de référence : « no »). Afin de modéliser la probabilité de l'évennement complémentaire (le client souscrit à un compte à terme). Cette modélisation permet de déterminer pour chacune des variables explicatives le risque relatif ou Odds ratio (OR) d'une modalité considérée (par rapport à une autre de référence), qu'un client accepte la souscription.

L'Odds ratio ou rapport des cotes se calcule de la façon suivante :

$$\text{OR} = \frac{\text{P1}}{\text{P2}}$$

Avec P1 : la probabilité qu'un client accepte la souscription si on a un facteur de risque considéré
P2 : probabilité qu'un client accepte la souscription si on a pas le facteur de risque considéré

Interprétations des OR :

- L'OR de la modalité de référence de la variable explicative est toujours égal à 1.
- Lorsque l'OR des autres modalités de la même variable est > 1, cela signifie que ces modalités augmentent le risque (ou ici les chances) qu'un client accepte de souscrire à un compte à terme en comparaison à la modalité de référence de la variable explicative.
- A l'inverse, lorsque l'OR des autres modalités de la même variable explicative est <1, cela signifie que ces modalités diminuent le risque de survenu d'une souscription, par rapport à la modalité prise comme référence de la variable explicative.

Voici dans la Figure 13 les résultats des tests de Wald avec les OR des régressions logistiques de y en fonction de chaque variable explicative liée au marketing :

variables	p-value	OR
contact (référence = "cellular")	<2e-16	
"telephone"	0.0133	0.86
"unknown"	< 2.2e-16	0.249
duration	<2e-16	1.004
campaign	<2e-16	0.879
previous	<2e-16	1.118
poutcome (référence = "failure")	< 2.2e-16	
"other"	2.31e-05	1.389
"sucess"	< 2e-16	12.70
"unknown"	3.45e-13	0.70
jan (référence = "no")	0.064	
"yes"	0.0644	0.842
feb (référence = "no")	< 2e-16	
"yes"	4.61e-15	1.548
apr (référence = "no")	< 2.2e-16	
"yes"	<2e-16	1.912
may (référence = "no")	<2e-16	
"yes"	<2e-16	0.452
jun (référence = "no")	0.009	
"yes"	0.009	0.880
aug (référence = "no")	0.064	
"yes"	0.064	0.921
sep (référence = "no")	<2e-16	
"yes"	<2e-16	7.096
oct (référence = "no")	<2e-16	
"yes"	<2e-16	6.044
nov (référence = "no")	0.003	
"yes"	0.003	0.847
dec (référence = "no")	<2e-16	
"yes"	<2e-16	6.714

FIGURE 13 – Résultats des régressions logistiques de y avec les variables explicatives liées au marketing direct

Synthèse :
✓ Pour ce qui est du type de l'appareil utilisé pour contacter les clients. La campagne à plus de chance d'avoir du succès si le client est contacté par cellulaire.
✓ Quand la durée d'appel augmente, il devient bien sûr plus probable qu'un client accepte de souscrire à un compte à terme.
✓ Ce qui n'est pas le cas, lorsque le nombre de contacts du client augmente lors de la même campagne.
✓ Mais plus il a été contacté lors de précédentes campagnes, plus la probabilité qu'il accepte la souscription augmente.
✓ De même si le client a répondu favorablement, dans de précédentes campagnes.
✓ Et enfin pour les mois, ceux qui favorisent le succès d'une campagne sont février, avril, septembre, octobre et décembre. Par exemple pour le mois de septembre, il y a environ 7 fois plus de chance qu'un client souscrivent à un compte à terme, comparé à un client non sollicité ce mois-ci.

Dans la suite, nous verrons comment se comportent ces variables dans le modèle multivarié. En supposant qu'il n'y a aucun lien entre ces variables.

5 Analyse multivariée

5.1 Analyse des correspondances multiples

Pour résumer l'information donnée par les variables liées au profil du client et déceler des corrélations existantes, nous allons établir une analyse des correspondances multiples (ACM). Cette méthode est adaptée pour décrire un ensemble d'individus, suivant des attributs de type qualitatif. Par exemple dans le cadre d'une enquête, on interrogera les individus avec des questions à choix multiples. Ce qui définira pour chacun d'entre eux un profil de réponses.
On cherche à savoir quelles sont les modalités de variables corrélées entre elles, et le profil des individus caractérisant la modalité d'une variable d'intérêt. Cela est représenté graphiquement dans un plan factoriel.
Dans ce plan :
-Les individus qui se ressemblent (beaucoup de réponses communes) sont proches. -Ceux qui sont différents sont éloignés. -Il en va de même pour les modalités des variables. Elles sont proches, si il existe un lien entre elles qui entrainne le fait qu'elles ont tendance à être prises ensemble.

Dans cette étude, nous nous intéresserons aux modalités des variables liées aux données personnelles du client. En prennant y comme variable supplémentaire. y ne participera pas à la construction des axes factoriels, mais servira à les caractériser.

En prenant le critère de coude appliqué aux 10 premières valeurs propres, on décide de choisir d'étudier les 3 premiers axes factoriels. Il y a une diminution significative de l'inertie, juste après la deuxième valeur propre. Puis suivant les dimensions, l'inertie diminue peu et de manière régulière :

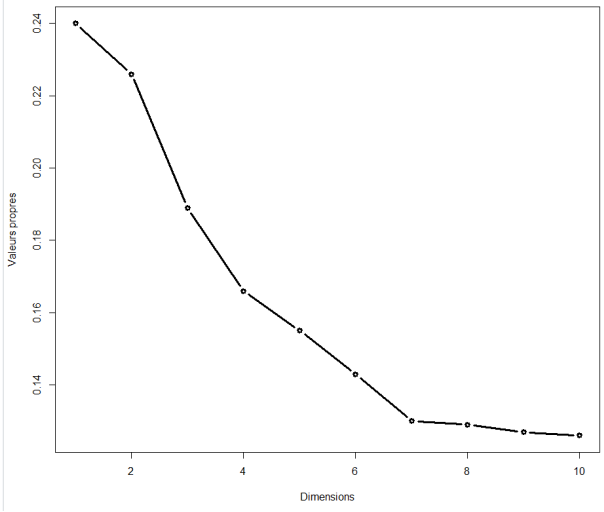


FIGURE 14 – Représentation des valeurs propres pour les 10 premières dimensions de l'ACM

Les 3 premier axes représentent 22.75% de l'inertie totale du nuage. C'est à dire que la représentation des variables dans le plan factoriel ne reflète que 22.75% de la réalité. Les

résultats de cette analyse fournira néanmoins les profils de clients acceptant ou refusant la souscription.
Les coordonnées de y sont nulles au niveau du premier axe, donc on n'en prendra pas compte par la suite.

Dans la Figure 15, on trouve les résultats de l'ACM pour les axes 2 et 3. On ne prendra compte que des modalités actives bien représentées, celles qui ont une contribution supérieure à 3.22% (100/31) et un $\cos^2$ (qui renseigne sur la qualité de représentation) supérieur à 5%. Le seuil associé au $\cos^2$, est choisi arbitrairement au vu du grand nombre de modalités actives. La valeur de la contribution moyenne par modalité est faible car il y a 31 modalités actives. Si la même modalité se retrouve bien représentée sur les 2 axes, on prendra que en compte l'axe où cette modalité a la contribution la plus forte. Le reste des détails concernant les détails de l'ACM, sont données en annexe.

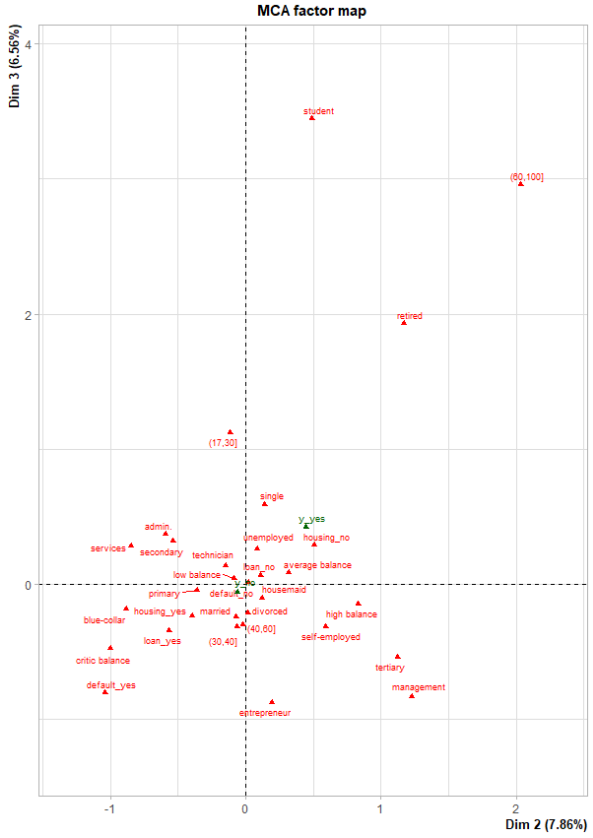


FIGURE 15 – Résultats de l'ACM suivant les axes factoriels 2 et 3

Maintenant, voyons quelles sont les modalités actives bien représentées suivant les axes 2 et 3 :

côté	axe 2	axe 3
positif	management (17.807, 40.9) retired (3.767, 7.2) tertiary (21.438, 55.9) housing_no (6.246, 20.1) (60,100] (5.662, 10.5)	retired (12.272, 19.5) student (14.124, 0.217) single (6.609, 13.9) secondary (3.646, 11.9) (17,30] (13.088, 0.234) (60,100] (14.334, 22.2)
négatif	blue-collar (9.296, 21.4) services (3.675, 7.3) secondary (8.581, 33.4) housing_yes (4.860, 20.1) critic balance (4.662, 9.2)	management (9.776, 18.8) tertiary (5.856, 12.8)

FIGURE 16 – Modalités les mieux représentées de l'ACM

Comme la modalité "yes" de y se trouve du côté positif des axes 2 et 3 (et inversement pour la modalité "no"). Les caractéristiques des profils des clients susceptibles de souscrire à un compte à terme, sont les modalités associées au côté positif dans la Figure 16 ci-dessus. Et inversement pour les caractéristiques des profils de clients n'étant pas susceptibles d'accepter la souscription.

Les résultats obtenus sont cohérents avec la synthèse faite lors de l'analyse bivariée, notamment en inférant selon le niveau du solde bancaire. On voit de plus que des étudiants, dans la tranche d'age de 17 à 30 ans pouraient aussi souscrire à un compte à terme. Dans le cas où ils auraient par exemple touché un héritage familial.

5.2 Regression logistique multivariée

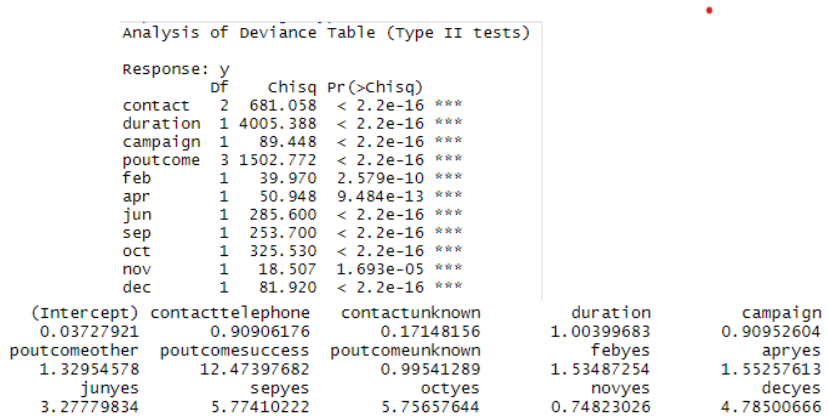
L'établissement d'un modèle de regression logistique multivarié, nous permettra d'exprimer y en fonction des variables explicatives liées au marketing. Toujours en voyant quelles sont les variables qui ont un impact signifactif sur y. Et selon les coefficients des modalités, comment influent-elles sur la probabilité de survenu de la souscription d'un client à un compte à terme.

Toujours avec le test de Wald, on effectue une sélection pas à pas descendente des variables. En éliminant une à une de façon décroissante, les variables ayant une p-value globale supérieure à 5% à ce test.
Voici sur la Figure 17, le modèle final obtenu issu de la sélection pas à pas descendante :

```
Call:
glm(formula = y ~ contact + duration + campaign + poutcome +
    feb + apr + jun + sep + oct + nov + dec, family = binomial(logit),
    data = bank)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.6290  -0.3928  -0.2886  -0.1662   3.3164

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.289e+00  6.255e-02 -52.591  < 2e-16 ***
contacttelephone -9.534e-02  7.309e-02  -1.304  0.19207
contactunknown   -1.763e+00  6.760e-02 -26.085  < 2e-16 ***
duration          3.989e-03  6.303e-05  63.288  < 2e-16 ***
campaign         -9.483e-02  1.003e-02  -9.458  < 2e-16 ***
poutcomeother     2.848e-01  8.819e-02   3.230  0.00124 **
poutcomesuccess   2.524e+00  7.898e-02  31.954  < 2e-16 ***
poutcomeunknown  -4.598e-03  5.555e-02  -0.083  0.93404
febyes            4.284e-01  6.777e-02   6.322 2.58e-10 ***
apryes            4.399e-01  6.163e-02   7.138 9.48e-13 ***
junyes            1.187e+00  7.025e-02  16.900  < 2e-16 ***
sepyes            1.753e+00  1.101e-01  15.928  < 2e-16 ***
octyes            1.750e+00  9.701e-02  18.042  < 2e-16 ***
novyes           -2.900e-01  6.742e-02  -4.302 1.69e-05 ***
decyes            1.565e+00  1.730e-01   9.051  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31045  on 43192  degrees of freedom
Residual deviance: 21702  on 43178  degrees of freedom
AIC: 21732

Number of Fisher Scoring iterations: 6
```

FIGURE 17 – Résultats de la regression logistique multivariée

On ajoute à cela les p-value globale de chaque variable et les valeurs des OR de chaque modalité :



```
Analysis of Deviance Table (Type II tests)

Response: y
           Df    Chisq Pr(>Chisq)
contact     2  681.058  < 2.2e-16 ***
duration    1 4005.388  < 2.2e-16 ***
campaign    1   89.448  < 2.2e-16 ***
poutcome    3 1502.772  < 2.2e-16 ***
feb         1   39.970  2.579e-10 ***
apr         1   50.948  9.484e-13 ***
jun         1  285.600  < 2.2e-16 ***
sep         1  253.700  < 2.2e-16 ***
oct         1  325.530  < 2.2e-16 ***
nov         1   18.507  1.693e-05 ***
dec         1   81.920  < 2.2e-16 ***
```

(Intercept)	contacttelephone	contactunknown	duration	campaign
0.03727921	0.90906176	0.17148156	1.00399683	0.90952604
poutcomeother	**poutcomesuccess**	**poutcomeunknown**	**febyes**	**apryes**
1.32954578	12.47397682	0.99541289	1.53487254	1.55257613
junyes	**sepyes**	**octyes**	**novyes**	**decyes**
3.27779834	5.77410222	5.75657644	0.74823026	4.78500666

FIGURE 18 – P-value globale de chaque variable et les valeurs des OR de chaque modalité

Au niveau des OR, on obtient des résultats qui amènent des raisonnements similaires à la synthèse de la regression logistique bivariée. A l'exception de la variable jun, être au mois de juin augmente la probabilité qu'un client souscrive à un compte à terme d'après le modèle multivarié.

6 Conclusion et recommandations

L'objectif de l'analyse de données issuent de campagnes de marketing direct permet en plus d'améliorer celles du futur, de renforcer les liens entre une entreprise et sa clientèle.
D'où l'intérêt de dissocier les variables à disposition, liées aux données personnelles du client et celles liées aux paramètres de la stratégie de marketing direct.

Dans notre étude, les profils des clients susceptibles de souscrire à un compte à terme sont ceux qui disposent d'un capital propre suffisant et avec une situation professionnel stable. Ce qui arrive souvent s'ils ont fait de longue étude et au bout d'un certain age (la soixantaine donc retraité). Dans le cas de personne plus jeune, la souscription peut se faire s'ils ont hérité d'une somme importante et s'il sont célibataire.
Les personnes disposant de prêts ou d'un solde bancaire faible voir critique, ne sont pas celle qui accepteront de bloquer temporairement une partie non négligeable de leur capital.

Concernant la stratégie du marketing direct, la campagne pour laquelle nous disposons les données s'est concentré sur une communication majoritairement sur le mois de mai. Mais ce mois a été le moins lucratif comparé aux autres. En terme de saisonnalité, les clients semblent répondre favorablement à la campagne au printemps et en automne. La banque devrait donc miser plus sur ces saisons en terme d'appels aux clients.
Pour le moyen de communication, on préconisera en priorité le téléphone portable (cellulaire) surtout pour les clients plus jeune (en dessous de la cinquantaine par exemple). Par contre, des sollicitations de clients répétées (pour ne pas dire abusives) réduisent les chancent qu'ils acceptent la souscription.
Et évidemment, les personnes ayant répondu favorablement à de précédentes campagnes pourront accepter de nouveau de souscrire à un compte à terme.

Références

[1] S. Moro P. Cortez P. Rita., A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, fichier csv, Juin 2014, http ://archive.ics.uci.edu/ml/datasets/Bank+Marketing (consulté en novembre 2020)

Annexes

variable	description
age (numérique)	age du client
job (modalités : "admin.","unknown", "unemployed","management" "housemaid","entrepreneur","student", "blue-collar","self-employed", "retired","technician","services")	métier du client
marital (modalités : "married","divorced","single")	statut marital du client
education (modalités : "unknown","secondary" ,"primary","tertiary")	niveau d'étude du client
default (modalités : "yes","no")	statut du défaut de crédit du client
balance (numérique)	solde bancaire du client
housing (modalités : "yes","no")	possession de prêt(s) immobilier du client
loan (modalités : "yes","no")	possession de prêt(s) du client
contact (modalités : "unknown","telephone" ,"cellular")	appareil de communication du client
day (numérique)	numéro du jour du mois où le client à été contacté la dernière fois
month (modalités : "jan", "feb", ..., "nov", "dec")	dernier mois où le client à été contacté
campaign (numérique)	nombre de fois où le client à été contacté durant cette campagne
pdays (numérique)	nombre de jours passés après le dernier contact du client pour la précédente campagne
previous (numérique)	nombre de contacts effectués pour le client avant cette campagne
poutcome (modalités : "unknown","other" ,"failure","success")	résultat de la précédente campagne avec le client
y (modalités : "yes","no")	souscription du client à un compte à terme

FIGURE 19 – Description des variables du jeu de données

Diagrammes en barres des variables explicatives restantes (données clients) en fonction de c_balance :

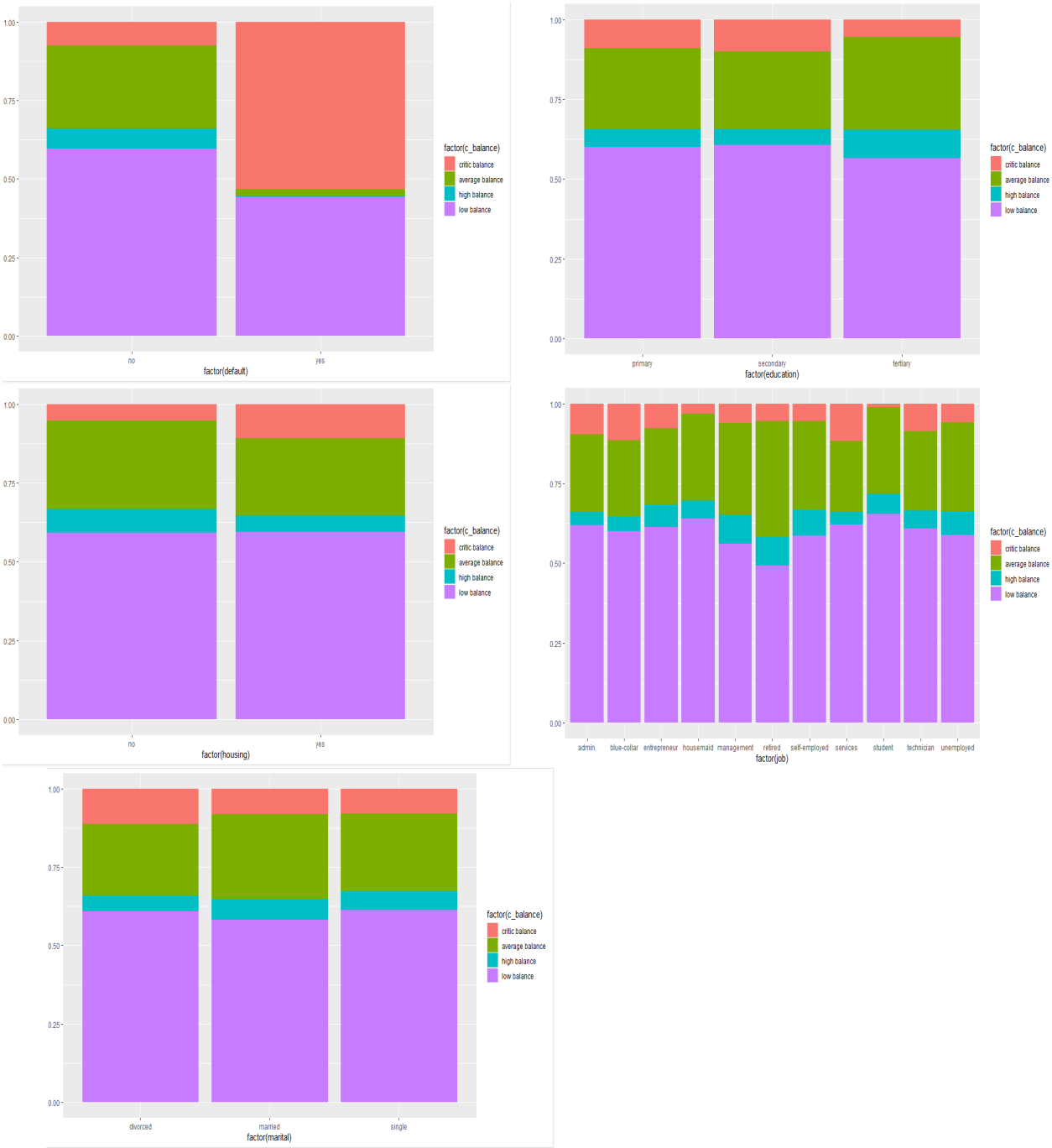


FIGURE 20 – Diagrammes en barres des variables explicatives (données clients) en fonction de c_balance

Sorties R pour l'ACM :

```
Call:
MCA(X = bank[, c(2, 3, 4, 5, 7, 8, 17, 18, 19)], ncp = 3, quali.sup = 7)

Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13  Dim.14  Dim.15  Dim.16  Dim.17  Dim.18  Dim.19  Dim.20  Dim.21  Dim.22  Dim.23
Variance               0.240   0.226   0.189   0.166   0.155   0.143   0.130   0.129   0.127   0.126   0.125   0.124   0.123   0.120   0.118   0.108   0.100   0.096   0.091   0.077   0.068   0.054   0.040
% of var.              8.332   7.859   6.564   5.782   5.398   4.982   4.511   4.478   4.429   4.369   4.350   4.312   4.284   4.186   4.109   3.769   3.462   3.326   3.175   2.686   2.369   1.889   1.380
Cumulative % of var.   8.332  16.191  22.754  28.536  33.934  38.916  43.427  47.905  52.333  56.702  61.053  65.365  69.649  73.835  77.944  81.713  85.175  88.501  91.676  94.363  96.732  98.620 100.000

Individuals
   Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
1  | 0.018  0.000  0.000 |  0.610  0.004  0.252 | -0.565  0.004  0.216 |
2  | -0.296 0.001  0.062 | -0.240  0.001  0.041 |  0.188  0.000  0.025 |
3  | -0.033 0.000  0.000 | -0.396  0.002  0.032 | -0.465  0.003  0.044 |
6  | -0.278 0.001  0.063 |  0.492  0.002  0.198 | -0.582  0.004  0.277 |
7  | -0.821 0.007  0.264 |  0.355  0.001  0.049 | -0.047  0.000  0.001 |
8  |  0.073 0.000  0.000 | -0.025  0.000  0.000 | -0.817  0.008  0.055 |
9  |  1.142 0.013  0.370 |  0.098  0.000  0.003 |  0.360  0.002  0.037 |
10 | -0.296 0.001  0.062 | -0.240  0.001  0.041 |  0.188  0.000  0.025 |
11 |  0.075 0.000  0.002 | -0.390  0.002  0.063 |  0.023  0.000  0.000 |
12 | -0.704 0.005  0.219 | -0.383  0.002  0.065 |  0.663  0.005  0.195 |
13 |  0.057 0.000  0.003 | -0.295  0.001  0.074 | -0.051  0.000  0.002 |
15 |  0.095 0.000  0.005 | -0.479  0.002  0.128 | -0.009  0.000  0.000 |

Categories
                    Dim.1     ctr    cos2  v.test     Dim.2     ctr    cos2  v.test     Dim.3     ctr    cos2  v.test
admin.          |  -0.341   0.702   0.015 -25.640 |  -0.593   2.253   0.046 -44.602 |   0.371   1.057   0.018  27.919 |
blue-collar     |   0.471   2.486   0.061  51.188 |  -0.884   9.296   0.214 -96.139 |  -0.182   0.470   0.009 -19.765 |
entrepreneur    |   0.147   0.037   0.001   5.624 |   0.194   0.068   0.001   7.408 |  -0.873   1.651   0.026 -33.356 |
housemaid       |   1.294   2.419   0.048  45.378 |   0.122   0.023   0.000   4.262 |  -0.098   0.018   0.000  -3.433 |
management      |  -0.445   2.205   0.054 -48.169 |   1.228  17.807   0.409 132.939 |  -0.832   9.776   0.188 -90.019 |
retired         |   2.490  16.071   0.324 118.310 |   1.171   3.767   0.072  55.626 |   1.931  12.272   0.195  91.761 |
self-employed   |  -0.194   0.070   0.001  -7.759 |   0.591   0.689   0.013  23.612 |  -0.311   0.228   0.004 -12.428 |
services        |  -0.242   0.284   0.006 -16.085 |  -0.846   3.675   0.073 -56.230 |   0.287   0.506   0.008  19.068 |
student         |  -2.111   4.173   0.081 -59.307 |   0.488   0.237   0.004  13.719 |   3.447  14.124   0.217  96.837 |
technician      |  -0.392   1.369   0.032 -36.954 |  -0.148   0.205   0.004 -13.903 |   0.139   0.218   0.004  13.090 |
unemployed      |   0.104   0.017   0.000   3.764 |   0.085   0.012   0.000   3.093 |   0.265   0.137   0.002   9.587 |
divorced        |   0.408   1.011   0.022  30.778 |   0.017   0.002   0.000   1.296 |  -0.212   0.348   0.006 -16.028 |
married         |   0.388   4.715   0.226  98.857 |  -0.069   0.159   0.007 -17.629 |  -0.239   2.264   0.086 -60.796 |
single          |  -0.991  14.509   0.388 -129.410 |  0.140   0.306   0.008  18.246 |   0.594   6.609   0.139  77.520 |
primary         |   1.262  13.084   0.298 113.372 |  -0.360   1.131   0.024 -32.371 |  -0.044   0.020   0.000  -3.956 |
secondary       |  -0.104   0.304   0.013 -23.274 |  -0.538   8.581   0.334 -120.095 |  0.321   3.646   0.119  71.543 |
tertiary        |  -0.465   3.467   0.096 -64.352 |   1.123  21.438   0.559 155.410 |  -0.537   5.856   0.128 -74.231 |
default_no      |   0.002   0.000   0.000   3.733 |   0.019   0.020   0.020  29.376 |   0.015   0.014   0.012  22.634 |
default_yes     |  -0.132   0.017   0.000  -3.733 |  -1.041   1.085   0.020 -29.376 |  -0.802   0.771   0.012 -22.634 |
housing_no      |   0.176   0.709   0.024  32.302 |   0.508   6.246   0.201  93.115 |   0.295   2.525   0.068  54.105 |
housing_yes     |  -0.137   0.552   0.024 -32.302 |  -0.395   4.860   0.201 -93.115 |  -0.230   1.965   0.068 -54.105 |
loan_no         |  -0.001   0.000   0.000  -0.547 |   0.112   0.577   0.063  52.322 |   0.067   0.249   0.023  31.408 |
loan_yes        |   0.006   0.000   0.000   0.547 |  -0.567   2.930   0.063 -52.322 |  -0.341   1.264   0.023 -31.408 |
(17,30]         |  -1.095   9.828   0.223 -98.230 |  -0.114   0.113   0.002 -10.246 |   1.122  13.088   0.234 100.609 |
(30,40]         |  -0.343   2.448   0.078 -58.007 |  -0.062   0.084   0.003 -10.460 |  -0.314   2.594   0.065 -52.997 |
(40,60]         |   0.554   6.741   0.223  98.133 |  -0.018   0.008   0.000  -3.268 |  -0.296   2.440   0.064 -52.400 |
(60,100]        |   3.048  11.998   0.236 100.909 |   2.034   5.662   0.105  67.326 |   2.957  14.334   0.222  97.896 |
critic balance  |  -0.062   0.017   0.000  -3.892 |  -1.001   4.662   0.092 -63.035 |  -0.475   1.259   0.021 -29.941 |
average balance |   0.176   0.421   0.011  21.700 |   0.319   1.466   0.036  39.320 |   0.087   0.129   0.003  10.656 |
high balance    |   0.183   0.110   0.002   9.833 |   0.833   2.397   0.046  44.677 |  -0.146   0.088   0.001  -7.806 |
low balance     |  -0.088   0.238   0.011 -22.020 |  -0.086   0.241   0.011 -21.494 |   0.045   0.079   0.003  11.254 |

Categorical variables (eta2)
              Dim.1 Dim.2 Dim.3
job         | 0.572 0.687 0.611 |
marital     | 0.388 0.008 0.139 |
education   | 0.323 0.563 0.144 |
default     | 0.000 0.020 0.012 |
housing     | 0.024 0.201 0.068 |
loan        | 0.000 0.063 0.023 |
cat_age     | 0.594 0.106 0.490 |
c_balance   | 0.015 0.158 0.023 |

Supplementary categories
                Dim.1    cos2  v.test     Dim.2    cos2  v.test     Dim.3    cos2  v.test
y_no        |   0.000   0.000  -0.033 |  -0.059   0.026 -33.733 |  -0.056   0.024 -31.877 |
y_yes       |   0.000   0.000   0.033 |   0.448   0.026  33.733 |   0.423   0.024  31.877 |

Supplementary categorical variables (eta2)
              Dim.1 Dim.2 Dim.3
y           | 0.000 0.026 0.024 |
```

FIGURE 21 – Sorties R associées à l'ACM

Table des figures