AUGMENTER LES BÉNÉFICES D'UNE CAMPAGNE DE MARKETING

Data Analyst: Project 8 -
Communiquez vos résultats

OpenClassrooms

Facundo ALCALA

Sommaire

1 Contexte

Dans le cadre du projet 8 (Communiquez vos résultats) de la formation de Data Analyst d'OpenClassrooms, il y avait trois options à choisir, détaillées ci-dessous:

- Option A : réaliser un stage (si vous remplissez les conditions).
- Option B : réaliser un projet libre à caractère social.
- Option C : réaliser un projet libre dans le domaine professionnel qui m'intéresse.

J'ai choisi l'option C, réaliser un projet libre dans le domaine professionnel qui m'intéresse.
Comme je suis diplômé en gestion avec une spécialisation en marketing, je voulais continuer l'apprentissage dans ce domaine. Comprendre les problématiques du secteur est essentiel pour faire le travail de data analyst. Dans ce projet, je ferai l'interface entre des données spécifiques et un représentant du métier.

L'objectif sera d'analyser les données pour mieux comprendre nos clients et améliorer le ciblage marketing dans les campagnes publicitaires. Cela nous aidera dans le futur à prendre de meilleures décisions marketing, qui se traduisent par des diminutions de coûts de campagnes marketing, l'optimisation des canaux de vente et un chiffre d'affaires élevé.

2 Présentation du problème

Lorsqu'on parle de campagne, on parle de communication et de marketing avec des outils qui permettent de diffuser un message ou promouvoir un produit par exemple.
Il est important de constater que beaucoup d'entreprises ne pensent pas à évaluer l'impact de ces campagnes et pourtant c'est un passage obligatoire pour voir si les objectifs fixés en amont ont été atteints.
Avec les nouvelles technologies, nous avons la chance de pouvoir étudier ces campagnes, autant par une étude quantitative que qualitative. La question à se poser est la suivante : quelle est la campagne la plus pertinente pour atteindre nos objectifs ? Répondre à cette question va nous permettre d'orienter nos efforts marketing et d'améliorer le ciblage de nos campagnes.
Cela pose une deuxième question : qu'est-ce que le ciblage dans le marketing ?
Le ciblage en marketing est une stratégie qui divise un grand marché en segments plus petits pour se concentrer sur un groupe spécifique de clients au sein de cet auditoire. Il définit un segment de clients en fonction de leurs caractéristiques uniques et se concentre uniquement sur leur service.
Au lieu d'essayer d'atteindre l'ensemble d'un marché, une marque utilise le marketing ciblé pour mettre son énergie à se connecter sur un groupe spécifique et défini dans ce marché.
Les types de marchés cibles sont souvent segmentés par des caractéristiques telles que :

- Données démographiques : âge, sexe, scolarité, état civil, race, religion, etc.
- Valeurs : croyances, intérêts, personnalité, style de vie, etc.
- Comportements d'achats: comment, quoi, quand, dans quelle fréquence, etc
- Zones géographiques : quartier, ville, région, pays, etc.

Grâce à cette stratégie de segmentation du marché, les marques deviennent plus spécifiques sur leur marché cible. Elles peuvent se concentrer sur un petit groupe de clients qui seront les plus susceptibles de bénéficier de leurs produits et d'en profiter.
Nous pouvons donc nous concentrer sur un marché cible plus restreint et spécifique. Plutôt que de faire du marketing auprès des masses, les campagnes peuvent se concentrer uniquement sur la vente de produits spécifiques aux clients ciblés.

Afin d' améliorer notre ciblage, nous pouvons utiliser le marketing prédictif.
Le marketing prédictif est l'utilisation stratégique d'ensemble de données sur les clients existants pour identifier des modèles et anticiper les comportements futurs des clients, les tendances des ventes et les résultats du marketing. Son intégration dans une stratégie moderne de ciblage marketing est ainsi parfaite. En effet, les organisations tirant parti du marketing prédictif ont plus de chances d'attirer le public qu'elles souhaitent.
La collecte de données sur les clients pour améliorer les futurs efforts marketing n'est pas une pratique nouvelle. Cependant, les stratégies modernes utilisent des algorithmes de machine learning et d'intelligence artificielle. Ainsi, il est possible de traiter des volumes d'informations jusqu'alors inimaginables. Aujourd'hui, les spécialistes du marketing peuvent tirer parti de bases de données agrégées de profils de clients pour élaborer des modèles prédictifs, enrichir les pistes de recherche et analyser les performances des différents programmes.

Dans ce projet il y aura:
- l'analyse de la performance de nos canaux de vente, de nos produits et de nos campagnes;
- la connaissance de nos clients et leur regroupement pour améliorer notre ciblage;
- la prédiction de la réponse positive ou négative à notre dernière campagne de marketing ainsi que la sélection des variables significatives et l'impact qu'elles ont sur nos modèles.

3 Préparation et exploration des données

3.1 Présentation des données

Dans ce rapport, nous analysons les données d'une entreprise fictive pour sélectionner des clients accessibles et augmenter le bénéfice d'une campagne de marketing.

Le dataset vient du site Kaggle. Vous pouvez le retrouver dans le lien qui se trouve ci-dessous:

https://www.kaggle.com/rodsaldanha/arketing-campaign?select=marketing_campaign.csv

Cet ensemble de données appelé Marketing Campaign: Boost the profit of a marketing campaign (augmenter les bénéfices d'une campagne de marketing) provient d'un adaptation du CRM de l'entreprise Ifood, lequel a été fourni aux étudiants pour leurs projets finaux dans un cours appelé Advanced Data Analytics dans le cadre du programme de Master of Science Business Analytics à Hult International Business School pour l'AY 20/21.

Il se compose de 2,240 clients de la société XYZ avec des données sur :

Variable	Description	Type
Succès/échecs de la campagne		
AcceptedCmp1/2/3/4/5	1 si le client a accepté l'offre dans le cadre de cette campagne, 0 autrement	Qualitatives
Response (target)	1 si le client a accepté l'offre dans la dernière campagne, 0 autrement	Qualitative
Profils client		
Complain	1 si le client s'est plaint au cours des 2 dernières années	Qualitative
Days_Engaged	Jours d'adhésion du client à l'entreprise	Quantitative
Education	Niveau de formation du client	Qualitative
Marital	Situation maritale du client	Qualitative
Kidhome	Nombre de jeunes enfants dans le ménage du client	Quantitative
Teenhome	Nombre d'adolescents dans le ménage du client	Quantitative
Income	Revenu annuel du ménage du client	Quantitative
Age	Âge du client	Quantitative
Préférences de produits		
MntFishProducts /MeatProducts / Fruits / SweetProducts / Wines / GoldProds	Montant consacré aux différents produits (Poisson, Viande, Fruits, Aliments Sucrés, Vins et produits d'Or) au cours des 2 dernières années	Quantitatives
Performance des canaux		
NumDealsPurchases	Nombre d'achats effectués avec remise	Quantitative
NumCatalogPurchases / StorePurchases / WebPurchases	Nombre d'achats effectués à l'aide du catalogue, en magasin et sur le site web	Quantitatives
NumWebVisitsMonth	Nombre de visites au site Web de l'entreprise au cours du dernier mois	Quantitative
Recency	Nombre de jours depuis le dernier achat	Quantitative

Pour plus de détail, n'hésitez pas à consulter les Jupyter Notebook “P8_02_code_p1”, “P8_02_code_p2” et “P8_02_code_p3” compris dans les livrables.

3.2 Nettoyage des données

Le nettoyage de données est l'opération de détection et de correction (ou suppression) d'erreurs présentes sur des données stockées dans des bases de données ou dans des fichiers.

Le dataset marketing_campaign contient:

- 24 valeurs nulles dans la variable “Income”. Comme on ne veut pas biaiser l’analyse avec l'imputation de valeurs, les supprimer semble l’option la plus adéquate. Dans l'annexe du Notebook P8_02_code_p1, il y a une régression linéaire qui montre une autre possibilité (imputation) pour la résolution de ce problème.

- 11 individus qui ont été supprimés pour être considérés comme des outliers (valeurs aberrantes ou atypiques):
  - Income: 8 individus considérés comme des outliers selon la technique Z-score.
  - Age: 3 clients qui ont plus de 100 ans.

- 3 individus supprimés parce que la quantité achetée avec remise dépasse la quantité totale d’achats dans les canaux de vente.

3.3 Exploration des données

Comment vend-on? Canaux de vente

Comme il a été mentionné dans le paragraphe précédent, la compréhension de nos canaux de distributions est importante. Dans la Fig-1 on peut constater que notre principal canal de vente est le magasin avec 46% du volume, suivi par les ventes en ligne avec 33% du volume et en dernier les ventes faites par catalogue avec 21% du volume total.

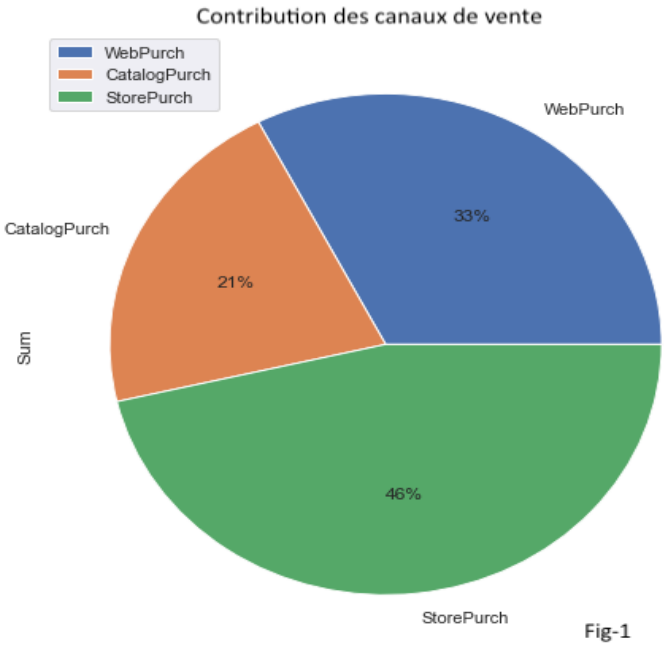


Fig-1

5067 d’achats étaient remisés, soit 18% du total de ventes. Pour approfondir les analyses et améliorer l’impact des réductions, il faudra avoir

des précisions sur la nature de l'origine des réductions telles que la satisfaction du client, le déstockage, le ré-achat, la fidélisation ou la recommandation client ainsi que le coût. Finalement nous verrons comment elles contribuent à l'amélioration des ventes.

Que vend-on? Produits

Le vin est le produit le plus acheté avec 50% des ventes, suivi par la viande avec presque 30% des ventes. Le reste est réparti entre les fruits, les aliments sucrés, les produits en or et le poisson (Fig-2).

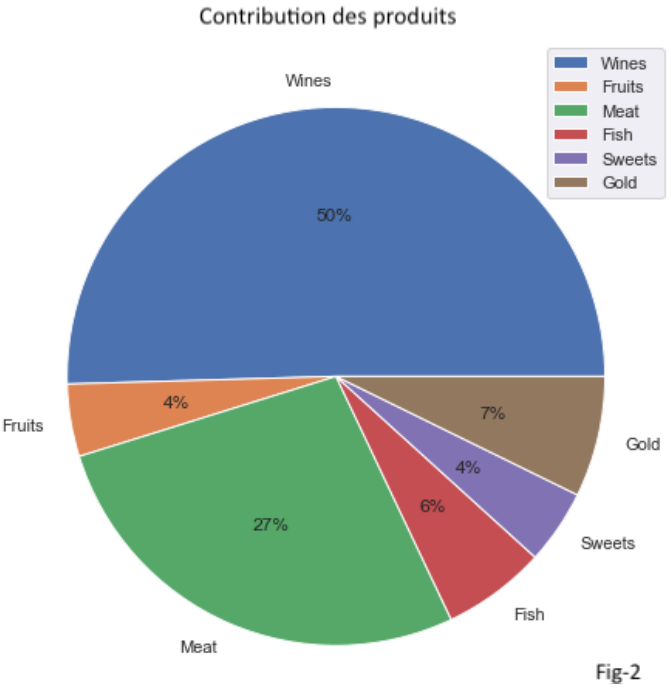


Fig-2

Pour enrichir notre analyse, dans une prochaine étape nous analyserons quels frais génèrent la vente de chacun de ces produits, par les dépenses liées au marketing ou au niveau de distribution. Avec ces données on pourra créer la combinaison de marketing mix la plus adéquate, celle qui maximise les revenus de l'entreprise.

Comment améliorer les ventes? Campagnes Marketing

Un de nos objectifs principaux de cette étude est de connaître l'évolution des campagnes.

Dans la Fig-3 on se rend compte que les campagnes 1, 3, 4 et 5 ont connu une acceptation d'environ 150 clients. La campagne 2 a été la moins performante avec un peu plus de 30 conversions.

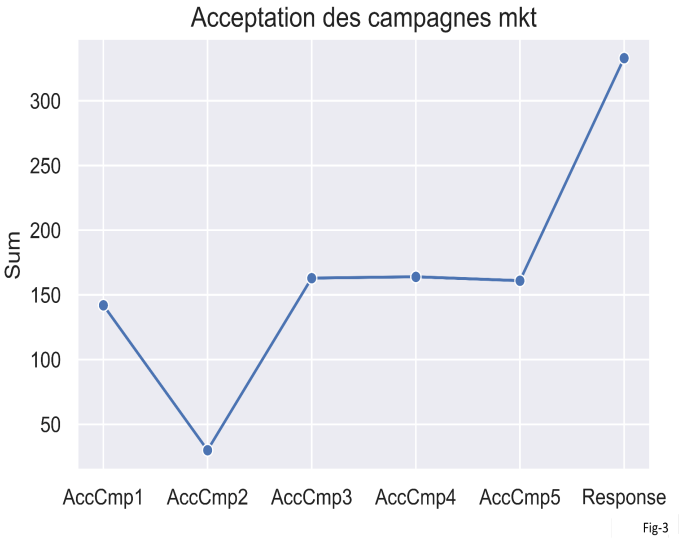


Fig-3

Dans la dernière campagne, la quantité de clients a doublé le nombre d'acceptations. La taux d'acceptation de cette dernière est de 15% par rapport à notre base de clients.

Étudier les échecs et succès des campagnes précédentes, pourra avoir des impacts positifs sur les prochaines. S'interroger sur les leviers de performance de la campagne 2 et de la dernière campagne pour créer des campagnes plus efficaces.

Pour mieux comprendre les campagnes et les produits, on a fait une analyse bivariée. Voici quelques conclusions (les graphiques se trouvent dans l'annexe 1 et dans le Jupyter Notebook):

- Les consommateurs de vin ont préféré les campagnes 1, 2, 4, 5.
- Les campagnes 1 et 5 ont eu plus d'acceptation pour les consommateurs de viande, fruits, produits sucrés et poisson.
- Les acheteurs d'or n'ont pas spécialement une campagne qui les différencie du reste.

4 Analyse de Clusters

La cluster analysis (également appelée analyse en cluster) est une méthode statistique de traitement des données qui organise les éléments étudiés en groupes en fonction de leur degré de similitude. Son objectif est d'identifier et visualiser des ensembles d'éléments similaires en fonction de critères définis.

Reconnaître des éléments similaires, va nous permettre de créer des profils de clients qui vont nous aider dans nos stratégies de ciblage.

L'algorithme utilisé pour cette analyse est K-means. Selon les méthodes "Elbow Method et Analyse de Silhouette" la quantité de clusters optimal est 2. Pour commencer l'analyse nous procéderons donc à une répartition en deux groupes. Prochainement lorsque nous aurons collecté davantage de données, et que notre connaissance de nos clients sera plus précise, nous pourrons certainement augmenter le nombre de clusters pour avoir plus de précision.

Pour le premier graphique de cette partie, nous utilisons une méthode de réduction de dimension appelée Analyse en Composantes Principales (ACP). L'ACP permet de transformer des variables très corrélées en nouvelles variables décorrélées les unes des autres. Le principe est simple : il s'agit de résumer l'information qui est contenue dans une large base de données en un certain nombre de variables synthétiques appelées : Composantes principales (CP).

Dans la Fig-4 nous trouvons la projection des deux groupes dans le deux premiers CP de notre ACP après avoir fait le partitionnement de nos clients avec l'algorithme K-means. Il semble que notre composante principale (F1) qui explique 41,2% de la variation de nos données, sert à différencier le partitionnement des données. On peut donc s'interroger sur quelles variables sont les mieux représentées sur cet axe? Pour répondre à cette question, analysons la Fig-5:

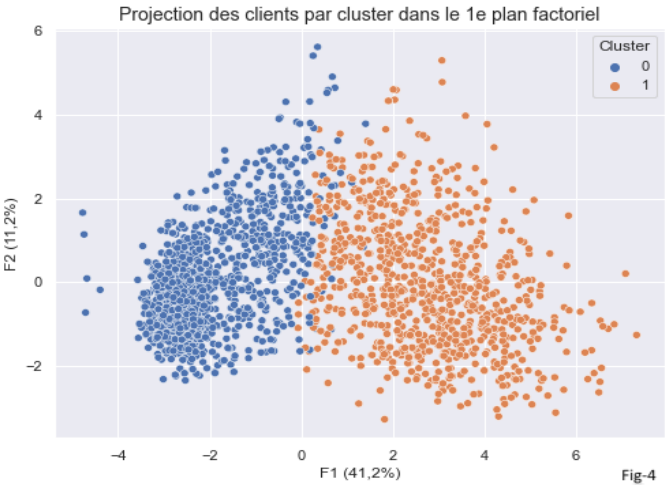


Fig-4

Corrélation des variables dans le F1

Il existe une corrélation positive entre cette composante principale et la quantité totale achetée dans nos canaux de vente, ainsi que avec les revenus de nos clients et la consommation de viande.
Pour le traduire, nos clients qui ont un revenu plus important et qui consomment plus (surtout de la viande) vont appartenir au C1, dans le cas contraire au C0. Comme la variance expliquée dans notre premier plan factoriel est autour de 52%, nous allons continuer l'analyse de nos groupes de clients avec les variables originales.

Variable	Corrélation
TotalPurch	0.34607
Income	0.33523
CatalogPurch	0.32527
Meat	0.30579

Fig-5

Selon cette première analyse on peut conclure que les clients du C1 consomment davantage. Cette tendance se reproduit-elle dans l'acceptation des campagnes?

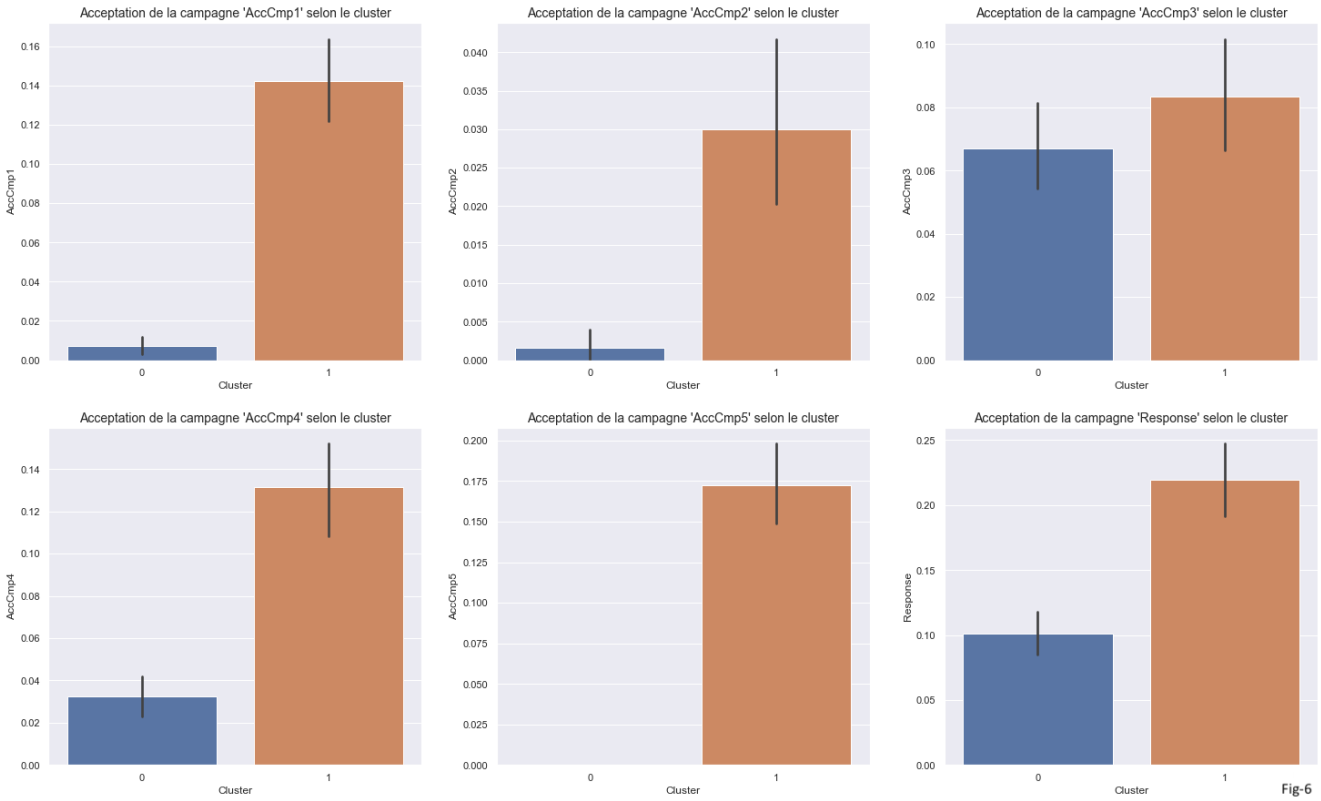


Fig-6

Dans les campagnes 1, 2, et principalement 5 on trouve une différence significative entre les clusters, la 4 un peu moins et la 3 semble ne pas avoir beaucoup de différence. La campagne que l'on cherche à expliquer "Response" ne montre pas de différences significatives entre les clusters. Le C1 toujours réceptif à nos campagnes et le C0 semble avoir un préférence supérieure aux autres campagnes (sauf la 3ème).

Selon l'analyse bivariée de campagnes et produits et celle des clusters (graphique en annexe 2), il est possible de conclure que les clients qui appartiennent au C1 préfèrent le vin et surtout les produits alimentaires.

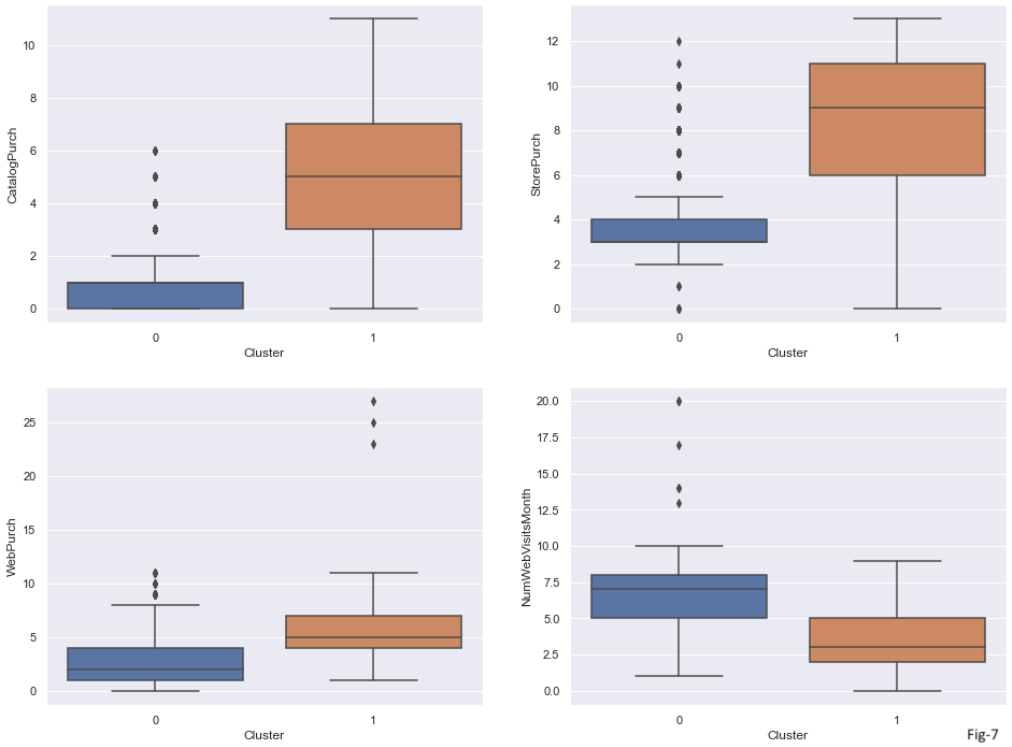


Fig-7

Autant que les clients du C1 achètent plus, selon la Fig-7 il paraît que ses canaux de vente préférés sont les catalogues et les magasins. Par contre, les clients du C0 font plus de visites sur le site par mois.

Pour finir cette section, on fera une description du client moyen de chaque cluster:

Cluster 0 : Ce client a 44 ans et gagne 37860 $ par an. Il a fini sa licence, s'est marié et il a un enfant. Cela fait moins d'un an et demi qu'il est notre client mais il n'achète pas beaucoup (environ 7 fois cette dernière année). Il a dépensé en moyenne 172 $ ces deux dernières années et ses produits favoris sont le vin, la viande et les produits en or. Il fait plus de 6 visites par mois sur le site et il profite davantage des remises que le client du C1. Il est très peu sensible à nos campagnes marketing, surtout aux campagnes 1, 2, 4 et 5. La campagne 3 et la dernière ont eu plus de succès avec lui.

Cluster 1: Ce client a 48 ans et gagne 70500 $ par an. Il a fini sa licence, s'est marié mais n'a pas d' enfants. Cela fait un peu plus d'un an et demi qu'il est notre client. Il achète régulièrement, presque 20 fois en dépensant en moyenne 1200 $ ces deux dernières années. Ses produits favoris sont le vin, la viande et les poissons. Par rapport aux visites mensuelles du site, il est moins actif que celui du C0 (3, 4 fois par mois). Même s'il achète plus que celui du C0 il utilise moins de remises. Il répond suivant à nos campagnes marketing, surtout les deux derniers.

5 Régression logistique

5.1 Introduction au problème

Dans la description du problème nous avions évoqué le marketing prédictif et son importance pour identifier des modèles et anticiper les comportements futurs des clients, les tendances des ventes et les résultats du marketing.
L'algorithme utilisé à cette occasion est la Régression Logistique, un modèle statistique d'apprentissage supervisé permettant d'étudier les relations entre un ensemble de variables Xi et une variable qualitative Y. Il s'agit d'un modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien.
Un modèle de régression logistique permet aussi de prédire la probabilité qu'un événement arrive (valeur de 1) ou non (valeur de 0) à partir de l'optimisation des coefficients de régression. Ce résultat varie toujours entre 0 et 1. Lorsque la valeur prédite est supérieure à un seuil, l'événement est susceptible de se produire, alors que lorsque cette valeur est inférieure au même seuil, il ne l'est pas.

De ce fait, on va essayer de prédire la variable qualitative Y qui sera la réponse positive ou négative à notre dernière campagne de marketing.

La Fig-8 montre la distribution de l'acceptation de notre variable Y:
- 15% ont répondu positivement (1) et 85% négativement (0).

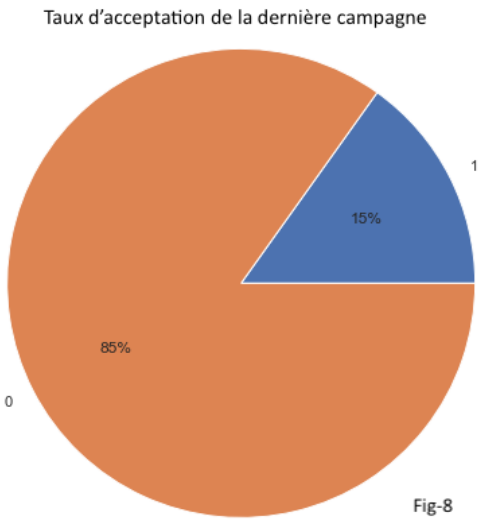


Fig-8

Une des classes est minoritaire par rapport à la population globale. Cela peut être un problème car la plupart des algorithmes de classification se basent sur l'exactitude (ou l'accuracy) pour construire leurs modèles. Voyant que la grande majorité des observations appartient à la même catégorie, on risque de se retrouver avec un modèle peu intelligent qui va toujours prédire la classe dominante.

5.2 Modélisation

Préparation des variables

Comme il a été vu précédemment, il existe deux types de variables: quantitatives et qualitatives.

- Pour les variables numériques nous proposons d'établir une normalisation, c'est-à-dire redimensionner ces variables pour qu'elles soient comparables sur une échelle commune. La

normalisation standardise la moyenne et l'écart-type de tout type de distribution de données, ce qui permet de simplifier le problème d'apprentissage en s'affranchissant de ces deux paramètres.

- On a fait un encodage à chaud sur nos données catégorielles. C'est un processus de traitement de ce type de données, pour les convertir en une représentation vectorielle binaire. Cette méthode crée simplement une colonne pour chaque valeur possible et place un 1 dans la colonne approprié. Le nombre de variables codées fictives nécessaires est inférieur de un au nombre de valeurs possibles, ce qui est K-1.

Pour évaluer la performance de notre modèle, on a utilisé une méthode statistique appelée validation croisée. Lorsque l'on entraîne un modèle sur des données étiquetées, on émet l'hypothèse qu'il doit également fonctionner sur de nouvelles données. Une confirmation supplémentaire sera tout de même nécessaire pour s'assurer de l'exactitude ou non de ses prédictions. La validation croisée permet justement de vérifier si cette hypothèse est valide ou non. Pour l'effectuer, on a utilisé la technique Train-Test split qui consiste à décomposer l'ensemble de données de façon aléatoire. Une partie permettra d'entraîner notre RL, tandis que l'autre servira pour le test de validation. Dans notre cas, 80% des données du dataset ont été utilisées pour l'entraînement. Le reste a été exploité dans le cadre de la Cross-Validation.

Model Summary

Après avoir utilisé plusieurs techniques qui aident à trouver les variables qui sont significatives pour le modèle et avoir fait des combinaisons différentes, nous n'avons finalement gardé que 10 variables. Nous procédons désormais à l'analyse du model summary de Statsmodels:

Logit Regression Results

Dep. Variable:	Response	No. Observations:	1761
Model:	Logit	Df Residuals:	1750
Method:	MLE	Df Model:	10
Date:	Thu, 17 Feb 2022	Pseudo R-squ.:	0.3930
Time:	12:29:37	Log-Likelihood:	-450.64
converged:	True	LL-Null:	-742.39
Covariance Type:	nonrobust	LLR p-value:	6.043e-119

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.1329	0.153	-13.966	0.000	-2.432	-1.834
TotalAcc	1.1293	0.088	12.855	0.000	0.957	1.301
Days_Engaged	0.8745	0.098	8.881	0.000	0.682	1.068
Recency	-0.8716	0.098	-8.933	0.000	-1.063	-0.680
Meat	0.6119	0.096	6.406	0.000	0.425	0.799
StorePurch	-0.6229	0.113	-5.489	0.000	-0.845	-0.401
Education_PhD	1.0172	0.198	5.140	0.000	0.629	1.405
Teenhome	-0.5155	0.112	-4.598	0.000	-0.735	-0.296
DealsPurch	0.5136	0.095	5.401	0.000	0.327	0.700
Marital_Status_Married	-1.4565	0.209	-6.961	0.000	-1.867	-1.046
Marital_Status_Together	-1.1715	0.227	-5.168	0.000	-1.616	-0.727

Fig-9

Notre objectif est d'améliorer notre ciblage pour les prochaines efforts du marketing, il faut donc interpréter les variables qui définissent leur caractéristiques démographiques et le comportement de nos clients vis à vis de la dernière campagne:

Variable	Coef	Description
TotalAcc	1,129	Cette variable qui explique le comportement de nos consommateurs indique que si un client a accepté les campagnes précédentes, il a plus de chances d'accepter la dernière campagne.
Days_Engaged	0,874	Cette variable issue du profil du client nous explique que plus un client est ancien, il a plus de chance d'accepter.
Recency	- 0.872	Pour cette campagne, il faut cibler les clients qui ont acheté plus récemment.
Meat	0,612	Les clients qui achètent plus de la viande (notre deuxième produit le plus contributeur à notre CA) ont plus de chances d'accepter cette campagne. Pourtant le vin qui représente le 50% n'est pas un élément discriminant pour cette campagne.
StorePurch	- 0.623	La relation négative avec les achats en magasin nous fait penser que même si les autres canaux de vente n'ont pas une incidence significative dans le modèle peut être cette campagne a été digitale ou par courrier par exemple avec les catalogues.
Education_PhD	1.017	Pour améliorer le ciblage de nos clients, cette caractéristique sociodémographique nous montre que ceux qui ont ce niveau d'éducation ont plus accepté cette campagne.
Teenhome	- 0.515	Les individus qui ont un adolecent dans le foyer ont moins accepté cette offre.
DealsPurch	0.5136	La personne qui est sensible à la promotion l'est aussi à cette campagne.
Marital_Status_Married	- 1.456	Nos clients mariés ont été moins sensibles à cette campagne.
Marital_Status_Together	- 1.171	Les clients en concubinage ont moins accepté cette offre.

Ces données serviront à améliorer le ciblage de nos clients pour une campagne similaire. Nous pouvons désormais répondre à la question suivante: Qui a accepté notre dernière campagne marketing?

Il s'agit principalement de nos anciens clients célibataires sans adolescents dans le foyer qui ont fait un doctorat. Ils ont accepté les dernières campagnes et ont acheté récemment, surtout de la viande via notre site internet ou par catalogue. Ils profitent aussi des remises !

5.3 Évaluation de la régression

Le pseudo-R2

Le pseudo R2(de McFadden) est basé sur la comparaison des déviances respectives du modèle étudié et du modèle par défaut (modèle trivial). Lorsque la régression ne sert à rien, les variables explicatives n'expliquent rien, l'indicateur vaut 0 ; lorsque la régression est parfaite, l'indicateur vaut 1.
Dans notre modèle, avec R2 MF = 0.3930, il semble que notre modèle se démarque du modèle trivial.

La Matrice de Confusion

Elle confronte toujours les valeurs observées de la variable dépendante avec celles qui sont prédites, puis comptabilise les bonnes et les mauvaises prédictions. Son intérêt est qu'elle permet à la fois d'appréhender la quantité de l'erreur (le taux d'erreur) et de rendre compte de la structure de l'erreur (la manière de se tromper du modèle).

Dans un problème à 2 classes (+ vs. -), à partir de la forme générique de la matrice de confusion, plusieurs indicateurs peuvent être déduits pour rendre compte de la concordance entre les
valeurs observées et les valeurs prédites.
a- TL 0 / PL 0 **Vrais Négatifs (VN)**: les observations qui ont été classées négatives et le sont réellement.
b- TL 0 / PL 1 **Faux Positifs (FP)**: les individus classés positifs et qui sont en réalité des négatifs.
c- TL 1 / PL 0 **Faux Négatifs (FN)**: les observations classées négatives qui sont en réalité positives.
d- TL 1 / PL 1 **Vrais Positifs (VP)**: les individus qui ont été classés positifs et qui le sont réellement.

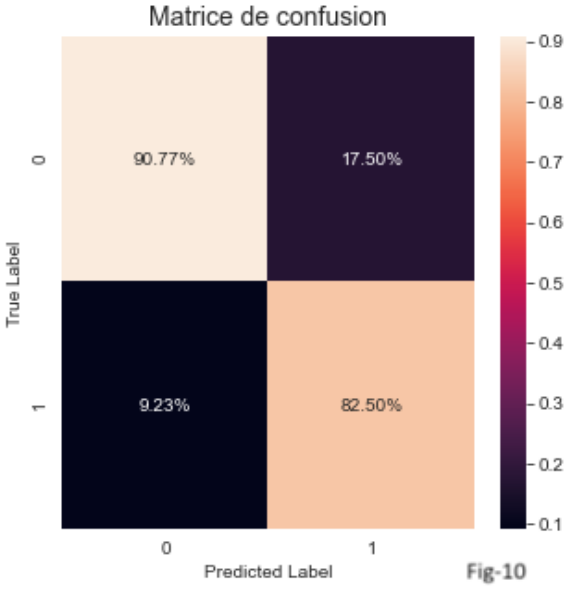


Fig-10

	precision	recall	f1-score	support
0	0.91	0.98	0.94	371
1	0.82	0.47	0.60	70
accuracy			0.90	441
macro avg	0.87	0.73	0.77	441
weighted avg	0.89	0.90	0.89	441

Fig-11

La **précision** indique la proportion de VP et VN parmi les individus qui ont été classés positifs ou négatifs respectivement. Dans ce cas, 90,77% pour les 0 et 82,50% pour les 1.

La **sensibilité** (ou le rappel) indique la capacité du modèle à retrouver les individus positifs. Dans notre modèle c'est le point le plus faible avec un pourcentage de 47%.

La **spécificité**, à l'inverse de la sensibilité, indique la proportion de négatifs détectés. Dans ce cas, il est de 98%.

La **F-Mesure** synthétise (moyenne harmonique) le rappel et la précision, l'importance accordée à l'une ou à l'autre est paramétrable avec β.

Valeur AUC

Avant de définir cette valeur, nous devons connaître la signification de la courbe ROC. C'est une représentation graphique de la relation existante entre la sensibilité et la spécificité d'un test pour toutes les valeurs seuils possibles.
L'aire sous la courbe (ou Area Under the Curve – AUC) est un indice synthétique calculé pour les courbes ROC. L'AUC correspond à la probabilité pour qu'un événement positif soit classé comme positif par le test sur l'étendue des valeurs seuil possibles. Pour un modèle idéal, on a AUC=1, pour un modèle aléatoire, on a AUC=0.5. On considère habituellement que le modèle est bon dès lors que la valeur de l'AUC est supérieure à 0.7.

Notre modèle a un AUC de 0.7263.

Limitations

Selon l'analyse faite, la principale limite de notre modèle est le rappel, la capacité de retrouver tous les positifs. Comme la variable objectif est débalancée, ce n'est pas étonnant que l'algorithme favorise la classe dominante au moment de classer les prédictions.
On a utilisé des techniques comme SMOTE pour équilibrer la variable Y mais les résultats n'ont pas été les meilleurs.
Récolter plus d'informations, surtout celles provenant des individus qui répondent positivement à nos campagnes, va nous aider dans le futur pour améliorer ce modèle et les suivants.

6 Conclusion

Au début de ce rapport nous avions fixé comme principal objectif la création d'un modèle qui nous permette de prédire l'acceptation de notre dernière campagne. Cela nous permettra de mieux cibler nos campagnes marketing dans le futur ce qui se traduira dans des diminutions de coûts, plus de revenus et la fidélisation de nos clients. Connaître l'activité de l'entreprise et savoir qui achète nos produits va contribuer à cet objectif, à l'importance des analyses faites et la compréhension des variables qui contribuent au modèle prédicteur.

La première approche faite avec des analyses univariées et bivariées nous a appris que:

- Les ventes en magasin sont majoritaires, notre produit le plus contributeur à notre CA est le vin et la dernière campagne a été la plus performante.

Deuxièmement l'analyse de clusters nous a permis de différencier deux types de clients séparés en:

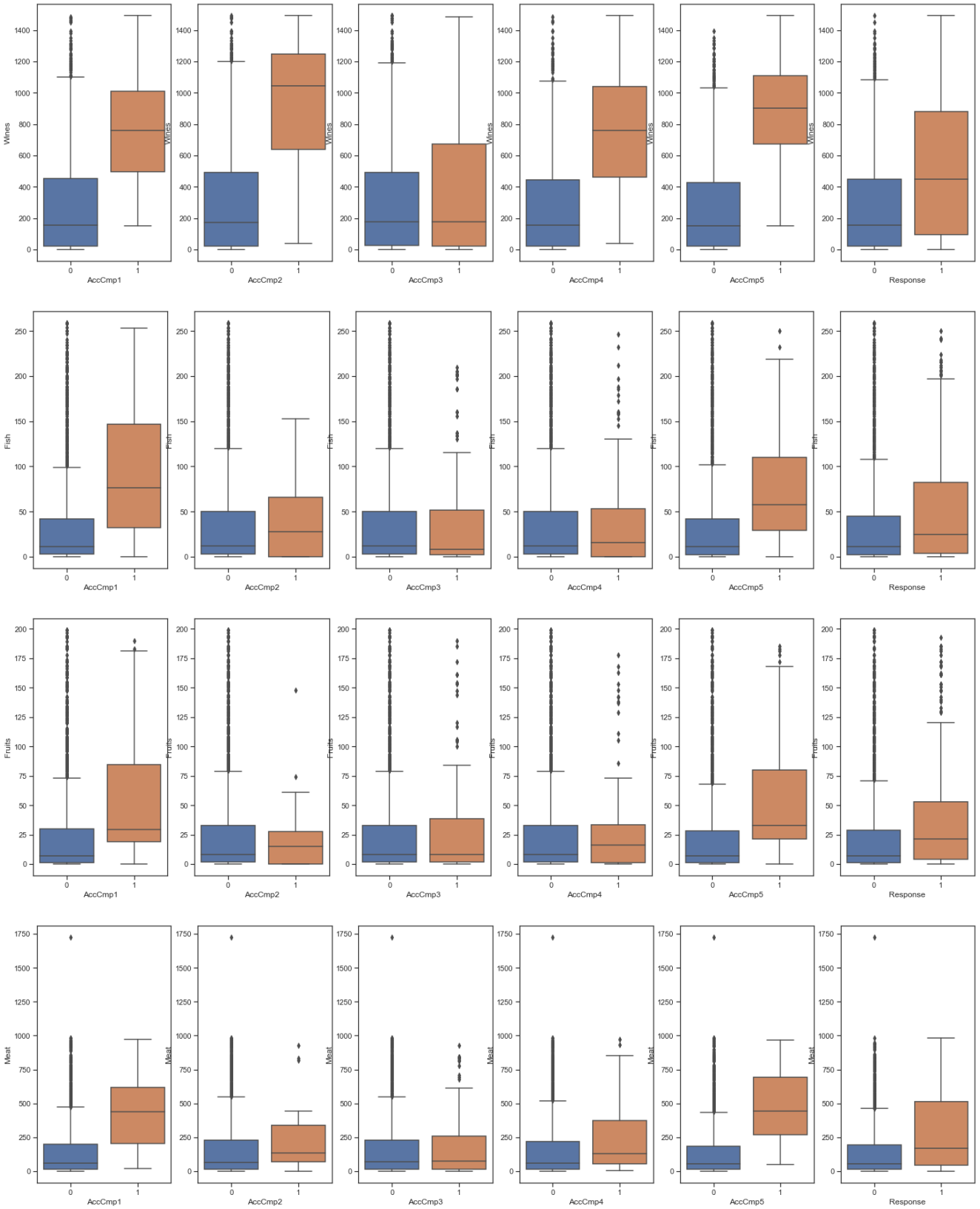
- C0 : Même si c'est un client occasionnel, il visite suivant notre site et profite des remises. Il est peu sensible à nos campagnes marketing.
- C1: Client fidèle avec des revenus plus importants que celles du C0, acheteur de viande et sensible à nos campagnes.

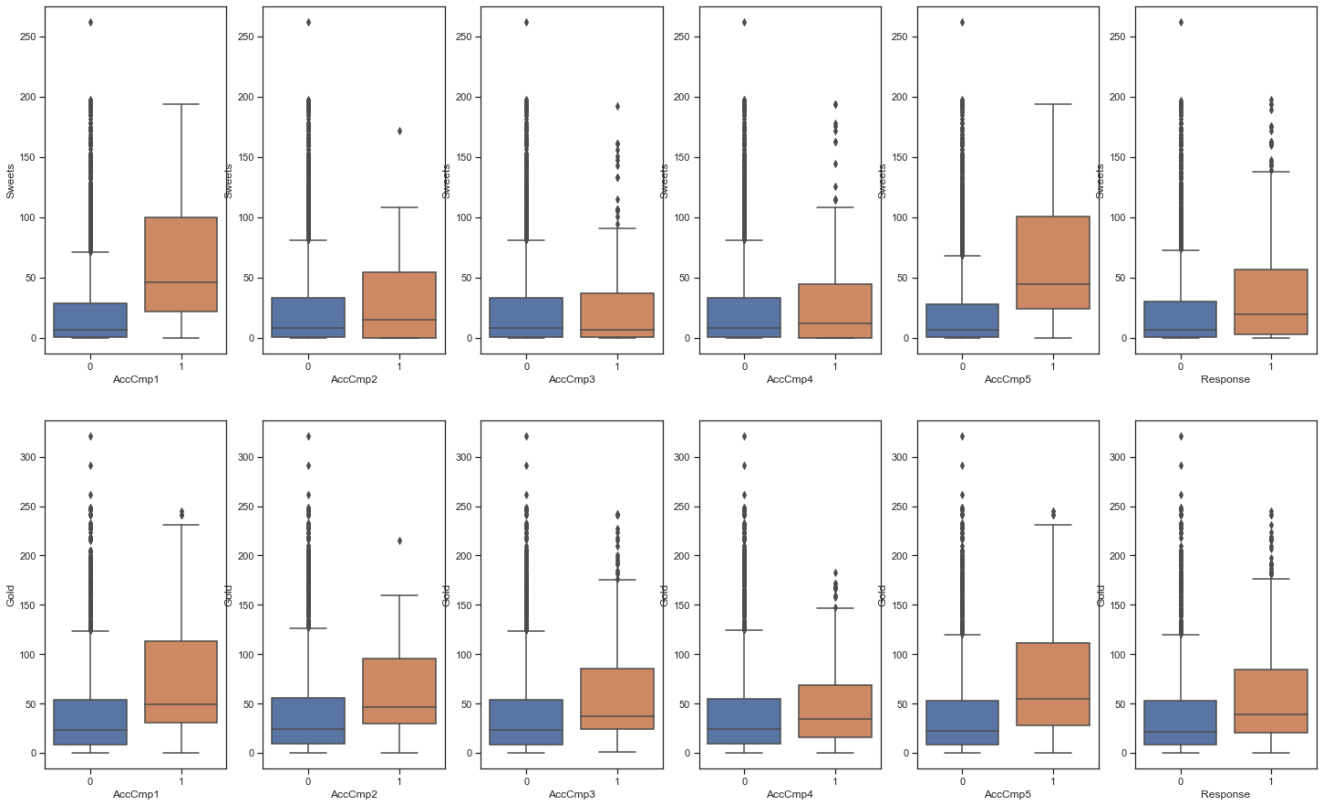
Finalement avec la régression logistique faite on a pu créer un modèle qui prédit les clients qui vont accepter la dernière campagne. On a identifié les variables pour créer un profil type "accepteur" comme l'ancienneté, le fait de ne pas être en couple ni avec des enfants adolescents, et le fait d'être détenteur d'un doctorat. Des autres variables ont été identifiées (liées au comportement du client), comme l'acceptation aux dernières campagnes, l'achat récent, l'achat de viande, l'achat via notre site internet ou par catalogue et les achats en promotion.

Pour continuer l'analyse et l'améliorer, il faudra essayer de balancer notre dataset, mais aussi avoir les informations de la campagne et ses objectifs. Quels sont les facteurs du succès de cette campagne? Le fait que les remises soient une variable explicative et qu'autant de clients du C0 l'ont aussi accepté nous fait nous interroger; cette campagne comprenait-elle des remises?

7 Annexes

7.1 Annexe 1: Campagnes et produits.





7.2 Annexe 2: Clusters et préférences de produits.