DATA CHALLENGE

« service client, innovation et mobilité »

Teddy LEANDRE Jean SOURIS Amine BENOUCIEF

Synthèse :

Dans ce devoir, nous avons déterminé quels types de client serait le plus à même de consommer les services proposés par le leader du télépéage. Pour ce faire, nous avons utilisé BigML afin de créer des clusters et ainsi cibler le profil type de clients qui serait intéressé par la nouvelle offre que nous leur proposons.

1) Introduction :

Dans ce dossier, il nous est demandé d'effectuer une segmentation afin de trouver les clients ayant le meilleur potentiel à consommer ou utiliser des services de télépéage puis trouver quels services permettraient d'augmenter le panier moyen de ces clients.
Pour ce faire, nous allons commencer par trouver combien de groupes de personnes, nous pouvons créer afin de les séparer de manière stratégique. Par la suite, nous allons étudier les habitudes d'achat de ce groupe de personnes afin de leur proposer un service optimal.

2) La méthode et les étapes pour réaliser l'étude :

Premièrement, nous devons nettoyer nos données. Cela consiste à enlever les colonnes et les lignes de notre base de données contenant des valeurs totalement nulles ou invraisemblables.

Par exemple, nous avons enlevé les colonnes : "DPT4", "DEMAT3", "ID", "MENSUEL", et bien d'autres (que vous retrouverez listées dans l'annexe) puisqu'elle ne contenaient pas de données et alourdissait donc notre fichier inutilement en plus de le rendre plus difficilement compréhensible.

Au niveau des lignes enlevées, nous avons retiré toutes les lignes où le Code Postal était incorrect. Il y avait des lignes avec des codes postaux contenant des lettres ou encore ne contenant que 4 chiffres.

En faisant ce tri, nous sommes passés de 13 278 lignes et 61 colonnes à 9 300 lignes et 46 colonnes.

Mais il nous reste tout de même trop de colonnes si nous voulons faire une étude précise.

C'est pourquoi, à l'aide de Python, nous avons créé une heatmap, afin de voir les corrélations entre ces différentes colonnes : (*voir annexe 1*)

Nous pouvons comparer notre première heatmap avec notre heatmap finale, contenant les colonnes qui nous intéressent le plus (*voir annexe 2*)

3) La présentation des principaux modèles utilisés et leurs caractéristiques

- Réduction de dimension

En ce qui concerne la réduction des données, nous avions en premier lieu l'objectif d'utiliser la fonction "whiten()" de la librairie "scipy.hierarachy.vq" mais la qualité de la base de données ne nous a pas permis. Nous avons donc opté pour la solution de facilité qu'est BigML.

- Clustering :

En ce qui concerne le clustering, nous avions 2 options de départ en utilisant Python :

- La méthode hiérarchique, en créant un dendrogramme afin de déterminer le chiffre "k" qui est le nombre de clusters optimal pour notre base de données, puis en visualisant ces clusters avec la fonction "scatterplot" de la library seaborn
- La méthode KMeans, idéale pour des bases de données contenant beaucoup de dimensions (ce qui est notre cas). Cela se présente sous cette forme :

kmeans(standardized_data, nombre_clusters, limite_de distorsion_des_clusters).

Mais ne pouvant pas standardiser nos données dues à la qualité de la base de données et ayant besoin de ce résultat dans les 2 cas de figure cités précédemment, nous avons traité les clusters sur BigML comme montré dans l'annexe 3. De plus, la heatmap ne montrant pas d'option pertinente dans la suppression des colonnes, nous les avons supprimés à la main en suivant notre logique. Finalement, nous avons donc 19 colonnes et 9299 lignes.

Après avoir uploadé notre fichier en tant que source dans BigML, nous avons ensuite créé un dataset, puis nous l'avons filtré en utilisant les colonnes "Civilite" ainsi que "PaysCompte".

4) La présentation du jeu de données avec les statistiques descriptive et exploratoire

Grâce à BigML, voici les informations que nous avons pu obtenir sur ce jeu de données :

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Minimum	0.04769	0.04608	0.06258	0.04495
Mean	0.10766	0.09952	0.18618	0.09418
Median	0.08996	0.08929	0.17652	0.08714
Maximum	1.80279	0.641	0.44948	0.64644

Standard deviation	0.09457	0.04797	0.09313	0.04297
Sum	330.0979	255.67306	8.37818	340.81934
Sum squares	62.95186	31.35408	1.94148	38.77734
Variances	0.00894	0.0023	0.00867	0.00185

5) Les principaux résultats :

a) Segmentation

Au niveau de la segmentation, nous avons simplement pris en compte le nombre de colonnes qu'il nous restait après que les données aient été nettoyées, donc 19, puis nous avons essayé plusieurs nombres de clusters avec Kmeans. Nous avons donc essayé avec 3, 4 et 5 clusters.

Puis, nous avons choisi que le modèle le plus pertinent était 4 Clusters.

Il y a bien évidemment une méthode plus fiable (mais pas absolue) pour trouver le nombre de clusters idéal ; cette méthode est décrite plus haut dans la partie 3).

Le détail de chaque cluster nous a été fourni par BigML dans le "cluster summary report".

b) Ciblage

Au niveau du ciblage, il nous est demandé de déterminer quel profil de clients serait plus à même de consommer plus en termes de télépéage ou de service.

Afin de trouver quel cluster contient le profil type du client pouvant consommer plus, dans le "Cluster Summary Report" de BigML, nous allons nous concentrer sur la colonne "Prix" du fichier .csv fourni.

Les étapes de collecte et mise en forme des données du "Cluster Summary Report" est dans l'annexe.

Comme vu dans l'annexe, le calcul de la moyenne ainsi que la somme des Prix nous montre que le cluster numéro 2 peut contenir le plus grand nombre des personnes appartenant au profil visé.

Ainsi, en nous focalisant sur le Cluster 2, nous pouvons **définir un persona** :

Notre persona est donc un homme, d'environ 58 ans, particulier, prenant un abonnement annuel, et restant en moyenne 10 mois, mais pas de pass Premium, possédant un seul badge et est le profil dépensant le plus d'argent sur l'année.

6) Limites de l'étude

L'étude de cas présentée a été limitée par la qualité de la base de données qui contenait un nombre de données vide type N/A ce qui nous a donc compliqué la tâche lors de la phase de nettoyage et de traitement de la base de données.

Deuxièmement, une autre limite peut être soulevée par rapport à la taille de la base de données. En effet, nous avons eu affaire à un échantillonnage et non à la base de données dans son entièreté ce qui limite la fiabilité du clustering. Par exemple, dans notre bilan de clustering, 2 colonnes se contredisent : la colonne "NBMOISCIRC_CD09_Parking_2018" représentant "Nombre de mois dans l'année ou les clients ont utilisé leur badge pour payer du parking" et "NbMoisCircules" représentant "Nombre de mois ou le badge a été utilisé dans l'année". En effet, dans le choix décisif de notre cluster cible, 2 facteurs nous ont influencés : le nombre de mois et le prix total dépensé par les clients dans les clusters.

Ainsi, les 2 colonnes citées plus haut étant contradictoires, la colonne "Prix" a été décisive. Le détail des calculs du Prix se trouve dans l'annexe. Mais, une fois de plus nous avons eu un souci, car le cluster numéro 3 ne possédait aucune valeur dans cette colonne car le dataset était en partie vide.

7) Conclusion

Ce travail nous demandant de trouver un service pouvant inciter notre persona à augmenter la taille de son panier, nous pourrions essayer de nous focaliser sur le fait qu'il n'ai pas pris de pass Premium pour commencer.

De plus, compte tenu de l'âge de notre persona, les offres contenant des services de mobilité à partager sont à proscrire.

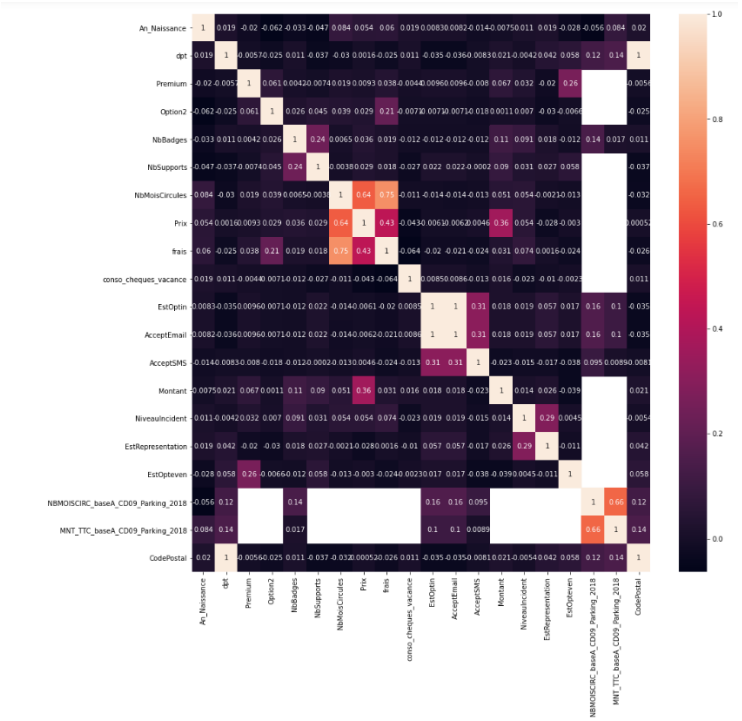
Le profil ciblé faisant partie de la branche payant le plus chaque année, nous pouvons partir du principe que ce genre de personnes possèdent des voitures haut de gamme et électriques selon un récent sondage réalisé par Adot.

Ainsi, nous pourrions leur proposer un service qui leur permettrait d'avoir des réductions sur les parkings de stationnement sur l'ensemble du territoire français contenant des bornes électriques et qui, de plus, lorsqu'ils garent leur voiture, auraient l'occasion de se la faire nettoyer, et cirer par une société de nettoyage premium. De plus, pour mettre en valeur l'offre Premium, nous pourrions faire en sorte que cette offre contient un crédit " lavage haut-de-gamme" par mois, et cela, pour un tarif réduit pour les personnes souscrivant à un abonnement annuel plutôt que mensuel.

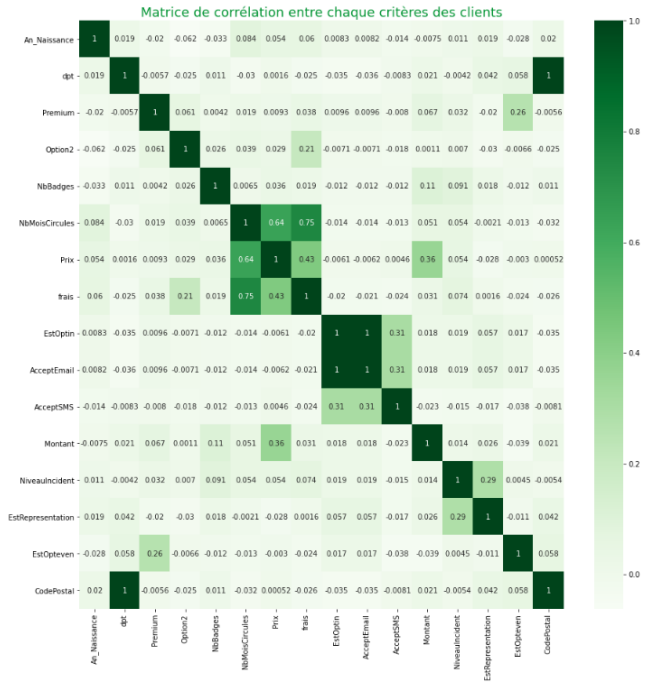
Annexe :

1) Images :

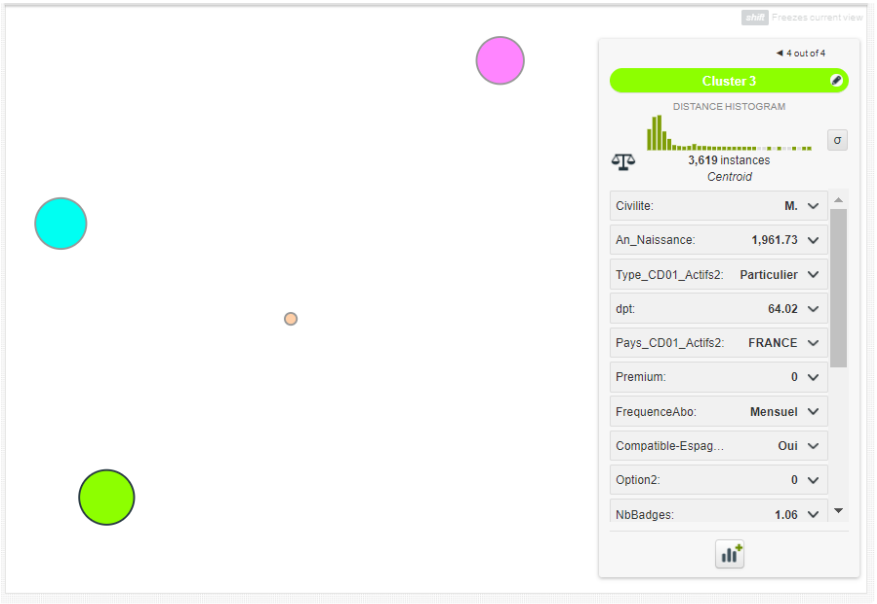
Annexe 1 :



Annexe 2 :



Annexe 3 :



2) <u>Nettoyage des données :</u>

a) <u>Liste des colonnes enlevées :</u>

- Voici les colonnes que nous avons retiré de notre fichier car elles ne contenaient aucune données :

“”TYPE_CLIENT”, “CIVILITE2”, “POSTAL”, “PAYS”, “DATECREATION”, “NBPORTEUR”, “DATEABO”, “ID”, “DEMAT3”, “MENSUEL”, “DPT4”, “nbmois”, “TTC”, “NBMOIS_CIRC_baseB_R_Parking2018”, “MNT_TTC_baseB_R_Parking2018”

- Voici les colonnes enlevées à notre fichier car elles n’étaient pas pertinentes :

“Demat”, “conso_cheques_vacances”, “Nbsupports”, “Option2”, “NBMOISCIRC_baseA_CD09_Parking_2018”, “MNT_TTC_baseA_CD09_Parking_2018”, “MNT_TTC_baseA_CD09_Parking_2018”, “NBMOISCIRC_baseA_CD09_Parking_2018”, “Pays_CD01_Actifs2”, “DateDebut_CD01_Actifs2”, “FrequenceAbo”, “Compatible-Espagne”, “ModeLivraison”, “DateDebut_CD02_Souscriptions”, “TypeFacturation”, “Pays_CD02_Souscriptions”, “DateDebut_CD03_Premium”, “DateFin_CD03_Premium”, “

b) <u>Les fonctions Excel utilisées pour nettoyer la base de données :</u>

- La fonction “=SI()”

Afin d'enlever les données contenant des codes postaux erronés, nous avons créé une nouvelle colonne dans notre tableau, en lui inculquant la fonction =SI([@CodePostal]<10000;"oui";"non")

Ainsi, les codes postaux contenant seulement 4 chiffres ont été affichés par "oui" dans notre tableau, nous avons ensuite filtré ces "oui" puis supprimés de notre base de donnée.

Pour en finir avec les codes postaux, les ayant triés par ordre croissant, nous avons vu en fin de filtre que certains d'entre eux contenaient des chiffres. Nous les avons donc sélectionnés puis supprimés.

- La fonction "=SUPPRESPACE()"

Une seconde fonction utilisée a été =SUPPRESPACE(), afin de supprimer les potentiels espaces dans nos colonnes, pouvant compromettre la compréhension de la machine de certaines données.

3) Utilisation du fichier .csv "Cluster Summary Report"

a) Mise en forme du fichier :

Une fois notre fichier téléchargé, il nous a fallu le convertir en Excel, afin d'avoir une vision plus claire des données fournies.

Après la conversion sur Excel, qui consistait à enlever les "," qui séparaient nos données pour former des colonnes propres, nous avons eu accès à nos données propres.

b) Choix des données importante et mise en valeur de celles-ci :

Après nous être rappelés du profil de client recherché, nous avons pu conclure que la colonne "Prix" de notre fichier "Cluster Summary Report" était la plus importante.

Nous avons donc décidé d'en calculer la somme, puis la moyenne, le tout sur Python.

Mais avant cela, nous avons dû supprimer les " ' " qui encadraient les chiffres de la liste des prix pour chaque ligne.

Pour ce faire, nous avons simplement effectué un "Contrôle + H", puis remplacé les apostrophes par du vide. Notre liste est maintenant utilisable sur Python.

c) Calculs détaillés de la somme et de la moyenne de la colonne "Prix" des clusters :

Sur Python, une fois notre liste prête à l'emploi, nous avons effectué les calculs suivant :

- Calcul de la **somme** du premier cluster :

(la liste contenant 519 chiffres, nous avons donc décidé de la réduire)

x1 = [10, 100, 1018, 102, 103.2, 1031, 106, 107, 107.3, [...], 99, 990]

d = sum(x1)

print(d)

d = 189214.40000000005

- Calcul de la **moyenne** du prix du premier cluster :

x1 = [10, 100, 1018, 102, 103.2, 1031, 106, 107, 107.3, [...], 99, 990]

d = sum(x1)/len(x1)

print(d)

d = 364.57495183044324

- Tableau récapitulatif des sommes et moyennes de la colonne “Prix” des clusters :

Clusters	Somme	Moyenne
1	127393.39999999998	394.4068111455108
2	**207852.2**	**494.8861904761905**
3	/	/
4	49348.59999999998	158.16858974358968

4) Bibliographie :

Site du sondage Adot :
https://l.facebook.com/l.php?u=https%3A%2F%2Fwww.larevueautomobile.com%2FActu%2Fqui-sont-les-possesseurs-et-acheteurs-de-voiture-electrique-en-france.html%3Ffbclid%3DIwAR0NHy__Ks4ji7DByO5N26NV83FjsWJgf3KRgs_v0MwlChnkBDDcPofFHDY&h=AT3l8lIH5UfjyKBcm32KTf_UBx92o_InMOCXhV9Wxf9FN9mGJV98Aei7b8mggiI0TQvwlVzcqsjfyt76zg0ULbdRpBqRRuHaALf26EXbk7rnVZS_Vq3XSwk4xZ5x5R00KZ2big