Soutenance projet 5
Segmenter le comportement des clients

Etudiant: Pierre-Emmanuel Beaumale

Mentor: Mohammed Sedki

Date: 24/04/2018

OpenClassrooms: Parcours data scientist

Soutenance projet 5

- Plan de la soutenance:
  - Présentation de la problématique.
  - Présentation globale de la démarche
  - Traitement effectué sur le jeu de données.
  - Analyse exploratoire : Jeu de données et RFM
  - Segmentation clients obtenus
  - Modélisation: 1er modèle, classification suivant la première transaction.
  - Modélisation: 2nd modèle pour compléter le 1er modèle.
  - Synthèse des résultats et conclusion.
  - Q&A.

Problématique client

Problématique client

Notre travail consiste au sein d'une équipe marketing à mieux comprendre le comportement de clients de l'entreprise afin que celle-ci puisse adapter sa stratégie commerciale afin de maximiser la fréquence d'achat et d'augmenter son chiffre d'affaire. Notre travail se focalisera sur la compréhension du comportement des clients dans la durée afin de détecter ceux qui sont le plus susceptible de passer à l'achat à travers la réalisation d'une segmentation suivant des critères à définir.

Approche du problème

Notre approche (partie I)

Etape de segmentation :

- Transformation des données d'entrées en un tableau résumant l'ensemble des caractéristiques **clients observées**. Une ligne représentera l'ensemble des caractéristiques du client.
- Utilisation d'un algorithme de clustering afin de déterminer des catégories de clients ayant un sens marketing.
- Interprétation du clustering obtenu.

Notre approche (Partie II)

Prédiction de l'appartenance de clients à ces catégories :

- En utilisant les catégories créées, nous allons essayer de prédire le plus rapidement possible l'appartenance d'un client à une classe en développant 2 modèles:
  - Un modèle basé sur les **features du 1er achat.**
  - Un modèle complémentaire disposant des transactions clients après la première transaction après une période **après le 1er achat.**

  Conclusion et développement d'un programme de sélection aléatoire de séquences clients pour faire la démonstration de notre travail.

Traitement préliminaire du jeu de données

Jeu de données

Description sommaire du jeu de données:

- Nous disposons pour cela d'une année de transactions datant du 01/12/2010 au 09/12/2011 disponible depuis le site de l'UCI.

- Ce jeu de données contient 8 variables et 541909 lignes:
  - 1 ligne = 1 article d'une transaction
  - Variables: le numéro de facture, le code article en stock, la description synthétique de l'article, la quantité commandée pour l'article, la date de facture, le prix, un identifiant du client et le pays de résidence du client.

- **Premier traitement sur le jeu de données:**
  - Suppression du jeu de données des transactions avec une variable « Customer_ID » manquante (135080 articles supprimés).
  - Suppression du jeu de données des transactions dont la variable « InvoiceNo » commence par un c (transactions annulées) à l'exception des transactions discount.

Jeu de données

Premier traitement sur le jeu de données (suite):

- Nettoyage des données de type chaine de caractères et de type entière.
- Traitement des outliers (sensibilité de l'algorithme K-means).

Création des features suivantes:

- « Amount » produit du nombre d'articles par le prix unitaire.
- « Manual » pour les transactions ayant fait lieu à un retrait de la marchandise au dépôt.
- « POST » pour les transactions étant envoyé par la poste.
- « Discount » pour les transactions ayant fait l'objet d'un rabais.
- « Is_UK » pour déterminer si un client réside au Royaume-Uni ou non.
- « Year_Month » pour déterminer le mois et l'année d'achat.

Jeu de données: Information de base

Quelques informations sur les données disponibles:

- Chiffres d'affaires sur la période observée : 6 321 374 pounds.
- Nombre de transactions conservées: 17714 transactions.
- Nombre de clients: 4209 clients conservées.
- Montant moyen d'une transaction : 356,85 pounds pour une moyenne de 208 articles achetés par transactions.
- Nombre d'articles échangé 3610 articles.
- 37 pays de résidence pour les clients.

Jeu de données: Quelques explorations

- Quelques observations:

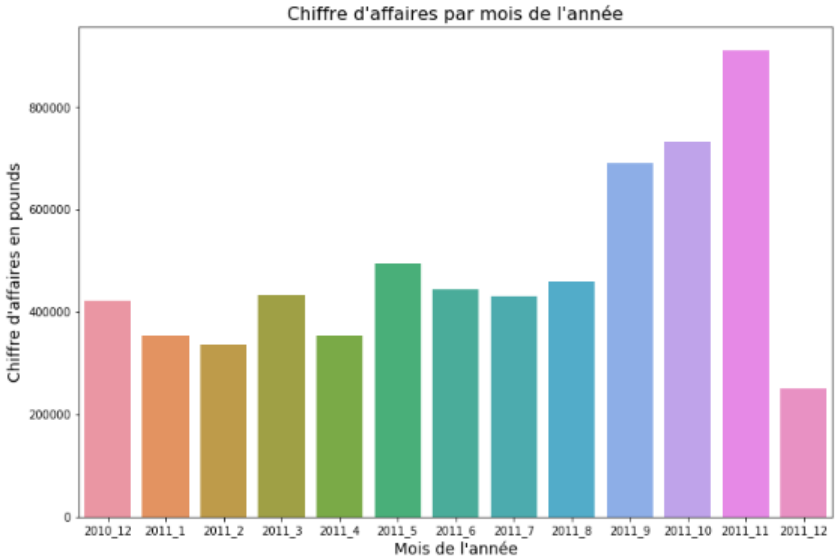

Figure 4 : Evolution du CA par mois.

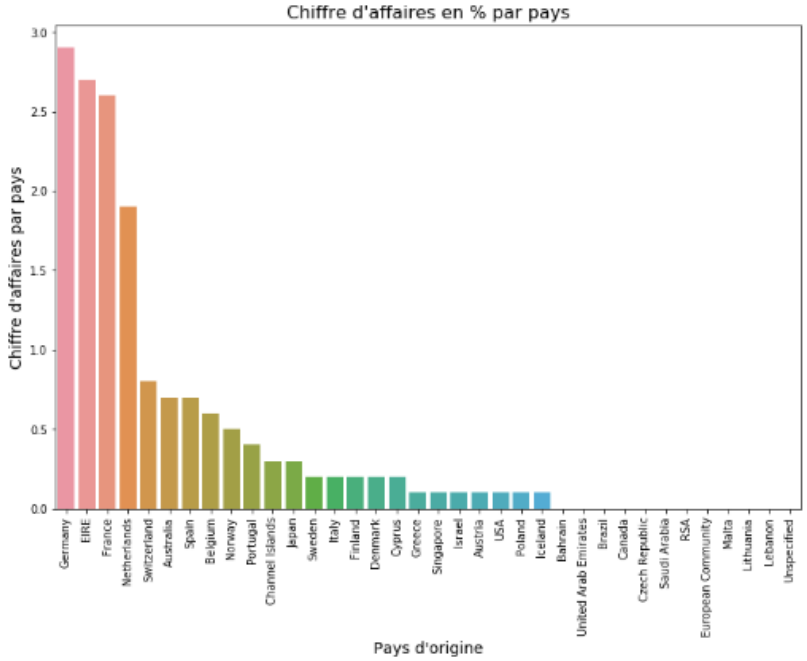

Figure 5 : Répartition du CA par pays en dehors de UK (84% des ventes).

Jeu de données: Quelques explorations

- Quelques observations:

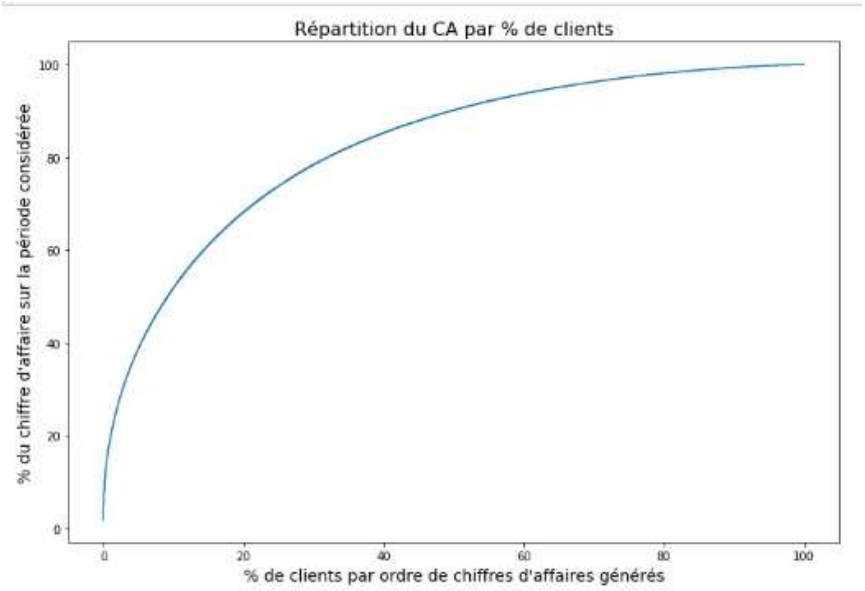

Figure 6 : Evolution du CA par clients cumulés.

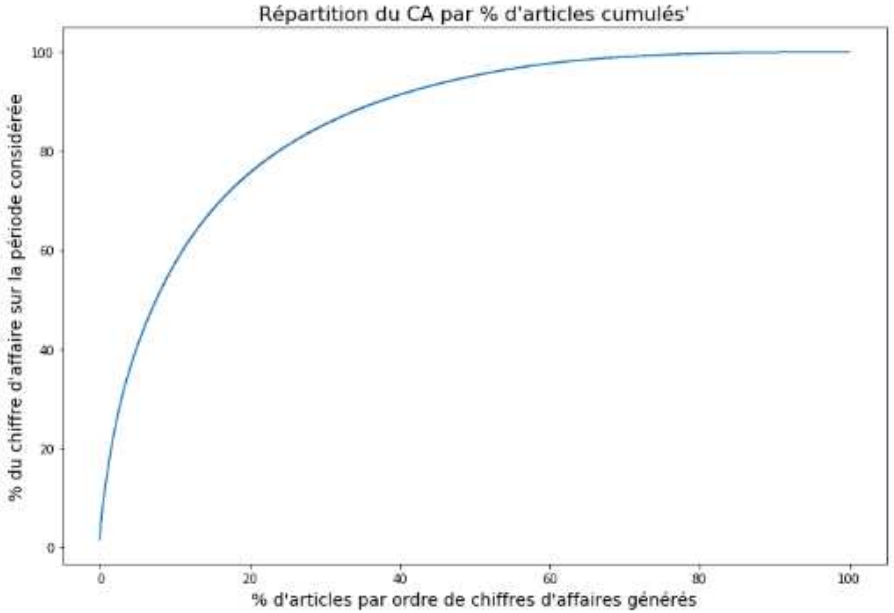

Figure 7 : Répartition du CA par articles cumulés.

Tableau comportement client

Création d'un tableau résumant les caractéristiques client avec les variables suivantes:

- « Amount » le montant total des transactions.
- « Amount_first » le montant de la première transaction.
- « Post » « Manual », « Discount » : La somme de ces valeurs cumulées par client.
- La moyenne par transaction des features suivantes:
  - Quantité minimale, moyenne, maximale.
  - Prix unitaire minimal, moyen, maximal.
  - Le prix d'achat minimal, moyen et maximal.
- « Fréquence » le nombre de transactions sur la période.
- « Recency » la distance en jours au dernier achat client.
- « Latency » la distance en jours au premier achat client.
- « Is_UK » indiquant si un client réside ou non au Royaume-Uni.

Nombreuses tentatives de clustering sur l'ensemble des données

→ Features finales retenues pour notre clustering : Frequency, Recency, Latency, Amount, Amount_first, Is_UK.

Tableau comportement client

Quelques observations sur la RFM

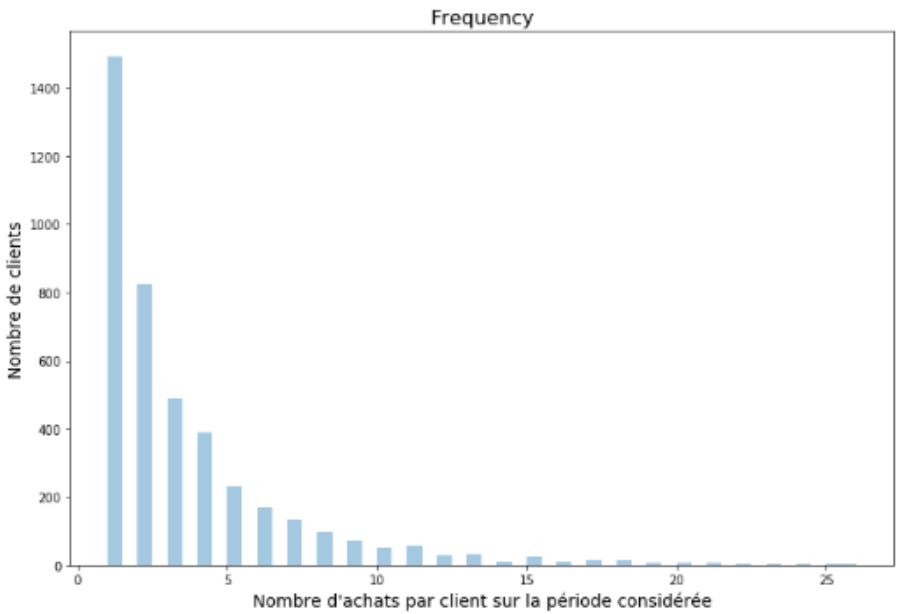

Figure 9: Distribution du nombre d'achat par client.

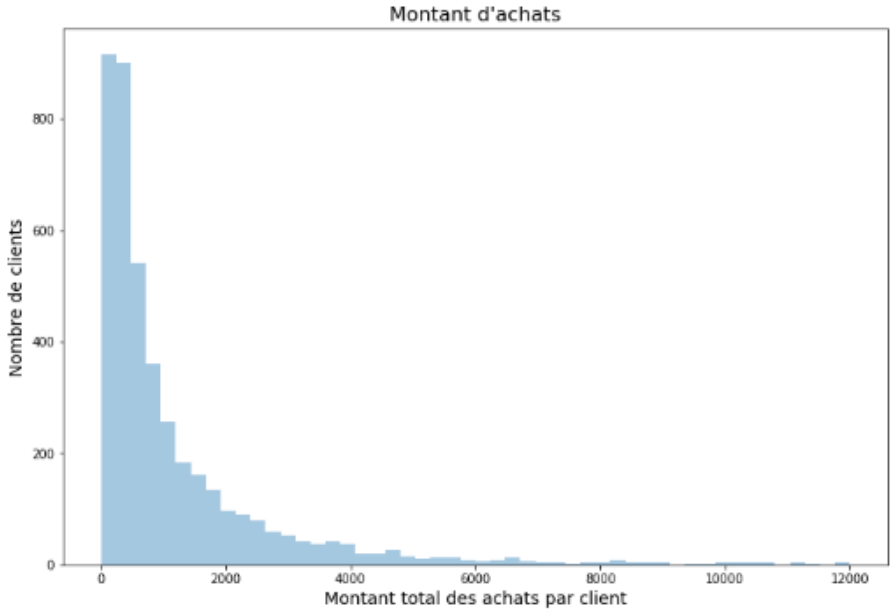

Figure 10: Distribution du montant d'achat par client.

Clustering obtenu

Applications de l'algorithme K-means à notre table de données

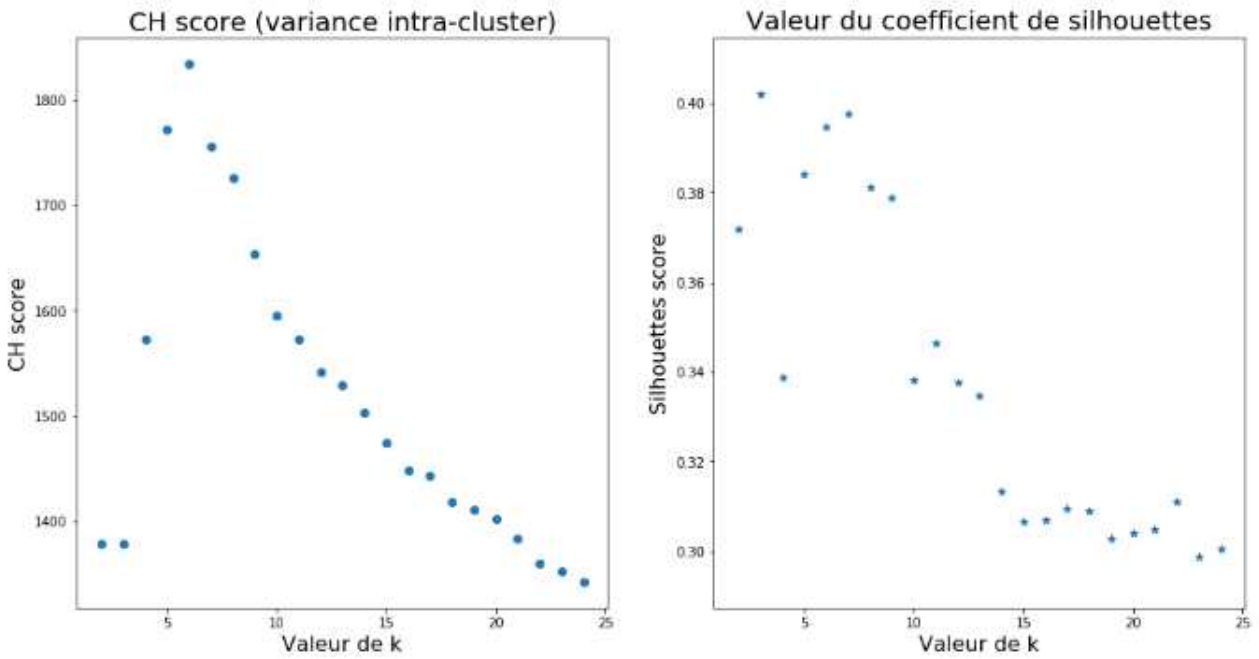

Figure 016: Evolution du coefficient CH et du coefficient de silhouettes en fonction des valeurs k de l'algorithme k-means appliqué à notre table de données.

Clustering obtenu

Sélection de k = 5 pour notre clustering:

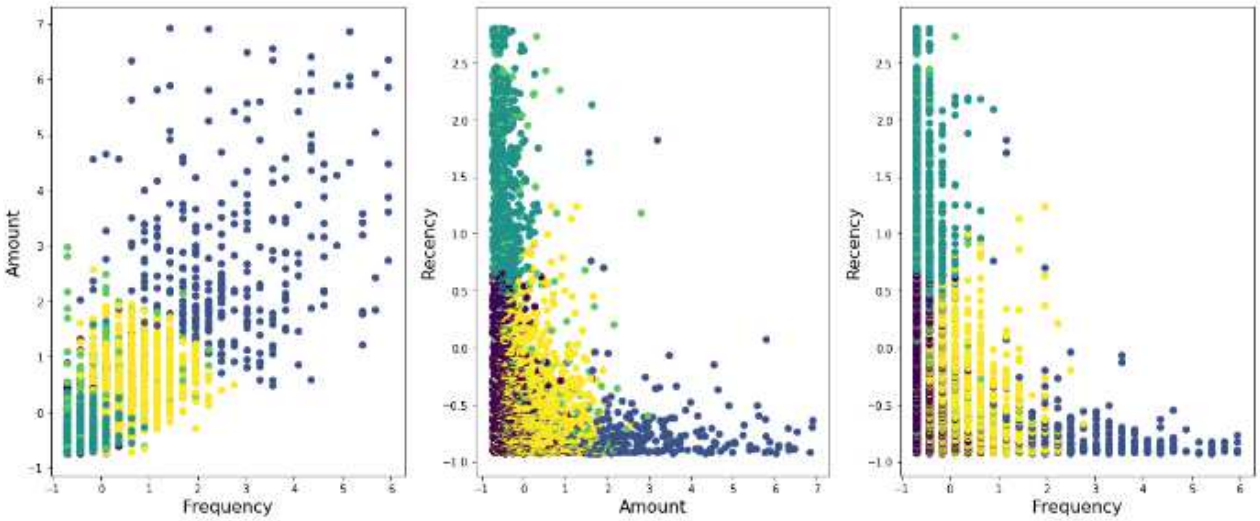

Figure 017: Représentation du clustering obtenu dans les 3 dimensions RFM.

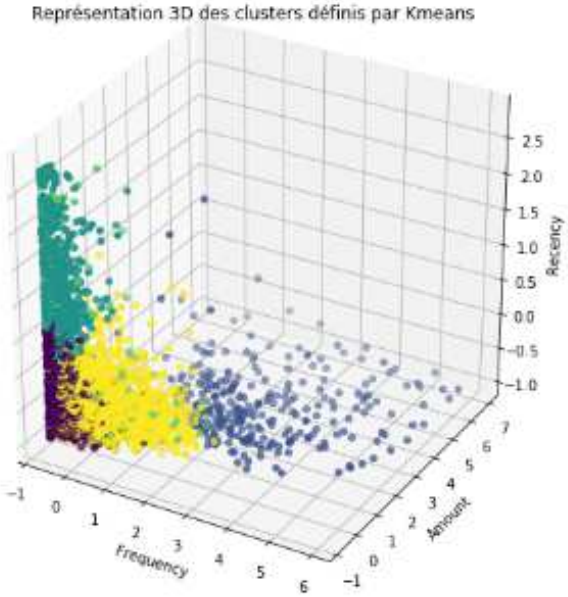

Figure 018: Représentation 3D du clustering obtenu sur les données dans les dimensions RFM.

Clustering obtenu

Sélection de k = 5 pour notre clustering:

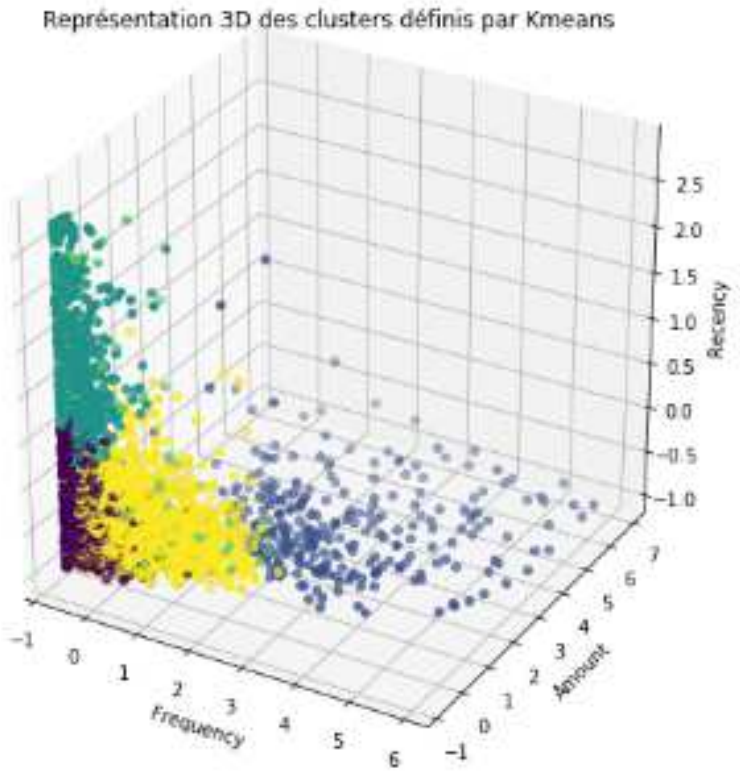

Figure 018: Représentation 3D du clustering obtenu sur les données dans les dimensions RFM.

Clustering obtenu

Sélection de k = 5 pour notre clustering:

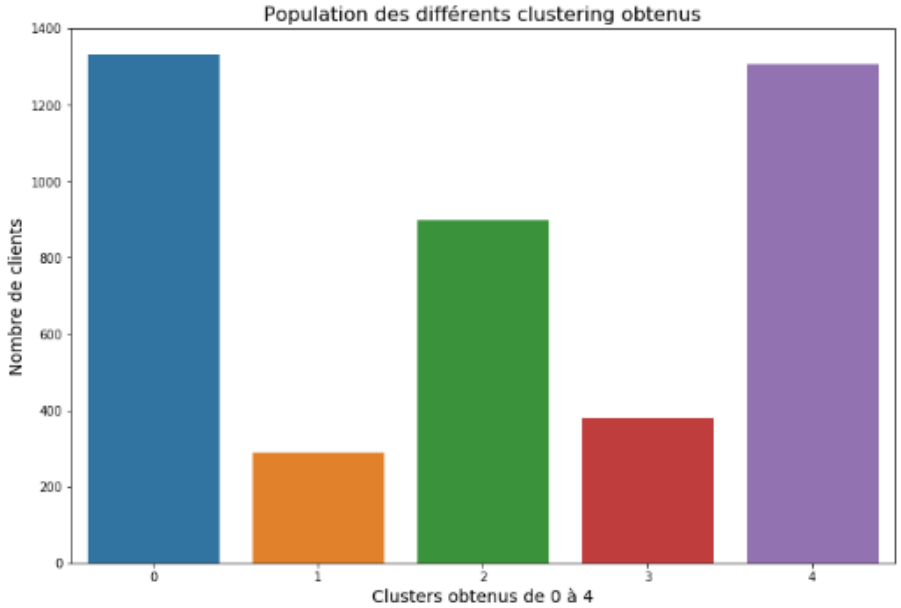

Figure 019: Représentation de la répartition de population entre les clusters obtenus. Le plus petit cluster contient 366 clients.

Interprétation du clustering

Sélection de k = 5 pour notre clustering:

- **Classe 0**: Ce sont les clients récents, ils représentent un peu plus de 10% du chiffre d'affaires et ont réalisé en moyenne 1 à 2 achats durant les 3 premiers mois après le premier achat. Il s'agit d'un segment client à développer et à fidéliser car ils représentent le potentiel de croisance du CA pour l'entreprise. Ils sont de plus assez nombreux (32% des clients).
- **Classe 1**: Ce sont les clients fidèles qui ont commencé très tôt à acheter sur le site et qui commandent le plus fréquemment (1 fois par mois).Ils représentent 25 % du CA pour seulement 7% de la clientèle. Il faut absolumment les conserver en leur faisant profiter de promotions (discount).

Interprétation du clustering

Sélection de k = 5 pour notre clustering:

- **Classe 2**: Ce sont les clients perdus, ils représentent un peu moins de 5 % du chiffre d'affaires et ont réalisé en moyenne 1 à 2 achats durant les 3 premiers mois après le premier achat. Il s'agit d'un segment de clientèle perdu depuis plus de 6 mois, il n'y a plus d'intérêt à aller faire du démarchage auprès de ces clients.Il représente 21 % des clients.
- **Classe 3**: Ce sont la plus grande partie des clients étrangers, ils sont peu nombreux (9% de la clientèle) et représente 7% du CA.
- **Classe 4**: Ce sont là aussi des clients fidèles, qui commandent régulièrement (1 fois tous les 2 mois), ils représentent 30% du chiffre d'affaire et 31% des clients. Sans campagne marketing, ces clients risquent d'évoluer vers groupe 1.Il faut les conserver par une campagne marketing régulière.

Modélisation : Prédiction de l'appartenance d'un client à une classe

Modélisation : 1er modèle

Prédiction le plus rapidement possible de l'appartenance d'un client à une classe:

- Modèle 1: Features obtenus sur les achats réalisés MAX_DAY après le 1er achat:
- Sur le premier achat, nous construisons les features suivantes::
  - « Unit Price » : moyenne, min, max.
  - « Amount »: Somme, moyenne, min, max.
  - « Quantité »: Somme, moyenne, min, max.
  - Discount, Manuel ,POST
  - Nombre de produits achetés.
  - Is_UK l'appartenance du lieu de résidence au Royaume-Uni.
  - Month: le mois de la première transaction.

- Entrainement par séparation du jeu de données et validation croisée (5 fold)

Modélisation : 1er modèle

Prédiction le plus rapidement possible de l'appartenance d'un client à une classe:

Modèle 1 - Classifier	Résultats
Dummy	24,9 %
KNN	66,2 %
Régression Logistique	67,5 %
SVC	67,8 %
Gradient Boosting	68,2 %
Vote combiné (3 derniers)	67,4 %

Modélisation : 1er modèle

Prédiction le plus rapidement possible de l'appartenance d'un client à une classe:

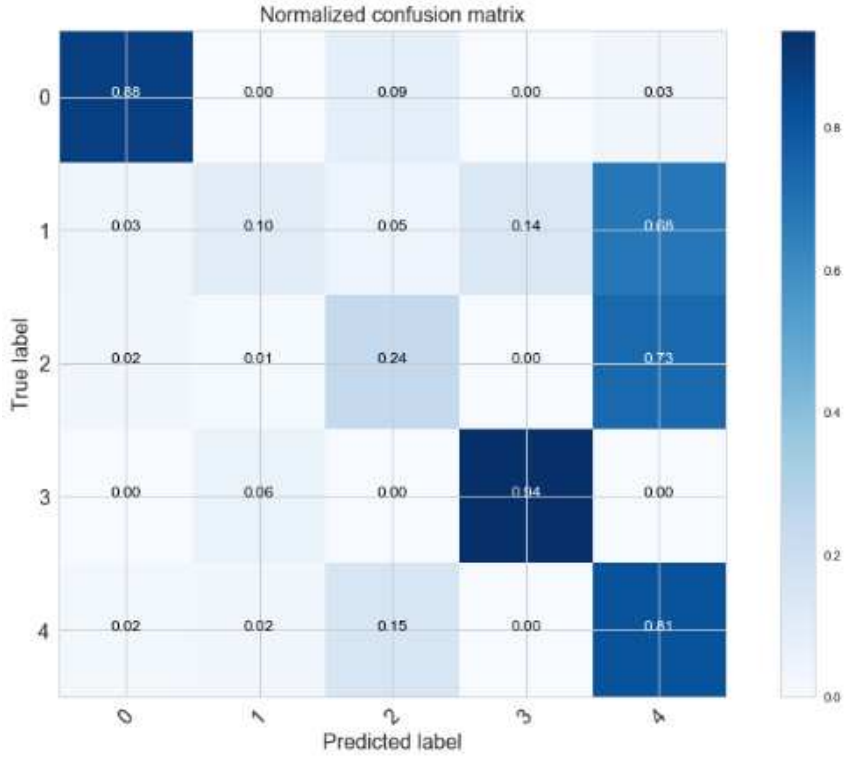

Figure 020: Matrice de confusion des résultats obtenus par notre classifieur combiné à partir des features du premier achat.

Modélisation : 1er modèle

Prédiction le plus rapidement possible de l'appartenance d'un client à une classe:

- Résultats intéressant pour:
  - La classe 3 (clients étrangers peu nombreux et représentant une faible partie du CA, score à 94%)
  - La classe 0 (les clients récents ayant fait peu d'achats, score à 88%).

- Difficultés observées:

  La classe 1 (les clients qui dépensent le plus et le plus souvent) ne sont pas du tout bien prédit avec ce modèle et qu'ils sont confondus avec la catégorie majoritaire, la catégorie 4 (les acheteurs relativement fréquents mais qui dépensent peu par individu au global).

  La classe 2 (les clients perdus) sont confondus avec les clients 4, cela est tout à fait normal car notre premier modèle ne tient pas compte de la fréquence des achats ou de la distance dans le temps au dernier achat.

Modélisation : 2nd modèle

Prédiction le plus rapidement possible de l'appartenance d'un client à une classe:

- Modèle 2: Variables obtenus sur les achats réalisés MAX_DAY après le 1er achat:
- A partir du premier achat, nous conservons les mêmes variables que le modèle 1.
- Sur l'ensembles achats sur la période, nous construisons en complément des variables précédentes:
  - « AmountTotal »: Le montant total des achats
  - « Frequency »: La fréquence des achats.
  - « Periode » :La période d'achat étant défini comme la distance en jours entre le premier achat et le dernier achat.

- Entrainement par séparation du jeu de données et validation croisée (5 fold)

Modélisation : 2nd modèle

Prédiction le plus rapidement possible de l'appartenance d'un client à une classe: MAX_DAY = 180 jours

Modèle 2 - Classifier	Résultats
Dummy	25,6%
KNN	75,2 %
Régression Logistique	82,7 %
SVC	82,4 %
Gradient Boosting	83,2 %
Vote combiné (3 derniers)	82,7 %

Modélisation : 2nd modèle

Prédiction le plus rapidement possible de l'appartenance d'un client à une classe:

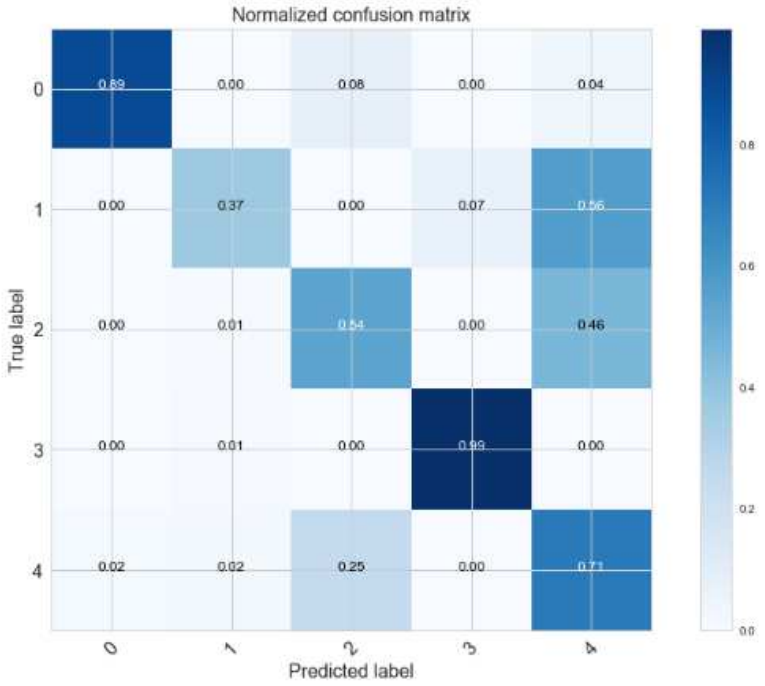

Figure 021: Matrice de confusion des résultats obtenus par notre classifieur combiné à partir du second modèle pour MAX_DAY =180.

Modélisation : 2nd modèle

Prédiction le plus rapidement possible de l'appartenance d'un client à une classe:

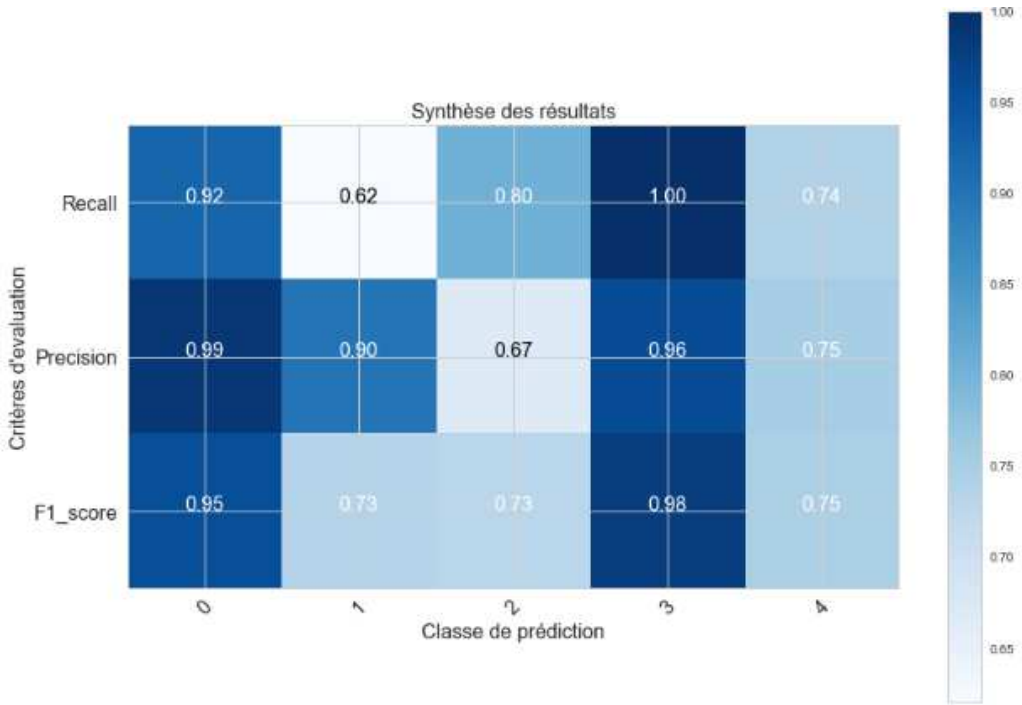

Figure 022: Critères d'évaluation de notre classifieur combinée avec MAX_DAY = 180.

Modélisation : 2nd modèle

Prédiction le plus rapidement possible de l'appartenance d'un client à une classe:

- On améliore le score global de notre prédiction mais notre algorithme ne parvient pas à obtenir de scores de prédiction supérieur à 90% pour les classes 1, 2 et 4.
- Cependant, on peut cependant noter que la précision sur la classe 1 est bonne (supérieur à 90%) ce qui veut dire que pour les clients identifiés de classe 1, nous pouvons engager un démarchage commercial avec efficacité (90% des clients prédits de classe 1 sont bien de la classe 1).

Influence du paramètre MAX_DAY

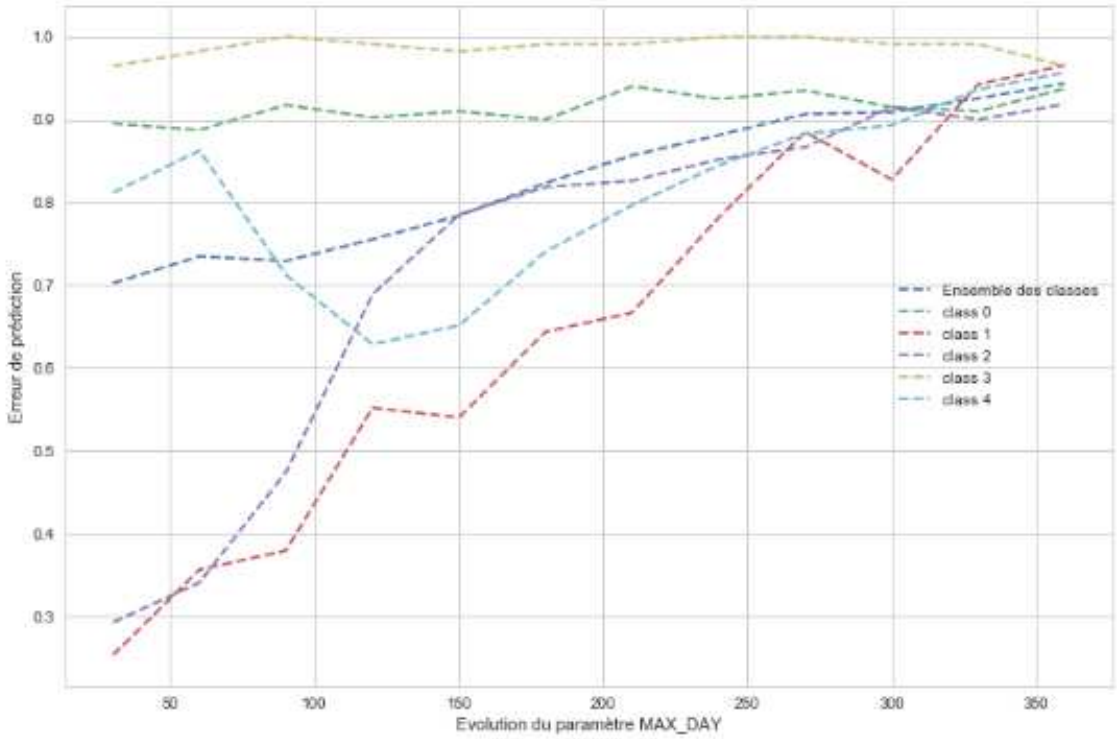

Figure 007: Evolution des erreurs de prédiction de notre modèle par classe et au total pour différentes valeurs de MAX_DAY.

Conclusion et perspectives

Conclusion et perspectives

- De nombreuses tentatives ont été effectuées en utilisant des algorithmes de clustering sur notre jeu de données initial afin d'obtenir notre segmentation, le modèle donnant les meilleurs résultats est celui intégrant la récence et la latence des achats. Ce modèle différencie alors les segments clients en fonction de l'évolution de leur comportement dans le temps ce qui complique notre tâche de prédiction dès le premier achat.

- Nos deux modèles sont complémentaires:
  - Le premier pourra servir à prédire rapidement les nouveaux clients (sans conserver les données privées du client).
  - Le second pourra être utile afin de lancer une démarche de ciblage marketing sur les clients de classe 1 identifier après 6 mois de transactions.

Conclusion

Afin de continuer notre travail, il conviendrait d'explorer :

- Plus en détails les possibilités de segmenter à partir de catégories d'articles les produis achetés.
- De continuer à introduire de nouvelles features afin de mieux prédire l'appartenance des clients à chaque classe.

Annexe : Description de l'algorithme du gradient boosting

Formulation du GBT

Gradient Boosting : Gradient descent + boosting

- Appartient à la famille des « forward stage-wise additives model »:
- Gradient Boosting est une généralisation de l'algorithme Adaboost avec des fonctions objectives généralisées.
- Principe : L'idée est de rajouter successivement des « apprenants faibles » à une fonction (de prédiction ou de régression) afin d'affiner le modèle.
  - A chaque ajout d'un additive (un apprenant faible), l'idée est d'améliorer le modèle
    - En adaptant le poids des données à chaque étape pour l'adaboost.
    - En ajoutant un apprenant faible paramétré par descente de gradient de l'erreur (de prédiction ou de régression) pour le gradient boosting. A chaque étape, un apprenant faible est rajouté et les apprenants précédent sont figés (forward stage-wise).

Formulation du GBT

Algorithme générale du gradient boosting multiclasse K:

Algorithm 10.4 *Gradient Boosting for K-class Classification.*

1. Initialize $f_{k0}(x) = 0,\ k = 1, 2, \ldots, K$.
2. For m=1 to M:
   (a) Set
   $$p_k(x) = \frac{e^{f_k(x)}}{\sum_{\ell=1}^{K} e^{f_\ell(x)}},\ k = 1, 2, \ldots, K.$$
   (b) For $k = 1$ to K:
   i. Compute $r_{ikm} = y_{ik} - p_k(x_i),\ i = 1, 2, \ldots, N$.
   ii. Fit a regression tree to the targets $r_{ikm},\ i = 1, 2, \ldots, N$, giving terminal regions $R_{jkm},\ j = 1, 2, \ldots, J_m$.
   iii. Compute
   $$\gamma_{jkm} = \frac{K-1}{K} \frac{\sum_{x_i \in R_{jkm}} r_{ikm}}{\sum_{x_i \in R_{jkm}} |r_{ikm}|(1 - |r_{ikm}|)},\ j = 1, 2, \ldots, J_m.$$
   iv. Update $f_{km}(x) = f_{k,m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jkm} I(x \in R_{jkm})$.
3. Output $\hat{f}_k(x) = f_{kM}(x),\ k = 1, 2, \ldots, K$.

Source : Element of statistical learning, Hastie et Tibshirani, p387, 2nd edition.

Probabilités de x d'appartenir à la classe k, k = 1,2..,K.

M: Nombre d'apprenant choisi
Jm: Taille des arbres choisis

Yik = 1 si yi appartient à la classe k, 0 sinon.
r_{ikm} est la déviance, écart entre la classe réelle et la classe prédite par chacun des classifieurs.
On construit un arbre de régression sur r_{ikm}

Formulation du GBT

Variations autour de l'algorithme à considérer:

- Ajout d'un terme de shrinkage v afin de contrôler la vitesse d'apprentissage, influence du paramètre v sur le nombre d'apprenants faible M.

$$f_m(x) = f_{m-1}(x) + \nu \cdot \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm}).$$

- Subsampling: Extension du gradient boosting au stochastic gradient boosting en exploitant le principe du bagging (boostrap aggregation). Les apprenants faibles sont entrainés sur une partie aléatoire seulement du jeu de données d'entrainement à chaque étape.

Formulation du GBT

Paramètres à considerer lors de la validation croisée:

- Jm: la typologie d'arbre utilisé, en règle général la profondeur des arbres.

- M: le nombre d'apprenant total utilisé.

- Shrinkage v: la vitesse d'apprentissage.

- Fonction de perte : exponentiel, binomiale ou mutlinomiale deviance.