Projet 4 : Segmentation des clients d'un site e-commerce

Formation IML 2020

Plan de présentation :

i. Le contexte du projet
ii. Les données à disposition
iii. La préparation des données
iv. Les différentes méthodes de segmentation
v. La stabilité des segments dans le temps
vi. Les axes d'améliorations

Le contexte du projet :

- Olist :
  - Une entreprise brésilienne
  - Activité de Marketplace
  - Nécessite une bonne connaissance des clients

- Objectifs :
  - Proposer une segmentation client exploitable
  - Evaluer la fréquence de mise à jour
  - Respecter PEP8 pour rendre réutilisable les codes

Les données :

- Les données 10/2016 à 10/2018

9 tables qui contiennent :

- Les informations acheteurs et vendeurs (adresse, ville, …)

99 441 acheteurs, 3 095 vendeurs, 70% des acheteurs sont à SP et RJ

- Les informations produits (catégorie, dimensions, nombre des photos, …)

32 951 produits, 71 catégories

- Les informations sur les commandes (dates, valeurs, frais de livraison, satisfactions, type de payement, …)

La préparation des données :

Nettoyage des données

- Ne garder que le statut « delivered » dans les commandes(97%)
- Ne garder que la période d'achat à partir de 01/01/2017(début réel d'activité)
- Ne pas garder la géolocalisation mais juste les états et les villes
- Supprimer les caractéristiques physique des produits

Fusion des différentes tables

```
customers :   (99441, 5)
geolocation :   (1000163, 5)
order_items :   (112650, 7)
order_payments :   (103886, 5)
orders :   (99441, 8)
order_reviews :   (100000, 7)
products :   (32951, 9)
sellers :   (3095, 4)
category_name_translation :   (71, 2)
```

AllData : (110 718, 26)

Constitution des nouvelles variables

- Scores Récence-Fréquence-Montant
- Score de satisfaction moyenne, Nombre de type de paiement
- Ancienneté du client, Montant d'un panier moyen
- CA du client par groupe de catégorie de produit

Data : (94 862 , 25)

Segmentation

Problème de segmentation pour un apprentissage non supervisé

Les méthodes :

- La méthode RFM classique :
- La méthode k-means :
  - K-means avec les variables RFM
  - k-means avec d'autres variables

Comparer la qualité de segmentation :

- Homogénéité des points dans un cluster
- Distance entre les clusters
- Exploitabilité selon les besoins métiers

} Silhouette_score *

* Silhouette_score : métrique avec une valeur comprise entre [-1,1], plus on se rapproche de 1, la prédiction est meilleure

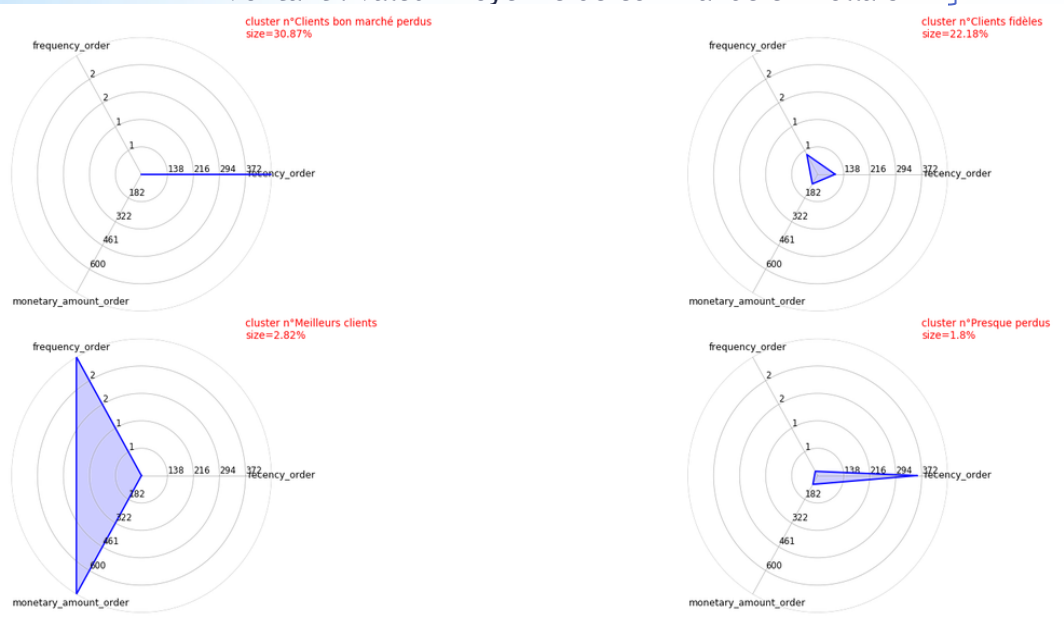
Segmentation: RFM Classique

Scoring à base de quartile*

Variables :

Récence : nbr de jour depuis la dernière commande
Fréquence : Nombre de commandes effectuées
Montant : valeur moyenne de commande en Dollars

Attribution de score de 1 à 4



6 clusters :

- Meilleurs clients
- Clients bon marchés perdus
- Clients fidèles
- Clients presque perdus
- Gros dépensiers
- Autres

Avantage : C'est une segmentation adaptable et facile à comprendre.

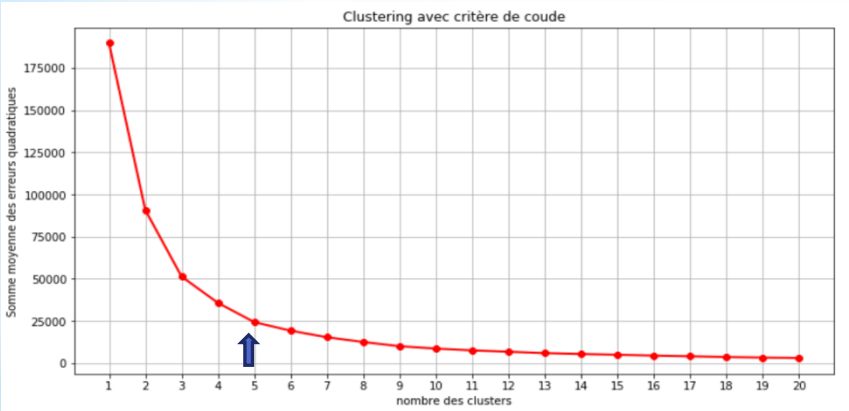
Inconvénient : Nombre de variables très limité, la plupart des clients n'ont acheté qu'une fois.

Quartile *: Q1,Q2,Q3,Q4 division des données en 4 parts égaux

Segmentation: k-means

Les variables R-F-M

Recherche de nombre de cluster :

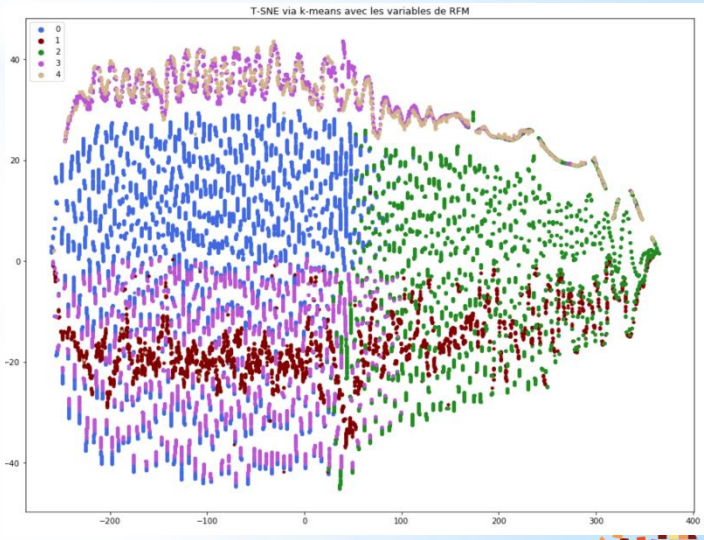


K-means :
- le centroide , distances euclidiennes entre les points

Recherche des meilleures valeurs de K [0,1,..,20] la somme des moyennes des erreurs quadratiques est stable à k = 5

Qualité : homogénéité, distance entre clusters
- > silhouette_score = 0,361

Représentation graphique avec T-SNE * :

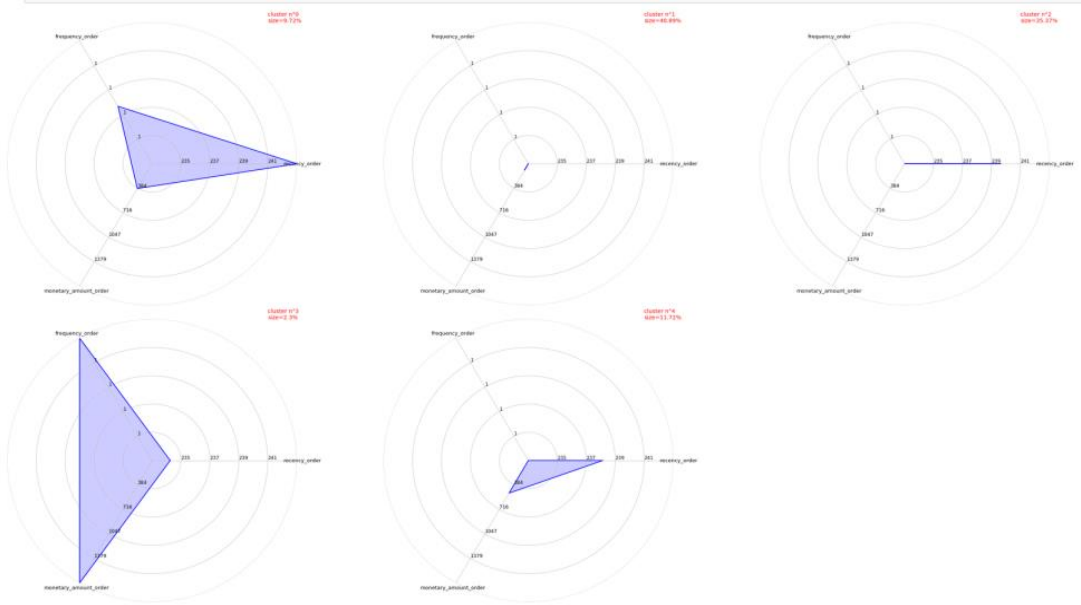


T-SNE *: technique de réduction dimensionnelle, pour visualiser des données à grandes dimensions

Segmentation: k-means

Les variables R-F-M

Répartition des variables dans chaque cluster :



5 clusters
Silhouette_score = 0,361

Constat : Forte ressemblance aux clusters du premier modèle,

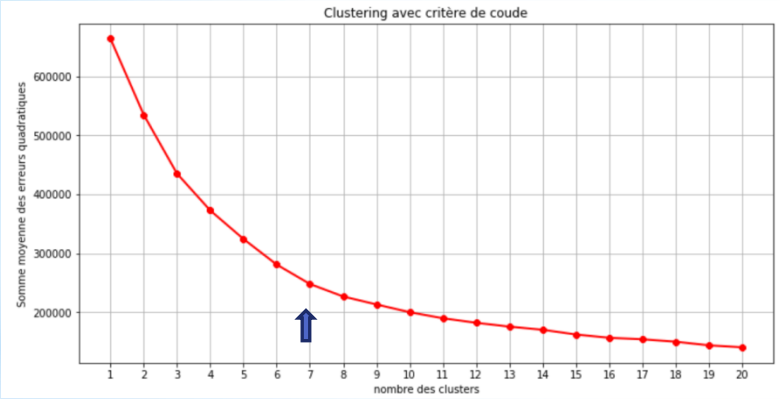
Non-adapté à car : 96 % des clients ont commandé une fois, insuffisance des variables d'études malgré le silhouette_score à 0.361

Segmentation: k-means

Des variables supplémentaires :

Satisfaction, ancienneté, cout d'achat moyen, nb type de paiement, quantité des photos

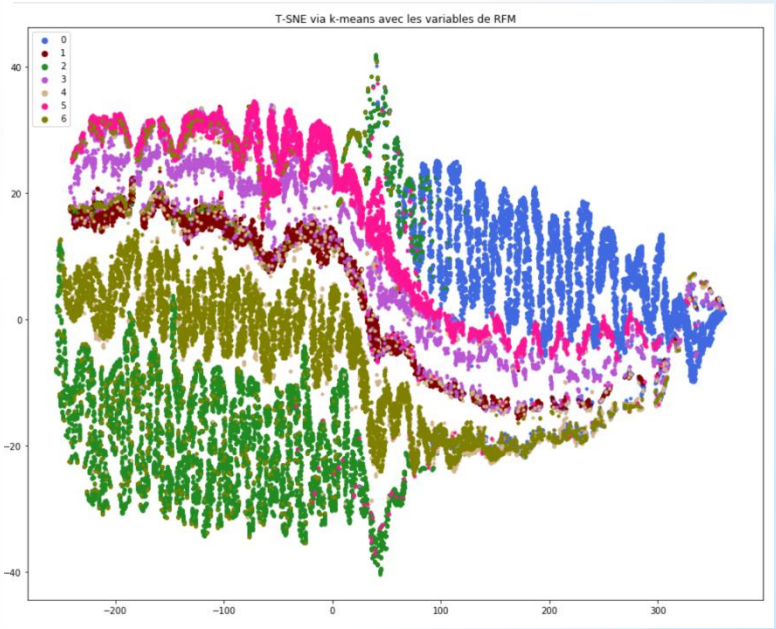
Recherche de nombre de cluster :



- > 7 clusters

- > Silhouette_score = 0,254

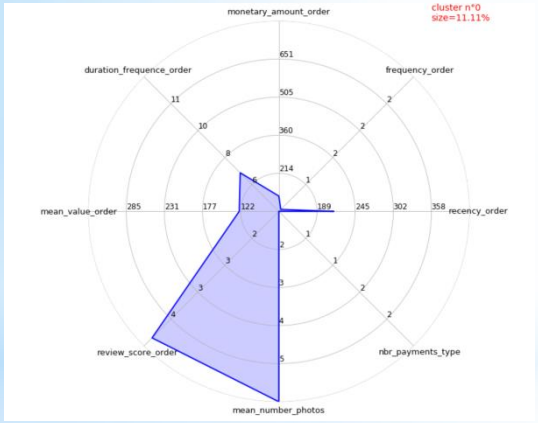
Représentation graphique avec T-SNE* :

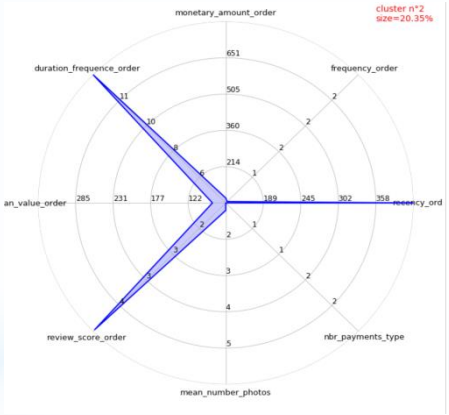


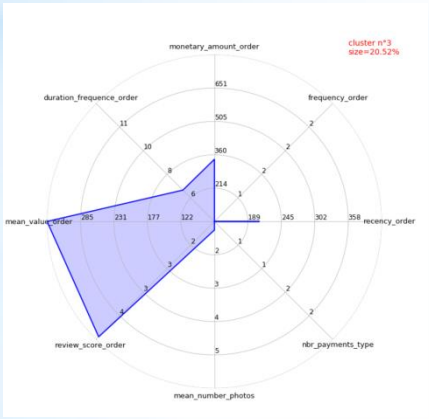
Segmentation: k-means

Des variables supplémentaires :

Répartition des variables dans chaque cluster :







Cluster 0 : très satisfaits de leur(s) achat(s)
avec des produits qui ont beaucoup de photos sur le site
et un cout d'achat moyen

Cluster 1 : très satisfaits de leur(s) achat(s)

Cluster 2 : ont fait leur première achat depuis +de 11 mois
et n'ont pas commandé depuis très longtemps
avec un montant pas cher mais sont satisfaits

Cluster 3 : Satisfaits avec un montant moyen assez élevé
et dernière commande depuis 3 mois

Cluster 4 :acheteurs assez récents et pas content

Cluster 5 : très bons clients, contents de leurs achats et
achètent pour des montants élevé et fréquemment

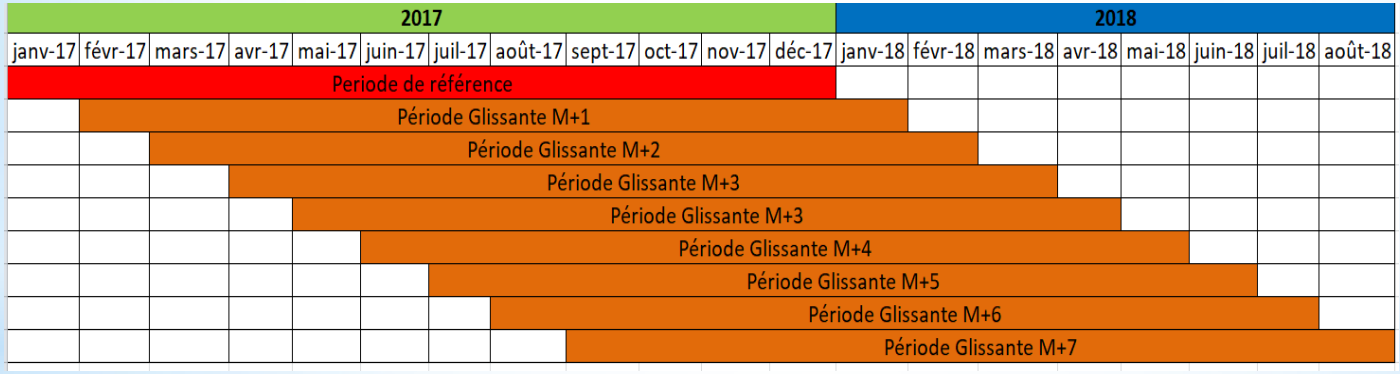
Cluster 6 : clients assez récents, content de leurs achats,
ont acheté au - 2 fois et utilisent +sieurs type de
paiements

La méthode la plus adaptée
car elle nous permet d'analyser
les comportements des clients sur plusieurs angles

Etudes de stabilité :

Stabilité d'un algorithme de segmentation :

- Consistance des résultats de prévision avec différentes données en entrée



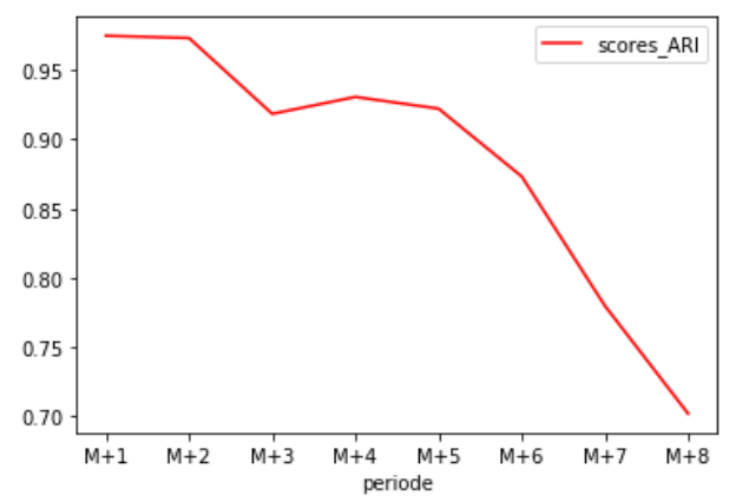
1er : Faire une prévision des clusters sur périodes glissantes avec les données de la période de référence en entrainement.
2e : Calculer les clusters des périodes glissantes avec leurs données.

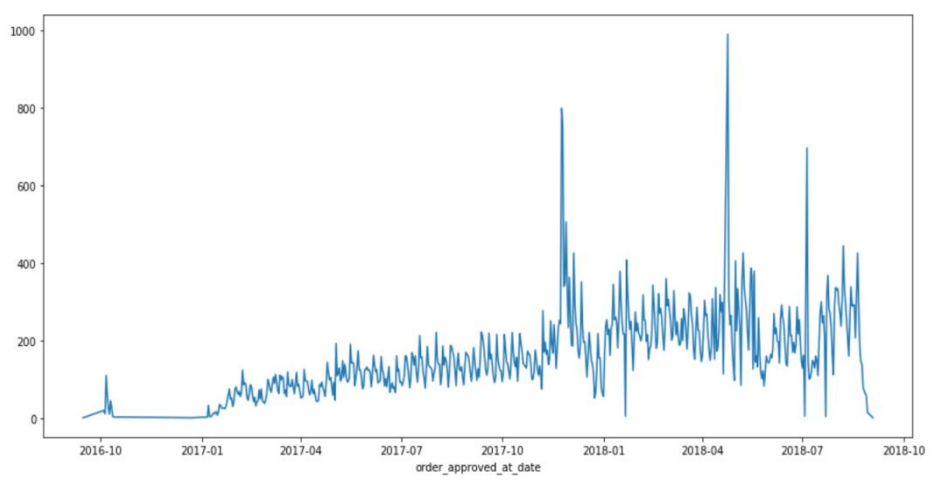
Comparer les résultats grâce à l'ARI *

ARI * : Indice de Rand Ajusté, sa valeur comprise entre [0,1],
Si proche de 0 c'est un clustering aléatoire et proche de 1 clustering parfait.

Etudes de stabilité :

Comparaison des résultats ARI :





Le score se dégrade dans le temps et s'accélère à partir de M+6

-> Proposition de MAJ des données tous les 6 mois avant les pics d'activité

Conclusion :

- BDD complètes avec beaucoup de possibilités d'analyse
- Choix des variables influe sur les clusters
- Segmentation à 7 clusters avec des « personnas marketing » identifiés
- Mise à jour des « clusters » tous les 6 mois
- Axes d'améliorations possibles :
    - identifier les poids des variables dans k-means
    - Interface pour rendre les utilisateurs autonomes

Merci pour votre attention