

Segmenter les clients d'un site de e-commerce

Soutenance de projet

olist

Nicolas FAUCONNIER
Parcours Ingénieur ML
25/03/2022

Plan

1.

Problématique

Projet et missions

2.

Data Cleaning

Nettoyage et agrégation des données à l'individu

3.

Analyse exploratoire

Analyse du dataset original et de l'agrégation à l'individu

4.

Modélisation

Différents algorithmes essayés et analyse des clusters

5.

Simulation

Estimation du délai de maintenance du modèle



1.

Problématique

Projet et missions

1. Problématique

Contexte

Mission de consulting pour la plateforme Olist, qui offre une solution de vente sur les marketplaces au Brésil. Olist souhaite segmenter les clients de sa base de données, et utiliser les résultats pour mener des campagnes marketing.

Missions


- Comprendre les différents types de clients d'Olist
 - Fournir des clusters de clients similaires obtenus avec du machine learning non supervisé, devant être actionnables par l'équipe marketing
 - Estimer le délai de réentraînement du modèle de segmentation choisi
-



2.

Data Cleaning

Nettoyage et agrégation des données à l'individu



Dataset

Le dataset est un extrait de la BDD des commandes anonymisées.

Le dataset spécifiquement utilisé ici est hébergé sur Kaggle au lien suivant:

<https://www.kaggle.com/olistbr/brazilian-ecommerce>

kaggle

9 .csv pour un total de 126Mb

Outils utilisés:

colab



Jointure et data cleaning

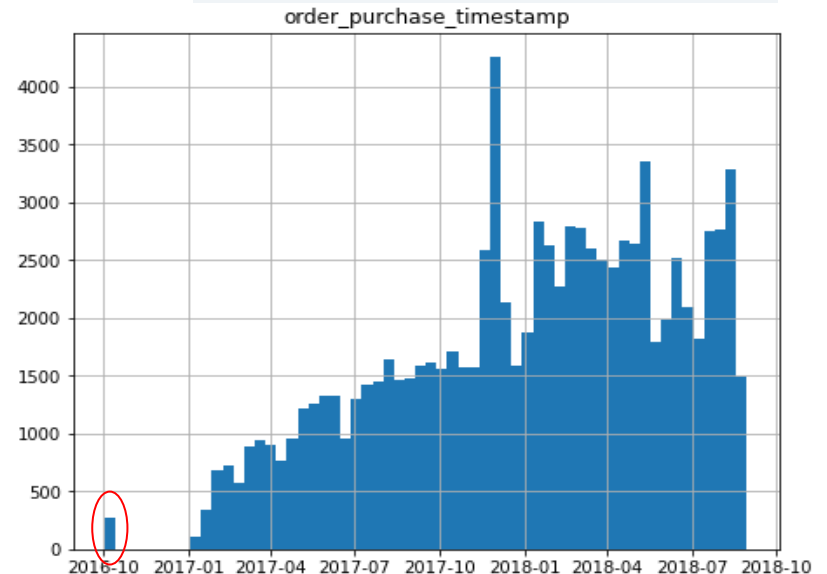
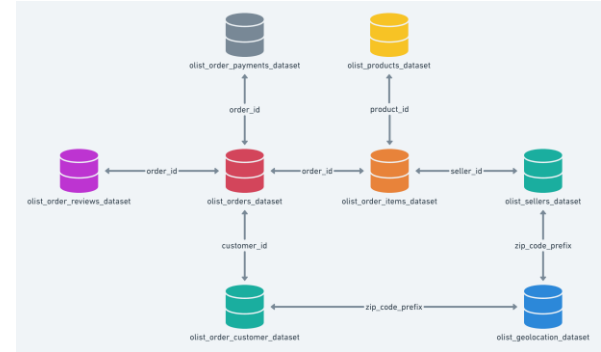
Jointure des différents fichiers

Suppression des observations sans données
de localisation et date de livraison

Suppression des duplicats de
Payment_sequential

Suppression des commandes avant 2017

→ Il reste 94002 commandes uniques



Aggrégation à l'individu (sur une période choisie)

Variables créées :

number_orders	average_product_length_cm
total_amount	average_product_height_cm
average_cart	average_product_width_cm
average_price	average_distance
average_qty	customer_state
average_review_score	order_different_state
number_reviews	first_payment_type
review_rate	first_product_category
average_review_lenght	average_product_name_lenght
average_payment_installments	average_product_description_lenght
average_freight_value	average_product_photos_qty
average_product_weight_g	days_since_last_order



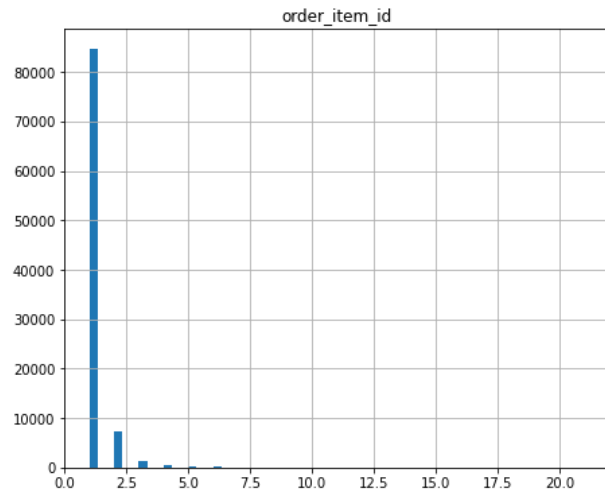
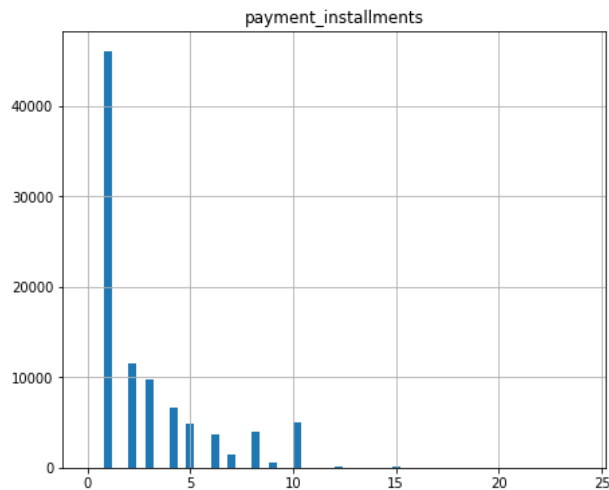
3.

Analyse exploratoire

Analyse du dataset original et de l'agrégation à l'individu

Analyse des commandes

- 88% des commandes contiennent un seul produit
- 98% des commandes contiennent une seule famille de produits
- Le paiement par carte de crédit est le plus commun
- Écrasante majorité des ventes partant de l'état de São Paulo
- La majorité des variables ont une distribution log-normale:

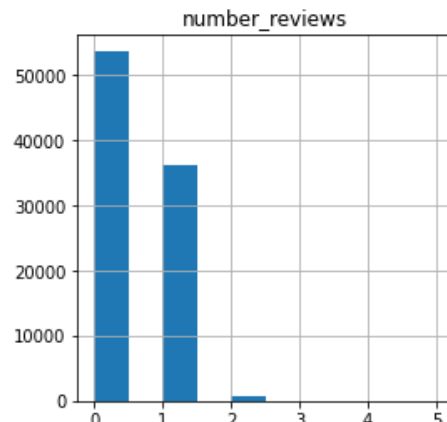
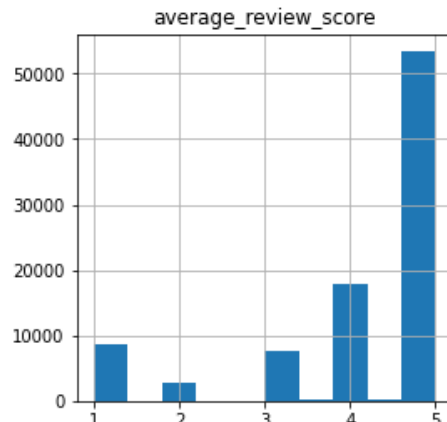
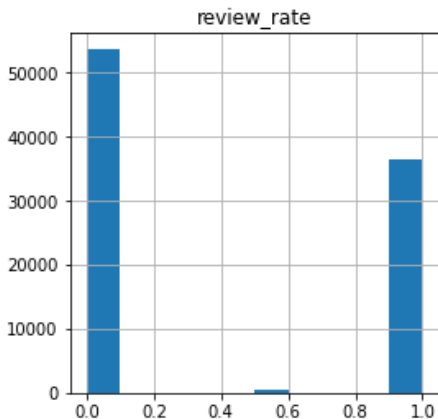
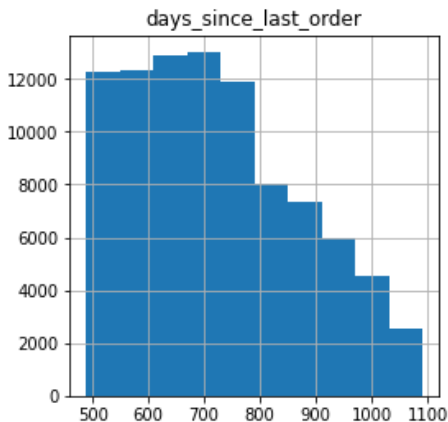


Analyse des données agrégées à l'individu

3% des individus ont effectué plus d'une commande

Distribution similaire de la plupart des variables après agrégation

2/3 des commandes sont passées uniquement auprès de vendeurs situés dans le même état





4.

Modélisation

Différents algorithmes essayés et analyse des clusters



Algorithmes et Features

Algorithmes:

- Kmeans
- DBScan
- Clustering agglomératif

Features:

- Récence: nombre de jours depuis la dernière commande
- Fréquence: nombre d'achats sur la période
- Montant: montant total dépensé sur la période
- Moyenne des notes sur la période
- *(Bonus: distance moyenne avec les vendeurs)*
- *(Bonus: nombre de paiements moyen)*

→ StandardScaler() sur les données en entrée

Alogrithmes non retenus

DBScan

Hyperparamètres testés:

- eps: 0.2 à 0.8
- min_samples: 50 à 400

→ Valeurs maximisant le score silhouette 0.5 et 100

Certains individus ne sont pas assignés à un cluster

Ne segmente que sur la note moyenne

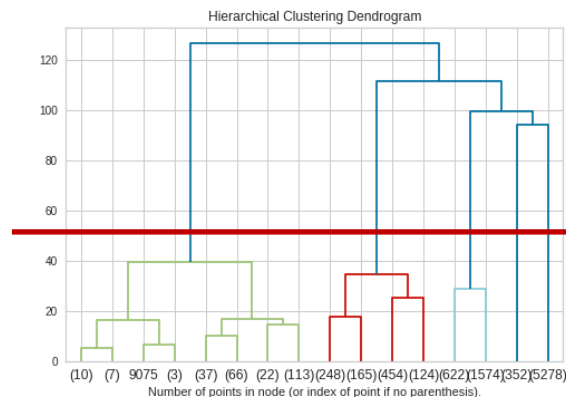
→ **6 clusters sans intérêt métier**

Clustering agglomératif

Entraînement sur 10% du dataset; stratifié sur le nombre d'achats

5 clusters:

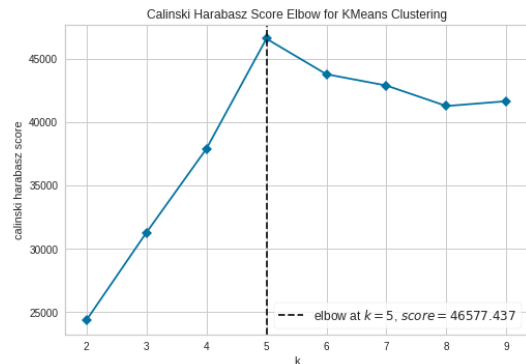
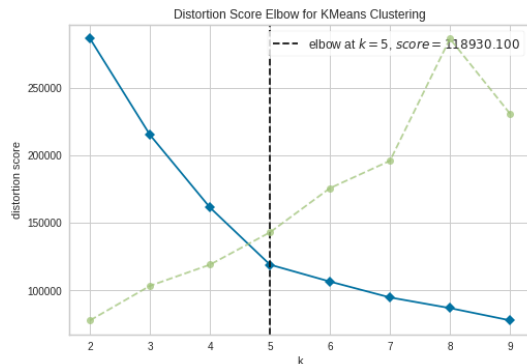
- Acheteurs multiples
- Acheteurs anciens
- Clients dépensiers
- Clients notant négativement (~1/5)
- Clients « moyens » sans particularités



Algorithme retenu: KMeans

5 clusters: *nombre déterminé par la méthode du coude*

- Acheteurs plus « **anciens** »
- Clients « **mécontents** »
laissant des reviews négatives et longues
- Clients « **engagés** »: achètent plusieurs fois, et laissent plus souvent des reviews
- Clients « **nouveaux** »: acheteurs plus récents, étalant moins leurs paiements
- Clients « **dépensiers** », qui achètent de gros produits chers et mieux décrits, qui étalent leurs paiements



Analyse du profil type de chaque cluster

0. Dépensiers

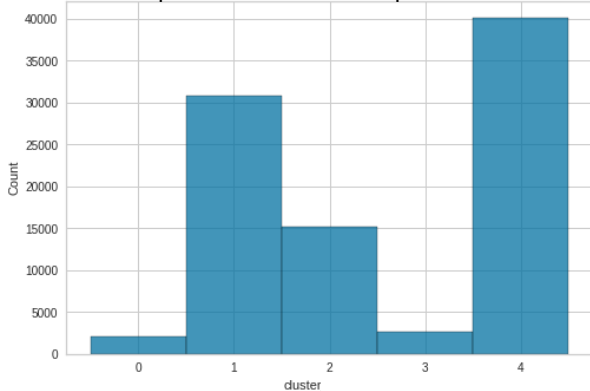
1. Anciens

2. Mécontents

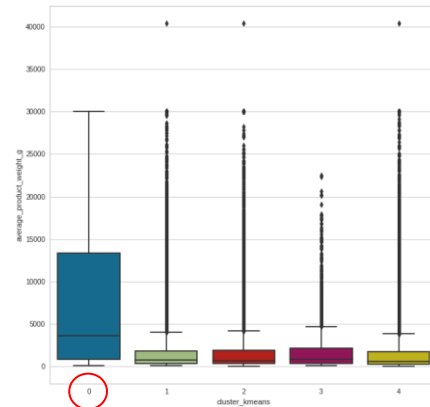
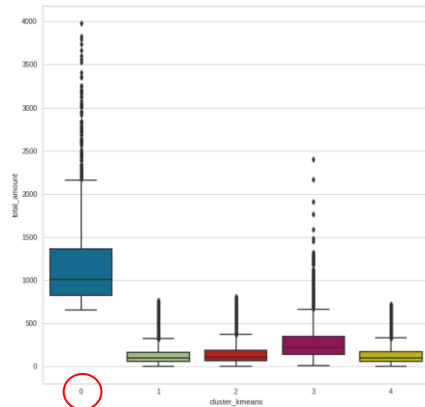
3. Engagés

4. Nouveaux

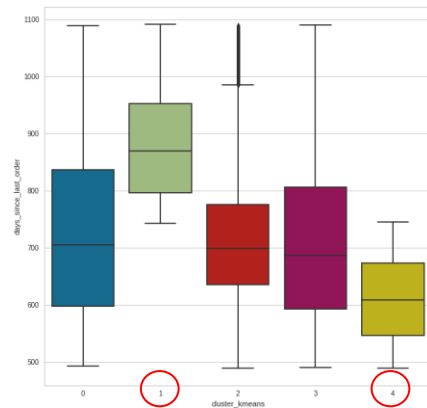
Répartition des individus par cluster



Les **0. Dépensiers** effectuent des commandes plus chères, comportant des produits plus volumineux et lourds:

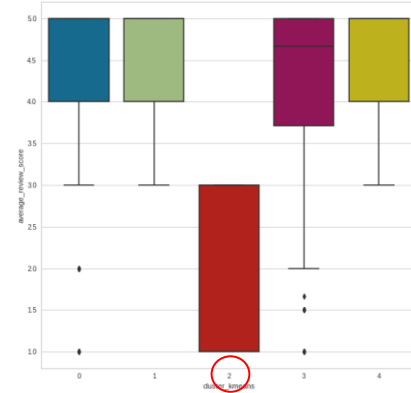
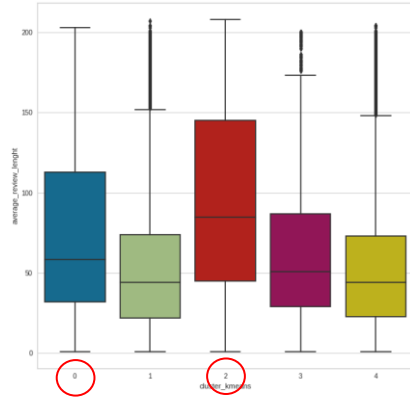


Les **1. Anciens** et **4. Nouveaux** se distinguent par le nombre de jours depuis leurs derniers achats:

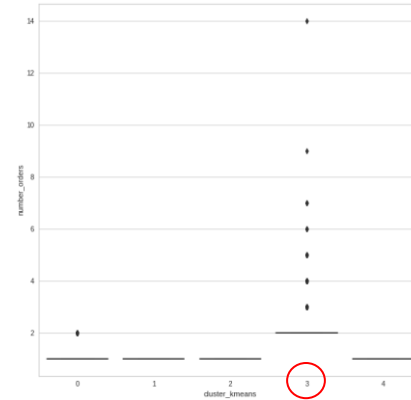
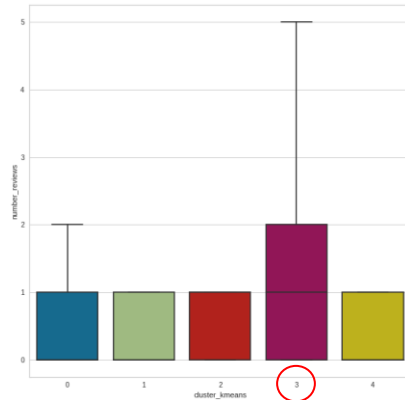


Analyse du profil type de chaque cluster

Les **2.Mécontents** laissent des reviews plus longues (similairement aux **0.Dépensiers**) avec de mauvaises notes



Les **3.Engagés** laissent plus de reviews et achètent plus souvent



“Bonus”: Kmeans avec plus de features

Ajout de la distance moyenne avec le vendeur et le nombre de paiements

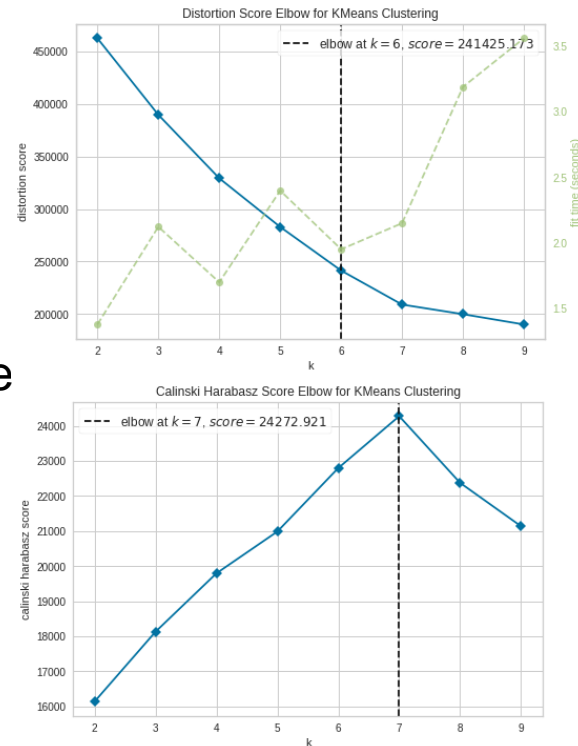
7 clusters:

Nombre de clusters plus difficile à déterminer avec la méthode du coude

Profils types plus fins et moins définis par une seule feature

Clusters se chevauchent davantage sur certaines features importantes d'un point de vue métier (eg. dépense totale)

→ Clusters moins bien séparés



5. Simulation

Estimation du délai de maintenance du modèle

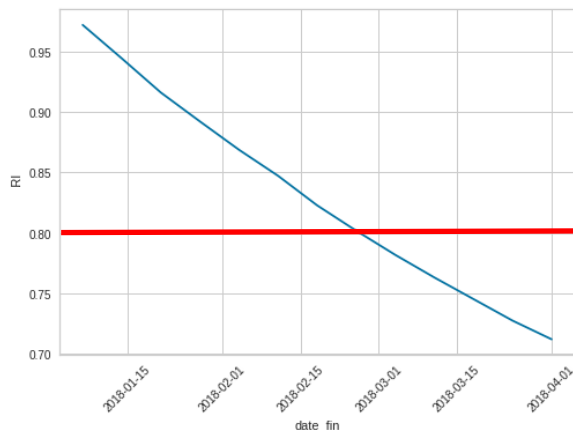
Fréquence de mise-à-jour de la ségmentation

Méthode:

- Agrégation des commandes à l'individu sur une période **T0**
- Agrégation des commandes à l'individu sur **Tn = T0 + (n * 7 jours)**
- **StandardScaler()** entraîné sur **T0** et appliqué sur **T0** et chaque **Tn**
- Calcul du Rand Index sur les clusters des individus de **T0** vs. leurs cluster dans **Tn**

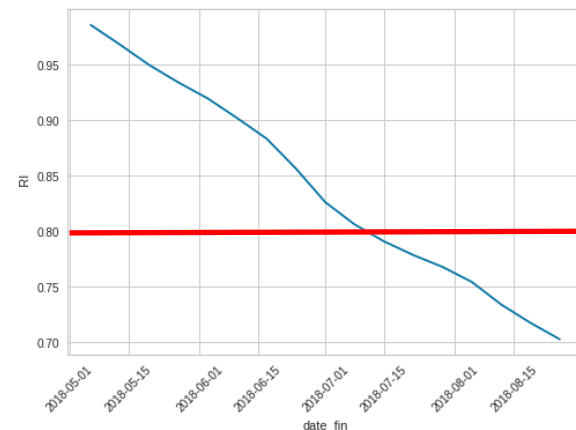
Essai 1:

T0 = année 2017



Essai 2:

T0 = janvier
2017 à avril
2018



→ Fréquence de mise-à-jour suggérée: 2 mois

Pistes d'amélioration

- Obtenir plus d'observations (eg. données plus récentes)
- Obtenir plus de features (notamment sur les clients et produits)
- Traiter et utiliser les verbatims
- Evaluer la pertinence du machine learning vs. segmentation basée sur des connaissances métier

Merci

Avez-vous des questions?
