Segmentez des clients d'un site e-commerce

Charlotte DUBUS

Soutenance du 29 mars 2024 - Parcours Data Scientist - Projet 5

olist

MISSIONS

Support Initial pour le Dashboard Customer Experience

Aider à l'implémentation de requêtes SQL **urgentes**

Segmentation des Clients pour l'Équipe E-commerce

Comprendre les différents types d'utilisateurs

Fournir une description actionable des segments clients

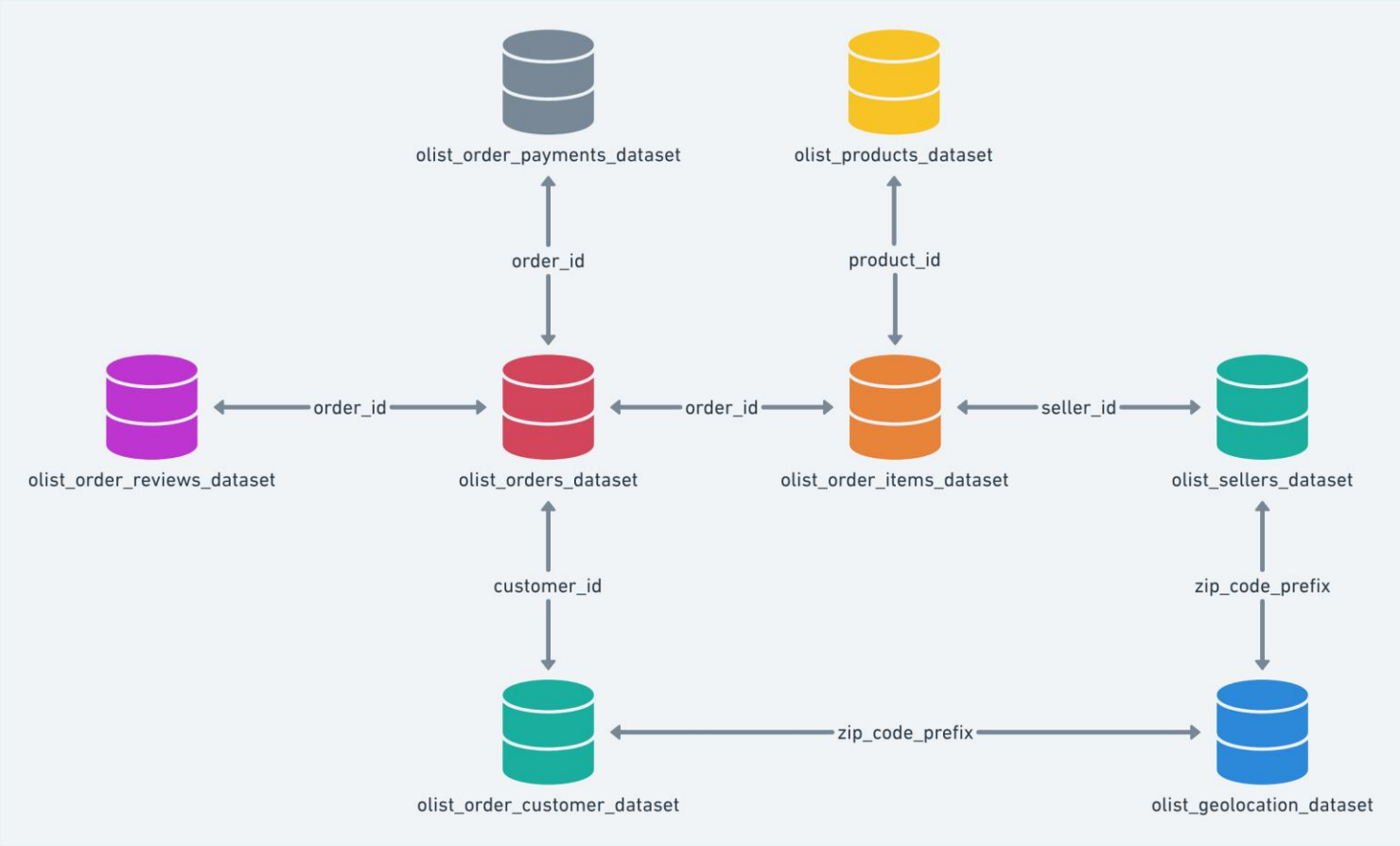
Proposition de Contrat de Maintenance

Analyser la stabilité des segments au cours du temps.

Recommander une fréquence de mise à jour

SOMMAIRE

Analyse et préparation des données

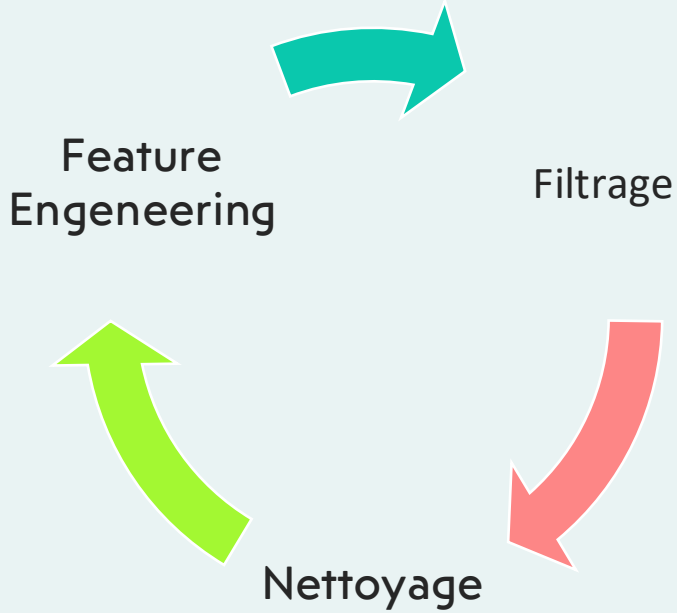


9 bases de données

Quelles informations représentent un CHOIX client ?

Création d'un Dataframe de 16 colonnes et plus de 100 000 données

- **Déterminer le cadre temporelle des données**




```
order_id
order_item_id
product_id
seller_id
shipping_limit_date
price
freight_value
product_category
payment_type
payment_installments
payment_value
review_score
order_status
order_purchase_timestamp
customer_id
customer_unique_id
```

- **Création d'un Dataframe : df_analyse**

Analyse temporelle

- Tendance des achats
- Récence des achats
- Moyenne des durée de paiements
- Evolution des scores de satisfaction dans le temps

Analyse comportementale

- Répartition de la fréquence des achats
- Distribution des dépenses des clients
- Relation bivariée entre les variables
- Recherche de correlation

Architecture des segments (Clustering)

	average_installments	average_review_score	recency	frequency	monetary	most_frequent_category
0	1.0	4.0	699	1	39.09	sports_leisure
1	1.0	3.0	699	1	53.73	sports_leisure
2	6.0	1.0	699	1	133.46	furniture_decor
3	4.0	5.0	699	1	40.95	fashion_shoes
4	2.0	5.0	699	1	154.57	toys
...	...	...	...	...	...	...
92749	7.0	1.0	4	1	73.10	toys
92750	8.0	5.0	4	1	510.96	kitchen_dining_laundry_garden_furniture
92751	1.0	5.0	4	1	61.29	health_beauty
92752	1.0	3.0	4	1	33.23	party_supplies
92753	1.0	5.0	4	1	93.75	computers_accessories

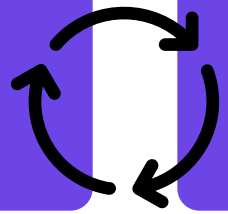
92754 rows × 6 columns

Sélection des features → Transformation → Normalisation et OneHotEncoder → Création d'un échantillon

Méthodologie

4 essaies de clustering

RFM



+ 1 feature complémentaire

- Kmeans
- Clustering Hiérarchique
- DBSCAN

→ Evaluation des modèles

Modèles envisagés

Kmeans

- Méthode de clustering partitionnelle qui divise un ensemble de données en K clusters distincts. Chaque point est attribué au cluster dont le centroïdes est le plus proche.

Clustering Hiérarchique

- Méthode qui construit soit de manière agglomérative (fusionnant progressivement les clusters), soit de manière divisive (séparant progressivement les clusters).

DBSCAN

- Méthode de clustering basée sur la densité qui identifie les 'clusters' comme des régions de haute densité séparées par des régions de faible densité.

Métriques d'évaluation

La méthode du coude

- Technique visuelle pour déterminer le nombre optimal de clusters

Silhouette Score

- Mesure la cohésion et la séparation des clusters.
- Les valeurs vont de –1 à 1. Une valeur élevée indique que les points sont bien adaptés à leur propre cluster.

Davies-Bouldin Index

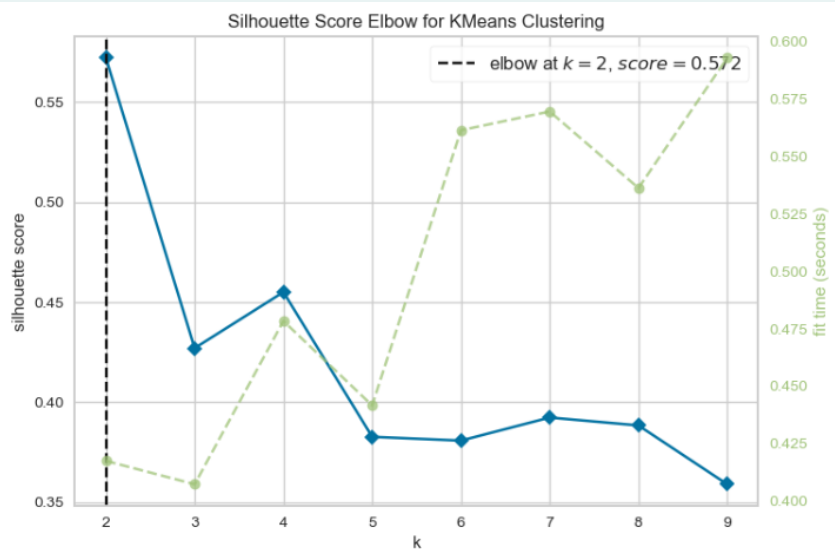
- Elle évalue le ratio entre la dispersion au sein des clusters et la séparation entre eux.
- Métrique où des valeurs plus faibles indique une meilleures séparation des clusters

Calinski-Harabasz Index

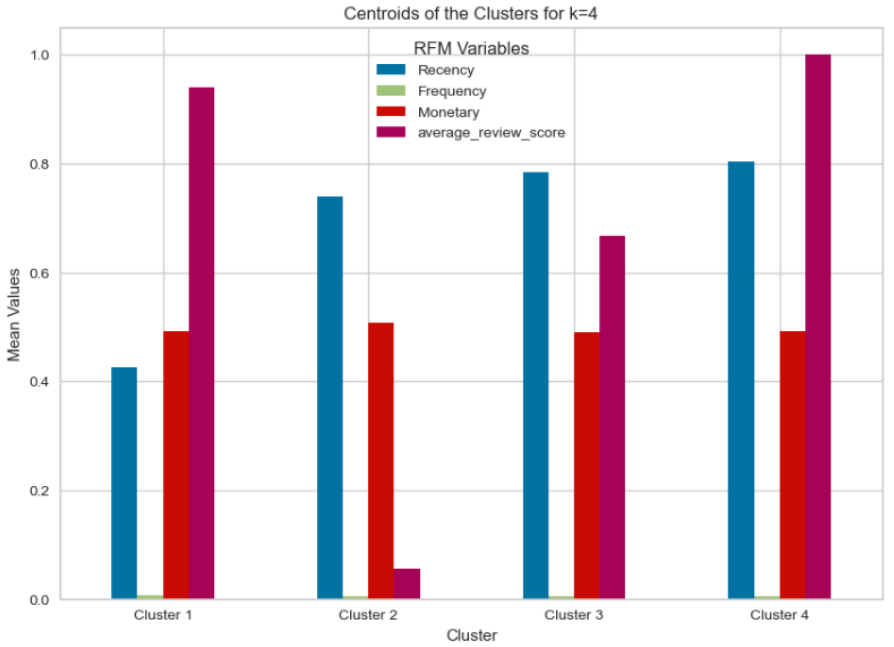
- Score qui mesure la dispersion entre les clusters par rapport à la dispersion à l'intérieur des clusters.
- Des valeurs plus élevées indiquent des clusters plus denses et mieux séparés

- L'ajoute d'average_score_review a amélioré les scores Kmeans
- Option intéressante : 4 clusters (pertinence métiers)
- Essai RFMS = Meilleures résultats métriques avec ces éléments
- **Features : RFMS**
- **Modèle : Kmeans**

Résultat : Choix du Modèle



	Method	Davies-Bouldin	Calinski-Harabasz	Silhouette	Execution Time (seconds)
0	KMeans avec k4	0.798899	23672.129721	0.454959	0.410159



Résultat : Segmentation client

Clients occasionnels satisfaits

- Satisfaits
- Achat récent
- Fréquence faible
- Petit montant

Clients perdus/à reconquérir

- Insatisfaits
- Achat récent
- Fréquence faible
- Montants faibles à élevés

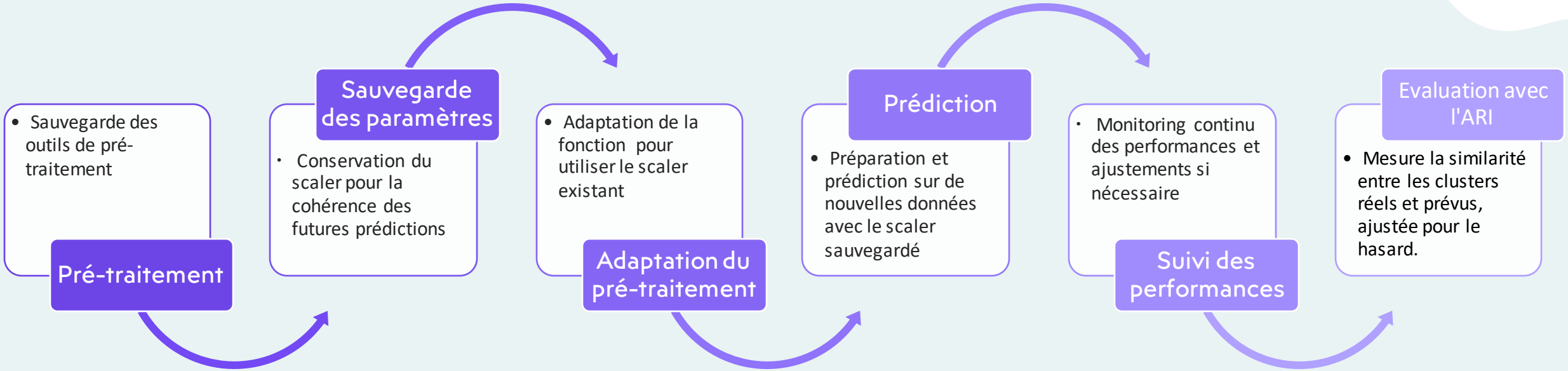
Clients potentiels

- Satisfaction modérée
- Récence moyenne
- Engagement d'achat
- Montants faibles à moyens

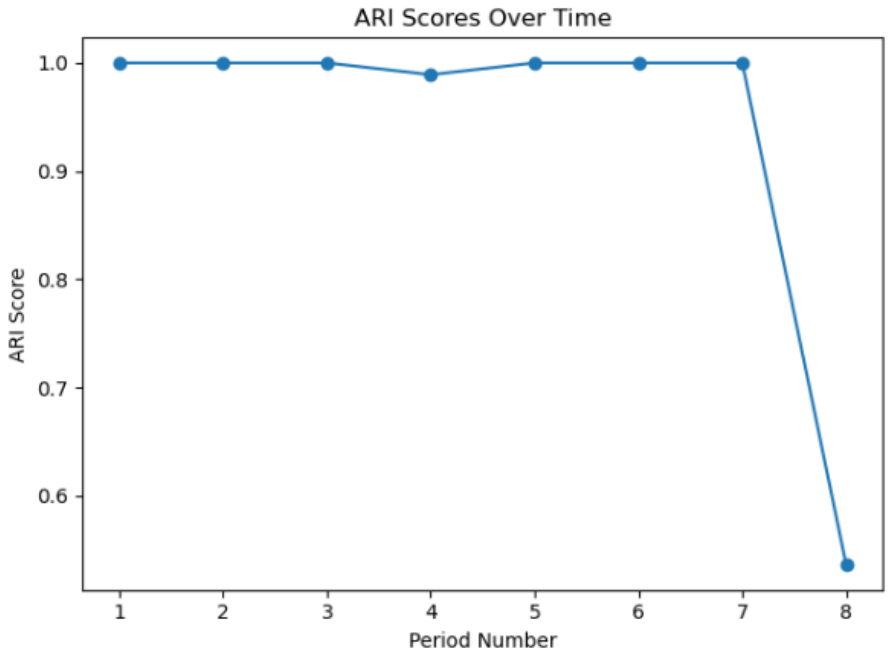
Clients ambassadeurs

- Satisfaits
- Achat très récent
- Fréquence plus élevée
- Montants faibles à élevés

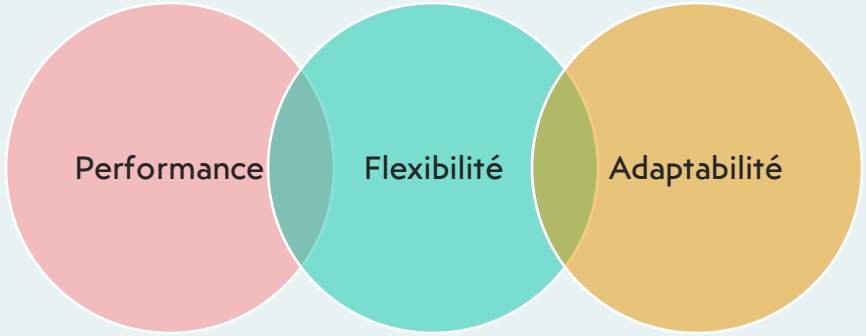
Contrat de maintenance : étapes clés



Contrat de maintenance : Proposition



- Evaluation Trimestrielle des score ARI
- Réévaluation semestrielle du modèle
- Ajustement annuel



Conclusion et Recommandations

Clustering	Identification de segments clients pertinents avec RFMS et Kmeans	Surveillance continue et intégration régulière de nouvelles données pour affiner les segments.
Maintenance	Maintenance régulière pour préserver la pertinence des segments	Révision trimestrielle basée sur les scores ARI, puis semestrielle du modèle et ajustement annuel

Merci pour votre attention !