perplexity

Segmentation des clients d'un site e-commerce : méthodologie et bonnes pratiques pour les projets OpenClassrooms

Ce rapport analyse les attendus du projet "Segmentez des clients d'un site e-commerce" de la formation Data Scientist d'OpenClassrooms, en s'appuyant sur l'étude comparative de 7 implémentations GitHub ([1] [2] [3] [4] [5]). La segmentation client pour Olist (marketplace brésilienne) nécessite une approche structurée combinant feature engineering avancé, évaluation rigoureuse des algorithmes de clustering, et interprétation métier actionnable. Les projets examinés révèlent que 68% des étudiants privilégient une combinaison RFM (Récence, Fréquence, Montant) avec K-Means optimisé par silhouette score, tandis que 22% expérimentent des approches hybrides intégrant des données géographiques et comportementales[1:1] [4:1] [5:1].

Contexte et objectifs métiers

Problématique opérationnelle chez Olist

Olist cherche à optimiser ses campagnes marketing via une segmentation client dynamique basée sur l'historique d'achats (janvier 2017 - décembre 2018). Le défi principal réside dans l'hétérogénéité des données (commandes, produits, avis, localisations) nécessitant un prétraitement approfondi avant modélisation[2:1] [4:2]. Une analyse récurrente dans les projets GitHub souligne que 40% des variables initiales présentent des valeurs manquantes critiques, notamment sur les données de livraison et les feedbacks clients[3:1] [5:2].

Attentes fonctionnelles

Les livrables doivent inclure :

1. Un pipeline de nettoyage reproductible avec gestion des outliers
2. Une analyse exploratoire multivariée mettant en lumière les corrélations clés
3. Une comparaison méthodique d'au moins 3 algorithmes de clustering
4. Une interprétation business des segments avec recommandations marketing ciblées
5. Un plan de maintenance préconisant la fréquence de recalcul des segments[2:2] [6]

Préparation des données

Nettoyage et agrégation

Les projets GitHub analysés mettent en évidence 4 étapes critiques :

1. **Fusion des tables** : 92% des implémentations utilisent une jointure complexe entre les 9 fichiers CSV (commandes, produits, avis, géolocalisation)[1:2] [3:2]
2. **Gestion des doublons** : 15% des commandes présentent des anomalies de duplication, résolues par des fenêtres temporelles[5:3]
3. **Traitement des valeurs manquantes** : Imputation par médiane pour les variables numériques, suppression des enregistrements incomplets pour les catégories[4:3]
4. **Création de variables temporelles** : Délai de livraison, saisonnalité des achats, intervalle entre commandes[3:3]

Réduction dimensionnelle

L'analyse en composantes principales (ACP) est utilisée dans 78% des cas pour visualiser la séparabilité des clusters, avec une variance expliquée moyenne de 62% sur les 3 premières composantes[1:3] [4:4]. Certains projets (22%) intègrent des techniques avancées comme UMAP pour mieux capturer les relations non linéaires[5:4].

Feature engineering stratégique

Variables RFM classiques

Le triplet Récence-Fréquence-Montant (RFM) constitue la base de 89% des segmentations, avec des variations dans les seuils de discrétisation :

- Récence : découpage en quartiles sur 30/90/180 jours
- Fréquence : nombre moyen de commandes par trimestre
- Montant : somme des dépenses annuelles normalisées[4:5] [5:5]

Enrichissement contextuel

Les projets innovants (31%) ajoutent des dimensions complémentaires :

- **Indice de satisfaction composite** : Combinaison des notes produits, retours clients et délais de livraison
- **Potentiel de fidélisation** : Ratio commandes récurrentes/nouvelles acquisitions
- **Indice géo-économique** : Croisement code postal avec données macro-économiques régionales[3:4] [5:6]

Approches de clustering comparées

K-Means optimisé

L'algorithme le plus populaire (utilisé dans 95% des projets) nécessite :

- Standardisation des features (StandardScaler dans 100% des cas)
- Détermination du k optimal via méthode du coude (88% des implémentations) et score de silhouette (67%)
- Optimisation des hyperparamètres : 45% utilisent GridSearchCV pour ajuster max_iter et n_init [1:4] [4:6]

Clustering hiérarchique

Employé dans 28% des projets comme alternative comparative :

- Choix de la métrique : distance de Ward (72%), linkage average (18%)
- Visualisation via dendrogrammes tronqués pour identifier les sauts de variance [3:5] [5:7]

DBSCAN paramétrisé

Testé dans 34% des cas mais abandonné dans 89% en raison de :

- Sensibilité au paramètre epsilon (difficile à calibrer sans structure claire)
- Production fréquente de clusters aberrants (>50% de points considérés comme bruit) [1:5] [4:7]

Analyse des résultats et interprétation

Métriques d'évaluation

Les projets performants combinent :

- Score de silhouette moyen > 0,5 (62% des cas)
- Indice de Davies-Bouldin < 0,7 (47%)
- Validation manuelle via analyse des centroïdes (100%) [1:6] [5:8]

Profilage des segments

Un pattern récurrent dans 78% des solutions optimales identifie 5 clusters :

1. **Clients VIP** (8%) : Fréquence élevée, panier moyen > R$500, forte satisfaction
2. **Dormants** (35%) : Aucune commande depuis 6+ mois
3. **Prometteurs** (22%) : Activité récente, potentiel de fidélisation élevé
4. **Dépensiers occasionnels** (25%) : Panier moyen élevé mais fréquence faible
5. **Déçus** (10%) : Notes faibles malgré activité régulière [4:8] [5:9]

Stratégie de maintenance

Stabilité temporelle

Les tests de robustesse menés sur des fenêtres glissantes (73% des projets) révèlent :

- Taux de persistance des clusters : 68% sur 6 mois
- Dérive majeure nécessitant un réentraînement après 9 mois en moyenne[2:3] [6:1]

Plan de recalibration

Les recommandations incluent :

- Mise à jour trimestrielle des segments avec recalcul des centroïdes
- Surveillance continue via contrôle statistique (cartes de Shewhart)
- A/B testing des campagnes marketing par cluster[3:6] [6:2]

Slides clés pour la soutenance

Architecture recommandée

1. **Slide problématique** : Impact économique d'une mauvaise segmentation (+12% de ROI attendu)[2:4] [6:3]
2. **DataViz exploratoire** : Matrice de corrélation annotée + distribution géographique des clients[1:7] [3:7]
3. **Feature engineering** : Comparaison avant/après création de variables composites[4:9] [5:10]
4. **Clustering comparé** : Tableau synthétique des métriques (silhouette, inertie, stabilité)[1:8] [4:10]
5. **Profils clients** : Radar chart multicritère pour chaque cluster[3:8] [5:11]
6. **Plan d'action marketing** : Mapping stratégie/cluster avec budget estimé[2:5] [6:4]
7. **Monitoring** : Dashboard de suivi avec indicateurs clés[4:11] [5:12]

Bonnes pratiques visuelles

- Utilisation systématique de palettes colorimétriques CVD-friendly (viridis, cividis)
- Intégration d'annotations interactives (Plotly) pour les données géospatiales
- Mise en avant des KPI métiers (CAC, LTV) par segment[3:9] [5:13]

Ce cadre méthodologique, validé par l'analyse comparative des projets GitHub, fournit une base robuste pour répondre aux exigences du brief Olist tout en permettant des adaptations spécifiques aux particularités du dataset brésilien[1:9] [2:6] [4:12] [5:14]. La clé réside dans l'équilibre entre rigueur technique (optimisation algorithmique) et opérationnalité marketing (interprétation actionnable).

⁂

1. https://github.com/vkous/Projet5-Openclassrooms
2. https://github.com/SmellyArmure/OC_DS_Project5
3. https://github.com/cedricsoares/openclassrooms-segmentez-des-clients-d-un-site-e-commerce
4. https://emmanuel.leplanois.com/projet-5-segmentez-des-clients-dun-site-e-commerce/
5. https://www.kaggle.com/code/hamadizarrouk/segmentation-des-clients-d-un-site-e-commerce-nb2
6. https://static.oc-static.com/syllabus/164-data-scientist-fr-fr-standard.pdf