Analyse de Données E-Commerce et Segmentation Client

Projet Final - Cloud Computing & Data Analytics

Tom Le Corre

Université Paris 1 Panthéon-Sorbonne
DU Data Analytics

18 février 2026

1. Contexte et Objectifs du Projet

Nous avons analysé un jeu de données massif issu d'une plateforme de e-commerce britannique pour optimiser la stratégie marketing.

Le Jeu de Données :

- **Source :** Transactions "Online Retail".
- **Volume Initial : 541 909 lignes**.
- **Technologie :** Traitement distribué avec **PySpark** (API DataFrame).

Les 3 Objectifs :

1. **Nettoyer** : Éliminer les données inexploitables (25% du volume).
2. **Segmenter** : Créer des groupes clients homogènes (RFM).
3. **Prédire** : Identifier les "Gros Dépensiers" via le Machine Learning.

2. Nettoyage et Fiabilisation (Data Cleaning)

L'analyse exploratoire a révélé des incohérences majeures nécessitant un filtrage strict avant toute modélisation.

Étape de Nettoyage	Action PySpark	Impact
1. Données Brutes	*Chargement CSV*	541 909 lignes
2. Clients Manquants	`dropna(subset="CustomerID")`	- 135 080 lignes
3. Annulations	`filter(Quantity > 0)`	- 8 905 lignes
DONNÉES FINALES	**Dataset Qualifié**	**397 924 lignes**

Transformation :

- Conversion de `InvoiceDate` en format Temporel.
- Création de la variable `TotalPrice` (Quantité $\times$ PrixUnit).

3. Segmentation : L'Approche RFM

Nous sommes passés d'une analyse par transaction à une analyse par **Client Unique** (4 339 clients identifiés).

Les 3 Dimensions Marketing

- **R - Récence :** Nombre de jours depuis le dernier achat.
- **F - Fréquence :** Nombre total de commandes validées.
- **M - Montant :** Somme totale dépensée (Chiffre d'Affaires).

Préparation Technique (Crucial) :

- Usage de `VectorAssembler` pour fusionner les colonnes.
- Application de `StandardScaler` pour normaliser les écarts d'échelle (Jours vs Euros).

4. Analyse des Segments (Bisecting KMeans)

L'algorithme a identifié **3 profils types** au sein de la clientèle.

Cluster	Récence Moy.	Fréquence Moy.	Montant Moy.	Profil Business
Groupe 0	**246 Jours**	1.1	280 €	**Clients Perdus / Inactifs**
(Majorité)	*Achat très ancien*	*Achat unique*	*Faible valeur*	*Besoin de relance*
Groupe 1	**42 Jours**	3.8	1 350 €	**Clients Réguliers**
(Cœur de cible)	*Achat récent*	*Régulier*	*Bon panier*	*À fidéliser*
Groupe 2	**6 Jours**	**19.5**	**10 500 €**	**VIP / Champions**
(Elite)	*Très Actifs*	*Intensifs*	*Très haute valeur*	*Traitement Premium*

5. Visualisation Stratégique des Segments

L'analyse des centroïdes nous permet de définir les actions prioritaires :

Les Perdus (Groupe 0)

- **Diagnostic :** Risque de Churn maximal.
- **Action :** Email de "Win-Back" avec promotion agressive (-20%).

Les Réguliers (Groupe 1)

- **Diagnostic :** Base stable mais potentiel inexploité.
- **Action :** Cross-selling (proposer des produits complémentaires).

Les VIP (Groupe 2)

- **Diagnostic :** 1% des clients font 30% du CA.
- **Action :** Programme de fidélité exclusif & frais de port offerts.

6. Prédiction Supervisée (Logistic Regression)

Objectif : Créer un modèle capable d'identifier automatiquement les clients à fort potentiel.

Configuration du Modèle

- **Cible (Label) : 1** si le client est un "Gros Dépensier" (Seuil > 500€), sinon **0**.
- **Features :** Variables RFM + Pays (encodé).
- **Protocole :** Séparation 70% Train (Apprentissage) / 30% Test (Validation).

Pourquoi la Régression Logistique ?

- Algorithme rapide sur les grands volumes de données (Big Data).
- Résultats facilement interprétables (probabilité d'appartenance à la classe).

7. Performance du Modèle

Résultats obtenus sur le jeu de test (données jamais vues par le modèle) :

Métrique	Résultat	Interprétation
Accuracy	**98.2 %**	Taux global de bonnes prédictions très élevé.
Précision	97.5 %	Peu de faux positifs (erreurs sur les VIP).
Rappel	95.0 %	Le modèle détecte la quasi-totalité des VIP.
F1-Score	**0.96**	Excellent équilibre global.

Conclusion Technique : Le modèle est fiable et robuste. Il peut être déployé en production pour scorer les nouveaux inscrits en temps réel.

8. Conclusion et Recommandations Finales

Bilan du Projet :

- **Technique :** Maîtrise de la chaîne PySpark complète (ETL $\rightarrow$ MLlib).
- **Business :** Segmentation claire en 3 groupes actionnables.

Pistes d'Amélioration :

- Intégrer des données démographiques (âge, genre).
- Tester des modèles non-linéaires (Random Forest) pour gagner les derniers points de précision.

Recommandation Clé

Prioriser l'effort marketing sur le **Groupe 0 (Les Perdus)** qui représente le plus gros volume de clients mais génère le moins de valeur actuelle.

Merci de votre attention.

Avez-vous des questions ?