FTML 2025 - Exercice 7

Segmentation Non-Supervisée des Clients de Cartes de Crédit

Marc GUILLEMOT, Emre ULUSOY, Rayan DRISSI, Gabriel MONTEILLARD

July 5, 2025

Objectif

Réaliser une segmentation non-supervisée des clients à partir d'un jeu de données réels, en analysant la structure latente des comportements financiers. Aucun label ne doit être utilisé pour l'entraînement des modèles de clustering. L'objectif métier est de définir des profils-clients et évaluer leur lien potentiel avec le risque de défaut.

Évaluation: Score de silhouette, analyse métier, visualisation, et taux de défaut par cluster.

Pipeline Général

1. Chargement et nettoyage des données
2. Feature engineering avancé (50+ variables dérivées)
3. Préprocessing intelligent (scaler optimal sélectionné automatiquement)
4. Recherche du nombre optimal de clusters (k)
5. Application de plusieurs algorithmes : KMeans, GMM, Hierarchical, DBSCAN
6. Réduction de dimension (UMAP/PCA)
7. Analyse métier des profils clients
8. Validation par rapport au taux de défaut
9. Visualisations finales

Choix Techniques

- **Données:** `default_of_credit_card_clients.csv`
- **Hardware:** 32 CPU cores + GPU RTX 5070Ti
- **Librarie GPU:** cuML, cuDF, cuKMeans, cuDBSCAN, cuUMAP
- **Parallélisation:** joblib

Clustering Optimisé

Méthodologie:

- Sélection automatique du meilleur scaler (RobustScaler vs StandardScaler)
- Recherche parallèle de k optimal (2 à 15), selon plusieurs métriques :
  - Silhouette
  - Calinski-Harabasz
  - Davies-Bouldin
- Consensus médian pour choix final

Algorithmes appliqués:

- **KMeans (GPU)** – Meilleur score silhouette
- **GMM** – Résultats cohérents
- **Agglomerative** – Clusters très déséquilibrés
- **DBSCAN (GPU)** – Outliers détectés, peu de structure exploitable

Réduction Dimensionnelle

- **UMAP (GPU)** utilisé pour visualisation 2D
- **PCA** en fallback

Analyse Métier

Chaque cluster a été analysé selon :

- Âge moyen
- Limite de crédit
- Utilisation du crédit
- Délai moyen de paiement
- Score de risque (indicateurs binaires)

Profils obtenus :

- **Premium**: haute limite, faible utilisation, aucun retard
- **Standard**: comportement régulier
- **Risque Modéré**: utilisation ¿ 60%, paiements irréguliers
- **Haut Risque**: utilisation ¿ 80%, retards fréquents

Validation par la Cible

- La variable `default payment next month` a été utilisée uniquement pour validation.
- Test du χ^2 : **p-value ¡ 0.05** $\rightarrow$ corrélation significative entre cluster et défaut
- Visualisation des taux de défaut par cluster

Visualisations Clés

- `clustering_results_advanced.png` : 9 graphiques (barplot, scatter UMAP, pie chart, histogrammes)
- Profils bien séparés, peu de recouvrement

Conclusion

- L'analyse non-supervisée a permis d'identifier **4 clusters significatifs**
- KMeans GPU combiné à un feature engineering robuste a permis de surpasser les approches naïves
- Résultats validés par la variable cible
- Visualisations riches pour communication métier