

# Segmentation Clients – Challenge Data Science (Jakala, 2025)

Ammar MOISE

Novembre 2025

*Analyse exploratoire, apprentissage supervisé et clustering non supervisé*

## Résumé

Le challenge Kaggle “*Customer Segmentation*” visait à classer des clients en quatre segments (A, B, C, D) à partir de données socio-démographiques et comportementales. L’objectif final était d’obtenir la meilleure **accuracy** possible sur le jeu de test. La démarche adoptée comprend : (1) une **analyse exploratoire** pour comprendre la structure et les biais des données, (2) la **construction et comparaison de modèles supervisés**, (3) et une **analyse non supervisée** par clustering (K-Prototypes) pour évaluer la cohérence des segments.

## Analyse exploratoire des données

### Structure et jeu de données

Le jeu d’entraînement contient 8068 individus et 10 variables explicatives ; le test en contient 2627. Variables mixtes :

- **Numériques** : Age, Work\_Experience, Family\_Size
- **Catégorielles** : Gender, Ever\_Married, Graduated, Profession, Spending\_Score, Var\_1

Anomalie majeure : **2332 identifiants sont présents à la fois dans le train et le test** ( 89 % du test). Cela révèle un **data leakage temporel** : mêmes clients observés à deux instants, seules Age et Work\_Experience évoluant.

**Conséquence** : risque d’overfit sur des individus déjà vus, biaisant la généralisation et la validation.

## Problèmes de qualité et valeurs manquantes

Variables les plus manquantes : `Work_Experience` (10.3%), `Family_Size` (4.2%), `Ever_Married` (1.7%). Les NaN ne sont pas aléatoires (surreprésentés chez D).

**Choix méthodologique** : conserver les NaN ; **LightGBM** les gère nativement et le *manquant* porte de l'information.

## Observations clés

Population majoritairement hommes (55 %), mariés (59 %), diplômés (62 %). Profession dominante : *Artist* (32 %). Nombreuses "Low" dépenses (60 %). Incohérences : ex. 18 ans avec 14 ans d'expérience. **Signal fort** : non-mariés → D ; mariés → C.

## Synthèse exploratoire

**Variables les plus discriminantes** : Age, Profession, `Spending_Score`, `Ever_Married`.

**Problèmes majeurs** : fuite temporelle, incohérences logiques, déséquilibres latents.

**Hypothèse** : segmentation basée sur des règles métier non purement géométriques.

## Modélisation comparative

### Pourquoi ces modèles ?

- **LightGBM** : robuste sur données hétérogènes, gère les NaN, bon ratio biais/variance.
- **Régression Logistique** : baseline linéaire interprétable (référence minimale).
- **SVM (RBF)** : frontière non linéaire pour capter des interactions.
- **Optuna** : réglage automatique pour limiter le surapprentissage.
- **Voting Ensemble** : agrégation pour réduire la variance (stabilité).
- **Lookback (très bref)** : si un ID du test existe dans le train, prédire le *segment majoritaire historique* de cet ID ; sinon, utiliser LightGBM. Des garde-fous simples (cohérence d'âge/expérience, seuil de confiance) évitent les erreurs temporelles.

## Préparation et stratégie

Split 80/20 stratifié (train/validation). Évaluation : **accuracy** et **CV 5-fold**.

## Comparaison des performances

Modèle	Train	Validation	Kaggle (test)	Commentaire
LGBM Simple	0.72	0.53	0.328	Bon point de départ
LGBM Optimisé	0.61	0.53	0.327	Gain de stabilité
LGBM FeatureSel	0.56	0.55	<b>0.337</b>	Meilleur compromis
LGBM Lookback	–	–	0.300	Récurrence ID peu utile ici
Régression Logistique	0.52	0.53	0.316	Baseline
SVM RBF	0.66	0.53	0.318	Non linéaire, peu de gain
Voting Ensemble	0.61	<b>0.56</b>	0.328	Plus équilibré
Hybrid	–	–	0.310	Combinaison sans gain

**Observation** : les scores test [0.30, 0.34] restent **supérieurs au hasard** (0.25) mais **faibles**. Ils indiquent que les tendances générales sont apprises, tandis que le bruit, la fuite temporelle et le chevauchement B/C limitent la généralisation.

## Analyse

Les modèles à arbres surpassent les linéaires, mais aucun ne dépasse 0.34 sur le test Kaggle. **Écart validation (0.56) vs test (0.33)** : probables effets combinés de fuite temporelle, distribution test différente et segmentation potentiellement instable. De meilleures *features* (règles métier, interactions, temporalité explicite) pourraient améliorer l'accuracy.

## Clustering non supervisé (K-Prototypes)

### Objectif et principe

K-Prototypes cherche des regroupements naturels sur données mixtes (euclidienne pour numériques + appariement catégoriel). But : voir si ces clusters recouvrent A/B/C/D.

### Résultats

- 4 clusters équilibrés (1000–2600 individus).
- Faible correspondance avec segments réels : **ARI = 0.11**, **NMI = 0.10**.
- Seul un cluster rapproche D (59 %).

**Interprétation** : les clusters révèlent des logiques démographiques, mais ne reproduisent pas la segmentation métier. La structure “naturelle” des données ne correspond pas aux labels fournis.

## Conclusion générale

### Bilan global :

- Données avec *data leakage*, incohérences et bruit élevé.
- Les modèles supervisés apprennent mieux que le hasard mais plafonnent à  $\approx 0.33$  sur test.
- Le clustering ne coïncide pas avec la segmentation business, confirmant l'absence de structure claire.
- L'écart **val** (0.56) vs **test** (0.33) suggère une segmentation instable et/ou un décalage de distribution ; un **feature engineering métier** et un **split temporel strict** seraient nécessaires pour progresser.

*Projet réalisé dans le cadre du Challenge Data Science – Jakala (2025).*