

PROJECT OF ANALYTICS EDGE

Achieved by :

Abderrahmane CHBIB
Bilal JAOUAD
Mohammed SITEL
Nizar DAHRABOU
Armel TOKALO BI TOKALO
Wendmanegda KABORÉ

Encadered by:

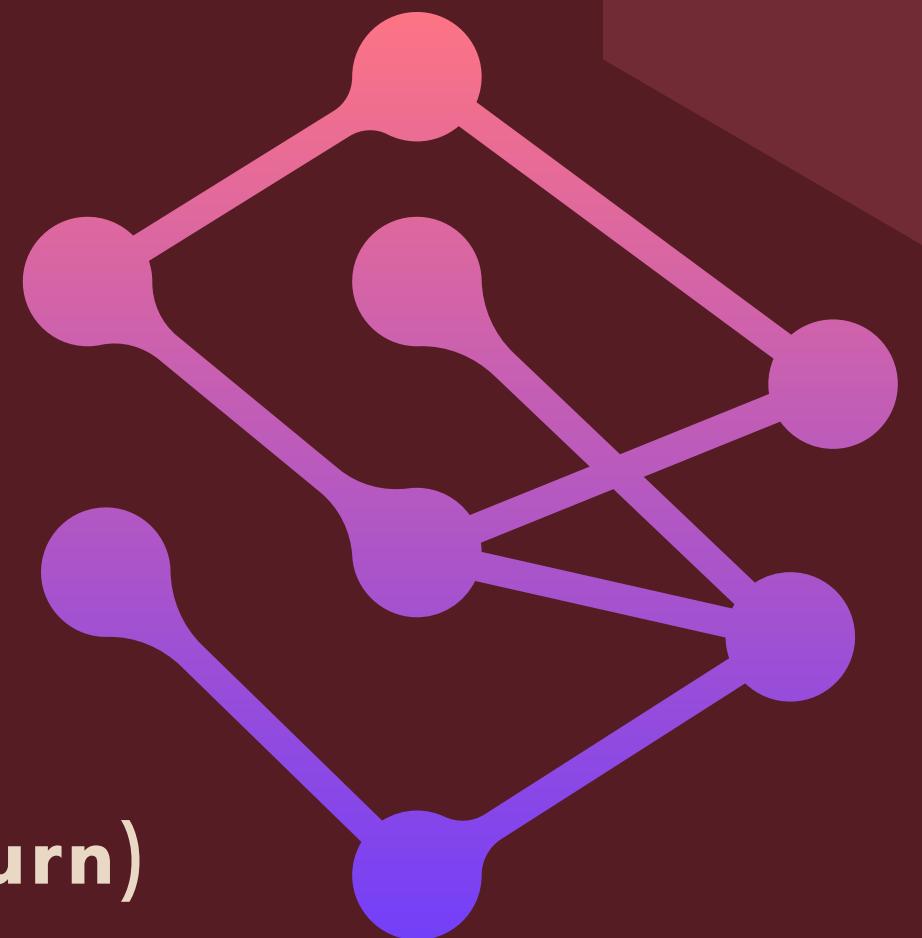
Maryam GUESSOUS

19 May, 2026



AGENDA

- 01 Introduction**
- 02 Analyse Exploratoire des Données (AED)**
- 03 Développement d'un Modèle de Segmentation RFM**
- 04 Développement d'un Modèle de Désabonnement (Churn)**
- 05 Conclusion**



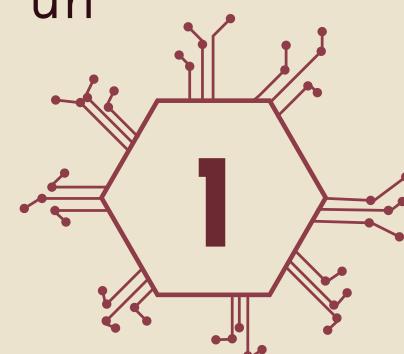


Introduction

Dans un contexte e-commerce ultra-concurrentiel où le coût d'acquisition client explose, le 'marketing de masse' (envoyer le même email à tout le monde) est inefficace et coûteux.

Comment passer d'une approche générique à une stratégie de fidélisation hyper-personnalisée en exploitant les données transactionnelles brutes pour identifier automatiquement les différents profils de comportement d'achat ?

L'une des méthodes actuellement utilisées consiste à utiliser les données brutes de transaction des clients, ainsi que des algorithmes de Machine Learning afin de concevoir un modèle capable de caractériser la clientèle.





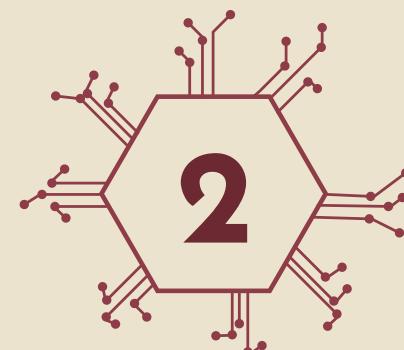
Introduction

D'où les questions centrales suivantes qui définissent les principaux paramètres d'analyse :

La Prédiction de Churn (Attrition) demande : "Quels clients sont susceptibles d'arrêter d'acheter chez nous ?"

La Segmentation Client demande : "Qui sont nos clients et comment devrions-nous adapter notre approche en fonction de leurs profils ?"

En combinant ces deux approches, nous sommes à même de déterminer nos clients et d'avoir des perspectives assez précises sur leurs comportements futurs.

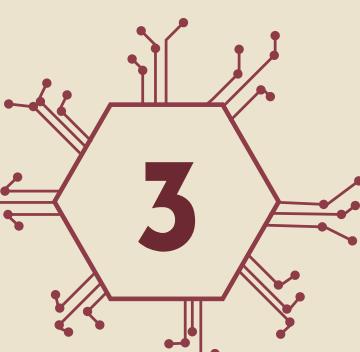




Analyse Exploratoire des Données (AED)

Périmètre et Nettoyage des Données

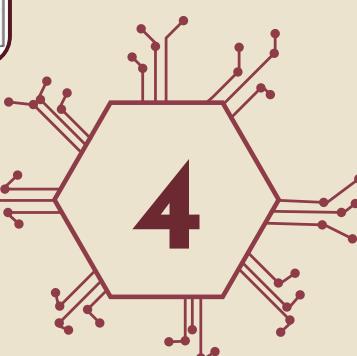
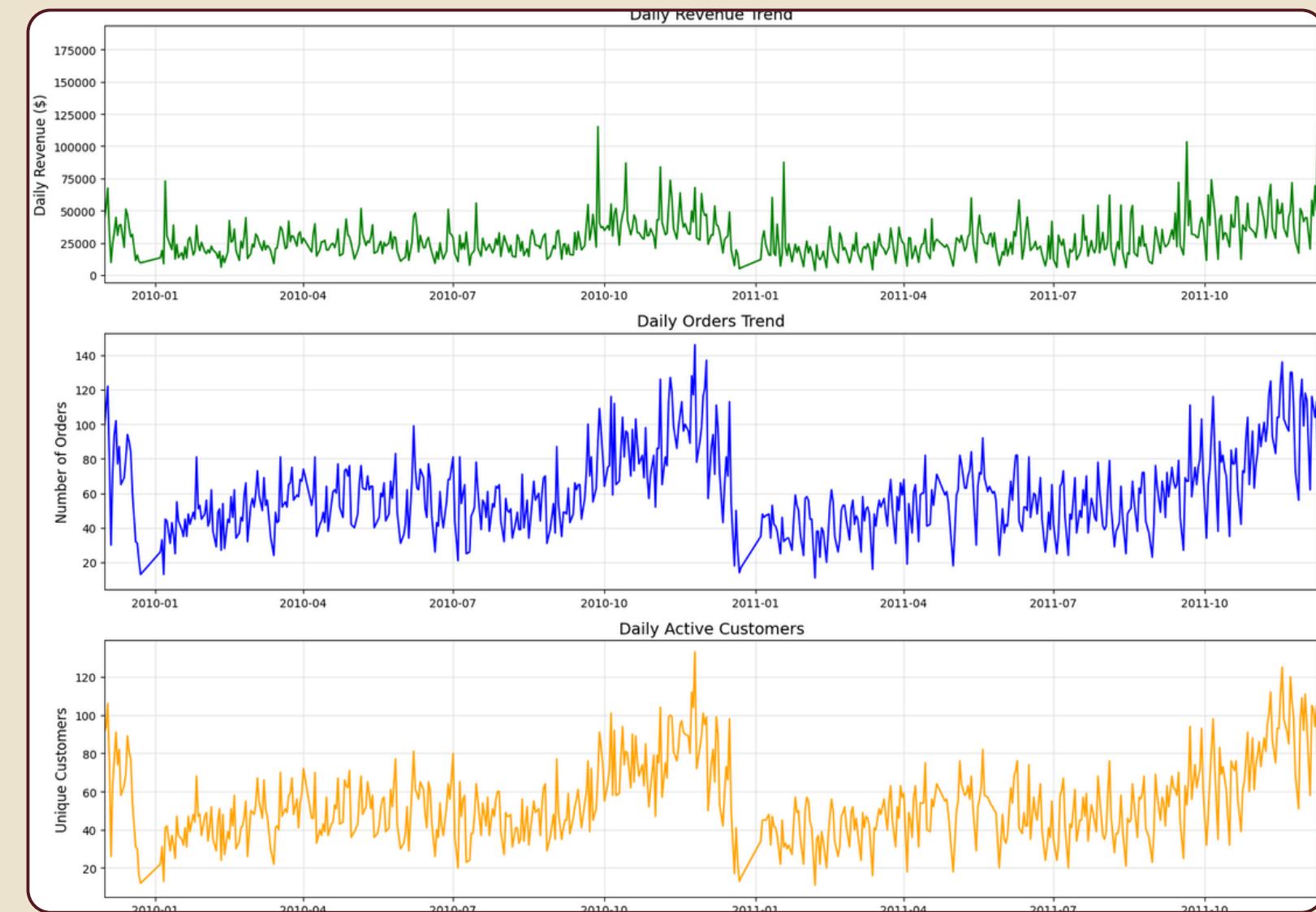
- **Source de données :** Transactions brutes du 01/12/2009 au 09/12/2011.
- **Pipeline de nettoyage :**
 - Suppression des transactions sans identifiant client (Indispensable pour le suivi).
 - Exclusion des retours (quantités négatives) et doublons
- **Impact du nettoyage :**
 - Données brutes : 1 067 371 lignes.
 - Données supprimées : 27,0 % (Principalement clients anonymes).
- **Dataset Final :**
 - 779 425 transactions valides.
 - 5 878 clients uniques suivis sur 738 jours.



Analyse Exploratoire des Données (AED)

Dynamique des Ventes et Saisonnalité

- **Couverture** : 2 années complètes, permettant de capturer les effets saisonniers annuels.
- **Tendance** :
 - Forte volatilité journalière des revenus.
 - Pics d'activité observables en fin d'année.
- **Implication pour la modélisation** :
 - Nécessité d'utiliser des fenêtres d'observation larges (90 jours) pour lisser le comportement client.
 - Validation de la stratégie de séparation temporelle (Train/Test) pour éviter le Data Leakage.



Analyse Exploratoire des Données (AED)

Profilage et Fidélité Client

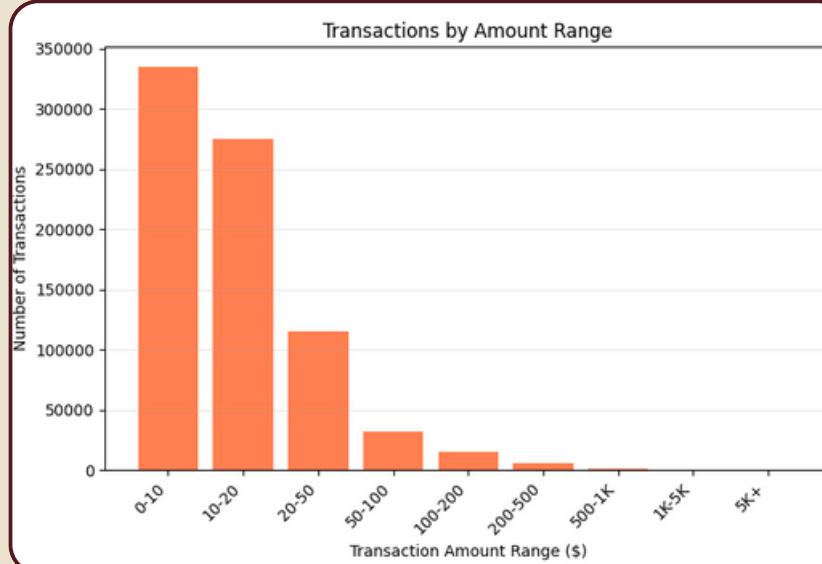
- **Distribution des commandes :**
 - Forte proportion de "One-time buyers" (acheteurs uniques).
 - Le taux de désabonnement naturel est élevé dès le premier achat.
- **Cycle de vie (Lifespan) :**
 - Disparité importante entre les clients éphémères (0 jour) et les clients fidèles (> 1 an).
- **Conclusion :**
 - L'enjeu du modèle est de distinguer un client volatile par nature d'un client fidèle sur le départ.



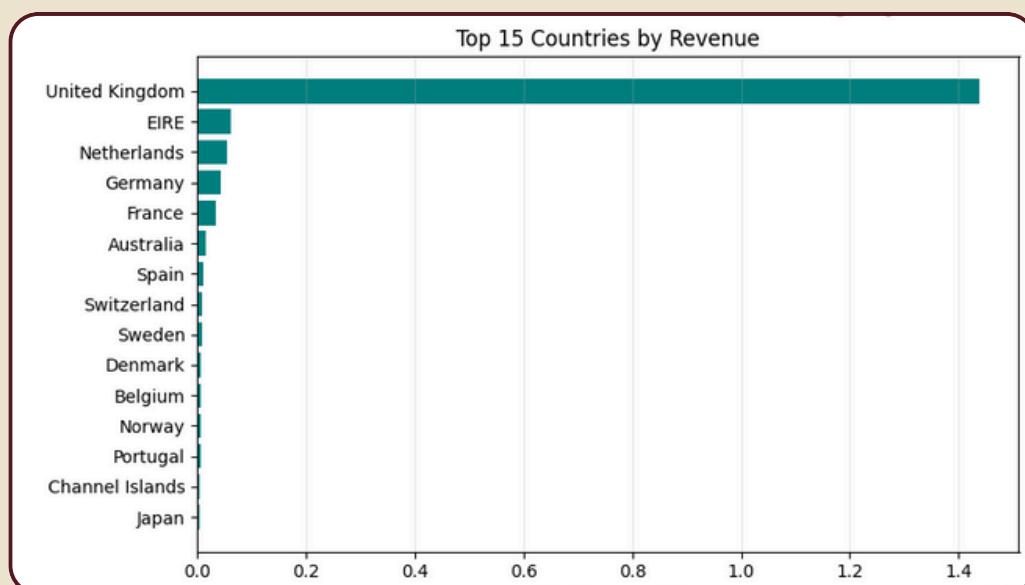


Analyse Exploratoire des Données (AED)

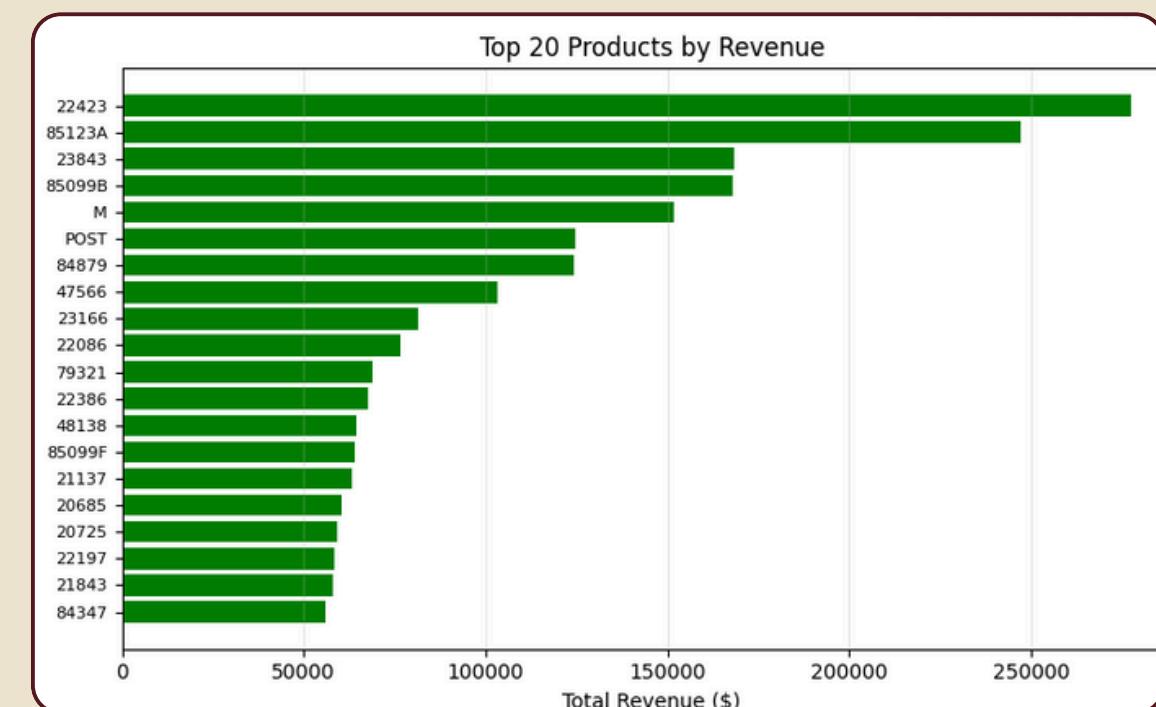
Valeur Client et Concentration



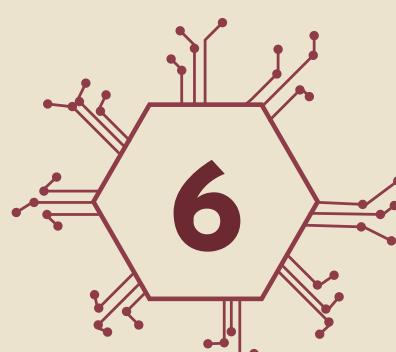
- **Distribution des Montants (Loi de Pareto) :**
 - La majorité des transactions sont de faible valeur.
 - Une minorité de clients ("VIP") génère une part significative du revenu.



- **Géographie :**
 - Forte concentration du chiffre d'affaires sur le marché principal (Royaume-Uni).



- **Analyse Produits :**
 - Identification des "Best-sellers" qui agissent comme moteurs de rétention.

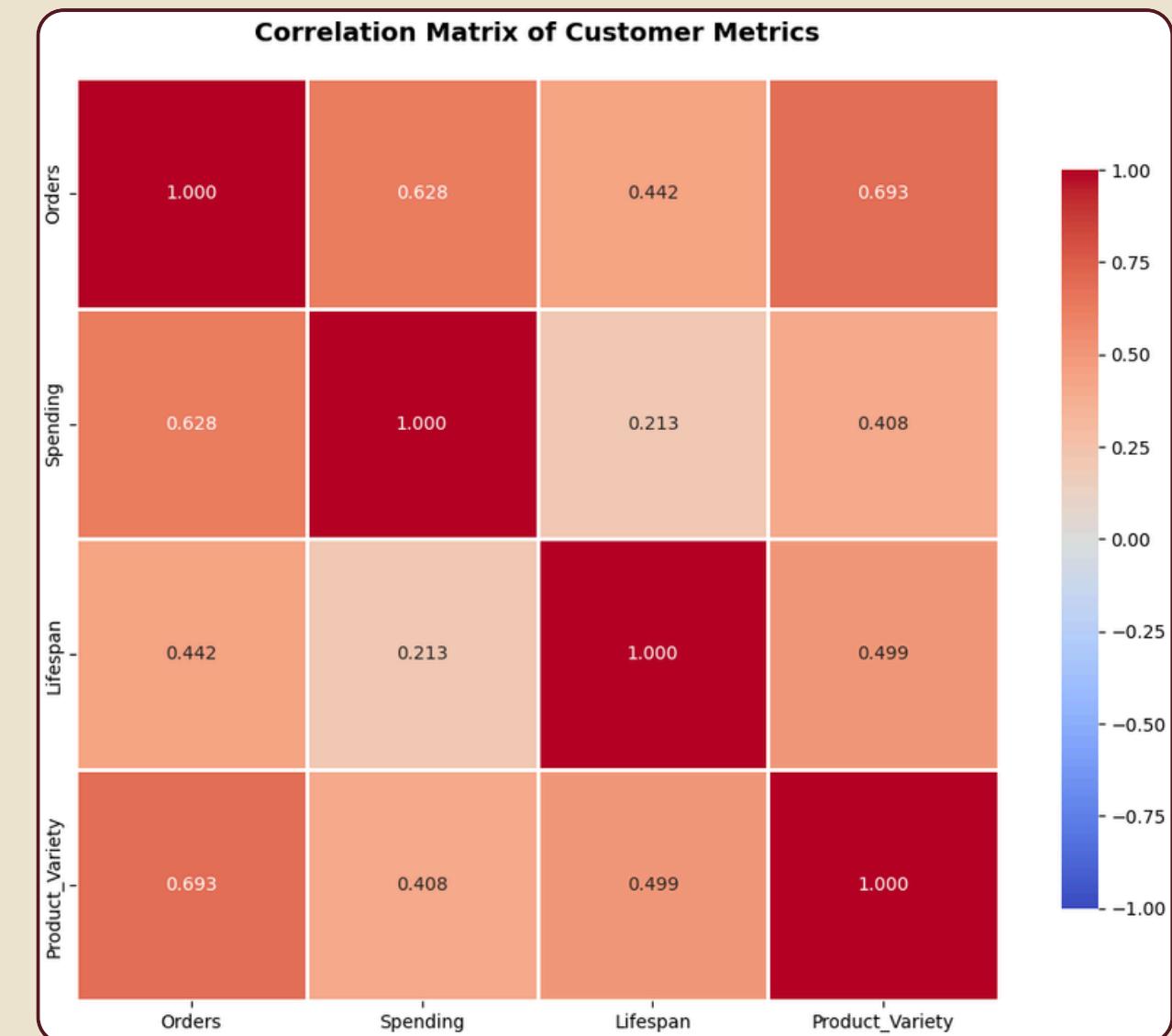




Analyse Exploratoire des Données (AED)

Analyse des Corrélations

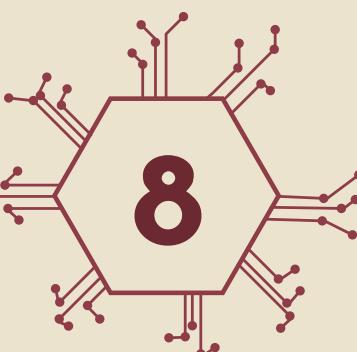
- **Matrice de Corrélation :**
 - Corrélation forte (> 0.8) entre la Fréquence (Nombre de commandes) et le Montant (Total dépensé).
 - Lien positif entre la Variété des produits achetés et la fidélité client.
- **Validation de l'approche RFM :**
 - Les variables Récence, Fréquence et Montant sont les indicateurs les plus fiables pour prédire le churn futur.



Analyse Exploratoire des Données (AED)

Conclusion de l'EDA et Prochaines Étapes

- **Validité des données :** Dataset propre et historique suffisant (2 ans).
- **Définition du Churn :**
 - Seuil fixé à 90 jours sans achat (basé sur l'analyse des cycles d'achat).
- **Stratégie retenue :**
 - a. Split Temporel : Entraînement sur le passé, test sur le futur (fenêtres glissantes).
 - b. Segmentation : Séparer l'analyse des nouveaux clients vs clients récurrents.
 - c. Modèles : Utilisation de Random Forest / XGBoost pour gérer les relations non-linéaires identifiées.



Développement d'un Modèle de Segmentation RFM

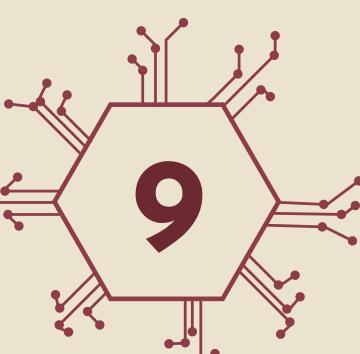
Qu'est-ce que le RFM ?

RFM = Recency, Frequency, Monetary

Recency : nombre de jours depuis le dernier achat

Frequency : nombre total de transactions

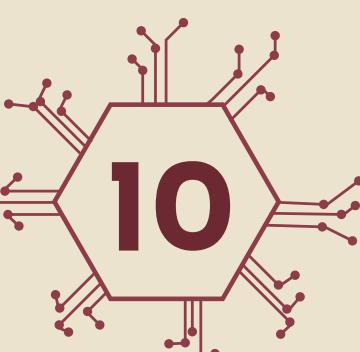
Monetary : montant total dépensé



Développement d'un Modèle de Segmentation RFM

Qu'est-ce que le RFM ?

- Recency (R)

$$\text{Date_ref} = \max(\text{InvoiceDate}) + 1$$
$$\text{Last_Purchase}_i = \max(\text{InvoiceDate}_i)$$
$$\text{Recency}_i = \text{Date_ref} - \text{Last_Purchase}_i$$


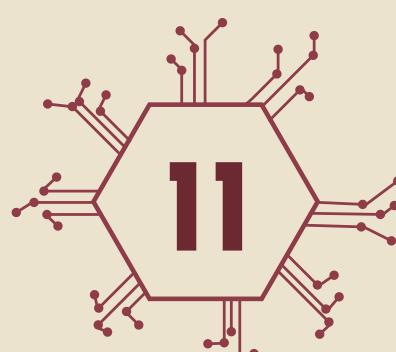
Développement d'un Modèle de Segmentation RFM

Qu'est-ce que le RFM ?

- **Frequency (F)**

Calculée comme la somme des quantités achetées par client

$$\text{Frequency} = \Sigma \text{ Quantity}$$



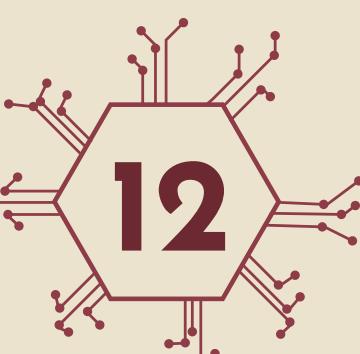
Développement d'un Modèle de Segmentation RFM

Qu'est-ce que le RFM ?

- **Monetary (M)**

Calculée comme la somme des montants dépensés par client

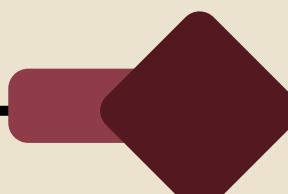
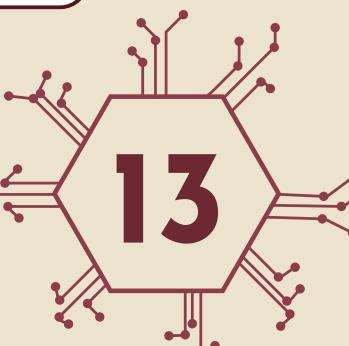
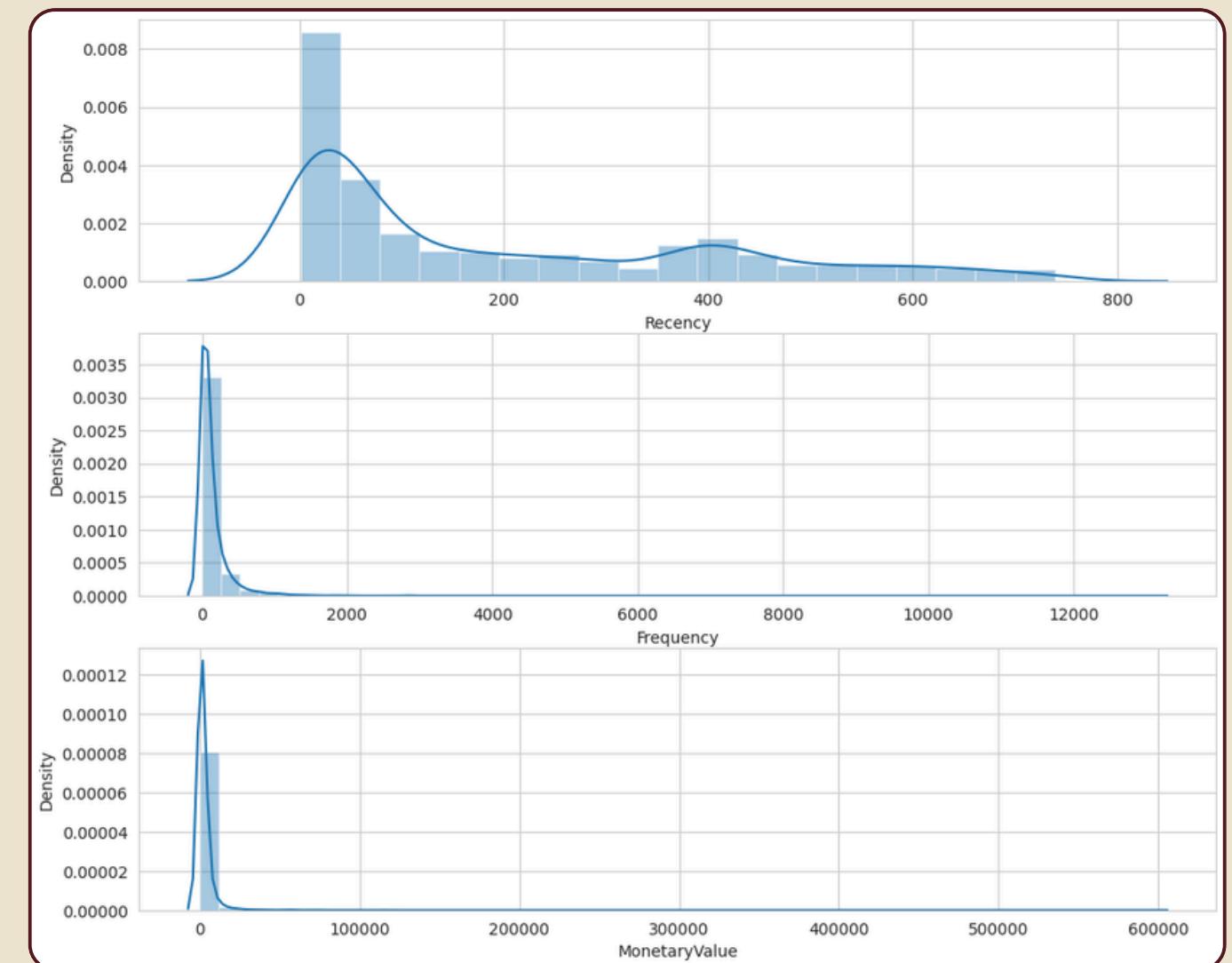
$$\text{Monetary} = \Sigma \text{ TotalPrice}$$



Développement d'un Modèle de Segmentation RFM

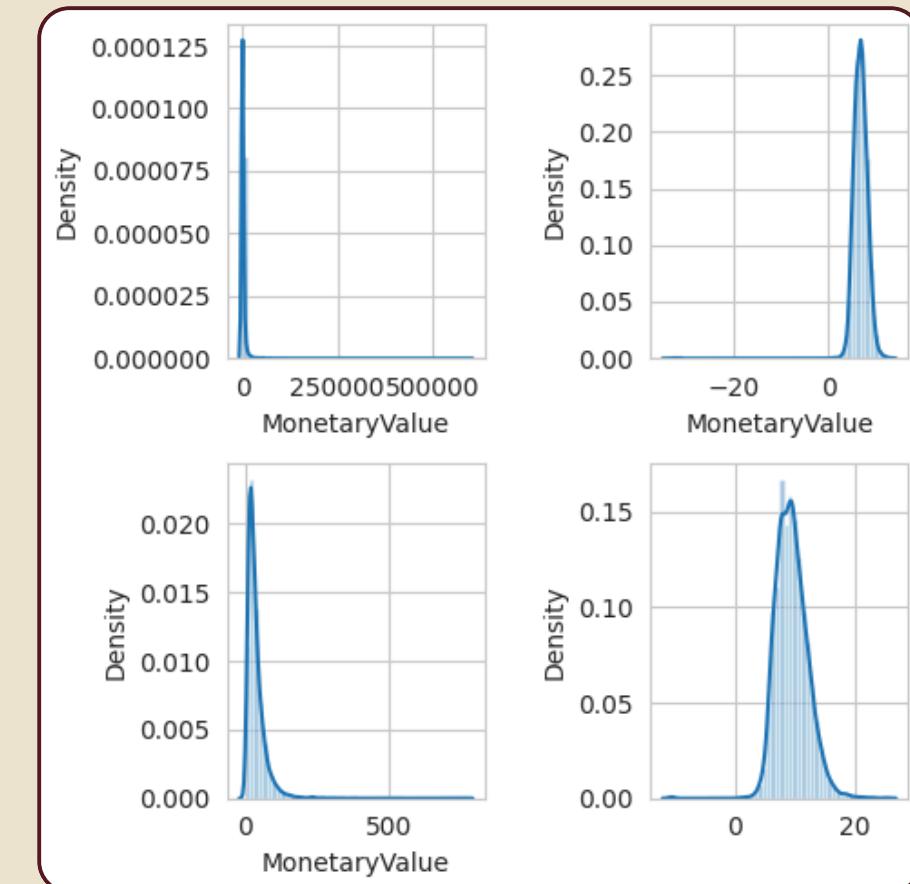
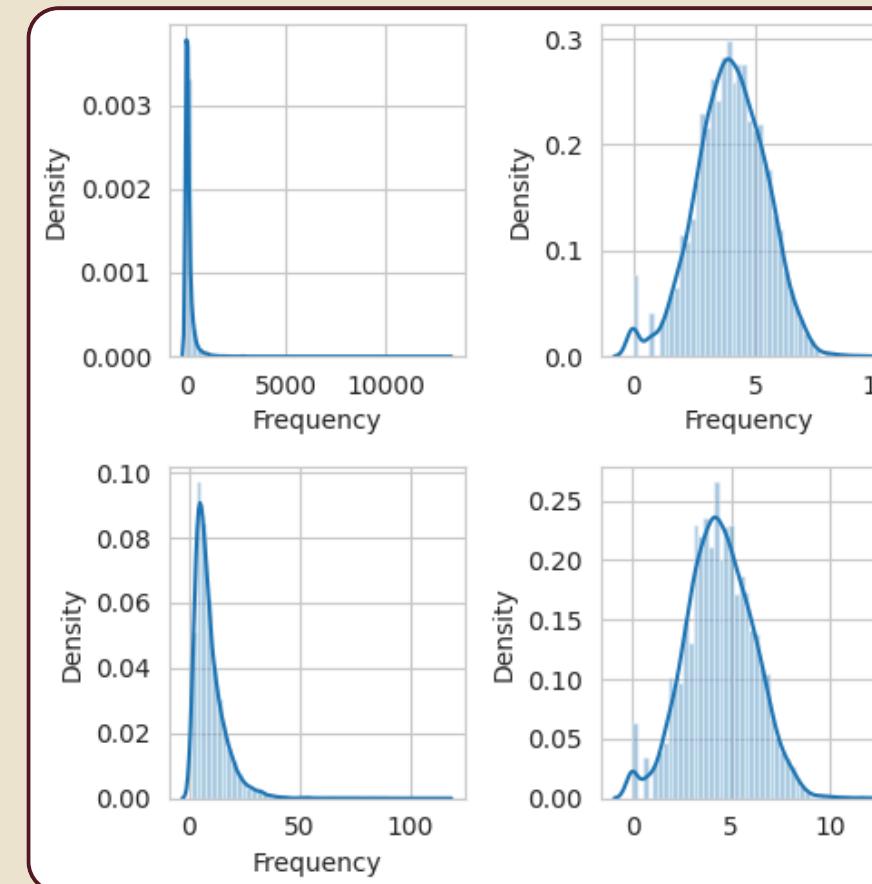
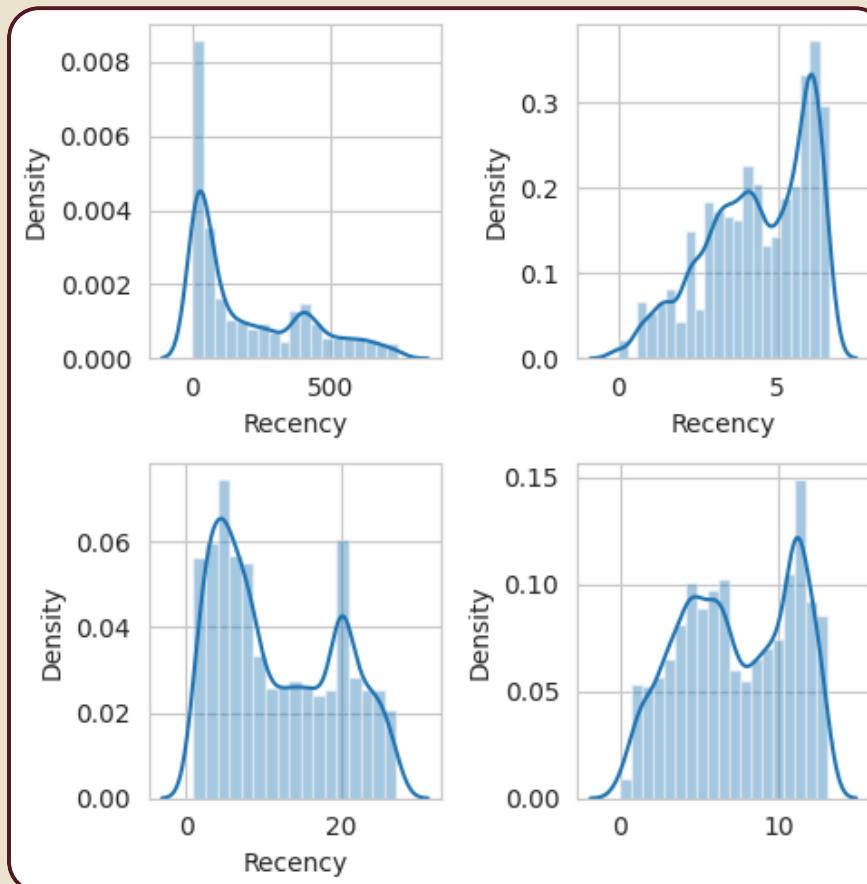
Analyse des distributions RFM

- Les distributions de R, F et M ne sont pas symétriques
- Les ordres de grandeur diffèrent entre récence, fréquence et valeur monétaire
- Une standardisation des données est nécessaire avant le clustering



Développement d'un Modèle de Segmentation RFM

Corrections des variables RFM (Skewness + Scaling)



- Comparer 4 transformations : Originale, Log, Racine carrée, Box-Cox
- Choisir la transformation la plus proche d'une distribution normale

Développement d'un Modèle de Segmentation RFM

Corrections des variables RFM (Skewness + Scaling)

→ Choix final : Box-Cox

- Réduction la plus forte de la skewness (par rapport à Log et Racine)
- Distributions plus proches d'une forme gaussienne
- Limite l'effet des valeurs extrêmes dans les distances
- Meilleure base pour standardisation et clustering

Développement d'un Modèle de Segmentation RFM

Corrections des variables RFM (Skewness + Scaling)

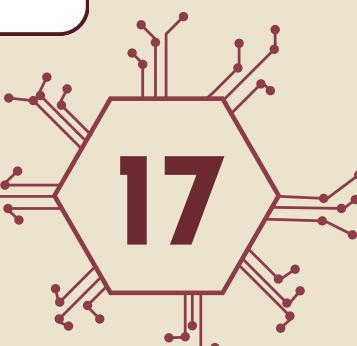
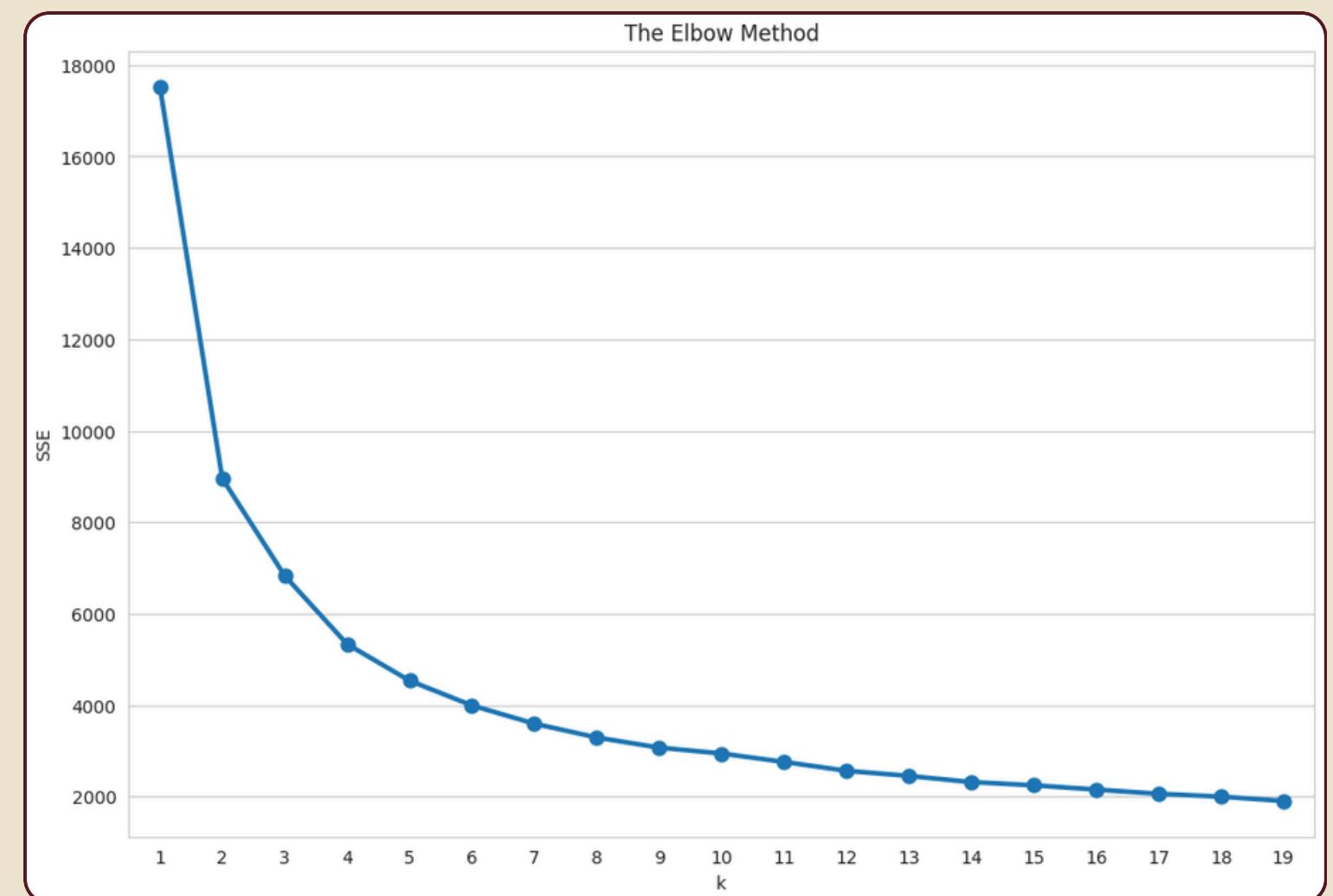
- **Standardisation : préparer les variables pour K-Means**
- Même après transformation, les échelles peuvent rester différentes
- StandardScaler : moyenne = 0 et écart-type = 1
- Chaque variable contribue de façon équilibrée au clustering
- Résultat : clusters plus stables et interprétables

Développement d'un Modèle de Segmentation RFM

Choix du nombre optimal de clusters – Méthode du Coude

- La SSE diminue lorsque le nombre de clusters augmente
- Le point optimal correspond à une rupture de pente de la courbe
- Au-delà de ce point, le gain devient marginal
- Le point de cassure est observé autour de $k = 4$

→ **Le clustering K-Means sera donc réalisé avec $k = 4$**



Développement d'un Modèle de Segmentation RFM

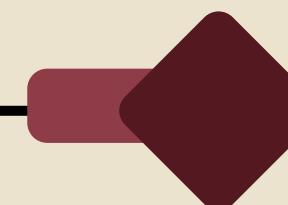
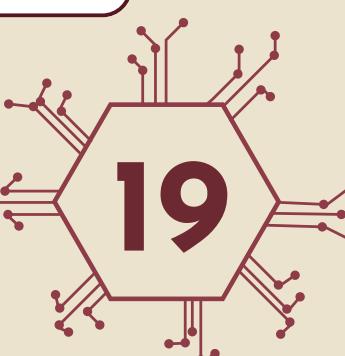
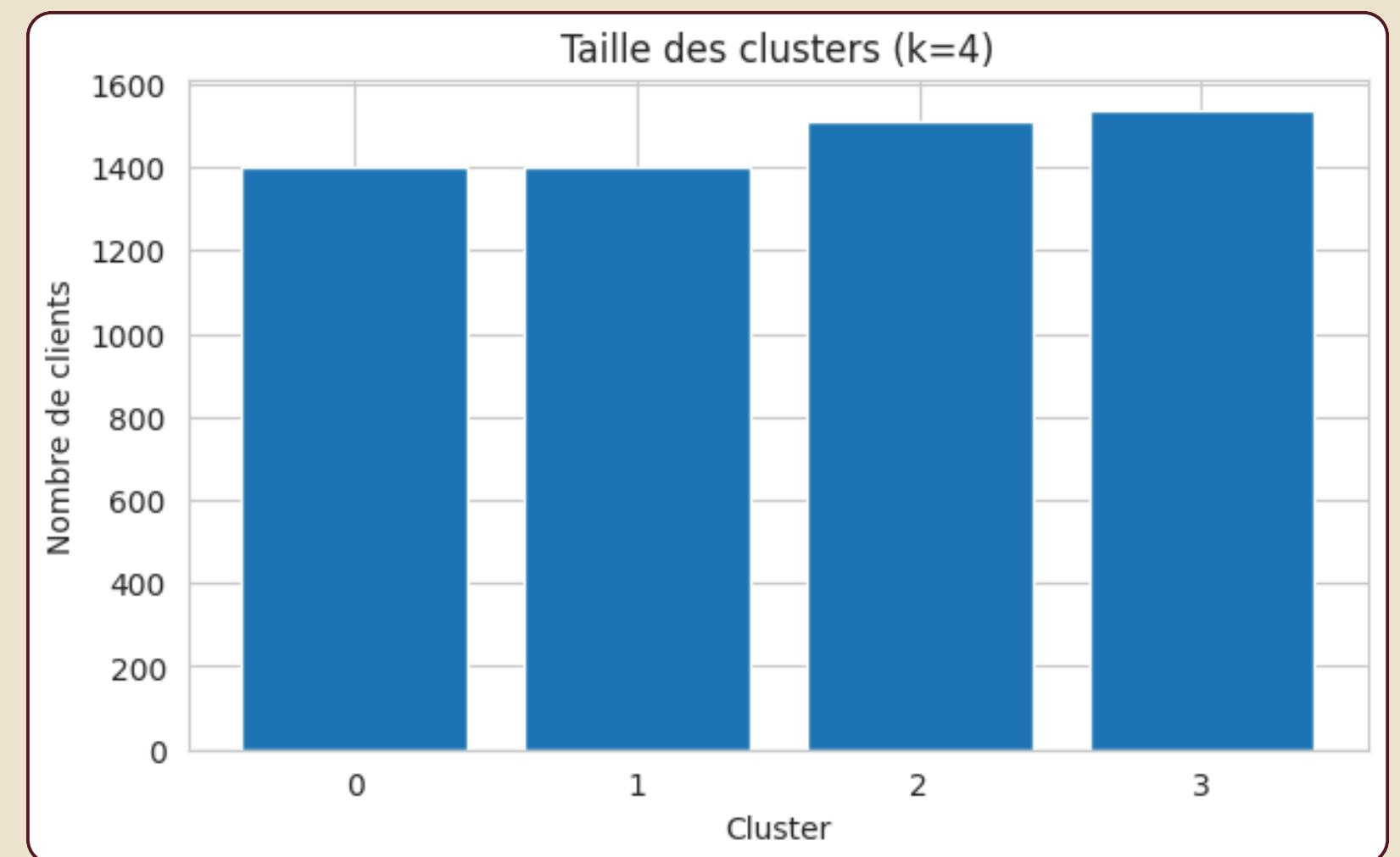
Algorithme de clustering K-Means

- Algorithme de clustering non supervisé basé sur la distance
- Objectif : regrouper les clients en segments homogènes
- Utilise la distance euclidienne entre les observations
- Chaque client est affecté au centroïde le plus proche
- Algorithme itératif visant à minimiser l'erreur intra-cluster (SSE)

Développement d'un Modèle de Segmentation RFM

Résultats du clustering K-Means ($k = 4$)

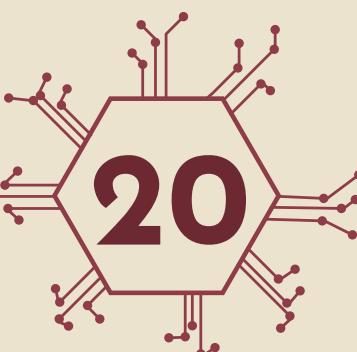
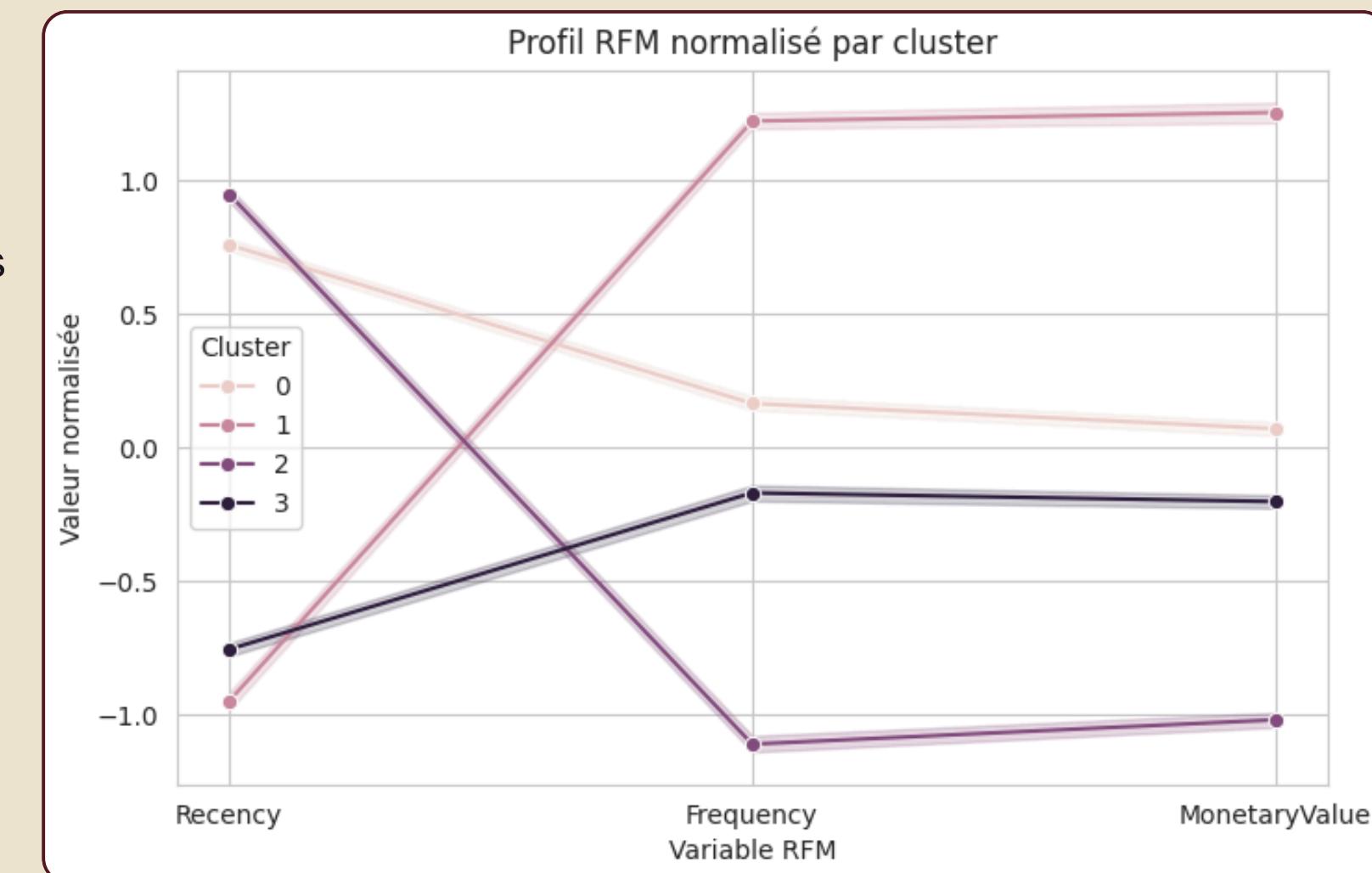
- K-Means entraîné sur les variables RFM standardisées
- Objectif : segmenter les clients en 4 groupes homogènes



Développement d'un Modèle de Segmentation RFM

Résultats du clustering K-Means ($k = 4$)

- Les Snake Plots représentent le profil moyen de chaque cluster sur les variables RFM
- Ils permettent de comparer visuellement plusieurs segments sur un ensemble de variables
- Chaque courbe met en évidence des comportements clients distincts



Développement d'un Modèle de Segmentation RFM

Résultats du clustering K-Means ($k = 4$)

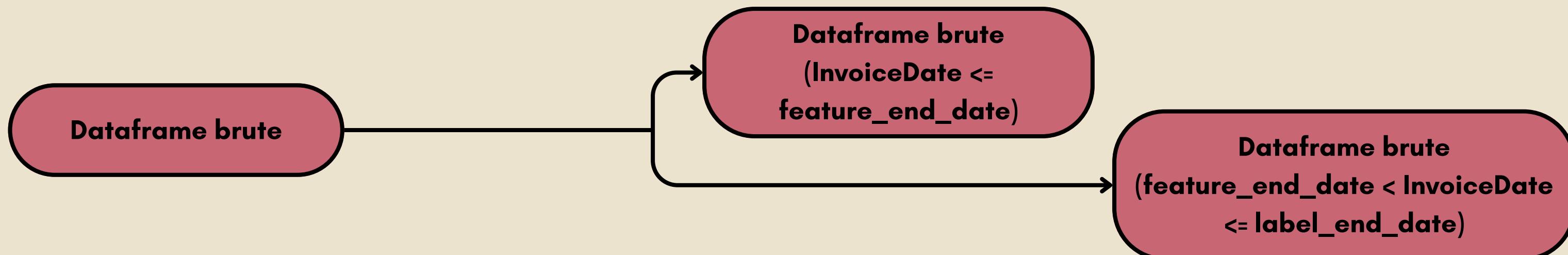
Cluster	Type de clients	Pourcentage (%)	Interprétation RFM
0	Clients à risque	24	clients dont le dernier achat remonte à un certain temps, mais qui ont réalisé par le passé des achats fréquents et des montants élevés
1	Clients fidèles	24	Clients les plus fréquents, avec les montants de dépenses les plus élevés et ayant effectué des achats très récemment
2	Clients perdus	26	clients dont le dernier achat remonte à longtemps et qui ont effectué peu d'achats. Ils correspondent donc à un segment de clients perdus ou churnés
3	Nouveaux clients	26	clients ayant effectué des achats récemment, avec une faible fréquence d'achat et un faible montant de dépenses

Développement d'un Modèle de Désabonnement

Feature Engineering



- Séparation des données de la Dataframe brute pour la conception de variables et de l'étiquetage



Développement d'un Modèle de Désabonnement

Feature Engineering

- Conception des variables



Variables liées aux produits :

Product_Diversity (Combien de produits différents achètent-ils ?)

Avg_Quantity (Combien d'articles par transaction ?)

Variables temporelles :

Days_Since_Last_Purchase (Combien de temps s'est écoulé depuis leur dernier achat ?)

Customer_Age_Days (Depuis combien de temps sont-ils clients ?)

Purchase_Span (Quelle est la durée totale de leur relation d'achat ?)

Variables de fréquence :

Purchase_Frequency (À quelle fréquence achètent-ils ?)

Avg_Days_Between_Purchases

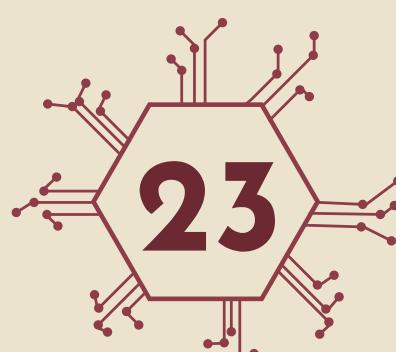
(Quel est le nombre moyen de jours entre leurs achats ?)

Variables monétaires :

Total_Spent (Dépenses totales)

Avg_Transaction (Montant moyen d'une transaction)

Spending_Per_Day (Régularité des dépenses)

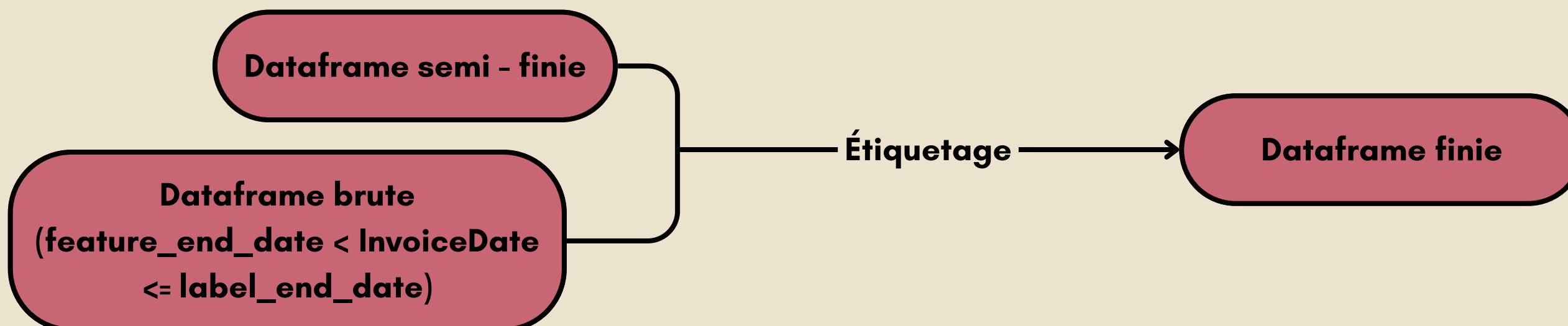




Développement d'un Modèle de Désabonnement

Feature Engineering

- Étiquetage des lignes de la Dataframe semi - finie



Nous déterminons si un client a "churné" (a déserté) en vérifiant s'il a effectué **AU MOINS** un achat pendant la fenêtre d'observation future.

En d'autres termes, nous vérifions que son "**Customer ID**" se trouve dans la Dataframe brute (`feature_end_date < InvoiceDate <= label_end_date`).

Si OUI → Churn = 0 (toujours actif)

Si NON → Churn = 1 (il nous a quittés)





Développement d'un Modèle de Désabonnement

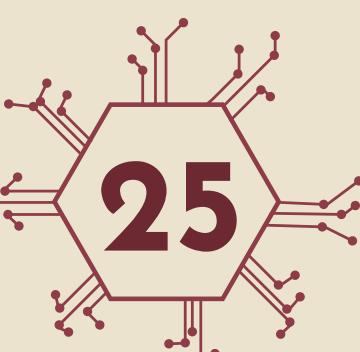
Entraînement et Évaluation du Modèle

Nous allons entraîner, après transformation des données brutes d'entraînement, plusieurs modèles d'apprentissage automatique différents et sélectionner le meilleur :

- **Régression Logique** – Une base simple et interprétable
- **Forêt Aléatoire (Random Forest)** – Un ensemble d'arbres de décision, performe généralement bien
- **Gradient Boosting** – Algorithme de boosting séquentiel
- **XGBoost** – Gradient boosting avancé (si disponible)

Détails importants

- Toutes les caractéristiques sont standardisées (mise à l'échelle pour une moyenne = 0 et un écart-type = 1).
- Nous utilisons une pondération des classes pour gérer le déséquilibre entre clients actifs et clients qui partent.
- Le meilleur modèle est sauvegardé pour une utilisation future.



Développement d'un Modèle de Désabonnement

Entraînement et Évaluation du Modèle

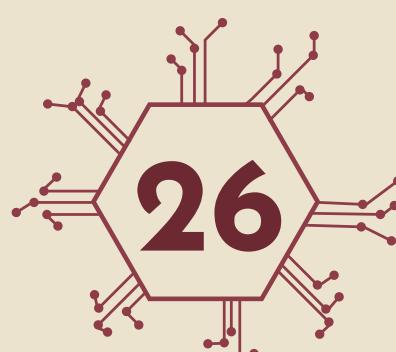
Pour l'évaluation des différents modèles, nous utilisons une validation croisée en 5 plis (5-fold cross-validation) sur l'ensemble d'entraînement pour estimer la performance.

La métrique qui nous intéresse est l'AUC-ROC (Aire Sous la Courbe ROC), qui mesure la capacité du modèle à distinguer les clients qui désertent de ceux qui restent.

```
=====  
Training: Logistic Regression  
=====  
Cross-Validation ROC-AUC: 0.780 (+/- 0.011)  
  
=====  
Training: Random Forest  
=====  
Cross-Validation ROC-AUC: 0.786 (+/- 0.011)  
  
=====  
Training: Gradient Boosting  
=====  
Cross-Validation ROC-AUC: 0.775 (+/- 0.013)  
  
=====  
Training: XGBoost  
=====  
Cross-Validation ROC-AUC: 0.762 (+/- 0.012)
```

Après entraînement →

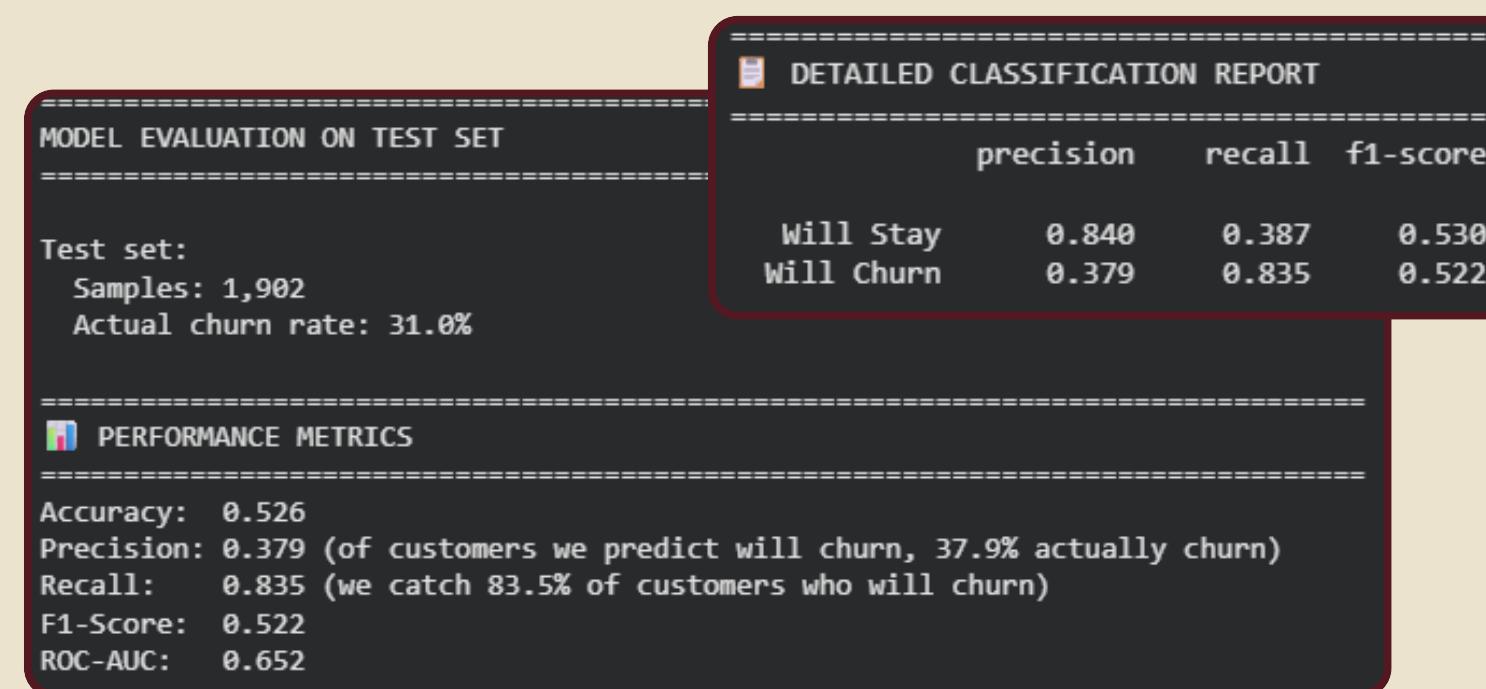
**Meilleur Modèle :
Random Forest**



Développement d'un Modèle de Désabonnement

Entraînement et Évaluation du Modèle

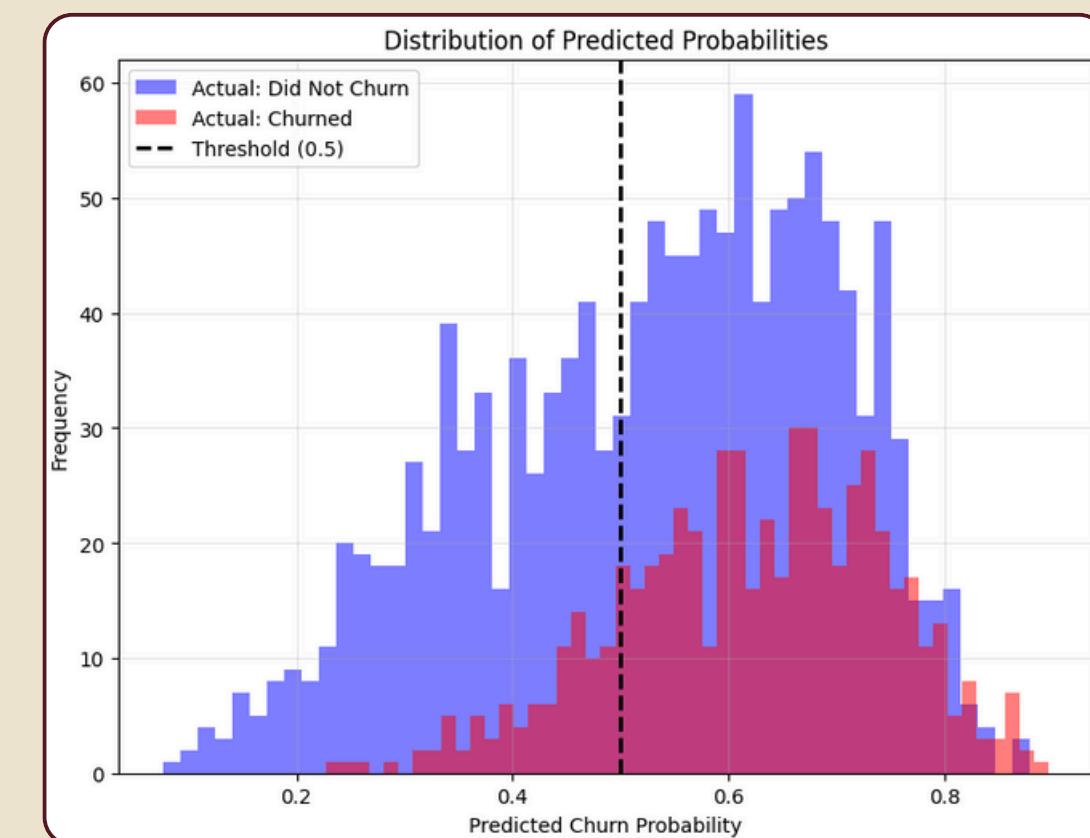
Après évaluation du modèle sur des données inédites (l'ensemble des données de test), consistant à se faire une idée sur son comportement en situation réel, nous nous sommes intéressés à certaines métriques que nous avons jugé pertinentes.



Ce modèle, sur un horizon de temps, nous permet donc :

- d'estimer à plus de 80 % les clients qui pourraient déserter, et
- de déterminer à plus de 80 % de précision ceux qui effectueront une transaction

Ce modèle nous permettrait donc d'affiner nos stratégies de marketing et de réapprovisionnement.



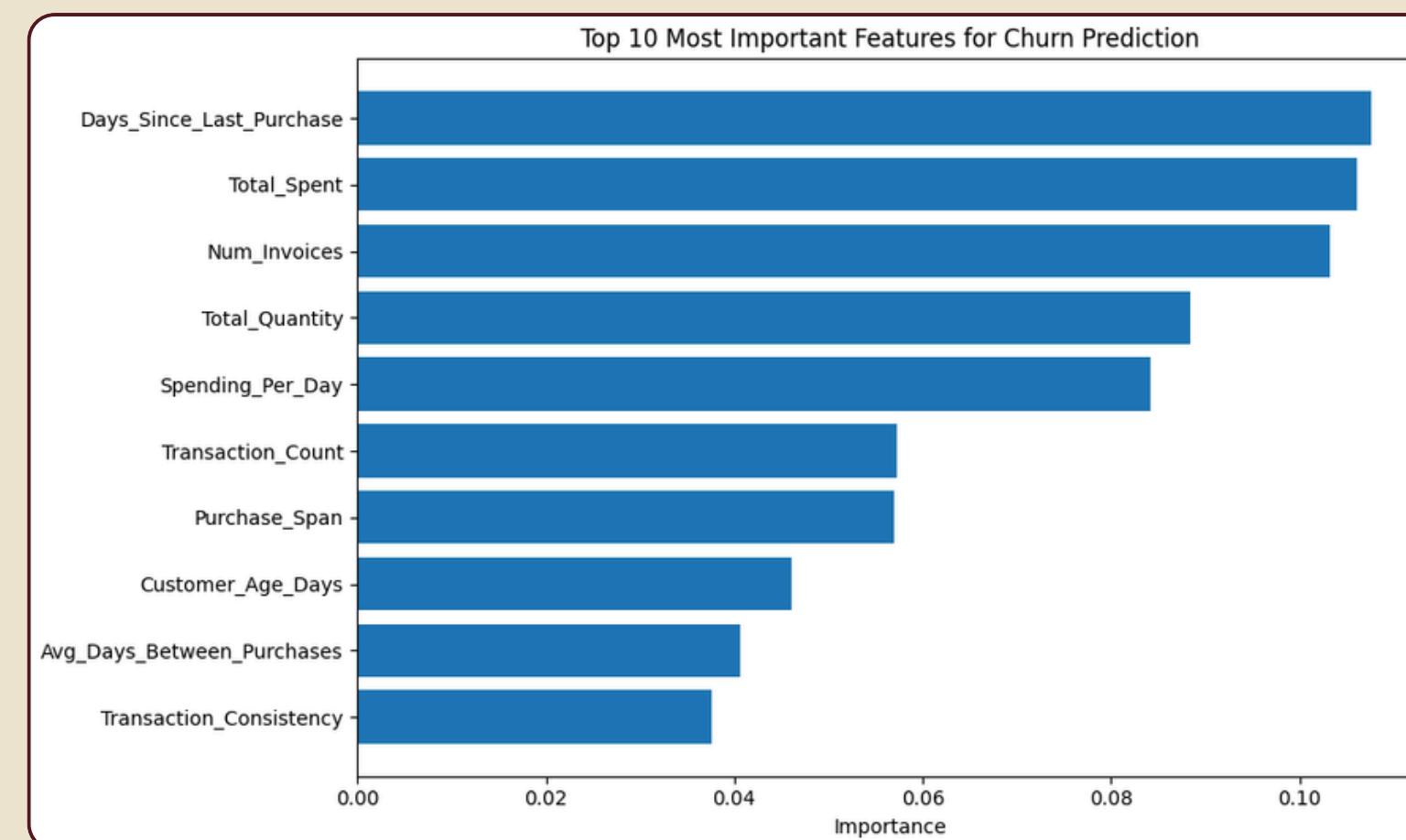


Développement d'un Modèle de Désabonnement

Analyse de Performance des Variables

Dans cette partie, nous nous sommes intéressés à la détermination des 10 variables les plus impactantes, ainsi qu'à la mesure de leur importance sur la prédiction de désabonnement.

IN . IB : Cela ne fonctionne que pour les modèles basés sur des arbres (Random Forest, Gradient Boosting, XGBoost) qui disposent de calculs d'importance des caractéristiques intégrés.





Conclusion

Points Clés

- Le modèle de prédiction du churn fonctionne : Nous pouvons prédire le futur départ des clients avec une précision raisonnable.
- La séparation temporelle est cruciale : Elle garantit une évaluation réaliste.
- La segmentation révèle des tendances : Des groupes de clients naturels émergent de leur comportement.
- Des outils différents pour des questions différentes : La prédiction et la description répondent à des besoins distincts.
- L'action est nécessaire : Les deux analyses pointent vers des interventions spécifiques.



Conclusion

Conseils d'utilisation

Pour la Prédiction du Churn :

- Réentraînement régulier : Recalibrez le modèle mensuellement avec de nouvelles données.
- Évaluation quotidienne : Exécutez les prédictions sur votre base de clients active.
- Segmenter et agir : Utilisez les niveaux de risque pour prioriser les interventions.
- Mesurer les résultats : Suivez le nombre de clients à risque que vous réussissez à fidéliser.
- Itérer : Ajustez les seuils et stratégies en fonction de ce qui fonctionne.



Conclusion

Conseils d'utilisation

Pour la Segmentation Client :

- Actualisation mensuelle : Re-segmentez les clients au fur et à mesure de leur évolution.
- Suivre les mouvements : Surveillez les clients changeant de segment.
- Personnaliser les campagnes : Utilisez les profils des segments pour adapter les messages.
- Tester et apprendre : Menez des tests A/B sur différentes approches au sein des segments.
- Combiner les insights : Superposez le risque de churn avec les segments pour un impact maximum.

Thank You
For Your
Attention

19 May, 2026

