BLOC 2 - MSPR - AMAZING

EPSI M1 CDPIA

Année 2023-24

Auteurs du rapport :

Ahmed Bouslama

Alexandre Canton Condes

Benjamin Hoarau

Samir Houbad

Eric Majri

Date de soutenance : le 24 Septembre 2024

Table des matières

Lien Code Source GitHub :

https://github.com/EricIrjam/e-amazing/tree/main

I. Contexte du projet

Amazing souhaite utiliser l'IA pour segmenter sa base clients selon leurs habitudes d'achat et leurs interactions avec le site (ex. : visites de pages produits, ajout/retrait du panier, achats).

Cette segmentation permettra de proposer des stratégies marketing plus efficaces et de personnaliser l'expérience client pour stimuler les ventes.

L'objectif est de créer un modèle capable de catégoriser les clients en fonction de leurs comportements d'achat et de navigation sur le site.

Ce modèle devra être capable d'évoluer et de s'intégrer dans l'infrastructure existante d'Amazing, en exploitant les données à grande échelle et en restant conforme aux réglementations telles que le RGPD.

II. Description des métriques pertinentes

A. Les métriques K-Means

Lors de l'utilisation de KMeans, les variables explicatives peuvent être divers indicateurs comportementaux et transactionnels. Nous avons probablement utilisé les types de variables suivants :

Nombre total d'achats : Indicateur simple du volume des transactions réalisées par un utilisateur.

Montant total des dépenses : Valeur totale dépensée par un utilisateur, indiquant sa contribution au chiffre d'affaires.

Fréquence des achats : Nombre de transactions effectuées sur une période définie.

Temps depuis la dernière transaction : Mesure la récence du dernier achat.

Diversité des produits : Nombre de catégories de produits achetés, reflétant l'engagement dans plusieurs gammes de produits.

B. Les métriques de la méthode RFM

Parmi les différentes méthodes de segmentation, la **méthode RFM** (Récence, Fréquence, Montant) associée au modèle de clustering **K-means** s'est révélée particulièrement efficace.

La méthode RFM permet de segmenter les clients selon trois dimensions clés :

Récence (R): Mesure le temps écoulé depuis la dernière interaction du client avec le site (en particulier les achats).

La donnée **event_time (**date et heure des événements d'achat) est essentielle pour mesurer la récence. En calculant la différence entre la date actuelle et la date du dernier achat ou de la dernière interaction, vous pouvez classer les clients selon le temps écoulé depuis leur dernier achat ou visite.

Fréquence (F): Mesure le nombre total d'achats effectués par chaque client sur une période donnée.

Pour calculer la fréquence, l'identifiant de l'utilisateur **user_id** et le type d'événement **event_type** sont nécessaires pour calculer la fréquence.

Vous pouvez compter le nombre total d'événements pertinents (achats ou visites) pour chaque utilisateur sur une période donnée, ce qui donne une idée de la fréquence d'interaction ou d'achat de ce client.

Montant (M): Représente le total des dépenses réalisées par chaque client.

Lors du calcul du montant, la colonne **price** est utilisée pour calculer le montant total dépensé par chaque client. Il est important d'utiliser le **event_type** pour vous assurer que vous ne comptez que les événements d'achats (et non les vues ou ajouts au panier). L'agrégation des montants pour chaque utilisateur fournit une estimation de la valeur totale dépensée.

Ces trois indicateurs permettent d'établir un profil comportemental pour chaque client, facilitant la mise en place d'actions marketing ciblées.

III. Analyse descriptive et nettoyage des données

Il est essentiel d'effectuer un nettoyage minutieux des données afin d'assurer la qualité et la pertinence de l'analyse. Ce processus de préparation des données est une étape cruciale, car des données mal nettoyées ou incohérentes pourraient biaiser les résultats du modèle et, par conséquent, les conclusions du projet.

Voici les différentes étapes du processus de nettoyage et de préparation des données, telles qu'elles ont été appliquées.

Dans un premier temps, nous avons vérifié la présence de valeurs manquantes dans les colonnes critiques. Ces colonnes comprennent : user_id, event_time, event_type, price.

Une autre étape critique consiste à identifier et supprimer les doublons dans les données. Les doublons peuvent se manifester lorsqu'un même événement est enregistré plusieurs fois (même user_id, product_id, event_time, etc.). Ces doublons peuvent provenir d'erreurs de journal ou de processus redondants dans la collecte des données. Leur présence aurait pu fausser les résultats en sur-représentant certains achats ou interactions. Après identification, ces doublons ont été supprimés pour garantir l'unicité des événements.

De plus, la gestion des valeurs aberrantes est une étape essentielle pour garantir la fiabilité des résultats. Deux types d'anomalies ont été identifiées et corrigées :

- Des montants anormalement élevés ou bas ont été recherchés, notamment des achats à 0 € ou des valeurs exceptionnellement élevées qui pourraient résulter d'erreurs de saisie. Ces valeurs aberrantes, non représentatives du comportement d'achat standard, ont été soit corrigées, soit supprimées en fonction de leur nature.
- Certains utilisateurs affichent une fréquence d'achats anormalement élevée, ce qui pourrait indiquer un bug dans la collecte des données ou une activité anormale. Une analyse plus approfondie de ces comportements a permis de déterminer s'ils devaient être exclus de l'analyse pour ne pas fausser les résultats.

Il convient de vérifier le format des données, et notamment de :

- Format des dates : La colonne event_time a été formatée correctement en type datetime afin de permettre le calcul précis de la récence (R), qui repose sur la date du dernier achat ou événement pertinent.
- Format des montants : Les valeurs de la colonne prix ont été converties en type float pour garantir une agrégation correcte lors du calcul du montant total dépensé (M). Une vérification a été effectuée pour s'assurer que ces montants n'étaient pas stockés sous forme de chaînes de caractères, ce qui aurait pu entraîner des erreurs dans les calculs.

La dernière étape du nettoyage des données a consisté à vérifier leur cohérence. Nous avons notamment porté attention aux éléments suivants :

- Chaque événement d'achat doit être associé à un identifiant utilisateur (user_id) et à un montant d'achat (price). Des incohérences, telles que des achats sans montant ou sans utilisateur, ont été identifiées et corrigées.
- Dates des événements : Une vérification des dates a permis de s'assurer qu'aucun événement d'achat n'était enregistré dans le futur (au-delà de la période d'analyse) ou avant la date officielle de lancement du projet. Des anomalies dans les dates pourraient indiquer des erreurs dans la collecte des données.

Standardisation des formats :

- Assurer une cohérence dans les formats de dates, montants et autres variables pour garantir une bonne interprétation lors des analyses.

Normalisation des variables :

- Si nécessaire, nous pouvons normaliser les variables telles que le montant et la fréquence pour éviter qu’une variable ne domine l’analyse en raison de son échelle.

Une fois cette étape terminée, les données seront prêtes pour les analyses suivantes, qu'il s'agisse d'une segmentation plus poussée ou de l'application de modèles prédictifs.

Toutes les données utilisées dans ce projet respectent scrupuleusement les normes du Règlement Général sur la Protection des Données (RGPD). Cela signifie que les informations sensibles et personnelles des clients sont protégées et traitées conformément aux directives en vigueur, garantissant ainsi la confidentialité et la sécurité des utilisateurs. Par exemple, la colonne **user_id** a été anonymisée pour ne pas contenir d'informations directement identifiables, mais plutôt des identifiants uniques générés, assurant ainsi que les données personnelles ne sont pas exposées tout en permettant des analyses comportementales fiables.

IV. Exploration des Catégories Finales et Clusters : Analyse des Caractéristiques

A. Conception et Optimisation de Modèles pour la Segmentation Client dans le Projet e-Amazing.

Nous avons testé deux modèles d'apprentissage : K-means et Random Forest

1. Modèle 1 : KMeans

- **Choix du nombre de clusters** : Utilisation de la méthode du coude ("Elbow method") pour déterminer le nombre optimal de clusters.
- **Standardisation des données** : Vérifier que les données RFM sont bien normalisées pour éviter la domination des variables à grande échelle.
- **Validation croisée** : Appliquer une validation croisée pour vérifier la robustesse des clusters.

2. Modèle 2 : Random Forest pour la Classification des Clients

Les caractéristiques RFM utilisées ainsi que d'autres variables (par exemple, catégories de produits achetés, réactivité aux promotions) pour classifier les utilisateurs.

- **Hyperparamètres** : Ajustement du nombre d'arbres, de la profondeur maximale des arbres, et du nombre minimum d'échantillons par feuille.
- **Importance des variables** : Analyser les variables les plus influentes pour la classification (ex. : "Fréquence" pourrait être plus importante que "Récence").

3. Analyse des Performances

Évaluation des Modèles :

1. **KMeans :**
    - **Score de silhouette** : Mesure la qualité de la séparation des clusters.
    - **Inertie** : Indicateur de la compacité des clusters.
    - **Difficulté** : Peut être sensible à la forme des clusters et aux outliers.

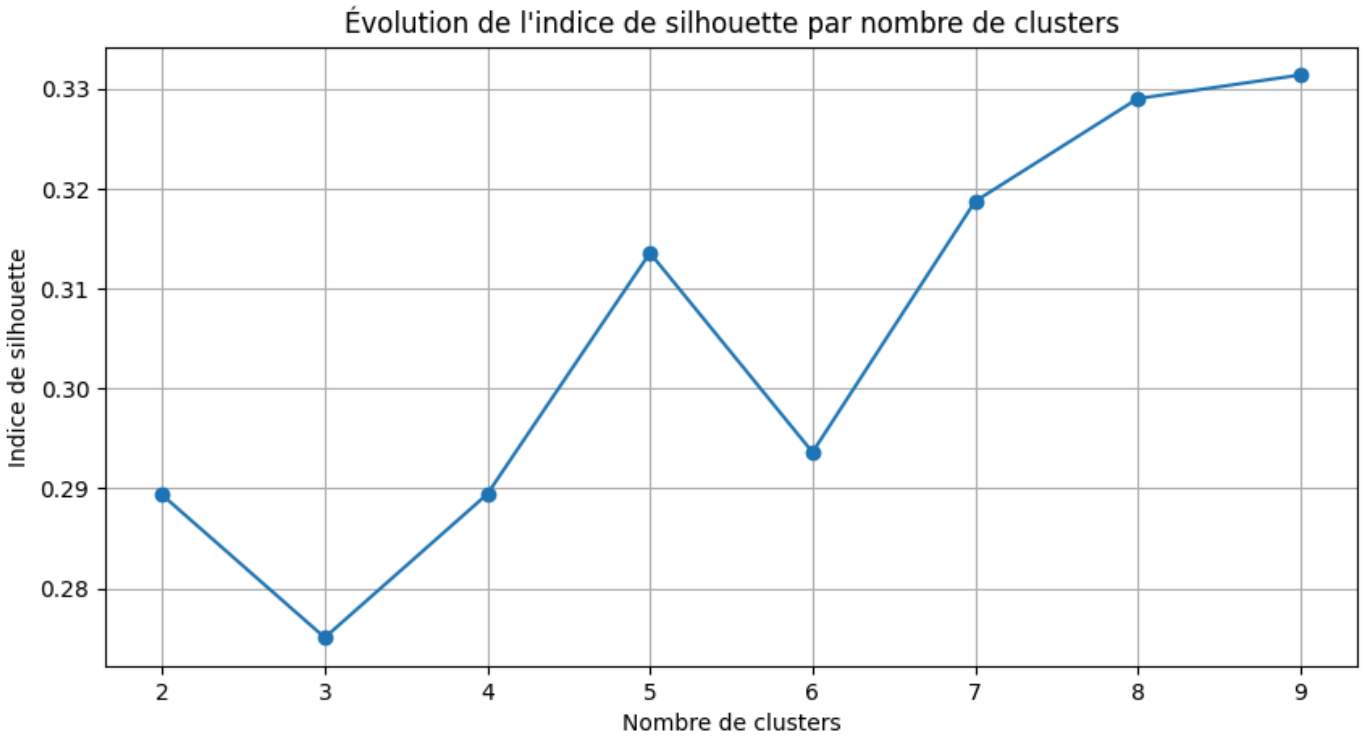
2. **Optimisation**
    - **Validation croisée** :Chaque modèle peut être optimisé à l'aide d'une validation croisée k-fold, garantissant une bonne généralisation des résultats et une meilleure stabilité.

4. Conclusion

- KMeans est efficace pour une première segmentation, mais peut manquer de finesse si les clusters ne sont pas bien séparés.
- Enrichir l'analyse RFM avec d'autres variables (catégories de produits, comportement de navigation, etc.) peut considérablement améliorer la segmentation.

L'optimisation se fera en fonction des résultats des métriques de performance et en ajustant les paramètres des modèles pour mieux capter la diversité des comportements client.

B. Exploration des Catégories Finales et Clusters : Analyse des Caractéristiques

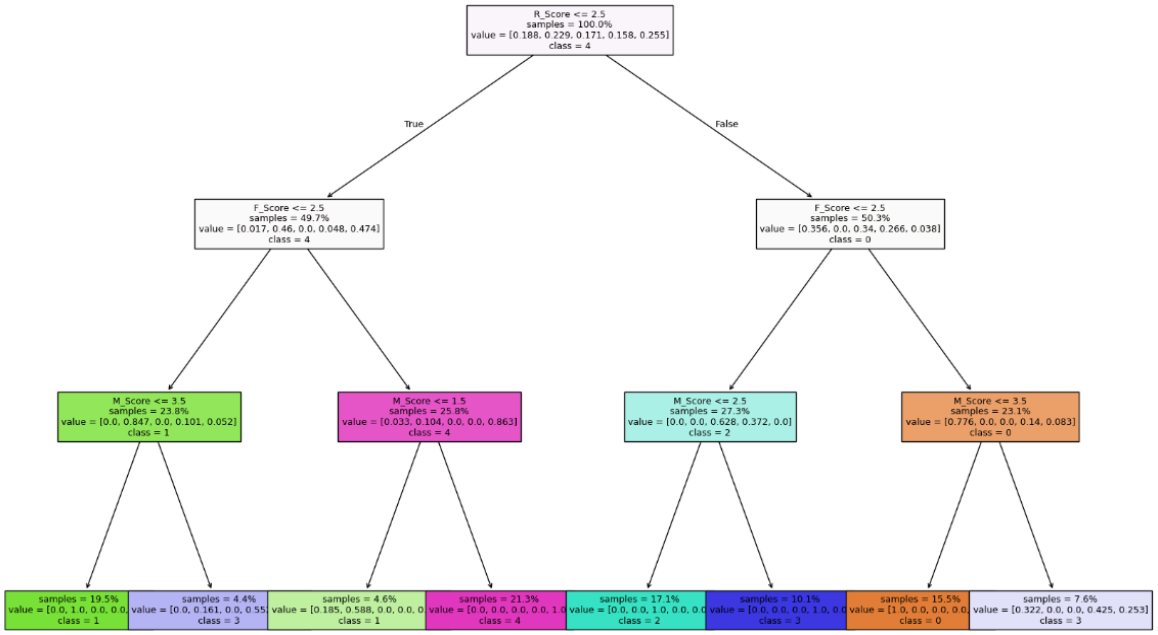


1. Indice de silhouette

- Un indice de silhouette proche de 1 indique que les clusters sont bien séparés et denses.
- Un indice proche de 0 signifie que les clusters se chevauchent ou ne sont pas clairement distincts.
- Des valeurs négatives (bien qu'il n'y en ait pas ici) indiquent que des points sont probablement affectés au mauvais cluster.

Nombre de clusters optimal : Nous observons un pic de l'indice de silhouette autour de 5 clusters, ce qui nous fait penser que 5 clusters pourraient être un bon choix pour une segmentation optimale selon cet indicateur. Cependant, Si nous souhaitons aller plus loin dans la granularité de l'analyse, tester avec 8 ou 9 clusters pourrait nous offrir une segmentation plus détaillée des clients, en prenant en compte des critères comme les gammes de produits ou la fréquence des achats.

2. Arbre de décision



(Schéma visible sur le notebook **4_RFM.ipynb -** https://github.com/EricIrjam/e-amazing/blob/main/4_RFM.ipynb)

- **R_score** fait référence à la **Récence**
- **F_score** fait référence à la **Fréquence**.
- **M_score** fait référence à la **Monétisation**

En regardant cet arbre de décision, nous pouvons associer différents groupes de clients à des animaux pour mieux visualiser leurs comportements :

- pour les clients avec un **R_Score ≤ 2.5**, ce qui signifie qu'ils ne sont pas très récents, nous pouvons les comparer à des **hiboux** 🦉. Ces clients sont plus discrets, moins actifs, et se trouvent majoritairement dans la classe 4.
- Certains, comme ceux avec un **F_Score ≤ 2.5** et un **M_Score ≤ 3.5**, peuvent être des **ours en hibernation** 🐻. Ils reviennent de temps en temps, mais ne dépensent pas énormément.
- Ensuite, pour les clients avec un **F_Score élevé** mais un faible **M_Score** (≤ 1.5), nous avons des **écureuils** 🐿️. Ces clients reviennent fréquemment, mais leurs achats sont modestes, un peu comme un écureuil qui accumule de petites quantités.
- Pour ceux avec un **R_Score élevé** (plus récents) et un **M_Score > 3.5**, nous sommes en présence de **lions** 🦁. Ces clients sont puissants et font des achats importants, bien qu'ils soient plus récents. Ils dominent la scène dans les classes 0 et 2. Ce sont les clients à chouchouter, car ils ont un potentiel de valeur élevé.
- Enfin, les clients dans les classes 3 et 2, avec un **F_Score ≤ 2.5**, peuvent être comparés à des **renards** 🦊. Ils ne reviennent pas très souvent, mais lorsqu'ils le font, ils dépensent un montant considérable, ce qui en fait des cibles intéressantes pour nos stratégies marketing.

En somme, cet arbre nous permet d'identifier et de comprendre la diversité des comportements des clients en fonction de leur récence, fréquence et montant. Les hiboux 🦉 et ours en hibernation 🐻 nécessitent peut-être des efforts de réactivation, tandis que les lions 🦁 et renards 🦊 sont des clients à haute valeur que nous devons fidéliser avec soin.

V. Architecture

Infrastructure de déploiement

A. **Conteneurisation** :
   - Utiliser **Docker** pour encapsuler le modèle dans des conteneurs afin de faciliter la portabilité entre les environnements de développement et de production.
   - Orchestration des conteneurs avec **Kubernetes** pour gérer la scalabilité et la résilience du service de prédiction.

B. **Environnement Serverless** (en option) : Utilisation de services tels qu'AWS Lambda ou Google Cloud Functions pour héberger le modèle dans un environnement serverless et ne consommer les ressources que lorsque nécessaire.

VI. Conclusion

L'analyse approfondie des modèles de segmentation pour le projet e-Amazing a permis de mieux comprendre les comportements clients et d'affiner les stratégies de segmentation. L'approche initiale basée sur plusieurs **KMeans** ont posé des bases solides pour une segmentation par clusters.

Principaux Enseignements :

KMeans reste un choix efficace pour une segmentation initiale, mais son optimisation (via la méthode du coude et la standardisation des données) est essentielle pour maximiser ses performances.

Optimisation et Performances :

L'optimisation des modèles via la **validation croisée** et l'ajustement systématique des hyperparamètres a permis d'améliorer la robustesse des résultats. Les mesures de performance, telles que le **score de silhouette**, **inertie**, **F1-score**, nous ont aidé à sélectionner les modèles les plus appropriés pour répondre aux besoins du projet.

- **Recommandations :**

  Approche hybride : Une combinaison des méthodes non supervisées (KMeans et DBSCAN) et supervisées (Random Forest, XGBoost) permettrait de capturer à la fois les patterns globaux et les spécificités locales des utilisateurs.

  Enrichissement des données : Intégrer davantage de variables comportementales (par exemple, navigation sur le site, réponses aux campagnes marketing) afin d'affiner les modèles.

Segmentation continue : Mettre en place un système de mise à jour régulière des segments clients pour s'adapter aux changements dans les comportements d'achat.
En conclusion, cette démarche offre une meilleure compréhension des catégories clients et permet une personnalisation des offres et des stratégies marketing, renforçant ainsi l'efficacité des campagnes et la fidélisation des clients.