Veille Modèle

# **Introduction :**

Dans ce projet, notre but est de mieux comprendre les différents types de clients qui utilisent la plateforme Amazing Basics. On ne cherche pas à prédire une valeur (comme le prix d’un achat) ou à classer les clients dans des catégories déjà connues. Au contraire, on veut découvrir automatiquement des groupes de clients qui se comportent de manière similaire, sans avoir besoin d’étiquettes au départ.

# **Présentation des données :**

**Type :** Les données auxquelles nous avons accès sont des fichiers au format CSV.

Il y a **7 fichiers CSV compressés**.

**Couverture temporelle :** Les données couvrent les mois d’octobre 2019 à avril 2020. Un fichier CSV correspond à **un mois** de données de cet ensemble.

| **Estimation du volume de données** | | | |
| --- | --- | --- | --- |
| **Fichier** | **Taille Compressé (Gb)** | **Estimation du Nombre de Lignes** | **Estimation Taille Décompressé (Gb)** |
| 2019-Oct | 1.62 | 42448765 | 5.52 |
| 2019-Nov | 2.69 | 70485912 | 9.16 |
| 2019-Dev | 2.74 | 71796059 | 9.33 |
| 2020-Jan | 2.23 | 58432559 | 7.60 |
| 2020-Feb | 2.19 | 57384442 | 7.46 |
| 2020-Mar | 2.25 | 58956618 | 7.66 |
| 2020-Apr | 2.73 | 71534030 | 9.3 |
| **Total** | 16.45 | 431038385 | 56.03 |

# **Typologie des modèles IA :**

## **La Régression :**

**Définition simple :**

La **régression** est une technique de machine learning **supervisée** qui permet de **prédire une valeur numérique continue** à partir d’autres variables (appelées *features* ou *caractéristiques*).

On l’utilise quand la **sortie attendue n’est pas une catégorie**, mais une **valeur chiffrée**.

**Modèles classiques de régression :**

Régression linéaire, Régression polynomiale, Ridge / Lasso, SVR, Arbres de décision régressifs, Random Forest Regressor et XGBoost,...

**Metriques :**

**MAE – Mean Absolute Error (Erreur absolue moyenne) :** Moyenne des valeurs absolues des écarts entre les vraies valeurs et les valeurs prédites. La MAE mesure l’erreur moyenne que fait le modèle, en unités unités réelles (ex : euros, minutes).

**MSE – Mean Squared Error (Racine de l’erreur quadratique moyenne) :** Moyenne des carrés des erreurs entre les prédictions et les vraies valeurs. Les grandes erreurs sont plus fortement pénalisées, ce qui peut rendre ce score plus sensible aux anomalies.

**R² – Coefficient de détermination :** Permet de voir si le modèle explique bien ce qu’il essaie de prédire.

**Compatibilité avec le projet Amazing :**

La régression ne répond pas à l’objectif de segmentation ou modélisation de profils clients car le but de la régression est de prédire une valeur numérique continue, tandis que Amazing cherche à mieux identifier des groupes de clients afin de proposer des offres personnalisées.

## **La classification:**

**Définition simple :**

La classification consiste à prédire une **catégorie ou un label discret**. Par exemple, dire si un client achètera un produit (oui/non) ou à quel segment il appartient (fidèle, occasionnel, à risque).

**Modèles classiques de régression :**

Naive Bayes, Linear SVC, KNeighbors, SGD Classifier,..

**Metriques :**

**Accuracy (Exactitude)** : Elle permet de montrer la performance globale du modèle. : Proportion de bonnes prédictions parmi toutes les prédictions.

**Precision :** Elle mesure la fiabilité des prédictions positives : Parmi les éléments prédits positifs, combien sont corrects (vrai positifs).

**Recall (Sensibilité) :** Mesure la capacité à retrouver tous les cas positifs : Parmi les éléments réellement positifs, combien sont correctement trouvés.

**F1-score :** Moyenne entre précision et rappel : Permet de trouver un équilibre entre faux positifs et faux négatifs.

**Matrice de confusion :**Permet de visualiser les vrais positifs, faux positifs, vrais négatifs, faux négatifs.

**ROC AUC :** Courbe qui mesure la capacité du modèle à distinguer entre classes (valeurs proches de 1 = excellent).

**Compatibilité avec le projet Amazing :**

Pour utiliser la classification, il faut **avoir des étiquettes** (labels) précises sur chaque client, ce qui n’est pas le cas ici (pas de catégories clients définies). Sans les données étiquetées, la classification ne peut pas être utilisée

## **La Réduction de dimensionnalité :**

**Définition simple :**

La réduction de dimensionnalité est une méthode visant à simplifier un jeu de données en réduisant le nombre de variables tout en conservant l’essentiel de l’information. Cela facilite l’analyse, la visualisation et le traitement des données complexes.

**Modèles classiques de régression :**

Analyse en composantes principales (ACP), IsoMap, Spectral Embbeding, …

**Metriques :**

**Explained Variance (notamment pour l’ACP)** : % de la variance conservée dans les composantes principales. Elle permet de vérifier si la réduction ne supprime pas trop d’informations utiles

**KL Divergence :** Mesure la perte d’information due à une simplification.

**Reconstruction Error :** Mesure ce que la réduction fait perdre.

**Compatibilité avec le projet Amazing :**

La réduction de dimensionnalité n’est pas une méthode de segmentation ou de modélisation de clients types. Elle ne permet ni de prédire une catégorie, ni de former des groupes à elle seule. Son rôle est de simplifier les données, en réduisant le nombre de variables, pour aider d'autres modèles comme le clustering ou la régression à mieux fonctionner.

## **Le Clustering :**

**Définition simple :**

Le clustering est une technique d’apprentissage **non supervisé** qui consiste à regrouper des objets similaires en groupes ou clusters, sans avoir besoin de labels préalables.

**Modèles classiques de régression :**

KMeans, MiniBatch KMeans, Spectral Clustering, DBSCAN, …

**Metriques :**

**Silhouette Score :** Mesure à quel point un point est proche de son propre groupe et loin des autres groupes.

**Davies-Bouldin Index :** Moyenne de la ressemblance entre chaque groupe et celui qui lui est le plus proche.

**Calinski-Harabasz Index :** Mesure combien les groupes sont concentrés en interne et séparés entre eux.

**Visualisation 2D/3D :** avec ACP ou t-SNE pour vérifier visuellement la qualité des clusters.

**Elbow Method :** Permet de trouver le **nombre optimal de groupes**.

**Compatibilité avec le projet Amazing :**

Le clustering est parfaitement adapté pour modéliser des clients types dans le cadre de Amazing. En l’absence d’étiquettes, il permet d’identifier des groupes de clients aux comportements similaires, ce qui va faciliter la personnalisation des offres.

# **Le choix du modèle de Clustering :**

Dans notre projet, on cherche à développer un modèle capable de catégoriser les clients par habitudes d’achat et de visites sur le site. On va donc définir un nombre de segment (cluster) de clients à partir des données disponibles.

Il existe deux catégories de modèles de clustering :

1. **Les modèles à nombre de clusters inconnu :**  
    Ces modèles déterminent automatiquement le nombre de clusters en fonction de la structure des données.  
    **Exemple :** DBSCAN, Mean Shift.
2. **Les modèles à nombre de clusters défini :**  
    Ces modèles nécessitent de spécifier à l’avance le nombre de segments (ou clusters) que l’on souhaite identifier dans les données.  
    **Exemple :** K-Means, MiniBatch K-Means, Spectral Clustering, GMM..

**Pourquoi MiniBatch K-means est le modèle le plus adapté à notre projet ?**

Dans notre cas, nous allons définir un nombre précis de clusters à l’avance (acheteurs réguliers, occasionnels, visiteurs passifs, etc.). Cette connaissance préalable du nombre de groupes cible nous pousse à choisir un modèle à nombre de clusters défini (K-means, MiniBatch K-means, GMM, etc..)

Entre ces options, nous avons choisi MiniBatch K-means qui est particulièrement adapté lorsque le nombre de clusters est relativement faible. MiniBatch K-means utilise des mini-lots de données pour effectuer les mises à jour des centres de clusters, ce qui le rend plus rapide et moins coûteux que K-means classique, tout en conservant une bonne qualité de segmentation.

**REMARQUE / POINTS A RAJOUTER DANS LE DOC :**

* Amelioration de la mise en page
* Explication en détail de chaque paragraphe
* Techniques d’évaluation