DOUNIA HULLOT
SAMUEL ETIENNE

Rapport Projet Machine Learning

19/12/2023

Sommaire

Introduction

Ce rapport se focalise sur l'analyse des préférences d'achat de produits par différentes tranches d'âge. L'objectif principal est de développer un modèle de prédiction capable de déterminer le type de produit susceptible d'être acheté par des individus appartenant à des tranches d'âge spécifiques. Pour ce faire, nous avons utilisé un jeu de données contenant des informations sur les habitudes d'achat de différentes tranches d'âge.

La Dataset et son aperçu

Customer Shopping Trends Dataset:

La database se base sur les préférences d'achat des clients, offre des informations sur le comportement des consommateurs et leurs habitudes d'achat.

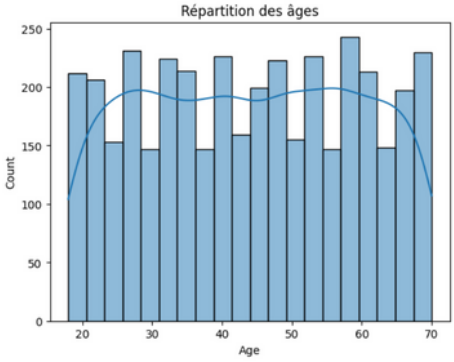
Lien de la Dataset : https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset
Le jeu de données utilisé comprend des informations détaillées sur les habitudes d'achat de différentes personnes réparties dans des tranches d'âge spécifiques.
Celui-ci inclut de nombreuses caractéristiques :

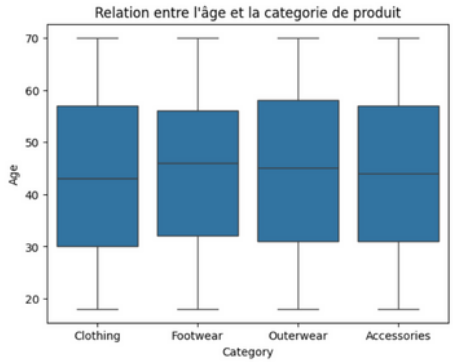
- Identifiant client
- Âge et genre du client
- Articles achetés et leur catégorie
- Montant des achats
- Localisation des achats
- Évaluation des achats par les clients
- Statut d'abonnement
- Méthode de paiement préférée
- Fréquence d'achat

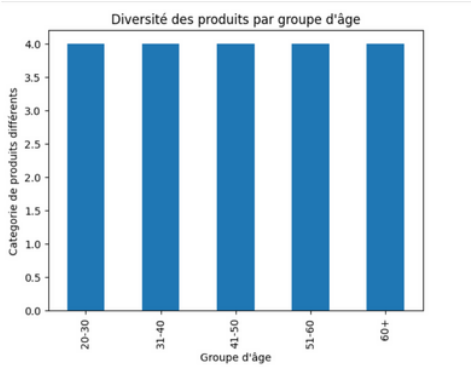
Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
2896	56	Female	Hoodie	Clothing	86	Montana	L	Green	Summer	4.60	No	Standard	No	No	29	Bank Transfer	Monthly
2752	27	Female	Dress	Clothing	52	Minnesota	S	Indigo	Fall	3.10	No	Free Shipping	No	No	50	Venmo	Monthly
1224	69	Male	Pants	Clothing	24	Kansas	L	Red	Winter	3.90	No	Free Shipping	Yes	Yes	21	Bank Transfer	Weekly
2485	60	Male	Hoodie	Clothing	97	New Hampshire	M	Green	Summer	4.80	No	2-Day Shipping	No	No	50	Cash	Every 3 Months
3286	58	Female	Hat	Accessories	31	Hawaii	XL	Magenta	Fall	4.60	No	Free Shipping	No	No	11	Cash	Weekly

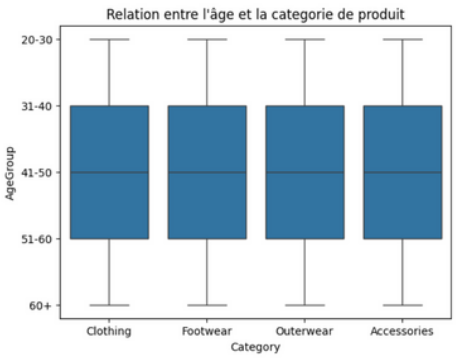
Pré-traitement de la Dataset

Avant le prétraitement, nous avons exploré la dataset en visualisant la relation entre plusieurs caractéristiques de la dataset sous forme de graphe .









Pré-traitement

Nous avons gardé uniquement les catégories âge et vêtement qui nous intéresse pour répondre à notre question.
Nous avons également affecté chaque âge à une tranche d'âge.

On lit le dataset et on trie les âges par tranches

```
raw_dataframe = pd.read_csv("shopping_trends_updated.csv")

raw_dataframe = raw_dataframe[["Age", "Category"]]

bins = [18, 25, 40, 50, 60, 100]
labels = ['18-25', '26-40', '41-50', '51-60', '60+']
raw_dataframe['Age'] = pd.cut(raw_dataframe['Age'], bins=bins, labels=labels, right=False)
```

Ensuite, nous avons encodé les données qui étaient sous forme de chaine de caractère, ce sont maintenant des nombres entiers qui représentent chacune des classes.

Encodage des données

```
from sklearn.preprocessing import LabelEncoder

label_encoder = LabelEncoder()
raw_dataframe['Age'] = label_encoder.fit_transform(raw_dataframe['Age'])
raw_dataframe['Category'] = label_encoder.fit_transform(raw_dataframe['Category'])
```

CHOIX DU MODÈLE

Notre approche a impliqué l'exploration de plusieurs modèles d'apprentissage automatique pour déterminer celui qui convient le mieux à notre tâche de prédiction des types de produits achetés par différentes tranches d'âge. Parmi les modèles explorés figurait la régression logistique, les arbres de décision et Forêts Aléatoires, les réseaux de Neurones, la méthodes de Classification Naïve Bayes et enfin les machines à Vecteurs de Support (SVM).

Nous avons opté pour la régression logistique. Cette décision s'est basée sur plusieurs raisons.
D'abord, ce modèle peut gérer plusieurs catégories de produits, ce qui correspondait parfaitement à notre objectif de prédire différents types d'achats.
Ensuite, il est simple à comprendre. Ça nous a aidé à voir comment chaque caractéristique influe sur les choix d'achat, ce qui est super pour interpréter les résultats.
Enfin, c'était plus rapide à entraîner par rapport à d'autres modèles plus compliqués.

Le modèle

Nous avons séparé les données en ensembles d'entraînement et de test pour former le modèle de régression logistique.
Nous avons aussi tenté de faire de l'affinage sur le modèle, mais cette tentative n'a pas donné des résultats concluants.

Séparation des données en données d'entrainement et de test

```
7]: from sklearn.model_selection import train_test_split

X = raw_dataframe.drop(['Category'], axis=1)
y = raw_dataframe['Category']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

Voici comment nous avons utilisé le modèle :

```
from sklearn.linear_model import LogisticRegression

model = LogisticRegression()
model.fit(X_train, y_train)
```

Evaluation du modèle

Nous avons atteint un accuracy score de 45%

Et un cross_val_score (On découpe la base de données en cinq segments de façon aléatoire. On en utilise 4 pour l'apprentissage et 1 pour tester. On recommence 5 fois. Si le modèle est robuste, les cinq scores de test seront sensiblement égaux) d'environ 45% également.

```
cv_scores
```

```
array([0.44230769, 0.44230769, 0.44230769, 0.44230769, 0.44070513])
```

```
accuracy_score
```

```
0.45897435897435895
```