Etude Marketing : Analyse de la clientèle d'un grand magasin et définition de profils clients.

Projet - UV ODATA

27 octobre 2022

Objectifs du projet

Pour réussir sa stratégie commerciale et augmenter ses ventes, une entreprise doit connaître les différents profils de ses clients. **L'objectif du projet est d'analyser la clientèle d'un grand magasin et de créer des profils clients en effectuant un clustering des clients à partir de données collectées via des questionnaires.** Ces profils permettront à l'entreprise de mieux comprendre les habitudes d'achat de ses clients et d'optimiser ses efforts commerciaux en proposant des produits et des services ciblés en cohérence avec les attentes des différents types de clients.

Les méthodes de clustering à considérer / tester sont les suivantes :
- K-means
- Classification ascendante hiérarchique (CAH)
- Modèle de mélange de Gaussiennes
- DBSCAN
- Partitionnement spectral (Spectral clustering)

D'abord, une étude préalable vous permettra de comparer ces 5 méthodes sur des données simulées. Vous étudierez le principe des méthodes DBSCAN et Spectral clustering par vous-même, vous analyserez les performances des différentes techniques sur plusieurs jeux de données simulées et vous préciserez quelles sont les techniques les mieux adaptées à chaque type de données.

Ensuite, vous vous intéresserez à l'analyse et au clustering des clients d'un grand magasin à partir d'un fichier de données regroupant un certain nombre d'informations sur ces clients : informations sociologiques, habitudes d'achat, utilisation d'offres de réduction, lieux d'achat.

Vous analyserez les partitions obtenues par les différentes méthodes à l'aide des métriques de votre choix. Enfin vous proposerez une liste de profils clients et rédigerez une carte d'identité pour chacun des profils types identifiés.

Livrables attendus

A la fin du projet, vous devrez fournir 2 fichiers :

1. Un rapport (format pdf) comprenant les éléments suivants :
   — un rappel des objectifs du projet,
   — une description du protocole expérimental mis en place : objectifs, métriques utilisées, ... Pour chaque méthode vous préciserez les paramètres choisis et vous justifierez vos choix,
   — les réponses aux différentes questions posées,
   — une analyse et une interprétation des résultats obtenus.
2. Le code, clair et commenté (archive au format zip).

1 Etude préalable : Comparaison des méthodes de clustering sur des données simulées

1.1 Etude des méthodes DBSCAN et Spectral Clustering

Parmi les méthodes considérées, seules les méthodes DBSCAN et Spectral clustering n'ont pas été étudiées en cours, il faut donc faire une recherche sur ces méthodes et comprendre leur principe.

Expliquez de façon succincte le principe de chaque méthode et indiquez ses avantages.

1.2 Etude des classes et modules relatifs aux différentes méthodes

Dans le projet, vous allez considérer les classes ou modules suivants des librairies `scikit-learn` et `scipy` :

— K-means : `sklearn.cluster.KMeans`
— CAH : `scipy.cluster.hierarchy`
— Modèle de mélange de Gaussiennes : `sklearn.mixture.GaussianMixture`
— DBSCAN : `sklearn.cluster.DBSCAN`
— Spectral clustering : `sklearn.cluster.SpectralClustering`

Etudiez ces différentes classes et modules : paramètres d'appel, attributs, méthodes, fonctions.

1.3 Expérimentations

Il s'agit d'étudier les performances des 5 méthodes sur des données simulées correspondant à des clusters de formes différentes. Ces données sont stockées dans les fichiers suivants : `jain.txt`, `aggregation.txt` et `pathbased.txt`.

Chaque fichier contient 3 colonnes. Pour chaque individu (ou point), les 2 premières colonnes correspondent aux valeurs de 2 caractéristiques, la 3ème colonne indique sa classe d'appartenance. Cette information pourra servir de référence pour évaluer les partitions obtenues par clustering.

Importez les données et visualisez les nuages de points correspondants.

Pour chaque méthode :
- Précisez les valeurs choisies pour les paramètres, par exemple pour l'initialisation, le type de distance, la méthode de linkage en classification hiérarchique....
- Proposez un partitionnement de chaque jeu de données. Pour le nombre de clusters K, vous choisirez le nombre réel de classes.

Evaluez les performances des 5 méthodes :
- de façon qualitative : représentez les points des clusters obtenus par des couleurs différentes et comparez visuellement avec les "vraies" classes.
- de façon quantitative : comparez le partitionnement obtenu avec la vraie classification en calculant l'indice de Rand ajusté (ARI).

Comparez les performances obtenues avec les 5 méthodes. Précisez quelles sont les méthodes les mieux adaptées aux différentes formes de clusters.

2 Analyse et partitionnement de la clientèle d'un grand magasin

L'objectif est maintenant d'utiliser ces 5 méthodes pour classer automatiquement les clients d'un grand magasin en fonction de leur profil à partir de données collectées au travers de questionnaires.

L'intérêt de cette classification (non supervisée) est de guider la stratégie marketing de l'entreprise pour proposer des produits et des services ciblés en cohérence avec les attentes des différents profils de clients.

Le fichier `customer_database.csv` stocke pour 2240 clients les données présentées ci-après.

Informations concernant la fiche client :
— ID : Identifiant unique du client
— Registration : Date d'inscription du client dans le fichier clients
— Group : Numéro du groupe dans lequel la fiche client est classée.
— SubGroup : Numéro du sous-groupe dans lequel la fiche client est classée.
Informations sociologiques :
— BirthYear : Année de naissance du client
— Education : Niveau d'éducation du client
— CivilStatus : Situation de famille du client
— Income : Revenu annuel du ménage du client
— Kids : Nombre d'enfants dans le foyer du client
— Teens : Nombre d'adolescents dans le foyer du client
Habitudes d'achat (au cours des 2 dernières années) :
— LastPurchase : Nombre de jours depuis le dernier achat du client
— Wines : Montant dépensé pour du vin
— Fruits : Montant dépensé pour des fruits
— Meat : Montant dépensé pour de la viande
— Fish : Montant dépensé pour du poisson
— Sweet : Montant dépensé pour des sucreries
— Luxury : Montant dépensé pour des produits de luxe
— Claims : 1 si le client s'est plaint au cours des 2 dernières années, 0 sinon.
Lieux d'achat :
— WebPurchases : Nombre d'achats effectués sur le site web du magasin
— CatalogPurchases : Nombre d'achats effectués par le biais d'un catalogue
— StorePurchases : Nombre d'achats effectués directement en magasin
— WebVisits : Nombre de visites sur le site web de l'enseigne au cours du dernier mois
Utilisation d'offres de réduction :
— DiscountPurchases : Nombre d'achats effectués avec une réduction
— Accept1 : 1 si le client a accepté une offre lors de la 1ère période de réductions, 0 sinon.
— Accept2 : 1 si le client a accepté une offre lors de la 2ème période, 0 sinon.
— Accept3 : 1 si le client a accepté une offre lors de la 3ème période, 0 sinon.
— Accept4 : 1 si le client a accepté une offre lors de la 4ème période, 0 sinon.
— Accept5 : 1 si le client a accepté une offre lors de la 5ème période, 0 sinon.
— AcceptLast : 1 si le client a accepté une offre lors de la dernière période, 0 sinon.

Visualisez le tableau de données, puis importez les données du fichier.

Les sections suivantes proposent une ligne directrice pour l'analyse et le clustering des données. Mais **le projet est ouvert, donc soyez curieux, n'hésitez pas à proposer, tester, expérimenter... Votre démarche, vos idées, votre analyse comptent plus que le résultat de clustering lui-même. Donnez toutes les informations qui vous paraissent pertinentes pour enrichir la connaissance client de l'entreprise.**

2.1 Examen des données

Après l'importation, il est important d'examiner plus en détail ces données, en particulier :

- la taille du jeu de données,
- le type des données (numérique : int, float ou qualitatif/catégoriel : object),
- la qualité des données (est-ce qu'il y a des données manquantes ?),
- la distribution des données (est-ce qu'il y a des données aberrantes ?).

En utilisant les méthodes de la classe `DataFrame`, procédez à l'examen des données et notez les informations qui vous paraissent pertinentes.

En particulier, il est important d'identifier les données qualitatives, les données manquantes (représentées par le symbole 'NA' : Not Available) et les données aberrantes qui devront toutes être pré-traitées (voir section suivante). En utilisant la méthode `isna()` de la classe `DataFrame` et la fonction `sum()`, vous pouvez obtenir le nombre de valeurs manquantes pour chacune des variables.

Il est aussi intéressant de connaître les statistiques des données à traiter. Pour cela, vous pouvez utiliser la méthode `describe()` de la classe `DataFrame` et construire une visualisation de type histogramme pour chaque variable numérique avec la méthode `hist()` de la classe `DataFrame`. Vous pouvez ainsi identifier les éventuelles données aberrantes, c'est-à-dire en dehors de l'échelle de valeurs prises habituellement par une variable.

Après ce premier examen des données, quelles variables proposez-vous de conserver pour l'analyse et le clustering des données ?

Pour supprimer les colonnes qui ne vous paraissent pas utiles, vous pouvez utiliser la méthode `drop()` de la classe `DataFrame`.

2.2 Pré-traitement des données

Pour faire fonctionner correctement les algorithmes de clustering, il est nécessaire d'avoir des données numériques de bonne qualité.

Données manquantes et aberrantes : Pour résoudre le problème des valeurs manquantes et des valeurs aberrantes, plusieurs solutions sont possibles :

— rechercher la vraie valeur via d'autres sources d'information,
— attribuer une valeur conforme à la distribution de la variable : moyenne, médiane, valeur la plus probable... (avec la méthode `fillna()` de la classe `DataFrame` pour les valeurs manquantes),
— supprimer la variable correspondante, si le nombre de valeurs manquantes ou aberrantes est très important (plus d'un tiers des données environ).

Transformation des variables qualitatives en variables numériques : Les algorithmes d'apprentissage ne traitent que des grandeurs numériques. Il faut donc transformer les variables qualitatives en variables numériques. Dans le cas de variables booléennes, le remplacement peut se faire directement avec la méthode `replace()`. Dans les autres cas, il est plus facile de faire appel à la classe `LabelEncoder`.

Normalisation des données : Un dernier point concernant la préparation des données est le recalibrage des variables. Lorsque les variables numériques ont des échelles différentes, il est nécessaire de centrer et réduire les données, en utilisant par exemple la classe `StandardScaler`.

Après avoir réalisé toutes ces transformations, il est intéressant d'examiner à nouveau toutes les variables qui seront utilisées pour le clustering.

2.3 Recherche de corrélations

Pour mieux comprendre les données, il faut s'intéresser aux relations qui existent entre les variables. Pour cela, il faut calculer le coefficient de corrélation entre chaque couple de variables numériques, par exemple avec la méthode `corr()` de la classe `DataFrame`.

Comme le nombre de variables est assez grand, la matrice de corrélation peut être entièrement représentée avec la fonction `heatmap()` de la librairie `seaborn`. De façon générale, cette librairie propose des fonctions intéressantes pour la visualisation des données.

Commentez les résultats obtenus.

2.4 Analyse exploratoire des données

Avant d'appliquer les méthodes de clustering, il est intéressant d'effectuer une ACP pour aller plus loin dans l'analyse des données. L'ACP va permettre de mieux comprendre les relations (corrélations) entre les variables, de faire des regroupements de clients similaires et éventuellement de réduire la dimension des données.

Effectuez une ACP sur les données. Combien d'axes proposez-vous de conserver ?

Représentez la projection des clients (la totalité ou une partie pour plus de lisibilité) dans le(s) premier(s) plan(s) principal(aux) ainsi que la projection des variables dans le(s) cercle(s) des corrélations. Donnez une interprétation des axes conservés.

2.5 Clustering des données

Effectuez le clustering proprement dit sur les données de la clientèle de l'entreprise avec les 5 méthodes proposées.

Pour chaque méthode :
- Précisez les valeurs choisies pour les paramètres.
- Evaluez la qualité des partitions obtenues pour différentes valeurs de K, le nombre de clusters, en utilisant les métriques de votre choix disponibles dans le module `sklearn.metrics`.
- Proposez une valeur de K, justifiez votre choix,
- Pour la valeur de K choisie, analysez les clusters obtenus au regard des données du fichier.
- Visualisez graphiquement les clusters obtenus sur le premier plan principal défini par les 2 premiers axes de l'ACP.

Vous pouvez appliquer les algorithmes sur les données complètes (sans ACP) et/ou les données réduites obtenues après l'ACP.

A partir du meilleur résultat de clustering obtenu, vous proposerez une liste de profils clients et rédigerez une carte d'identité pour chacun des profils types identifiés.