ESP 2020

T-DAT-901

REPORT

Crée : 01/01/2022

Dernière modification : 30/01/2022

Eric Gadbin / Emma Rouzaud / Thomas Fournier

Morgan Druesne / Guilhem Lacombe

SOMMAIRE

I. Phase de travail

1.1 Analyse de la forme

Identification de la cible : définir la relation entre les customers et des items

Identification des valeurs manquantes : peu de valeurs manquantes, le dataset est relativement propre

Types de variables : 7 qualitatives (catégories) , 1 quantitative

1.2 Analyse de fond

<u>Compréhension des différentes variables et subtilités:</u>

- Plus d'univers que de mailles
- Un univers peut contenir la même maille qu'un autre univers
- Il y a une saisonnalité des ventes
- Quelques outliers qui se démarquent énormément des autres clients
- Les clients les plus récents sont les plus actifs

1.2.1 Décisions prises suite à l'analyse du dataframe

- Nous utilisons la MAILLE ainsi que la FAMILLE plutôt que l'UNIVERS qui n'a pas d'utilité dans notre cas.
- Par simplicité nous n'avons pas pris en compte la saisonnalité des ventes
- Nous considérons que les outliers sont des profil d'acheteur à part entière
- Malgré qu'ils soient peu actifs nous gardons les clients les plus anciens pour notre étude.

1.3 Pre-processing

Le but de cette phase est de supprimer les colonnes et valeurs qui ne nous seront pas utiles pour notre Recommender system, rendre le dataset plus facilement manipulable, utiliser uniquement les données dont nous avons besoin.

Parmi les lignes supprimés on trouve les doublons qui ne nous sont pas utiles comme par exemple un même produit acheté plusieurs fois sur un même ticket. (Uniquement pour la recommandation, pas pour le clustering où cette donnée est à prendre en compte.)

Nous n'utilisons pas l'UNIVERS mais plutôt la MAILLE ainsi que le FAMILLE pour nos procédés de recommandation.

Cela a nécessité dans certains cas l'encodage de certaines valeurs ainsi que la division du dataframe.

Pour notre Recommender system final, celui ci s'appuie sur deux dataframes

Un premier se concentrant sur les commandes, il subit un encodage des libellés et à pour objectif d'être celui utilisé par l'algorithme pour recommander des produits.

Le second se concentre sur les produits et conserve le LIBELLE, il nous permet de pouvoir mettre un nom sur les produits que nous recommandons.

Selon les différentes techniques d'approche nous nous sommes retrouvés avec des cas où le dataset contient les colonnes suivantes :

<u>Exemples de cas</u>

Cas 1 :

TICKET_ID	FAMILLE	UNIVERS	MAILLE	LIBELLE	CLI_ID	PRIX_NET	MOIS_VENTE

Cas 2:

TICKET_ID	FAMILLE	UNIVERS	MAILLE	LIBELLE	CLI_ID	ALL_LIBELLES	CODE_LIBELLE

NB: ALL_LIBELLES étant une liste de tous les libellés que l'on retrouve dans un TICKET_ID, CODE_LIBELLE étant un LIBELLE encodé.

Notre Recommender system final utilise l'algorithme **FP GROWTH**, très conseillé pour des problématiques tel que la recommandation de produits.

Cet algorithme est une amélioration de la méthode Apriori. Un modèle fréquent est généré sans qu'il soit nécessaire de générer des candidats. L'algorithme de croissance FP représente la base de données sous la forme d'un arbre appelé arbre de modèle fréquent ou arbre FP.

Cette arborescence maintiendra l'association entre les ensembles d'éléments. La base de données est fragmentée en utilisant un élément fréquent. Cette partie fragmentée est appelée «fragment de motif». Les itemsets de ces modèles fragmentés sont analysés. Ainsi, avec cette méthode, la recherche d'ensembles d'éléments fréquents est comparativement réduite.

II. Choix des technologies et algorithmes

2.1 Librairies

- Pandas
- Numpy
- Mlxtend
- Seaborn
- Sklearn
- Networkx

Algorithmes

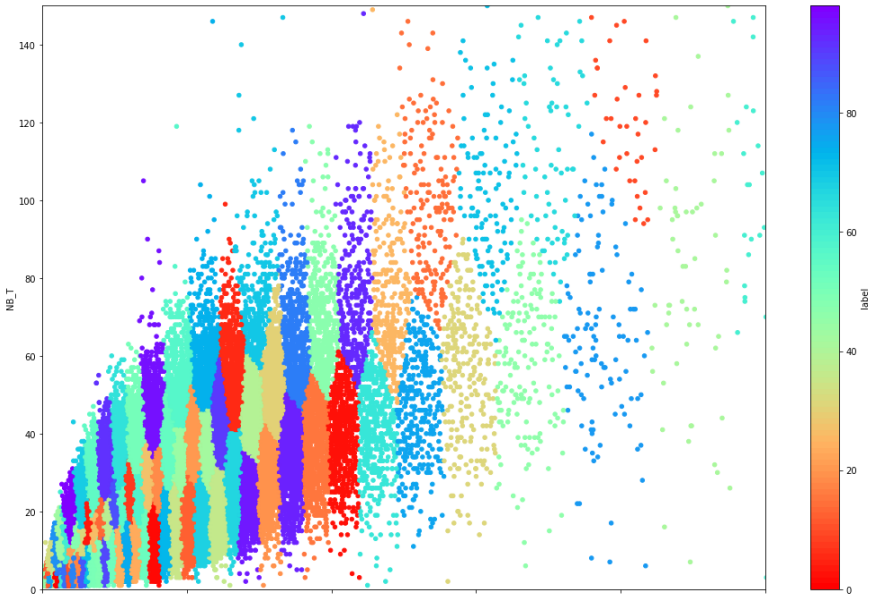
- JPGrowth
- Apriori
- Kmeans
- Kamada-Kawai

2.2 Clusterisation

Nous avons donc premièrement calculé le nombre d'items acheté ainsi que total dépensé sur l'année

Nous avons ensuite créé 100 clusters différents pour classer par nombre di'tems et total dépensé. Les clients seront donc mis dans un cluster qui nous servira de référence pour la suite.

Nombre d'item acheté par apport a total dépensé sur l'année



On remarque que la majorité des clients ont un profil plutôt peu dépensier et qu'il existe une relation linéaire, voir logarithmique entre le nombre d'item acheté et le total dépensé.

2.3 Systèmes de recommandation

- ➢ Recommandation de produits selon un CLI_ID ainsi qu'une maille choisie.
- ➢ Recommandation de produits en utilisant l'algorithme JPGrowth ainsi que les associations rules.
- ➢ Recommandation finale selon un CLI_ID donné, une liste de produits susceptibles de lui plaire.

III. Résultats obtenus

3.1 Résultats

Prenons le produit suivant : `GD JDM4 GRENADE FL200ML`

Les produits proposés à un client ayant acheté ce produit selon nos différents systèmes de recommandation:

<u>1er Recommender system:</u>

Basé sur un CLI_ID et une MAILLE selon les produits les plus achetés pour chaque maille.

```
Products the client bought :  {'CD JDM4 CAFE FL 200ML', 'GD JDM4 LAVANDIN DE PROVENCE 200ML', 'GD JDM4 GRENADE FL200ML', 'GD FLEUR DE TIARE JDM3 200ML', 'GD JDM4 ORANGE FL 200ML', 'GD JDM4 PAMPLEMOUSSE FL 200ML', 'GD FL200ML JDM PAMPLEMOUSSE', 'GD JDM4 TIARE FL 200ML', 'CD JDM4 COTON  FL 200ML'}
Top products for this maille :  {'GD JDM4 LOTUS FL200ML', 'GD JDM4 GRENADE FL200ML', 'GD JDM4 ORANGE FL 200ML', 'GD JDM4 PAMPLEMOUSSE FL 200ML', 'CD JDM4 MAGNOLIA FL 200ML', 'GD JDM4 TIARE FL 200ML', 'CD JDM4 COTON  FL 200ML', 'GD JDM4 CIT VERT FL 200ML', 'CD JDM4 RIZ VIOLET FL200 ML'}
Recommended products :  {'GD JDM4 LOTUS FL200ML', 'GD JDM4 CIT VERT FL 200ML', 'CD JDM4 MAGNOLIA FL 200ML', 'CD JDM4 RIZ VIOLET FL200 ML'}
```

Nous obtenons comme produits recommandés pour notre premier système de recommandation :

- **GD JDM4 LOTUS FL200ML**
- **GD JDM4 CIT VERT FL 200ML**
- **CD JDM4 MAGNOLIA FL 200ML**
- **CD JDM4 RIZ VIOLET FL200ML**

2ème Recommender system :

BASÉ SUR LE TICKET_ID ET LE LIBELLE (Encodé) - FP GROWTH

```
Original product : GD JDM4 GRENADE FL200ML

Recommended products :
GD FL200ML JDM PAMPLEMOUSSE
GD JDM4 PAMPLEMOUSSE FL 200ML
GD JDM4 CIT VERT FL 200ML
GD JDM4 LOTUS FL200ML
GD JDM4 ORANGE FL 200ML
```

Produits proposés :

- **GD FL200ML JDM PAMPLEMOUSSE**
- **GD JDM4 PAMPLEMOUSSE FL 200ML**
- **GD JDM4 CIT VERT FL 200ML**
- **GD JDM4 LOTUS FL200ML**
- **GD JDM4 ORANGE FL 200ML**

NB: Les produits en rouge sont ceux que l'ont retrouvent aussi dans les autres Recommender systems que nous avons mis en place

BONUS :

SPIDERGRAM PERMETTANT DE VISUALISER LES PRODUITS LES PLUS ACHETÉS EN RELATION PROCHE