SolutionStat
CONSULTATION ET FORMATION

Analyse du panier d'épicerie

Présentation d'une méthodologie d'analyse

Mise en contexte

- En data mining, on utilise la technique des règles d'association pour déterminer les éléments qui se retrouvent ensembles.
- L'analyse du panier d'épicerie (« market basket analysis ») est un terme plus spécifique au commerce au détail. Cette analyse utilise les règles d'association.
- Dans une épicerie, les règles d'association décrivent les produits qui se retrouvent dans le même panier.

Beurre d'arachides → **Pain en tranches**

Définitions

- Définitions
  - **Transactions** : achats fait par un seul client.
  - **Items** : produits achetés.
  - **Règle d'association** : énoncé de la forme (item X) ⇨ (item Y).
    - Item X = produit à analyser
    - Item Y = produit associé

- Règle d'association :
  - On choisira d'étudier des règles d'association permettant d'en apprendre davantage sur le comportement des clients. Les résultats de l'analyse devront être utiles et pratiques.
  - On choisira un niveau de granularité. On peut étudier l'association entre des ensembles de produits : ceux qui achètent des céréales achètent aussi du lait. Ou l'association entre des produits plus précis : ceux qui achètent du vin rouge bon marché achètent des cubes de bœuf pour ragoût.

Mesures

- La force d'association sera mesurée par :
  - **Support** : probabilité d'acheter le produit X et le produit Y.

$$\frac{\text{Nombre de transactions contenant les produits X et Y}}{\text{Nombre total de transactions}}$$

  - **Confiance** : probabilité d'acheter le produit Y étant donné que le produit X a été acheté.

$$\frac{\text{Nombre de transactions contenant les produits X et Y}}{\text{Nombre de transactions contenant le produit X}}$$

Règle	Support	Confiance
X ⇨ Y	25%	59%
X ⇨ Z	5%	33%

Lift

- Calcul du lift :
  - Le lift est une bonne mesure de performance de la règle d'association.
  - Le lift est la confiance de la règle divisée par la valeur espérée de la confiance.
  - Pour la règle d'association X ⇨ Y, le calcul de la valeur espérée de la confiance est le suivant :

$$\frac{\text{Nombre de transactions contenant le produit Y}}{\text{Nombre total de transactions}}$$

Règle	Support	Confiance	Confiance espérée	Lift
X ⇨ Y	25%	59%	42.5%	1.31
X ⇨ Z	5%	33%	45%	0.74

Lift (suite)

- Interprétation du lift
  - Un lift supérieur à 1 :
    - Indique une corrélation positive;
    - Parmi les paniers contenant le produit X, on retrouve plus souvent le produit Y que dans l'ensemble des paniers.
  - Un lift de 1 indique une corrélation nulle;
  - Un lift inférieur à 1
    - Indique une corrélation négative;
    - Dans cette situation, la règle négative aura plus d'intérêt : X ⇨ non Y.

Règle	Support	Confiance	Confiance espérée	Lift
X ⇨ Y	25%	0.59	42.5%	1.31
X ⇨ Z	5%	0.33	45%	0.74

Exemple pratique

- Une pharmacie est intéressée à connaître les associations entre les achats de cosmétiques et les autres produits (journaux, crayons, bonbons,…). Une base de données contient 400,000 transactions effectuées dans les derniers 3 mois. Les produits suivants sont étudiés :

Bar soap	Markers	Prescription medications
Cosmetic	Pain relievers	Shampoo
Candy bars	Pencils	Toothbrushes
Deodorant	Pens	Toothpaste
Greeting cards	Perfume	Wrapping paper
magazines	Photo processing	

Exemple pratique

- La base de données contient 4 variables
  - **Transaction ID** : identificateur unique de la transaction
  - **Product** : Catégorie de produits
  - **Quantity** : quantité de produits de cette catégorie acheté
  - **Store** : numéro du magasin

Quantity	Transaction ID	Store	Product
1	12359	2	CandyBar
2	12362	9	PainReliever
2	12362	9	PainReliever
1	12365	5	Toothpaste
2	12371	2	Cosmetics
1	12380	6	GreetingCards
3	12383	1	PainReliever
3	12383	1	PainReliever
1	12386	7	PainReliever
2	12386	7	PainReliever
1	12392	7	Shampoo
1	12392	7	Magazine

Les variables **transactionID** et **Product** seront utilisées pour les calculs.

Note : la base de données est construite de sorte que chaque item est sur une ligne différente. Le numéro d'une transaction peut se répéter sur plusieurs lignes si plus d'un item a été acheté.

Exemple pratique

- Calcul des indicateurs et du lift pour chacune des règles d'association :
  - Utilisation d'un programme SAS permettant de gérer les données, de faire les calculs et d'obtenir une table de résultats.
  - La table d'entrée doit avoir au minimum deux variables. Une variable doit identifier les transactions et une variables doit identifier les items.
  - La table de sortie contiendra les résultats pour toutes les associations de deux produits.

ANALYSIS_UNIT	ANALYSIS_UNIT_FREQ	ASSOC_ANALYSIS_UNIT	ASSOC_ANALYSIS_UNIT_FREQ	FREQ_CO_OCCU	SUPPORT	CONFIDENCE	EXPECTED_CONFIDENCE	LIFT
Cosmetics	10929	Toothbrush	13470	2268	0.0113	0.2075	0.0674	3.08
Cosmetics	10929	PrescriptionMed	2901	267	0.0013	0.0244	0.0145	1.68
Cosmetics	10929	Shampoo	6760	518	0.0026	0.0474	0.0338	1.4
Cosmetics	10929	WrappingPaper	10198	759	0.0038	0.0694	0.051	1.36
Cosmetics	10929	Perfume	17992	907	0.0045	0.083	0.09	0.92
Cosmetics	10929	PhotoProcessing	11696	474	0.0024	0.0434	0.0585	0.74
Cosmetics	10929	Soap	8605	344	0.0017	0.0315	0.043	0.73
Cosmetics	10929	Deodorant	1084	33	0.0002	0.003	0.0054	0.56
Cosmetics	10929	Magazine	48261	1447	0.0072	0.1324	0.2413	0.55
Cosmetics	10929	Toothpaste	32085	813	0.0041	0.0744	0.1604	0.46
Cosmetics	10929	Markers	1614	35	0.0002	0.0032	0.0081	0.4
Cosmetics	10929	CandyBar	34201	566	0.0028	0.0518	0.171	0.3
Cosmetics	10929	GreetingCards	29377	468	0.0023	0.0428	0.1469	0.29
Cosmetics	10929	Pens	28715	439	0.0022	0.0402	0.1436	0.28
Cosmetics	10929	PainReliever	5340	44	0.0002	0.004	0.0267	0.15
Cosmetics	10929	Pencils	26985	216	0.0011	0.0198	0.1349	0.15
CandyBar	34201	GreetingCards	29377	8732	0.0437	0.2553	0.1469	1.74
CandyBar	34201	Toothpaste	32085	7956	0.0398	0.2326	0.1604	1.45
CandyBar	34201	Pencils	26985	6603	0.033	0.1931	0.1349	1.43

Exemple pratique

- Table de sortie :
  - ANALYSIS_UNIT : produit A (condition)
  - ANALYSIS_UNIT_FREQ : nombre de transactions avec le produit A
  - ASSOCIATION_ANALYSIS_UNIT : produit B (résultat)
  - ASSOCIATION_ANALYSIS_UNIT _FREQ: nombre de transactions avec le produit B
  - FREQ_CO_OCCUR : nombre de transactions avec les produits A et B.

ANALYSIS_UNIT	ANALYSIS_UNIT_FREQ	ASSOC_ANALYSIS_UNIT	ASSOC_ANALYSIS_UNIT_FREQ	FREQ_CO_OCCU
Cosmetics	10929	Toothbrush	13470	2268
Cosmetics	10929	PrescriptionMed	2901	267
Cosmetics	10929	Shampoo	6760	518
Cosmetics	10929	WrappingPaper	10198	759
Cosmetics	10929	Perfume	17992	907
Cosmetics	10929	PhotoProcessing	11696	474
Cosmetics	10929	Soap	8605	344

Exemple pratique

- Table de sortie :
  - SUPPORT : support
  - CONFIDENCE : confiance
  - EXPECTED_CONFIDENCE : valeur espérée de la confiance
  - LIFT: calcul du lift

ANALYSIS_UNIT	ANALYSIS_	ASSOC_ANALYSIS	SUPPORT	CONFIDENCE	EXPECTED_CONFIDENCE	LIFT
Cosmetics	10929	Toothbrush	0.0113	0.2075	0.0674	3.08
Cosmetics	10929	PrescriptionMed	0.0013	0.0244	0.0145	1.68
Cosmetics	10929	Shampoo	0.0026	0.0474	0.0338	1.4
Cosmetics	10929	WrappingPaper	0.0038	0.0694	0.051	1.36
Cosmetics	10929	Perfume	0.0045	0.083	0.09	0.92
Cosmetics	10929	PhotoProcessing	0.0024	0.0434	0.0585	0.74
Cosmetics	10929	Soap	0.0017	0.0315	0.043	0.73
Cosmetics	10929	Deodorant	0.0002	0.003	0.0054	0.56
Cosmetics	10929	Magazine	0.0072	0.1324	0.2413	0.55
Cosmetics	10929	Toothpaste	0.0041	0.0744	0.1604	0.46

Exemple pratique

- Table de sortie :
  - La table est triée selon le produit A (ANALYSIS_UNIT).
  - Pour un même produit A, les résultats sont classés selon le LIFT.
  - La meilleure règle d'association sera en premier.

ANALYSIS_UNIT	ANALYSIS_	ASSOC_ANALYSIS	SUPPORT	CONFIDENCE	EXPECTED_CONFIDENCE	LIFT
Cosmetics	10929	Toothbrush	0.0113	0.2075	0.0674	3.08
Cosmetics	10929	PrescriptionMed	0.0013	0.0244	0.0145	1.68
Cosmetics	10929	Shampoo	0.0026	0.0474	0.0338	1.4
Cosmetics	10929	WrappingPaper	0.0038	0.0694	0.051	1.36
Cosmetics	10929	Perfume	0.0045	0.083	0.09	0.92
Cosmetics	10929	PhotoProcessing	0.0024	0.0434	0.0585	0.74
Cosmetics	10929	Soap	0.0017	0.0315	0.043	0.73
Cosmetics	10929	Deodorant	0.0002	0.003	0.0054	0.56
Cosmetics	10929	Magazine	0.0072	0.1324	0.2413	0.55
Cosmetics	10929	Toothpaste	0.0041	0.0744	0.1604	0.46

Exemple pratique

- Table de sortie :
  - Dans notre exemple la règle d'association PrescriptionMed ⇨ WrappingPaper obtient le lift le plus élevé.

ANALYSIS_UNIT	ANALYSIS_	ASSOC_ANALYSIS_	SUPPORT	CONFIDENCE	EXPECTED_CONFIDENCE	LIFT
PrescriptionMed	2901	WrappingPaper	0.0061	0.4188	0.051	8.21
PrescriptionMed	2901	Cosmetics	0.0013	0.092	0.0546	1.68
PrescriptionMed	2901	Markers	0.0002	0.0128	0.0081	1.58
PrescriptionMed	2901	Shampoo	0.0005	0.0365	0.0338	1.08
PrescriptionMed	2901	Toothbrush	0.001	0.07	0.0674	1.04

  - 0.6% des transactions contiennent les deux produits (support);
  - 5% des transactions contiennent du WrappingPaper;
  - 41.9% des transactions avec PrescriptionMed contiennent aussi du WrappingPaper.

Mises en garde

- Pour un commerce au détail, le nombre de règles d'association possibles est souvent énorme. Vouloir étudier toutes les associations entre des produits à un niveau très fin de granularité amènerait à des résultats non interprétables. Pour obtenir des résultats cohérents et utiles, il faut tout d'abord faire une liste pertinente des règles d'association d'intérêt.

- Si le support est petit, il faut se questionner sur l'intérêt de la règle d'association. En pratique, on peut fixer un support minimum requis et exclure les règles d'association n'ayant pas le support requis.

- Un niveau de confiance très élevé ou très faible peut aussi gonfler (ou réduire) artificiellement le lift.

Mises en garde

- L’objectif d’étudier les produits concomitants est de mieux comprendre une dynamique du comportement du client. En d’autres mots, on veut découvrir des associations non connues et prendre des décisions d’affaires basées sur ces nouvelles connaissances.

- Les règles qui obtiennent un bon support, une bonne confiance et un bon lift sont potentiellement utiles. Toutefois, ces règles peuvent être triviales, inexplicables ou difficiles à traduire en actions concrètes. Il faut au départ bien choisir les règles à étudier.