**Devoir de cours de l’introduction à l’exploitation de données**

**Niloofar Sokhandan Asl**

**11278920**

**Le 20 Juin 2020**

**Problème 1 :**

1. Veuillez interprétez la règle d’association numéro 3 (support, confiance, lift). D’après vous, comment vous caractériseriez cette règle et est-ce une règle intéressante pour vos futures promotions ? (Expliquez votre réponse).

* Support = 2% :

2% des clients achètent les gans hivernaux et le Manteau avec doublures.

* Confiance = 82% :

82% des clients qui achètent les gans hivernaux achètent aussi le Manteau avec doublures

* Lift = 3.4 :

Les clients qui achètent des gans hivernaux ont 3.4 fois plus de chance d’acheter le Manteau avec doublures que les clients en général.

La valeur de lift est supérieure à 1.15, donc il s’agit d’une association positive et donc d’une règle intéressante à considérer pour de future promotion.

1. Un de vos partenaires marketing se pose beaucoup de questions sur la règle d’association numéro 2. Est-ce une règle intéressante pour la prochaine promotion ? (Expliquez votre réponse en quelques lignes).

La valeur de lift de la règle 2 est 1.1. Cette valeur est supérieure à 1 et donc il y a une association positive entre l’achat de Bracelet en Or et l’achat de Chapeau de style MA, mais quand même cette valeur est inférieure à 1.15. Normalement, si la valeur de lift est inférieure à 1.15, l’association n’est pas forte et il ne vaut pas la peine d’en créer une promotion. Plus la valeur de lift (lift > 1) est proche de 1, moins la promotion sera performante. En fait, en comparant le lift de règle 2 avec le lift d’autres règles, on remarque que le lift de règle 2 est inférieur aux tous autres lifts. Aussi, la valeur de Support de règle 2 est seulement 13%, c.-à-d., seulement 13% des clients achètent les Bracelet en Or et le Chapeau de style MA. Donc ce n’est pas vraiment une bonne décision de créer une promotion basée sur règle 2.

1. Parmi l’ensemble des règles d’associations calculées, veuillez choisir la meilleure règle pour la prochaine promotion et veuillez décrire en quelques lignes en quoi consisterait cette promotion.

La règle 9 est la meilleure règle car son lift est plus grand que 1.15 et son support est 27% qui est le support les plus élevés des règles disponibles. C.-à-d. que déjà 27% des clients achètent des Londres et Parapluie et font achat en lignes. Selon cette règle, les clients qui achètent des Londres ont 1.9 fois plus de chance d’acheter le parapluie et de faire l’achat en ligne que les clients en général. La confiance de cette règle est aussi élevée, c.-à-d., 82% des clients achètent des Londres et le parapluie et ils font l’achat en ligne. Donc, en créant une promotion sur ces 2 items, on peut augmenter la chance d’achat de Parapluie et l’achat en lignes de 1.9 fois. Il est un bon investissement de créer une telle promotion.

1. Vous possédez un petit échantillon sur lequel vous voulez effectuer des règles d’association. Cet échantillon comporte des transactions uniques indiquant la présence d’un article (oui ou non). De plus, il comporte la colonne Montréal indiquant si la personne habite à Montréal. À l’aide du tableau, veuillez déterminer le support, la confiance et le lift des règles d’association suivantes :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 |
| tansaction ID | manteau | chaussures | pantalon | chemises | ceinture | montreal |
| #v25654 | oui | non | non | non | non | non |
| #v54542 | non | oui | oui | oui | non | non |
| #v57545 | non | non | oui | oui | non | oui |
| #v86745 | non | oui | oui | oui | oui | non |
| #v64542 | oui | oui | oui | oui | oui | non |
| #v85163 | oui | non | oui | oui | non | oui |
| #v54258 | oui | non | oui | non | non | non |
| #v85963page2image56518784 | oui | oui | oui | non | non | oui |
| #v54212page2image56510912page2image56509184page2image56514368page2image56509568page2image56518208page2image56513216page2image56503360page2image56503552 | oui | non | oui | oui | non | non |
| #v54265 | oui | non | non | oui | oui | oui |
| #v79745 | oui | oui | oui | oui | oui | oui |
| #v65342 | non | oui | oui | non | non | non |

# de transactions totales = 12

1. Chaussures 🡪 Pantalon

X=2, Y=3

|  |  |  |
| --- | --- | --- |
| Support (X=> Y) | Confiance (X => Y) | Lift |
| 50% | 100% | 1.2 |

1. Pantalon 🡪 Chaussures

X=3, Y=2

|  |  |  |
| --- | --- | --- |
| Support (X=> Y) | Confiance (X => Y) | Lift |
| 83% | 60% | 1.2 |

1. Montréal & Manteau 🡪 Chemises

X=1&6, Y=4

|  |  |  |
| --- | --- | --- |
| Support (X=> Y) | Confiance (X => Y) | Lift |
| 33% | 75% | 1.1 |

**Problème 2 :**

1. Les directeurs vous posent quelques questions afin de mieux comprendre le contexte :
   1. Quel est le produit qui se retrouve le plus au sein des transactions ?

StockCode=85123A, Frequency= 0.1067, Description= cream/white hanging heart t-light holder

* 1. Quelle est la provenance engendrant le plus de transactions ?

"United Kingdom"

* 1. Quel est le produit le plus rentable au sein des transactions ?

StockCode= DOT, Description= doctom postage

J’ai fait une recherche et voilà ce que j’ai trouvé concernant :

‘DOTCOM POSTAGE’ seems to indicate the amount spent by the customer on postage. Postage isn’t a direct indicator of sales and might skew the amount spent across cohorts.

Donc, si doctom postage n’est pas un produit, je prends le deuxième item rentable dans ma liste :

StockCode= 22423, Description= regency cakestand 3 tier

* 1. Quel est le client ayant fait le plus de visites ?

CustomerID= 12748

1. À l’aide des règles d’association et de l’algorithme Apriori, veuillez déterminer 3 règles que vous pensez intéressantes pour la compagnie. Veuillez donner une interprétation à ces règles (sur toutes les mesures) et justifier en quoi elles sont intéressantes. Enfin veuillez proposer une action avec chacune d’entre elles.

* Règle 1 :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Antecedent (X) | Consequent (Y) | Support | Confiance (X => Y) | Lift |
| {22698, 22699} | {22697} | 0.02104014 | 0.8944954 | 23.99 |

22698 = pink regency teacup and saucer

22699 = roses regency teacup and saucer

22697 = Green regency teacup and saucer

Interpretation:

* 2.1% des clients achètent le 22698, 22699 et 22697
* 89.4% des clients qui achètent 22698, 22699 achètent aussi 22697
* Les clients qui achètent 22698 et 22699 ont 23.99 fois plus de chance d’acheter 22697 que les clients en général.
* Règle 2 :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Antecedent (X) | Consequent (Y) | Support | Confiance (X => Y) | Lift |
| {22698} | {22697} | 0.02481657 | 0.8273381 | 22.19 |

22698 = pink regency teacup and saucer

22697 = Green regency teacup and saucer

Interpretation:

* 2.48% des clients achètent le 22698 et 22697
* 82.7% des clients qui achètent 22698 achètent aussi 22697
* Les clients qui achètent 22698 ont 22.19 fois plus de chance d’acheter 22697 que les clients en général.
* Règle 3 :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Antecedent (X) | Consequent (Y) | Support | Confiance (X => Y) | Lift |
| {23300} | {23301} | 0.02497842 | 0.7291339 | 17.87 |

23300 = gardener kneeling pad cup of tea

23301 = gardener kneeling pad keep calm

Interpretation:

* 2.49% des clients achètent le 23300 et 23301
* 72.9% des clients qui achètent 23300 achètent aussi 23301
* Les clients qui achètent 23300 ont 17.87 fois plus de chance d’acheter 23301 que les clients en général.

**Problème 3 :**

En vous basant sur la table de données présentée, veuillez déterminer la valeur de fidélité (la variable cible dans ce problème) d’Alice en analysant quelles sont les 6 femmes (K=6) les plus proches d’elle. Pour calculer, la distance nous vous demandons d’utiliser seulement les variables quantitatives.

**Problème 4 :**

1. En quelques lignes, veuillez expliquer pourquoi nous cherchons à minimiser la variance et donc à ne pas la maximiser.

On rappelle pour avoir un modèle de l’arbre de décision performant, on cherche l’homogénéité maximale dans un feuille. R2= SSB/SST =1- SSW/SST= SSB/SST

Dans le cas où on a les variables explicatives continues, ces variables doivent être discrétises avant d’être utilisées dans le modèle. R2 devient un indice d’homogénéité de chaque nœud de l’arbre.

Pour des valeurs plus petites de SSW, R2 est plus élevé, c.-à-d., l’homogénéité de la feuille est plus élevée. C’est ce qu’on cherche dans une feuille de l’arbre de régression : d’avoir une homogénéité maximale.

En regardant la formule de SSW, on se rend compte que SSW inclut le terme de variance. Donc, pour avoir une homogénéité maximale (R2 plus élevé), il faut minimiser le SSW, ce qui veut dire minimiser la variance.

1. En vous basant sur les données présentées, nous vous demandons de trouver la première variable importante afin de prédire la variable « Dépense ». Plus précisément, nous vous demandons de déterminer quelle sera la première variable à utiliser entre les variables :« Provenance », « Age » et « Sexe ».

Après avoir choisi la première variable importante, veuillez donner une interprétation pour chaque modalité.

La première variable c’est la variable âge, car sa réduction de variation est plus grande que celle des variables Provenance et Sexe. En d’autres termes, en discrétisant la variable cible de Dépense par variable de l’âge, on augmente l’homogénéité du modèle.

Le tableau ci-dessous compare la réduction de variabilité des variables Provenance, Age et Sexe.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable explicative | Variation Pondérée de Dépense | Variation Globale | Réduction de variation |
| Sexe | 1081 | 1118 | 36 |
| Provenance | 1072 | 45.8 |
| Age | 1071 | 46.4 |

IV(N)=1|S|2∑i∈S∑j∈S12(xi−xj)2−(1|St|2∑i∈St∑j∈St12(xi−xj)2+1|Sf|2∑i∈Sf∑j∈Sf12(xi−xj)2)

<http://www.appliedaisystems.com/papers/RegressionTrees.pdf>

<https://sefiks.com/2018/08/28/a-step-by-step-regression-decision-tree-example/>

Help :

Pour les variables explicatives (X) quantitatives, ça va être tjrs le Gini. Il y a un calcul supplémentaire qui va dire, pour cette variable-là c quoi le meilleur point de coupure. Et le point de coupure va être aussi calculé à l’intérieur de Gini etc. pour par exemple décider que le meilleur point de coupure est 30. Puis aux finales, on va dire est-ce que la valeur est < 30 ou > 30. Mais pour faciliter (comme dans le devoir) on peut facilement basé sur la moyenne.

Est-ce que la variable cible (Y) doit être tjrs qualitative ? on a 2 concepts : ADC et ADR

* Arbre de décision 🡪 classification 🡪 le fait de classer 🡪 on induit que Y est qualitative. Indice de Gini se fait que sur les variables qualitatives 🡪 évidement quand on parle de l’arbre de décision c’est sur les variables qualitatives nominales.
* Arbre de régression 🡪 Y prendre des variables quantitatives cette fois-ci. Ça c’est l’objectif de devoir. Il n’a rien avec la régression linéaire. C exactement même concept de l’arbre de classification sauf qu’au lien de Gini et entropie et gain informationnel, il y qqch que vous connaissez bien, vous allez devoir travailler avec pour arriver à faire le même concept d’homogénéité etc.

Devoir : Ex2 sur R seulement = méthode Kimen on n’a pas vu. Faire des recherches. k-mean apprentissage non-supervise, kinin est supervisé, utilisé pr prévision. Nettoyage 🡪 stack overflow

Devoir future : Trouver une commande pour passer de data frame a transaction

jusqu’a 20 juin

une semaine avant le finale

20%

4 problems:

2 : règles d’asso - application de réglé d’asso avec R

1: réglé de distance (calcul de distance - un algo)

1: arbre de regression (arbre de classification)