Eliott Loop,
Grégoire de Crombrugghe

TP2-Data Mining

Introduction

Lors du premier TP, de data mining, il nous a été demandé de faire une analyse “visuelle”. A la suite de ce TP nous en avions conclu qu’il n’était pas possible de déterminer, à l’œil nu, un attribut “décisionnel” (qui permet de prendre une décision sur le type de produit à proposer).

Pour ce deuxième TP, il nous est demandé d’analyser une seconde fois les différents attributs mais, cette fois-ci, avec une approche quantitative. En effet, il nous est demandé d’implémenter le code R calculant l’Entropie, l’Entropie Conditionnel et l’information gain. Et d’ensuite analyser certaine donnée.

Description du problème

Une Banque, nous demande de réaliser un modèle qui permettra de déterminer quel produit d’investissement à proposer à un client, suivant son profile. La Banque nous met à disposition une série de données, qui doit être analysé afin de déterminer les attributs dites “utile“ et de pouvoir construire ce modèle.

Attribut continu: SE1,BA1-BA7
Attribut discret: SE2,PE1-PE15, IA1-IA3
Variable cible: InvType

Eliott Loop,
Grégoire de Crombrugghe

Données Obtenus

Information gain=$\Delta H(y, x) = H(y) - H(y|x)$.
Information gain ratio = I(f,x)/H(f) (pour les attribut continu, f est la valeur pour laquelle l'information Gain est le plus grand. H(f) étant l'entropie normalisé).

Attribut	Information gain	Information gain ratio	Value (pour les attribut continu,la valeur pour laquelle l'information Gain est le plus grand)
BA3	0,144456733	0,14737084	18899
BA7	0,10164385	0,102630852	16371
BA5	0,055975917	0,059459887	16396
BA4	0,048956529	0,053012718	16396
SE1	0,03652936	0,041097321	53
SE2	0,021894631	0,026735939	/
BA1	0,009160626	0,038015086	24
IA3	0,005466622	0,013835623	/
PE10	0,004581686	0,00519914	/
PE12	0,003553709	0,004930615	/
PE15	0,002822064	0,008575454	/
PE9	0,002774747	0,011157604	/
PE5	0,002721656	0,003060283	/

PE8	0,00185796	0,004778081	/
PE4	0,001739443	0,032306844	/
PE6	0,001527011	0,002784296	/
BA6	0,001479024	0,001810853	/
PE13	0,001186667	0,001418057	/
PE11	0,001138118	0,001553856	/
IA1	0,000806219	0,04549336	/
BA2	0,000750181	0,008465243	12568,2
PE3	0,000704663	0,010932269	/
PE7	0,000531	0,001529305	/
IA2	0,000205024	0,071097941	/
PE14	0,000189486	0,000465755	/
PE2	7,43E-05	0,002407872	/
PE1	3,5113041490508800000E-05	0,000123651	/

Au vu des données ci-dessus, nous constatons que ce sont BA5,BA3,BA7 qui ont l'information gain le plus élevé. Toutefois, il est difficile de prédire si ces attributs sont utile à la "prise de décision". Nous les développons ci-dessous.

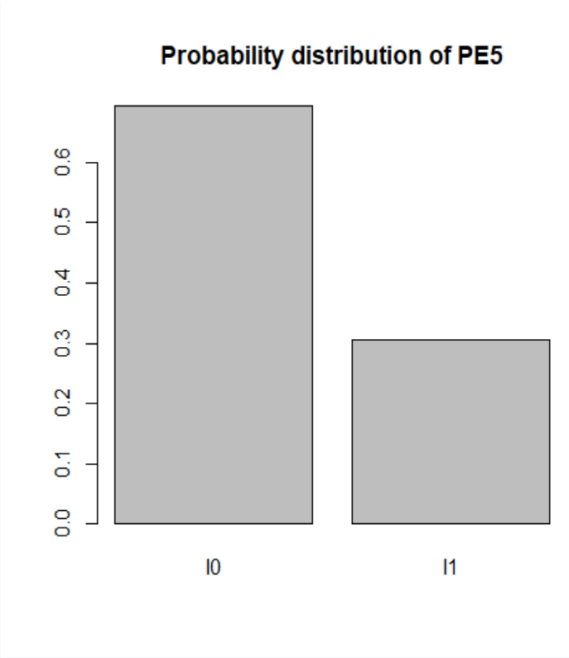
Eliott Loop,
Grégoire de Crombrugghe

Analyse entropie normalisée

Entropie la plus grande :

Variable : PE5

Valeur : 0.889347808279643

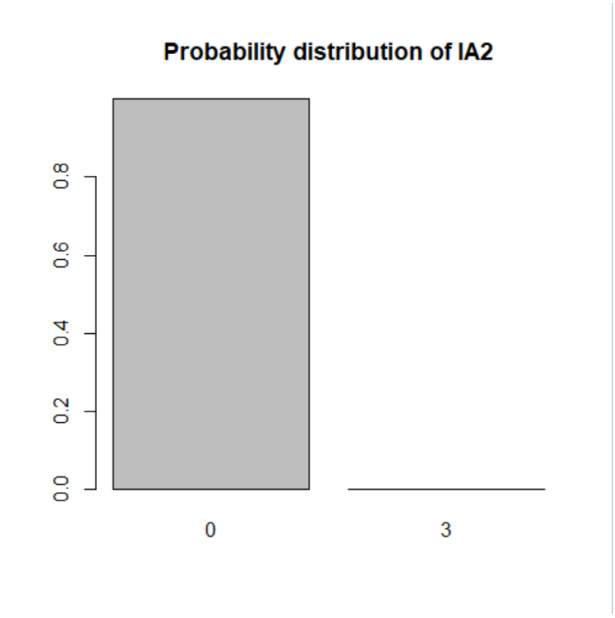


L'Attribut PE5 possède une entropie élevée. En effet il apparaît clairement, sur les graphes que, prédire la valeur que peut prendre cet attribut est plus compliqué que pour l'attribut IA2.(Elle est donc moins "purs" car plus de chance de se tromper). En effet, il y a 69% de chance de "tomber" sur la valeur I0 et 31% de chance sur la valeur I1.

Entropie la plus petite :

Variable : IA2

Valeur : 0.00288368961660139

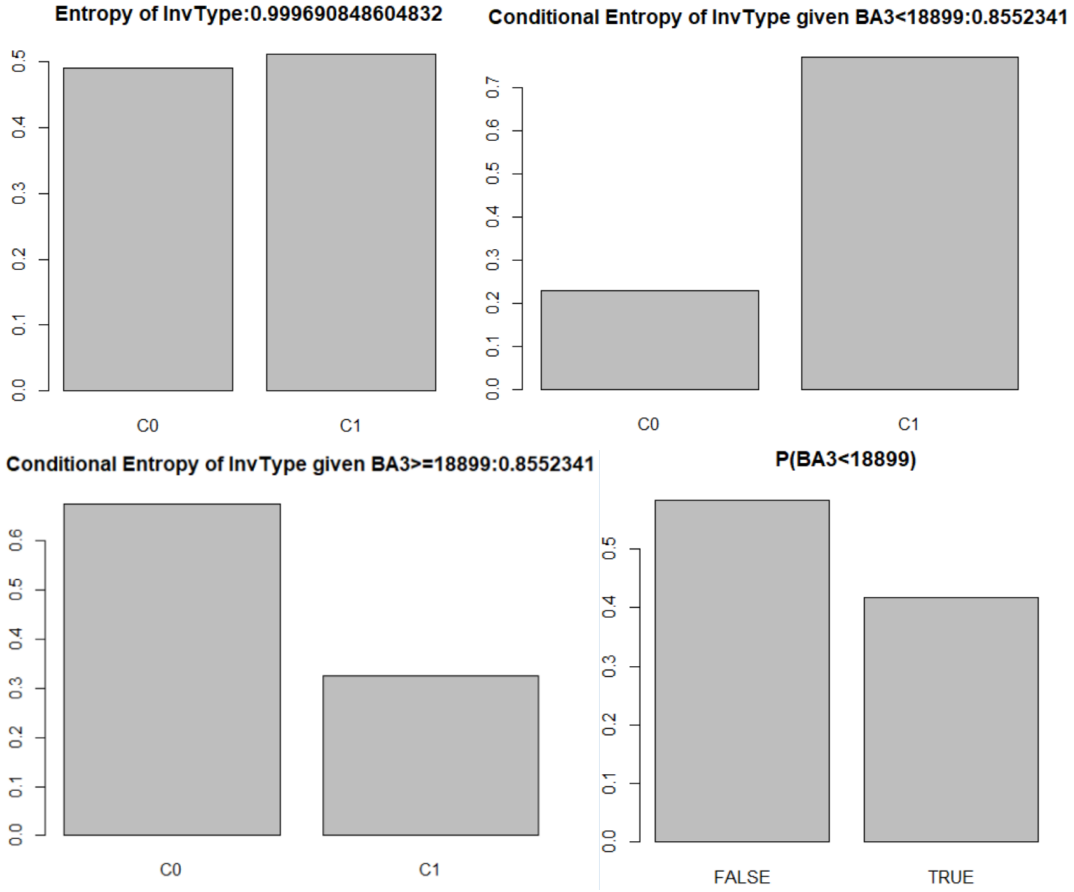


Eliott Loop,
Grégoire de Crombrugghe

Une petite entropie veut dire que les données de la variable sont purs. Pour IA2 on voit très bien que la très grande majorité des donnees,99% des données, ont la valeur 0.

Analyse information gain

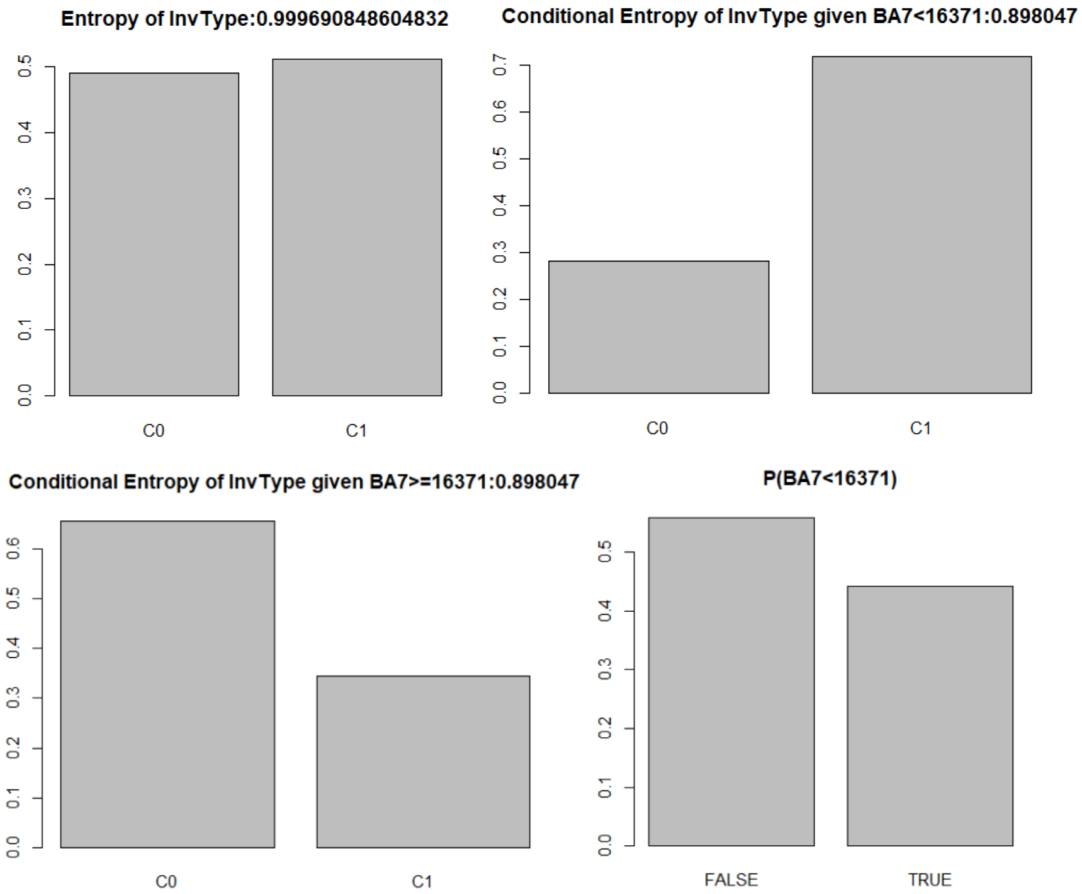
Max : BA3,BA7,BA5
Variable : BA3
Valeur : 0.144456733343609



Nous observons, pour l'attribut BA3, que la probabilité qu'une personne investisse dans le produit C1 est de 77%, lorsque sa valeur pour BA3 est plus petite que 18899 (et 23% pour C0). Lorsqu'elle est plus grande, la probabilité, qu'une personne ai investis dans C1, est de 32% (et 68% pour C0).

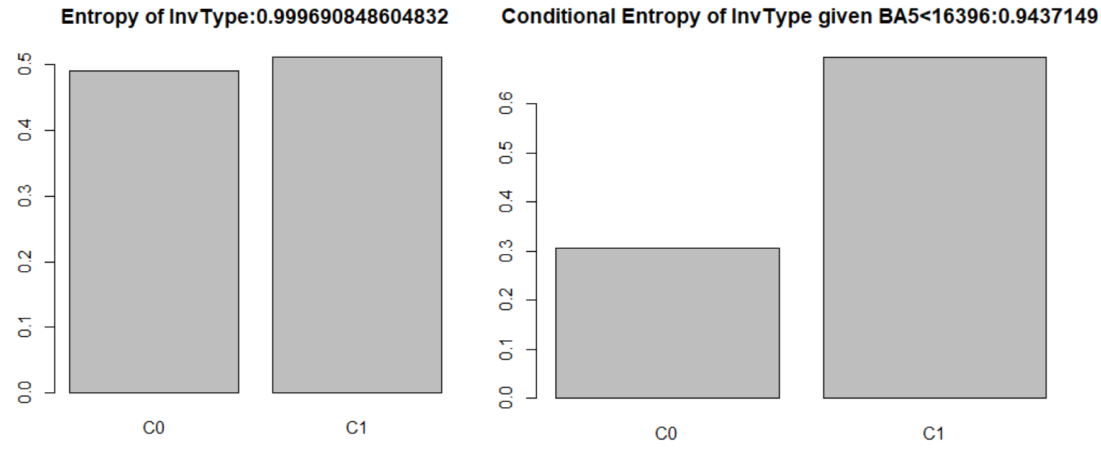
Eliott Loop,
Grégoire de Crombrugghe

Variable : BA7
Valeur : 0.101643849636697

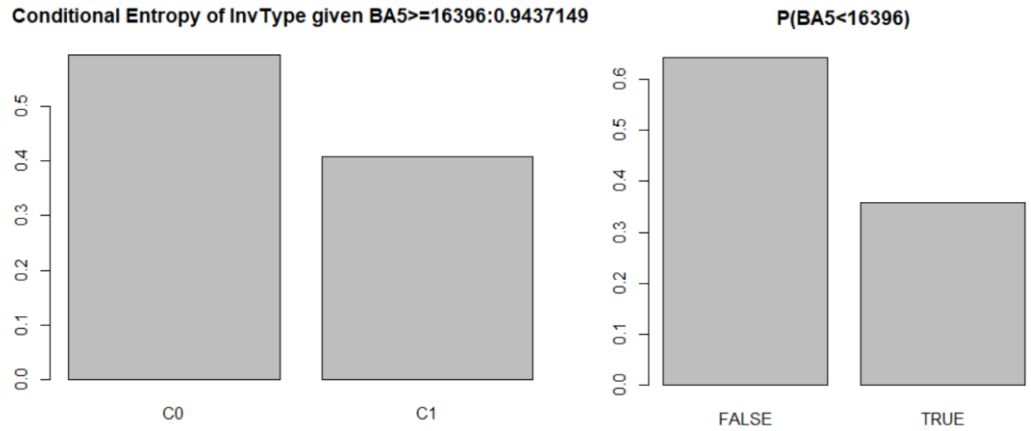


Nous observons, pour l'attribut BA7, que la probabilité qu'une personne investisse dans le produit C1 est de 72%, lorsque sa valeur pour BA7 est plus petite que 16371 (et 28% pour C0). Lorsqu'elle est plus grande, la probabilité, qu'une personne ai investis dans C1, est de 35% (et 65% pour C0).

Variable : BA5
Valeur : 0.0559759165004248



Eliott Loop,
Grégoire de Crombrugghe



Nous observons, pour l'attribut BA5, que la probabilité qu'une personne investisse dans le produit C1 est de 69%, lorsque sa valeur pour BA5 est plus petite que 16396 (et 31% pour C0). Lorsqu'elle est plus grande, la probabilité, qu'une personne ai investis dans C1, est de 41% (et 59% pour C0).

Selon nous, l'information gain se reflète dans 2 endroits. Premièrement, dans le graphe représentant la disparité entre les valeurs que prends la conditional probability distribution, pour une "une certaine valeur" données. Deuxièmement, dans la probability distribution, un attribut trop inégal au niveau de la probability distribution aura un faible information gain. En effet, nous constatons que pour BA3, lorsque sa "valeur" est égale à < 18899, l'attribut BA3 possède la plus grande différence (des 3 attributs) entre la probabilité de choisir C0 ou C1. Il en va de même lorsque BA3 prend comme "valeur" >=188899. Possédant aussi la probability distribution la moins inégales (des 3 attributs).

Conclusion

Pour conclure, lors de la conclusion du premier TP nous en avions conclu qu'aucun attribut ne permettait de pouvoir, avec certitude, prendre une décision sur le produit à proposer. De plus, nous avions mis l'accent sur le faite que les attributs BA6, BA1,SE1,IA2 était selon nous inutile.

Pour ce TP2, nous concluons que les attributs BA3,BA7,BA5 possèdent l'information gain la plus élevé, elles ont donc le plus de potentiel pour prédire la variable cible. La différence avec la conclusion faite lors du premier TP, est que pour ce rapport-ci nous avons des données sur lequel basé nos conclusions. Alors que, pour le premier TP la conclusion ce basais uniquement sur ce qui étais visible.