IST565 Data Mining - HW3 Association Rules

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/16/2019**

For this homework, we are going to explore the bank data, available on the LMS, and an accompanying description of the attributes and their values. The dataset contains attributes on each person's demographics and banking information in order to determine they will want to obtain the new PEP (Personal Equity Plan).

Your goal is to perform Association Rule discovery on the dataset using Weka or R.

First perform the necessary preprocessing steps required for association rule mining, specifically the id field needs to be removed and a number of numeric fields need discretization or otherwise converted to nominal.

Next perform association rule discovery on the preprocessed data. Experiment with different parameters and preprocessing so that you get on the order of 20-30 strong rules, e.g. rules with high lift and confidence which at the same time have relatively good support. Don't forget to report in details what you have tried.

Finally, set PEP as the right hand side of the rules, and see what rules are generated.

Select the top 5 most "interesting" rules and for each specify the following:

- Support, Confidence and Lift values
- An explanation of the pattern and why you believe it is interesting based on the business objectives of the company.
- Any recommendations based on the discovered rule that might help the company to better understand behavior of its customers or to develop a business opportunity.

Note that the top 5 most interesting rules are most likely not the top 5 in the strong rules. They are rules, that in addition to having high lift and confidence, also provide some non-trivial, actionable knowledge based on underlying business objectives.

To complete this assignment, write a short report describing your association rule mining process and the resulting 5 interesting rules, each with their three items of explanation and recommendations. For at least one of the rules, discuss the support, confidence and lift values and how they are interpreted in this data set.

You should write the report as if you are working for a client who knows little about data mining. Your report should give your client some insightful and reliable suggestions on what kinds of potential buyers your client should contact and convince your client that your suggestions are reliable based on the evidence gathered from your experiment results.

In more detail, your report should include:

- Description of preprocessing steps
- Description of parameters and experiments in order to obtain strong rules
- Give the top 5 most interesting rules and the 3 items listed above for each rule
- For one rule, discuss the support, confidence and lift numbers and how they were computed from the data

IST565 Data Mining - HW3 Association Rules

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/16/2019**

Solution:

Preprocessing Steps:

1. Load the data
2. Discretize age
3. Discretize income
4. Convert numeric to nominal for children
5. Apply Apriori Algorithm

```
library(plyr)
library(dplyr)
library(arules)

bd <- read.csv("C:/Users/Akshay/Desktop/Syracuse University - Information Management/Sem 3/IST 707/HW3/bankdata_csv_all.csv")
summary (bd)
#class(bd$pep)
#levels(bd$pep)
#class(bd$age)
#levels(bd$age)
#View(bd)
str(bd)
```

Following are the key observations obtained by using summary():

i) **id:** There are a total of 600 unique IDs in the dataset.

ii) **age:** The range of Age is from 18 years to 67 years, with the mean and median age being 42.4 years and 42.0 years respectively.

iii) **sex:** There are 600 customers in total out of which, 300 are Male and 300 are Female

iv) **region:** There are 600 customers in total out of which, 269 reside in the inner part of the city, 96 are from rural area, 62 are from suburb and 173from town.

v) **income:** The range of Income is from $5,014 to $63,130, with the mean and median salary being $27,524 and $24,925 respectively.

vi) **married:** There are 600 customers in total out of which 396 are married and 204 are not.

vii) **children:** The range of number of children is from 0 to 3, with the mean and median number of children being 1 and 1.01 respectively.

viii) **car:** There are 600 customers in total out of which 296 have car and 304 do not have car.

ix) **save_act:** There are 600 customers in total out of which 414 have a savings account and 186 do not.

x) **current_acct:** There are 600 customers in totalout of which 455 have a current account and 145 do not.

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/16/2019**

xi) **mortgage:** There are 600 customers in total out of which 391 have mortgage and 209 do not.
xii) **pep:** There are 600 customers in total out of which 326 do not have PEP and 274 have.

Following are the key observations obtained by using str():

i) **id:** Character class
ii) **age:** Integer class
iii) **sex:** Character class
iv) **region:** Character class
v) **income:** Numeric class
vi) **married:** Character class
vii) **children:** Integer class
viii) **car:** Character class
ix) **save_act:** Character class
x) **current_acct:** Character class
xi) **mortgage:** Character class
xii) **pep:** Character class

Through the overview, it can be concluded that there are categorical as well as numeric variables.

1. Discretize age by customized bin

```
bd$age <- cut(bd$age, breaks = c(0,10,20,30,40,50,60,Inf),labels=c("chi
ld","teens","twenties","thirties","fourties","fifties","old"))
```

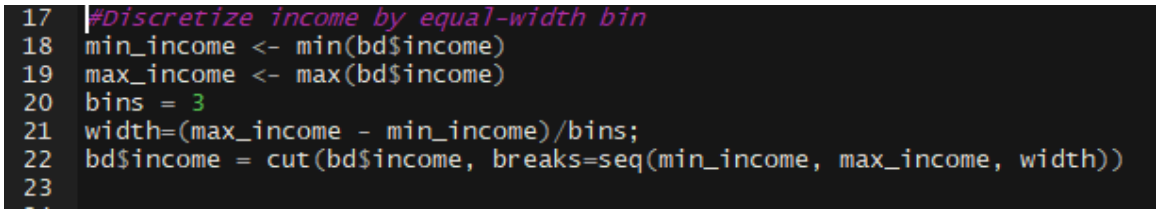
```
14 #Discretize age by customized bin
15 bd$age <- cut(bd$age, breaks = c(0,10,20,30,40,50,60,Inf),labels=c("child","teens","twenties","thirties","fourties","fifties","old"))
```

IST565 Data Mining - HW3 Association Rules

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/16/2019**

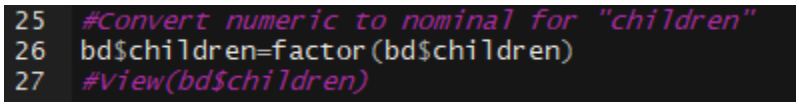
2. Discretize income by equal-width bin

```
min_income <- min(bd$income)
max_income <- max(bd$income)
bins = 3
width=(max_income - min_income)/bins;
bd$income = cut(bd$income, breaks=seq(min_income, max_income, width))
```



3. Convert numeric to nominal for "children"

```
bd$children=factor(bd$children)
```



4. Now the second step of conversion, changing "YES" to "[variable_name]=YES".

```
bd$married=dplyr::recode(bd$married, YES="married=YES", NO="married=NO
")
bd$car=dplyr::recode(bd$car, YES="car=YES", NO="car=NO")
bd$save_act=dplyr::recode(bd$save_act, YES="save_act=YES", NO="save_act
=NO")
bd$current_act=dplyr::recode(bd$current_act, YES="current_act=YES", NO=
"current_act=NO")
bd$mortgage=dplyr::recode(bd$mortgage, YES="mortgage=YES", NO="mortgage
```

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/16/2019**

```
=NO")
bd$pep=dplyr::recode(bd$pep, YES="pep=YES", NO="pep=NO")
```

```
#Now the second step of conversion, changing "YES" to "[variable_name]=YES".
bd$married=dplyr::recode(bd$married, YES="married=YES", NO="married=NO")
bd$car=dplyr::recode(bd$car, YES="car=YES", NO="car=NO")
bd$save_act=dplyr::recode(bd$save_act, YES="save_act=YES", NO="save_act=NO")
bd$current_act=dplyr::recode(bd$current_act, YES="current_act=YES", NO="current_act=NO")
bd$mortgage=dplyr::recode(bd$mortgage, YES="mortgage=YES", NO="mortgage=NO")
bd$pep=dplyr::recode(bd$pep, YES="pep=YES", NO="pep=NO")


bd <- bd[-c(1)]
head(bd)
view(bd)
```

5. Now load the transformed data into the Apriori algorithm

First Iteration: Setting confidence = 0.9 and support = 0.001

```
#Now load the transformed data into the apriori algorithm
myRules = apriori(bd, parameter = list(supp = 0.001, conf = 0.9))
summary(myRules)
```

```
> summary(myRules)
set of 896167 rules

rule length distribution (lhs + rhs):sizes
     2      3      4      5      6      7      8      9     10
     3    109   2234  24876 114285 253556 286235 166684  48185

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   7.000   8.000   7.662   8.000  10.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.001667   Min.   :0.9000   Min.   : 1.187   Min.   :  1.000
 1st Qu.:0.001667   1st Qu.:1.0000   1st Qu.: 1.535   1st Qu.:  1.000
 Median :0.001667   Median :1.0000   Median : 2.000   Median :  1.000
 Mean   :0.002523   Mean   :0.9996   Mean   : 2.450   Mean   :  1.514
 3rd Qu.:0.003333   3rd Qu.:1.0000   3rd Qu.: 2.553   3rd Qu.:  2.000
 Max.   :0.186667   Max.   :1.0000   Max.   :16.216   Max.   :112.000

mining info:
 data ntransactions support confidence
   bd           600   0.001        0.9
> |
```

➔ We can see that we got a set of 896167 rules. We need to decrease the number of rules. So, let's increase the confidence and support in the second iteration.

Second Iteration: Setting confidence = 1 and support = 0.025

IST565 Data Mining - HW3 Association Rules

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/16/2019**

```
#Now load the transformed data into the apriori algorithm
myRules = apriori(bd, parameter = list(supp = 0.025, conf = 1))
summary(myRules)
```

```
> summary(myRules)
set of 536 rules

rule length distribution (lhs + rhs):sizes
  2   3   4   5   6   7   8
  2  37 138 181 132  40   6

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00    4.00    5.00    5.02    6.00    8.00

summary of quality measures:
    support         confidence        lift            count
 Min.   :0.0250   Min.   :1   Min.   :1.32   Min.   :15.0
 1st Qu.:0.0267   1st Qu.:1   1st Qu.:1.45   1st Qu.:16.0
 Median :0.0300   Median :1   Median :1.51   Median :18.0
 Mean   :0.0335   Mean   :1   Mean   :1.72   Mean   :20.1
 3rd Qu.:0.0367   3rd Qu.:1   3rd Qu.:2.11   3rd Qu.:22.0
 Max.   :0.1333   Max.   :1   Max.   :2.28   Max.   :80.0

mining info:
 data ntransactions support confidence
   bd            600   0.025          1
>
```

➔ This time we got a set of 536 rues. Let's again increase the confidence and support in the third iteration.

Third Iteration: Setting confidence = 1 and support = 0.029

```
#Now load the transformed data into the apriori algorithm
myRules = apriori(bd, parameter = list(supp = 0.029, conf = 1))
summary(myRules)
```

```
> summary(myRules)
set of 292 rules

rule length distribution (lhs + rhs):sizes
  2   3   4   5   6   7   8
  2  27  85 104  56  16   2

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   4.000   5.000   4.825   6.000   8.000

summary of quality measures:
    support         confidence        lift            count
 Min.   :0.03000   Min.   :1   Min.   :1.319   Min.   :18.00
 1st Qu.:0.03167   1st Qu.:1   1st Qu.:1.449   1st Qu.:19.00
 Median :0.03500   Median :1   Median :1.515   Median :21.00
 Mean   :0.03959   Mean   :1   Mean   :1.704   Mean   :23.76
 3rd Qu.:0.04333   3rd Qu.:1   3rd Qu.:2.113   3rd Qu.:26.00
 Max.   :0.13333   Max.   :1   Max.   :2.281   Max.   :80.00

mining info:
 data ntransactions support confidence
   bd            600   0.029          1
```

➔ This time we got a set of 292 rules. Let us try a different combination for support and confidence in the next iteration.

Fourth Iteration: Setting confidence = 0.9 and support = 0.10

IST565 Data Mining - HW3 Association Rules

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/16/2019**

```
#Now load the transformed data into the apriori algorithm
myRules = apriori(bd, parameter = list(supp = 0.10, conf = 0.9))
summary(myRules)
inspect(myRules[1:10])
```

```
> summary(myRules)
set of 19 rules

rule length distribution (lhs + rhs):sizes
2 3 4 5
2 8 1 8

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   3.000   3.789   5.000   5.000

summary of quality measures:
    support         confidence          lift            count
 Min.   :0.1000   Min.   :0.9000   Min.   :1.449   Min.   : 60.00
 1st Qu.:0.1067   1st Qu.:0.9278   1st Qu.:1.485   1st Qu.: 64.00
 Median :0.1217   Median :0.9481   Median :1.679   Median : 73.00
 Mean   :0.1259   Mean   :0.9508   Mean   :1.733   Mean   : 75.53
 3rd Qu.:0.1333   3rd Qu.:0.9738   3rd Qu.:1.985   3rd Qu.: 80.00
 Max.   :0.1867   Max.   :1.0000   Max.   :2.019   Max.   :112.00

mining info:
 data ntransactions support confidence
   bd           600     0.1        0.9
>
```

➔ Now we have only 19 rules.

```
options(digits=3)
rules <- sort(myRules, decreasing=F, by="lift")
inspect(rules[1:10])
```

```
> inspect(myRules[1:10])
     lhs                                                rhs                          support   confidence lift     count
[1]  {income=(4.38e+04,6.31e+04]}                    => {save_act=save_act=YES}      0.1333333 1.0000000  1.449275  80
[2]  {age=twenties}                                  => {income=(5.01e+03,2.44e+04]} 0.1866667 0.9411765  1.988401 112
[3]  {income=(4.38e+04,6.31e+04],current_act=current_act=YES} => {save_act=save_act=YES}      0.1050000 1.0000000  1.449275  63
[4]  {age=twenties,region=INNER_CITY}                => {income=(5.01e+03,2.44e+04]} 0.1033333 0.9538462  2.015168  62
[5]  {age=twenties,car=car=NO}                       => {income=(5.01e+03,2.44e+04]} 0.1083333 0.9558824  2.019470  65
[6]  {age=twenties,pep=pep=NO}                       => {income=(5.01e+03,2.44e+04]} 0.1216667 0.9240506  1.952220  73
[7]  {age=twenties,mortgage=mortgage=NO}             => {income=(5.01e+03,2.44e+04]} 0.1266667 0.9382716  1.982264  76
[8]  {age=twenties,married=married=YES}              => {income=(5.01e+03,2.44e+04]} 0.1216667 0.9480519  2.002927  73
[9]  {age=twenties,save_act=save_act=YES}            => {income=(5.01e+03,2.44e+04]} 0.1133333 0.9315068  1.967972  68
[10] {age=twenties,current_act=current_act=YES}      => {income=(5.01e+03,2.44e+04]} 0.1450000 0.9456522  1.997857  87
>
```

IST565 Data Mining - HW3 Association Rules

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/16/2019**

```
> options(digits=3)
> rules <- sort(myRules, decreasing=T, by="lift")
> inspect(rules[1:10])
     lhs                                rhs                                support confidence lift count
[1]  {age=twenties,
      car=car=NO}                    => {income=(5.01e+03,2.44e+04]}   0.108      0.956 2.02    65
[2]  {age=twenties,
      region=INNER_CITY}             => {income=(5.01e+03,2.44e+04]}   0.103      0.954 2.02    62
[3]  {age=twenties,
      married=married=YES}           => {income=(5.01e+03,2.44e+04]}   0.122      0.948 2.00    73
[4]  {age=twenties,
      current_act=current_act=YES}   => {income=(5.01e+03,2.44e+04]}   0.145      0.946 2.00    87
[5]  {age=twenties}                  => {income=(5.01e+03,2.44e+04]}   0.187      0.941 1.99   112
[6]  {age=twenties,
      mortgage=mortgage=NO}          => {income=(5.01e+03,2.44e+04]}   0.127      0.938 1.98    76
[7]  {age=twenties,
      save_act=save_act=YES}         => {income=(5.01e+03,2.44e+04]}   0.113      0.932 1.97    68
[8]  {age=twenties,
      pep=pep=NO}                    => {income=(5.01e+03,2.44e+04]}   0.122      0.924 1.95    73
[9]  {married=married=YES,
      children=0,
      save_act=save_act=YES,
      current_act=current_act=YES}   => {pep=pep=NO}                    0.133      0.920 1.69    80
[10] {married=married=YES,
      children=0,
      save_act=save_act=YES,
      mortgage=mortgage=NO}          => {pep=pep=NO}                    0.122      0.912 1.68    73
>
```

```
options(digits=3)
rules <- sort(myRules, decreasing=T, by="lift")
inspect(rules[1:5])
```

```
> inspect(rules[1:5])
    lhs                                  rhs                                support confidence lift count
[1] {age=twenties,car=car=NO}            => {income=(5.01e+03,2.44e+04]} 0.108   0.956      2.02  65
[2] {age=twenties,region=INNER_CITY}     => {income=(5.01e+03,2.44e+04]} 0.103   0.954      2.02  62
[3] {age=twenties,married=married=YES}   => {income=(5.01e+03,2.44e+04]} 0.122   0.948      2.00  73
[4] {age=twenties,current_act=current_act=YES} => {income=(5.01e+03,2.44e+04]} 0.145   0.946      2.00  87
[5] {age=twenties}                       => {income=(5.01e+03,2.44e+04]} 0.187   0.941      1.99 112
>
```

After Setting PEP as RHS (PEP = RHS)

IST565 Data Mining - HW3 Association Rules

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/16/2019**

```
peprules <- apriori(bd, parameter = list(maxlen=5), appearance = list(rhs = c("pep=pep=YES", "pep=pep=NO")))

inspect(peprules[1:10])
```

```
> inspect(peprules[1:10])
     lhs                                                                rhs          support confidence lift count
[1]  {children=1}                                                    => {pep=pep=YES} 0.183   0.815      1.78 110
[2]  {children=1,mortgage=mortgage=NO}                               => {pep=pep=YES} 0.118   0.845      1.85  71
[3]  {married=married=YES,children=1}                                => {pep=pep=YES} 0.123   0.831      1.82  74
[4]  {children=1,save_act=save_act=YES}                              => {pep=pep=YES} 0.133   0.842      1.84  80
[5]  {children=1,current_act=current_act=YES}                        => {pep=pep=YES} 0.140   0.832      1.82  84
[6]  {children=1,save_act=save_act=YES,current_act=current_act=YES}  => {pep=pep=YES} 0.105   0.863      1.89  63
[7]  {sex=FEMALE,married=married=YES,children=0}                     => {pep=pep=NO}  0.130   0.830      1.53  78
[8]  {married=married=YES,children=0,car=car=NO}                     => {pep=pep=NO}  0.133   0.800      1.47  80
[9]  {married=married=YES,children=0,mortgage=mortgage=NO}           => {pep=pep=NO}  0.173   0.897      1.65 104
[10] {married=married=YES,children=0,save_act=save_act=YES}          => {pep=pep=NO}  0.178   0.899      1.65 107
>
```

Observations (Top 10):

Customers who did not bought PEP have the key attributes:

- Female
- No Children
- No Mortgage
- No Current account
- Have Savings Account

```
peprules <- apriori(bd, parameter = list(maxlen=5), appearance = list(rhs = c("pep=pep=YES", "pep=pep=NO")))

inspect(peprules[1:5])
```

IST565 Data Mining - HW3 Association Rules

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/16/2019**

```
> inspect(peprules[1:5])
    lhs                                  rhs          support confidence lift count
[1] {children=1}                      => {pep=pep=YES} 0.183   0.815      1.78 110
[2] {children=1,mortgage=mortgage=NO} => {pep=pep=YES} 0.118   0.845      1.85  71
[3] {married=married=YES,children=1}  => {pep=pep=YES} 0.123   0.831      1.82  74
[4] {children=1,save_act=save_act=YES} => {pep=pep=YES} 0.133   0.842      1.84  80
[5] {children=1,current_act=current_act=YES} => {pep=pep=YES} 0.140   0.832      1.82  84
>
```

Observations (Top 5):

Customers who bought PEP have the key attributes:

- Have Children
- Don't have Mortgage
- Married
- Have a savings account
- Have a current account

Conclusion:

We can see that when people have children, they have PEP. They are married and they also have a savings or a current account.
Here,

- ➔ Confidence is varying from 0.815 to 0.845
- ➔ Support is varying from 0.118 to 0.183 and
- ➔ Lift is varying from 1.78 to 1.85

Thus, we can say that there is a strong correlation between Children and Current account. This rule has support of 0.133, confidence of 0.842 and lift of 1.84. This is a strong rule. Also, the findings suggest that people who are married and have children, have a savings or a current account. This means that they have future planning and are more likely to buy PEP. Therefore, there is a business opportunity.

On the other hand, if the customers don't have children or they don't have a current account, they don't have PEP. Female customers are also less likely to buy PEP.