HW3

Jose Reyes

7.21.2020

#### Introduction

Marketing campaigns are most effective when the target audience is most receptive to the message. In this case, marketing a bank’s personal equity plan (PEP) would certainly appeal more to some than others.

To launch a successful marketing campaign, a data mining technique called Association Rule mining was performed on bank data containing information on customers and their demographics and banking information. This technique is particularly useful in finding key patterns in data such as discovering features that occur most commonly together. By identifying these features such as age, income, and mortgage status, the algorithm can predict the likelihood that a customer will apply for a PEP based on specific characteristics.

The analysis will provide key findings and suggestions for future marketing campaigns to target customers who are most likely to purchase a PEP. Conversely, the results will also show customers who are least likely to purchase a PEP and can therefore be avoided in these campaigns.

#### Analysis and Models

#### About the Data

The data contains information on 600 customers and their age, sex, region, income, marital status, number of children, car ownership, savings account ownership, current account ownership, mortgage status, and if they have purchased a Personal Equity Plan (PEP). An initial summary of this data is included below.

#### Load packages for Association Rules Mining

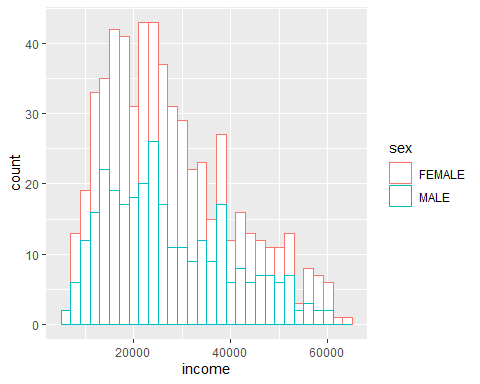
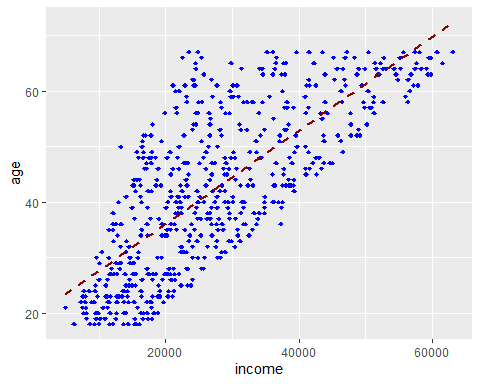
require(dplyr)  
require(RWeka)  
require(arules)  
require(arulesViz)  
require(ggplot2)

#### Load and view bank data str(bank)

## 'data.frame': 600 obs. of 12 variables:  
## $ id : chr "ID12101" "ID12102" "ID12103" "ID12104" ...  
## $ age : int 48 40 51 23 57 57 22 58 37 54 ...  
## $ sex : chr "FEMALE" "MALE" "FEMALE" "FEMALE" ...  
## $ region : chr "INNER\_CITY" "TOWN" "INNER\_CITY" "TOWN" ...  
## $ income : num 17546 30085 16575 20375 50576 ...  
## $ married : chr "NO" "YES" "YES" "YES" ...  
## $ children : int 1 3 0 3 0 2 0 0 2 2 ...  
## $ car : chr "NO" "YES" "YES" "NO" ...  
## $ save\_act : chr "NO" "NO" "YES" "NO" ...  
## $ current\_act: chr "NO" "YES" "YES" "YES" ...  
## $ mortgage : chr "NO" "YES" "NO" "NO" ...  
## $ pep : chr "YES" "NO" "NO" "NO" ...

summary(bank)

## id age sex region   
## Length:600 Min. :18.00 Length:600 Length:600   
## Class :character 1st Qu.:30.00 Class :character Class :character   
## Mode :character Median :42.00 Mode :character Mode :character   
## Mean :42.40   
## 3rd Qu.:55.25   
## Max. :67.00   
## income married children car   
## Min. : 5014 Length:600 Min. :0.000 Length:600   
## 1st Qu.:17265 Class :character 1st Qu.:0.000 Class :character   
## Median :24925 Mode :character Median :1.000 Mode :character   
## Mean :27524 Mean :1.012   
## 3rd Qu.:36173 3rd Qu.:2.000   
## Max. :63130 Max. :3.000   
## save\_act current\_act mortgage pep   
## Length:600 Length:600 Length:600 Length:600   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##



As expected, the income generally increases as age increases and the distribution of incomes is positively skewed.

To perform Association Rules mining, the data must be checked for missing values, cleansed, and transformed into the proper data types.

#### Check for missing data

anyNA(bank, recursive = FALSE)

## [1] FALSE

#### Clean and Transform Data

-Drop ID column which is not useful for analysis  
-Discretize age into 7 separate groups  
-Discretize income into 4 equal groups using the quartile values obtained in the data summary

-Convert “Yes” and “No” values into “[variable]=”Yes/No"" to allow for proper Association Rules analysis  
-Convert all other values to nominal values

#### Final check before analysis

str(bank)

## 'data.frame': 600 obs. of 11 variables:  
## $ age : Factor w/ 7 levels "children","teenagers",..: 5 4 6 3 6 6 3 6 4 6 ...  
## $ sex : Factor w/ 2 levels "FEMALE","MALE": 1 2 1 1 1 1 2 2 1 2 ...  
## $ region : Factor w/ 4 levels "INNER\_CITY","RURAL",..: 1 4 1 4 2 4 2 4 3 4 ...  
## $ income : Factor w/ 4 levels "LowestIncome",..: 2 3 1 2 4 4 1 3 3 2 ...  
## $ married : Factor w/ 2 levels "married=NO","married=YES": 1 2 2 2 2 2 1 2 2 2 ...  
## $ children : Factor w/ 4 levels "0","1","2","3": 2 4 1 4 1 3 1 1 3 3 ...  
## $ car : Factor w/ 2 levels "car=NO","car=YES": 1 2 2 1 1 1 1 2 2 2 ...  
## $ save\_act : Factor w/ 2 levels "save\_act=NO",..: 1 1 2 1 2 2 1 2 1 2 ...  
## $ current\_act: Factor w/ 2 levels "current\_act=NO",..: 1 2 2 2 1 2 2 2 1 2 ...  
## $ mortgage : Factor w/ 2 levels "mortgage=NO",..: 1 2 1 1 1 1 1 1 1 1 ...  
## $ pep : Factor w/ 2 levels "pep=NO","pep=YES": 2 1 1 1 1 2 2 1 1 1 ...

#### Association Rules Mining

The Apriori principle for association rules mining will be applied to reduce the overall computational complexity. This is based on the principle that if an itemset is frequent, then all subsets must also be frequent. This algorithm will produce a list of rules which each have a set of items that commonly occur together.

Each rule’s effectiveness is measured by Support, Confidence, and Lift.

**Support -** How often a rule is applicable to a given data set

**Confidence** - How frequently certain items appear with other items in a transactions

**Lift** - The ratio of the confidence of the rule and the expected confidence of the rule

All three values will be reviewed together holistically to properly assess the overall effectiveness of a rule.

#### Set Apriori rules

An initial analysis of all rules was performed. The minimum support is set to 0.2 and confidence to 0.76. When either of these values are increased, the rule set count becomes too low.

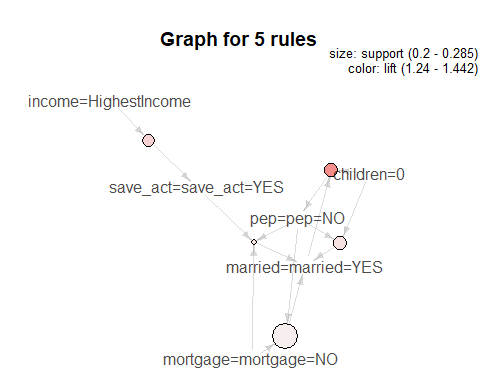
rules <- apriori(bank, parameter = list(supp = 0.2, conf = 0.76))

summary(rules)

## set of 35 rules  
##   
## rule length distribution (lhs + rhs):sizes  
## 2 3 4   
## 9 23 3   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.00 2.50 3.00 2.83 3.00 4.00   
##   
## summary of quality measures:  
## support confidence coverage lift count   
## Min. :0.200 Min. :0.760 Min. :0.237 Min. :1.00 Min. :120   
## 1st Qu.:0.227 1st Qu.:0.765 1st Qu.:0.283 1st Qu.:1.01 1st Qu.:136   
## Median :0.238 Median :0.776 Median :0.310 Median :1.02 Median :143   
## Mean :0.275 Mean :0.785 Mean :0.352 Mean :1.07 Mean :165   
## 3rd Qu.:0.292 3rd Qu.:0.786 3rd Qu.:0.370 3rd Qu.:1.05 3rd Qu.:175   
## Max. :0.532 Max. :0.913 Max. :0.690 Max. :1.44 Max. :319   
##   
## mining info:  
## data ntransactions support confidence  
## bank 600 0.2 0.76

This initial analysis resulted in a set of 35 rules; all which have a lift greater than 1 which is generally a solid indicator of a strong rule. From this, it can be assumed that there are several strong association rules that can be identified from this data set. The following 5 rules contain the highest lift values:

rules <- sort(rules, decreasing = TRUE, by = "lift")  
plot(rules[1:5],method="graph")



inspect(rules[1:5])

## lhs rhs support confidence coverage lift count  
## [1] {married=married=YES,   
## children=0} => {pep=pep=NO} 0.235 0.783 0.300 1.44 141  
## [2] {income=HighestIncome} => {save\_act=save\_act=YES} 0.228 0.913 0.250 1.32 137  
## [3] {save\_act=save\_act=YES,   
## mortgage=mortgage=NO,   
## pep=pep=NO} => {married=married=YES} 0.200 0.845 0.237 1.28 120  
## [4] {children=0,   
## pep=pep=NO} => {married=married=YES} 0.235 0.844 0.278 1.28 141  
## [5] {mortgage=mortgage=NO,   
## pep=pep=NO} => {married=married=YES} 0.285 0.818 0.348 1.24 171

Many of these strong association rules are rather intuitive. It can be assumed that married customers with no children will not purchase a PEP due to having dual incomes and lower expenses. Additionally, customers with relatively higher incomes would most likely have savings accounts to earn interest and defer spending for other big purchases such as a new house or a car.

#### PEP Analysis

Association rules is then performed to identify common customer demographics and banking information related to the purchasing of a PEP.

peprules <- apriori(bank, parameter = list(maxlen = 4), appearance = list(rhs = c("pep=pep=YES","pep=pep=NO")))

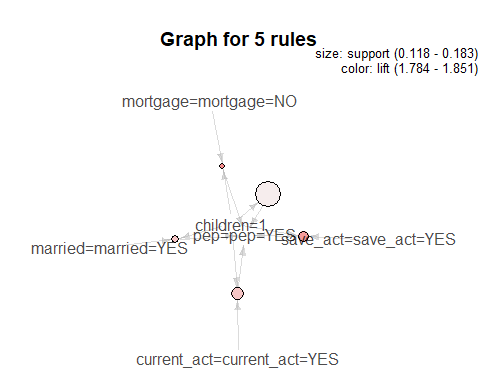
inspect(peprules)

## lhs rhs support confidence coverage lift count  
## [1] {children=1} => {pep=pep=YES} 0.183 0.815 0.225 1.78 110  
## [2] {children=1,   
## mortgage=mortgage=NO} => {pep=pep=YES} 0.118 0.845 0.140 1.85 71  
## [3] {married=married=YES,   
## children=1} => {pep=pep=YES} 0.123 0.831 0.148 1.82 74  
## [4] {children=1,   
## save\_act=save\_act=YES} => {pep=pep=YES} 0.133 0.842 0.158 1.84 80  
## [5] {children=1,   
## current\_act=current\_act=YES} => {pep=pep=YES} 0.140 0.832 0.168 1.82 84  
## [6] {children=1,   
## save\_act=save\_act=YES,   
## current\_act=current\_act=YES} => {pep=pep=YES} 0.105 0.863 0.122 1.89 63  
## [7] {sex=FEMALE,   
## married=married=YES,   
## children=0} => {pep=pep=NO} 0.130 0.830 0.157 1.53 78  
## [8] {married=married=YES,   
## children=0,   
## car=car=NO} => {pep=pep=NO} 0.133 0.800 0.167 1.47 80  
## [9] {married=married=YES,   
## children=0,   
## mortgage=mortgage=NO} => {pep=pep=NO} 0.173 0.897 0.193 1.65 104  
## [10] {married=married=YES,   
## children=0,   
## save\_act=save\_act=YES} => {pep=pep=NO} 0.178 0.899 0.198 1.65 107

This resulted in 10 strong association rules all of which have lift values that range from 1.65 to 1.78.

#### Results

To provide a summary of key findings and suggestions for a more targeted and successful campaign to increase PEP sales, the following 5 rules identified in the PEP analysis are highlighted and explained in greater detail:



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Lhs** | **Rhs** | **Support** | **Confidence** | **Lift** | **Count** |
| {children=1} | {pep=YES} | 0.183 | 0.815 | 1.78 | 110 |
| {married=YES, children=1} | {pep=YES} | 0.123 | 0.831 | 1.82 | 74 |
| {children=1, mortgage=NO} | {pep=YES} | 0.118 | 0.845 | 1.75 | 71 |
| {married=YES, children=0, car=NO} | {pep=NO} | 0.133 | 0.800 | 1.47 | 80 |
| {married=YES, children=0, mortgage=NO} | {pep=NO} | 0.173 | 0.897 | 1.93 | 104 |

Across all 5 of these rules, it is evident that having 0 or 1 child is a strong indicator of whether a customer will purchase a PEP.

In the first rule, the support value indicates that 18.3% of all customers have 1 child and purchased a PEP. This is a highly occurring association considering the total list of 600 customers. Additionally, the confidence value indicates that 81.5% of all customers with 1 child have purchased a PEP. And finally, the lift value is the ratio of the confidence of the rule and the expected confidence of the rule. With a lift value of 1.78, it can be surmised that the occurrence of having 1 child and purchasing a PEP occurs more often together than should be expected.

In summary, it is evident that having 1 child is strongly correlated with purchasing a PEP. This is indicated by the first 3 rules in the table which all contain high lift values and confidence values greater than 80%. It is also evident that having no children is strongly correlated with NOT purchasing a PEP. This is showed by the 4th and 5th rules with high lift values and confidence values also greater than 80%.

#### Conclusions

Association rules mining was performed on the bank data containing 600 customers and their demographics and banking information. This data mining technique is highly effective in identifying patterns in a dataset to produce actionable knowledge. In this case, the results were used to identify which specific customer characteristics occur most frequently with the purchasing of a PEP and vice versa.

The Apriori algorithm was performed on the data and customized to produce an initial set of 35 rules which, among other things, showed that married customers with no children are unlikely to purchase a PEP. This was further demonstrated by applying the algorithm to identify common demographics and banking information among customers who purchased or didn’t purchase a PEP. These results showed that customers with 1 child are highly likely to purchase a PEP with great confidence and again showed that married customers with no children are highly unlikely to purchase a PEP.

For a future marketing campaign to promote PEP sales, it is recommended that individuals with 1 child, regardless of marital status, should be targeted. Married individuals with no children do not need to be included in this campaign as they are highly unlikely to purchase a PEP.