MIS 510 Portfolio Project Option 2

Ty Cook

4/20/2020

## R Markdown

### Read Datset into Dataframe

crosssell.df <- read.csv("CatalogCrossSell.csv", header = TRUE) #Read CatalogCrossSell.csv into crosssell.df dataframe

### Dimensions/Subsetting Data

dim(crosssell.df) #View Dimensions of Dataframe

## [1] 60004 20

crosssell.df <- crosssell.df[c(1:10)] #Subset Data elimate empty NA rows  
crosssell.df <- crosssell.df[complete.cases(crosssell.df),] #Subset Data elimate empty NA Columns  
dim(crosssell.df)#View Dimensions of Subset Dataframe

## [1] 4998 10

### Summary

summary(crosssell.df[,2:10]) #Summary of columns minus the first customer id columns

## Clothing.Division Housewares.Division Health.Products.Division  
## Min. :0.00000 Min. :0.0000 Min. :1   
## 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:1   
## Median :0.00000 Median :0.0000 Median :1   
## Mean :0.03301 Mean :0.3936 Mean :1   
## 3rd Qu.:0.00000 3rd Qu.:1.0000 3rd Qu.:1   
## Max. :1.00000 Max. :1.0000 Max. :1   
## Automotive.Division Personal.Electronics.Division Computers.Division  
## Min. :0.0000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.00000   
## Median :0.0000 Median :0.0000 Median :0.00000   
## Mean :0.1349 Mean :0.4674 Mean :0.04682   
## 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:0.00000   
## Max. :1.0000 Max. :1.0000 Max. :1.00000   
## Garden.Division Novelty.Gift.Division Jewelry.Division  
## Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :0.2721 Mean :0.2275 Mean :0.3569   
## 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000

### Creating Data Matrix

count.df <- crosssell.df[,2:10] #Create count.df Dataframe using columns 2 through 10 of crosssell.df  
mat.df <- ifelse(count.df > 0, 1, 0) #Create mat.df using if else > 0 True = 1 and False = 0 of count.df  
cs.mat <- as.matrix(mat.df[, -1]) #Read dataframe mat.df into matrix cs.mat  
cs.mat[1:10, ] #Output top 10 Columns and Rows of the cs.mat

## Housewares.Division Health.Products.Division Automotive.Division  
## 1 1 1 1  
## 2 1 1 1  
## 3 1 1 1  
## 4 0 1 1  
## 5 0 1 0  
## 6 1 1 1  
## 7 1 1 0  
## 8 1 1 0  
## 9 0 1 1  
## 10 1 1 1  
## Personal.Electronics.Division Computers.Division Garden.Division  
## 1 1 0 0  
## 2 1 0 1  
## 3 1 0 1  
## 4 1 0 1  
## 5 1 0 1  
## 6 1 0 1  
## 7 1 0 1  
## 8 1 0 0  
## 9 1 0 1  
## 10 1 0 1  
## Novelty.Gift.Division Jewelry.Division  
## 1 1 0  
## 2 1 1  
## 3 1 1  
## 4 1 0  
## 5 1 0  
## 6 1 1  
## 7 0 0  
## 8 1 0  
## 9 0 1  
## 10 1 1

### Creating Transactions Database

library(arules) #Call library arules

## Warning: package 'arules' was built under R version 3.6.3

## Loading required package: Matrix

##   
## Attaching package: 'arules'

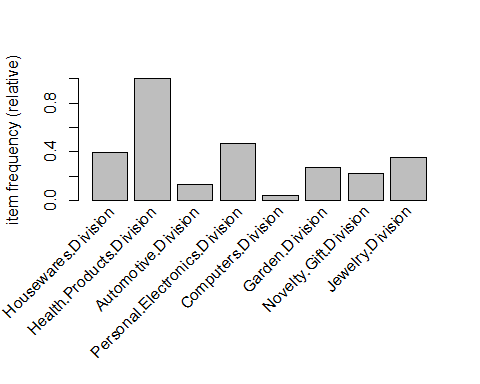
## The following objects are masked from 'package:base':  
##   
## abbreviate, write

cs.trans <- as(cs.mat, "transactions") #Read matrix cs.mat into a transactions database cs.trans  
inspect(cs.trans[1:10, ]) #Output top 10 Columns and Rows of the transactions database cs.trans

## items transactionID  
## [1] {Housewares.Division,   
## Health.Products.Division,   
## Automotive.Division,   
## Personal.Electronics.Division,   
## Novelty.Gift.Division} 1   
## [2] {Housewares.Division,   
## Health.Products.Division,   
## Automotive.Division,   
## Personal.Electronics.Division,   
## Garden.Division,   
## Novelty.Gift.Division,   
## Jewelry.Division} 2   
## [3] {Housewares.Division,   
## Health.Products.Division,   
## Automotive.Division,   
## Personal.Electronics.Division,   
## Garden.Division,   
## Novelty.Gift.Division,   
## Jewelry.Division} 3   
## [4] {Health.Products.Division,   
## Automotive.Division,   
## Personal.Electronics.Division,   
## Garden.Division,   
## Novelty.Gift.Division} 4   
## [5] {Health.Products.Division,   
## Personal.Electronics.Division,   
## Garden.Division,   
## Novelty.Gift.Division} 5   
## [6] {Housewares.Division,   
## Health.Products.Division,   
## Automotive.Division,   
## Personal.Electronics.Division,   
## Garden.Division,   
## Novelty.Gift.Division,   
## Jewelry.Division} 6   
## [7] {Housewares.Division,   
## Health.Products.Division,   
## Personal.Electronics.Division,   
## Garden.Division} 7   
## [8] {Housewares.Division,   
## Health.Products.Division,   
## Personal.Electronics.Division,   
## Novelty.Gift.Division} 8   
## [9] {Health.Products.Division,   
## Automotive.Division,   
## Personal.Electronics.Division,   
## Garden.Division,   
## Jewelry.Division} 9   
## [10] {Housewares.Division,   
## Health.Products.Division,   
## Automotive.Division,   
## Personal.Electronics.Division,   
## Garden.Division,   
## Novelty.Gift.Division,   
## Jewelry.Division} 10

### Item Frequency

itemFrequencyPlot(cs.trans) #Plot Item Frequency Graph using cs.trans



itemFrequency(cs.trans) #Item Frequency information

## Housewares.Division Health.Products.Division   
## 0.39355742 1.00000000   
## Automotive.Division Personal.Electronics.Division   
## 0.13485394 0.46738695   
## Computers.Division Garden.Division   
## 0.04681873 0.27210884   
## Novelty.Gift.Division Jewelry.Division   
## 0.22749100 0.35694278

### Apriori Function

rules <-apriori(cs.trans, parameter = list(supp = 0.01, conf = 0.5, maxlen = 2, target = "rules")) #Create Rules item using cs.trans with the parmaters of minimum support of 0.01 and confience interval of 0.5 max length of 2 by rules

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.5 0.1 1 none FALSE TRUE 5 0.01 1  
## maxlen target ext  
## 2 rules FALSE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 49   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[8 item(s), 4998 transaction(s)] done [0.00s].  
## sorting and recoding items ... [8 item(s)] done [0.00s].  
## creating transaction tree ... done [0.00s].  
## checking subsets of size 1 2

## Warning in apriori(cs.trans, parameter = list(supp = 0.01, conf = 0.5, maxlen  
## = 2, : Mining stopped (maxlen reached). Only patterns up to a length of 2  
## returned!

## done [0.00s].  
## writing ... [22 rule(s)] done [0.00s].  
## creating S4 object ... done [0.00s].

summary(rules) #Summarize rules

## set of 22 rules  
##   
## rule length distribution (lhs + rhs):sizes  
## 1 2   
## 1 21   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 2.000 2.000 1.955 2.000 2.000   
##   
## summary of quality measures:  
## support confidence lift count   
## Min. :0.02341 Min. :0.5000 Min. :1.000 Min. : 117.0   
## 1st Qu.:0.07718 1st Qu.:0.5666 1st Qu.:1.000 1st Qu.: 385.8   
## Median :0.16617 Median :0.6330 Median :1.277 Median : 830.5   
## Mean :0.21026 Mean :0.7388 Mean :1.298 Mean :1050.9   
## 3rd Qu.:0.23549 3rd Qu.:1.0000 3rd Qu.:1.476 3rd Qu.:1177.0   
## Max. :1.00000 Max. :1.0000 Max. :2.328 Max. :4998.0   
##   
## mining info:  
## data ntransactions support confidence  
## cs.trans 4998 0.01 0.5

inspect(head(sort(rules, by = "lift"))) #Inspect rules of first rules sorted by lift

## lhs rhs support confidence lift count  
## [1] {Automotive.Division} => {Garden.Division} 0.08543417 0.6335312 2.328227 427  
## [2] {Computers.Division} => {Jewelry.Division} 0.02961184 0.6324786 1.771933 148  
## [3] {Computers.Division} => {Housewares.Division} 0.03161265 0.6752137 1.715667 158  
## [4] {Novelty.Gift.Division} => {Personal.Electronics.Division} 0.16906763 0.7431838 1.590082 845  
## [5] {Automotive.Division} => {Jewelry.Division} 0.07523009 0.5578635 1.562893 376  
## [6] {Novelty.Gift.Division} => {Housewares.Division} 0.13485394 0.5927880 1.506230 674

#### Right Hand Side

rules.rhs<-apriori(cs.trans, parameter=list(supp=0.01,conf = 0.05, maxlen = 2), #Create Rules.rhs item using cs.trans with the parmaters of minimum support of 0.01 and confience interval of 0.5 max length of 2  
 appearance = list(default="lhs",rhs="Garden.Division"), #Target where lhs can be any value and rhs is equal Garden.Division  
 control = list(verbose=F)) #Turn off warnings and error logs  
summary(rules.rhs) #Summary of rules.rhs

## set of 8 rules  
##   
## rule length distribution (lhs + rhs):sizes  
## 1 2   
## 1 7   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 2.000 2.000 1.875 2.000 2.000   
##   
## summary of quality measures:  
## support confidence lift count   
## Min. :0.01641 Min. :0.2721 Min. :1.000 Min. : 82.0   
## 1st Qu.:0.10554 1st Qu.:0.3212 1st Qu.:1.180 1st Qu.: 527.5   
## Median :0.13005 Median :0.3499 Median :1.286 Median : 650.0   
## Mean :0.14771 Mean :0.3831 Mean :1.408 Mean : 738.2   
## 3rd Qu.:0.19048 3rd Qu.:0.3907 3rd Qu.:1.436 3rd Qu.: 952.0   
## Max. :0.27211 Max. :0.6335 Max. :2.328 Max. :1360.0   
##   
## mining info:  
## data ntransactions support confidence  
## cs.trans 4998 0.01 0.05

inspect(head(sort(rules.rhs, by = "confidence"))) #Inspect rules.rhs of first rules.rhs sorted by confidence

## lhs rhs support confidence  
## [1] {Automotive.Division} => {Garden.Division} 0.08543417 0.6335312   
## [2] {Novelty.Gift.Division} => {Garden.Division} 0.11224490 0.4934037   
## [3] {Jewelry.Division} => {Garden.Division} 0.12725090 0.3565022   
## [4] {Computers.Division} => {Garden.Division} 0.01640656 0.3504274   
## [5] {Personal.Electronics.Division} => {Garden.Division} 0.16326531 0.3493151   
## [6] {Housewares.Division} => {Garden.Division} 0.13285314 0.3375699   
## lift count  
## [1] 2.328227 427   
## [2] 1.813259 561   
## [3] 1.310146 636   
## [4] 1.287821 82   
## [5] 1.283733 816   
## [6] 1.240569 664

#### Left Hand Side

rules.lhs<-apriori(cs.trans, parameter=list(supp=0.01,conf = 0.05, minlen = 2), #Create Rules.lhs item using cs.trans with the parmaters of minimum support of 0.01 and confience interval of 0.5 max length of 2  
 appearance = list(default="rhs",lhs="Garden.Division"), #Target where rhs can be any value and lhs is equal Garden.Division  
 control = list(verbose=F)) #Turn off warnings and error logs  
summary(rules.lhs) #Summary of rules.lhs

## set of 7 rules  
##   
## rule length distribution (lhs + rhs):sizes  
## 2   
## 7   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2 2 2 2 2 2   
##   
## summary of quality measures:  
## support confidence lift count   
## Min. :0.01641 Min. :0.06029 Min. :1.000 Min. : 82.0   
## 1st Qu.:0.09884 1st Qu.:0.36324 1st Qu.:1.262 1st Qu.: 494.0   
## Median :0.12725 Median :0.46765 Median :1.288 Median : 636.0   
## Mean :0.12994 Mean :0.47752 Mean :1.466 Mean : 649.4   
## 3rd Qu.:0.14806 3rd Qu.:0.54412 3rd Qu.:1.562 3rd Qu.: 740.0   
## Max. :0.27211 Max. :1.00000 Max. :2.328 Max. :1360.0   
##   
## mining info:  
## data ntransactions support confidence  
## cs.trans 4998 0.01 0.05

inspect(head(sort(rules.lhs, by = "confidence"))) #Inspect rules.lhs of first rules.rhs sorted by confidence

## lhs rhs support confidence  
## [1] {Garden.Division} => {Health.Products.Division} 0.27210884 1.0000000   
## [2] {Garden.Division} => {Personal.Electronics.Division} 0.16326531 0.6000000   
## [3] {Garden.Division} => {Housewares.Division} 0.13285314 0.4882353   
## [4] {Garden.Division} => {Jewelry.Division} 0.12725090 0.4676471   
## [5] {Garden.Division} => {Novelty.Gift.Division} 0.11224490 0.4125000   
## [6] {Garden.Division} => {Automotive.Division} 0.08543417 0.3139706   
## lift count  
## [1] 1.000000 1360   
## [2] 1.283733 816   
## [3] 1.240569 664   
## [4] 1.310146 636   
## [5] 1.813259 561   
## [6] 2.328227 427

### Relationship Chart

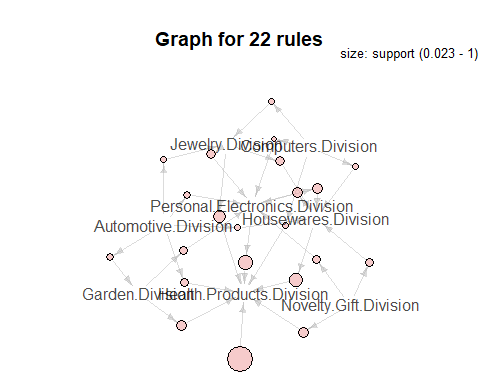
library(arulesViz) #Call to the library arulesViz

## Warning: package 'arulesViz' was built under R version 3.6.3

## Loading required package: grid

## Registered S3 method overwritten by 'seriation':  
## method from   
## reorder.hclust gclus

plot(rules,method="graph",shading=NA) #Plot relationship chart from rules



**Summary**

For my final project I choose to work on Option 2 using the CatalogCrossSell dataset to analyses catalog cross-selling using association rules.

The first step was to inspect the dataset and was able to determine that the data continue ten columns of 4999 rows continuing customer ids on the first column and the nine other columns where binary data indicating what type of catalog was bought. From there I proceeded to read the data set into the dataframe getting rid of the header data, column bringing the entries to 4998. Next, I would run the dim function to confirm the dimension data of the dataframe and noticed that it was returning with the values 60004 rows and 20 columns which didn’t match our initial evaluation. Upon looking at the view function I was able to determine that data set had an extended range in the csv file thus was return several NA values that would cause errors in data analytics. Thus, the data had to be sub setting to remove the NA values from our data frame by removing both the additional 10 columns and 55,006‬ other NA entries.

The next step was to run summary of the subset and from that data we weren’t able to really gain really insightful data. After that I needed to create a matrix from the data frame and from there create a transactions database from the matrix which I examined at each step. From there I wanted to create a visualization of the item frequency we can see that the Health Products catalog are the far most frequent product bought.

The last step was to process the association rules using the Apriori Function looking the for-support values of atleast 0.01, confidence of 0.5, and rules with a maximum length of 2. After that I sorted the rules by lift from there, we were given a set of 22 rules, with the strongest rule being people who bought Automotive catalogs are more likely to buy Garden catalogs with a support value of 0.08543417, a confidence value of 0.6335312, and a lift of 2.328227. From there I decided to look at further analysis of looking at all rules that predicted Garden catalogs on the right hand side and vice versa with the looking at the rules related to garden on the left hand side where as I observed that the lift and support of those who bought garden catalogs who also bought automotive was the same but the confidence value was lower indicating that while it may be true that those who buy automotive may also buy garden catalogs the inverse may not be as true. Following this I created a relationship chart of the 22 rules utilizing the arulesViz libraray.