Portfolio Project Option 2

Jessica Ham

11/03/2019

## Load/Explore the Data

### Load

sale.df <- read.csv('C:/Users/Jess/Documents/CSU File Storage/CatalogCrossSell.csv', header = TRUE)  
library(arules)

## Warning: package 'arules' was built under R version 3.5.3

## Loading required package: Matrix

##   
## Attaching package: 'arules'

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

library(arulesViz)

## Warning: package 'arulesViz' was built under R version 3.5.3

## Loading required package: grid

### Explore

dim(sale.df)

## [1] 4998 10

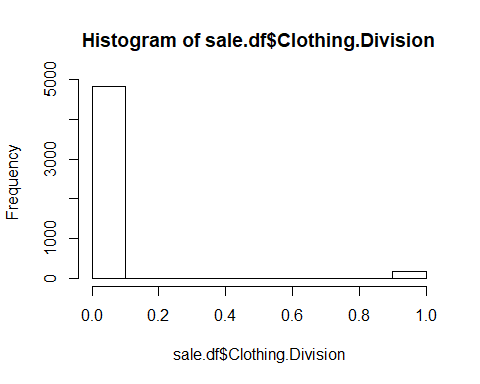
head(sale.df)

## Customer.Number Clothing.Division Housewares.Division  
## 1 11569 0 1  
## 2 13714 0 1  
## 3 46391 0 1  
## 4 67264 0 0  
## 5 67363 0 0  
## 6 72553 0 1  
## Health.Products.Division Automotive.Division  
## 1 1 1  
## 2 1 1  
## 3 1 1  
## 4 1 1  
## 5 1 0  
## 6 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  
## Novelty.Gift.Division Jewelry.Division  
## 1 1 0  
## 2 1 1  
## 3 1 1  
## 4 1 0  
## 5 1 0  
## 6 1 1

table(sale.df$Clothing.Division, dnn = "Clothing")

## Clothing  
## 0 1   
## 4833 165

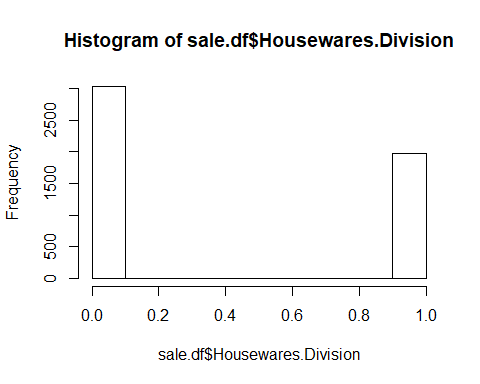
hist(sale.df$Clothing.Division)



table(sale.df$Housewares.Division, dnn = "Houseware")

## Houseware  
## 0 1   
## 3031 1967

hist(sale.df$Housewares.Division)



## Conversions

### Convert to Matrix and review

sale.mat <-as.matrix(sale.df[-1])  
head(sale.mat)

## Clothing.Division Housewares.Division Health.Products.Division  
## [1,] 0 1 1  
## [2,] 0 1 1  
## [3,] 0 1 1  
## [4,] 0 0 1  
## [5,] 0 0 1  
## [6,] 0 1 1  
## Automotive.Division Personal.Electronics.Division Computers.Division  
## [1,] 1 1 0  
## [2,] 1 1 0  
## [3,] 1 1 0  
## [4,] 1 1 0  
## [5,] 0 1 0  
## [6,] 1 1 0  
## Garden.Division Novelty.Gift.Division Jewelry.Division  
## [1,] 0 1 0  
## [2,] 1 1 1  
## [3,] 1 1 1  
## [4,] 1 1 0  
## [5,] 1 1 0  
## [6,] 1 1 1

### Convert to Transactional DB and review

sale.trans <- as(sale.mat, "transactions")  
inspect(head(sale.trans, n = 5))

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

## Rules

### Get Rules at 0.1 Conf and Inspect

rules <- apriori(sale.trans, parameter = list(supp = 0.01, conf = 0.1, target = "rules"))

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.1 0.1 1 none FALSE TRUE 5 0.01 1  
## maxlen target ext  
## 10 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 ...[9 item(s), 4998 transaction(s)] done [0.00s].  
## sorting and recoding items ... [9 item(s)] done [0.00s].  
## creating transaction tree ... done [0.00s].  
## checking subsets of size 1 2 3 4 5 6 7 done [0.00s].  
## writing ... [572 rule(s)] done [0.00s].  
## creating S4 object ... done [0.00s].

inspect(head(sort(rules, by="lift"), n=5))

## lhs rhs support confidence lift count  
## [1] {Personal.Electronics.Division,   
## Garden.Division,   
## Novelty.Gift.Division,   
## Jewelry.Division} => {Automotive.Division} 0.02400960 0.4511278 3.345307 120  
## [2] {Health.Products.Division,   
## Personal.Electronics.Division,   
## Garden.Division,   
## Novelty.Gift.Division,   
## Jewelry.Division} => {Automotive.Division} 0.02400960 0.4511278 3.345307 120  
## [3] {Housewares.Division,   
## Personal.Electronics.Division,   
## Garden.Division,   
## Novelty.Gift.Division,   
## Jewelry.Division} => {Automotive.Division} 0.01720688 0.4502618 3.338885 86  
## [4] {Housewares.Division,   
## Health.Products.Division,   
## Personal.Electronics.Division,   
## Garden.Division,   
## Novelty.Gift.Division,   
## Jewelry.Division} => {Automotive.Division} 0.01720688 0.4502618 3.338885 86  
## [5] {Garden.Division,   
## Novelty.Gift.Division,   
## Jewelry.Division} => {Automotive.Division} 0.02781112 0.4412698 3.272206 139

### Get Rules at 0.5 Conf and Inspect

rules <- apriori(sale.trans, parameter = list(supp = 0.01, conf = 0.5, target = "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  
## 10 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 ...[9 item(s), 4998 transaction(s)] done [0.00s].  
## sorting and recoding items ... [9 item(s)] done [0.00s].  
## creating transaction tree ... done [0.00s].  
## checking subsets of size 1 2 3 4 5 6 7 done [0.00s].  
## writing ... [424 rule(s)] done [0.00s].  
## creating S4 object ... done [0.00s].

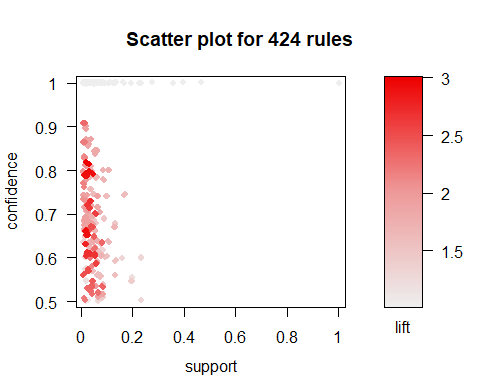
inspect(head(sort(rules, by="lift"), n=5))

## lhs rhs support confidence lift count  
## [1] {Automotive.Division,   
## Personal.Electronics.Division,   
## Novelty.Gift.Division,   
## Jewelry.Division} => {Garden.Division} 0.02400960 0.8163265 3.000000 120  
## [2] {Health.Products.Division,   
## Automotive.Division,   
## Personal.Electronics.Division,   
## Novelty.Gift.Division,   
## Jewelry.Division} => {Garden.Division} 0.02400960 0.8163265 3.000000 120  
## [3] {Automotive.Division,   
## Novelty.Gift.Division,   
## Jewelry.Division} => {Garden.Division} 0.02781112 0.8128655 2.987281 139  
## [4] {Health.Products.Division,   
## Automotive.Division,   
## Novelty.Gift.Division,   
## Jewelry.Division} => {Garden.Division} 0.02781112 0.8128655 2.987281 139  
## [5] {Automotive.Division,   
## Personal.Electronics.Division,   
## Novelty.Gift.Division} => {Garden.Division} 0.03541417 0.7972973 2.930068 177

## Visualization

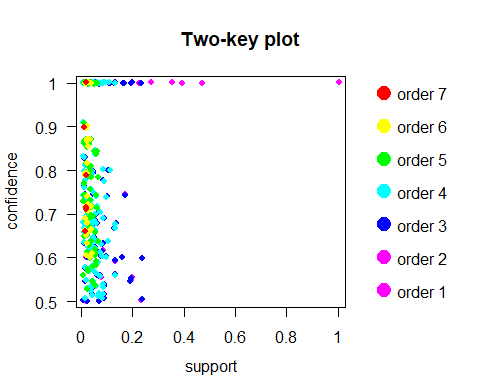
plot(rules)

## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.



plot(rules, method = "two-key plot")

## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.



### Adjust MaxLen

rules.3 <- apriori(sale.trans, parameter = list(supp = 0.01, conf = 0.5, maxlen=3, target = "rules"))

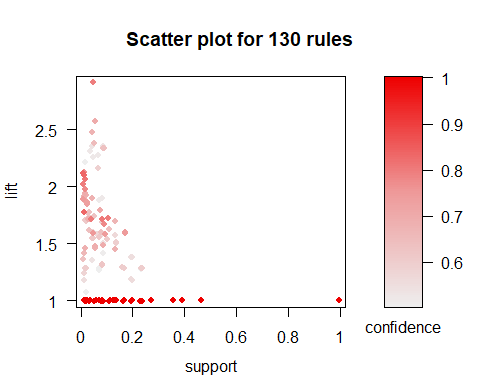
## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.5 0.1 1 none FALSE TRUE 5 0.01 1  
## maxlen target ext  
## 3 rules FALSE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 49   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[9 item(s), 4998 transaction(s)] done [0.00s].  
## sorting and recoding items ... [9 item(s)] done [0.00s].  
## creating transaction tree ... done [0.00s].  
## checking subsets of size 1 2 3

## Warning in apriori(sale.trans, parameter = list(supp = 0.01, conf =  
## 0.5, : Mining stopped (maxlen reached). Only patterns up to a length of 3  
## returned!

## done [0.00s].  
## writing ... [130 rule(s)] done [0.00s].  
## creating S4 object ... done [0.00s].

plot(rules.3, measure = c("support", "lift"), shading = "confidence")

## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.



inspect(head(sort(rules.3, by="lift"), n=5))

## lhs rhs support confidence lift count  
## [1] {Automotive.Division,   
## Novelty.Gift.Division} => {Garden.Division} 0.04401761 0.7913669 2.908273 220  
## [2] {Automotive.Division,   
## Personal.Electronics.Division} => {Garden.Division} 0.05802321 0.6987952 2.568072 290  
## [3] {Housewares.Division,   
## Automotive.Division} => {Garden.Division} 0.04541817 0.6715976 2.468121 227  
## [4] {Automotive.Division,   
## Jewelry.Division} => {Garden.Division} 0.04861945 0.6462766 2.375066 243  
## [5] {Automotive.Division,   
## Personal.Electronics.Division} => {Novelty.Gift.Division} 0.04441777 0.5349398 2.351477 222

### Export Rules to Manually Trim

write(rules.3,  
 file = "C:/Users/Jess/Documents/CSU File Storage/association\_rules.csv",  
 sep = ",",  
 quote = TRUE,  
 row.names = FALSE)

Portfolio Summary:

The purpose of this project was to identify rules that could be used by a mail order or online catalog sales company. The goal was to create rules for what catalog(s) or coupon(s) to include with shipment of orders from other catalogs. The data was first loaded and reviewed to understand the data: we are looking at 9 departments and using 4,998 transactions to determine our rules. While I looked at the table and histogram for each division to understand the popularity of that product group, for the purposes of brevity I only included a couple in the RMarkdown.

Conversion to a matrix (removing the transaction number) and then conversion to a transactional database are the next two steps. Next, we run the apriori function to find the association rules. Again for brevity we are only looking at the first five rows in the RMarkdown document. For the first pass, I ran a fairly broad algorithm (0.01 support and 0.1 confidence) to catch the most rules. With a 0.01 support, we need to see at least 49 transactions supporting the data to make a rule. At our low confidence, we generated 572 rules, and looking at some of the top lifts, they are not high confidence. The next stage was to increase the confidence, ruling out some of the lower confidence but high lift rules. At a minimum of 0.5 confidence, we have 424 rules. Still too many to use, but improving and honing in on the most reliable rules.

Here I utilized the arulesViz library to generate scatter plots that can help get an overview of large numbers of rules. The first plot shows that I was likely smart to leave the support low, as there are only a few rules higher than 0.1 confidence at many high confidence/high lift rules don’t have more than 50 transactions to support. With the larger data set, I would consider that volume of transactions sufficient to indicate a trend. The next plot is a two key, which is colored based on the orders (number of antecedents) and one thing we can see in this graph as well as the top 5 from the algorithm is our top lifts are high antecedents. First, they are likely repetitive of smaller antecedents, and secondly, they don’t serve our purpose, as the rule would require multiple purchases before a shopper qualified for a catalog in their shipment. These are too specific for our purposes, so I chose to re-run the apriori algorithm with a max length of 3, focusing only on the rules with 1 or two antecedents. This generated 130 rules which, if we run a new scatter plot, show a group that have a lift of 1 (no better than random) and should be removed and more concentrated group of rules.

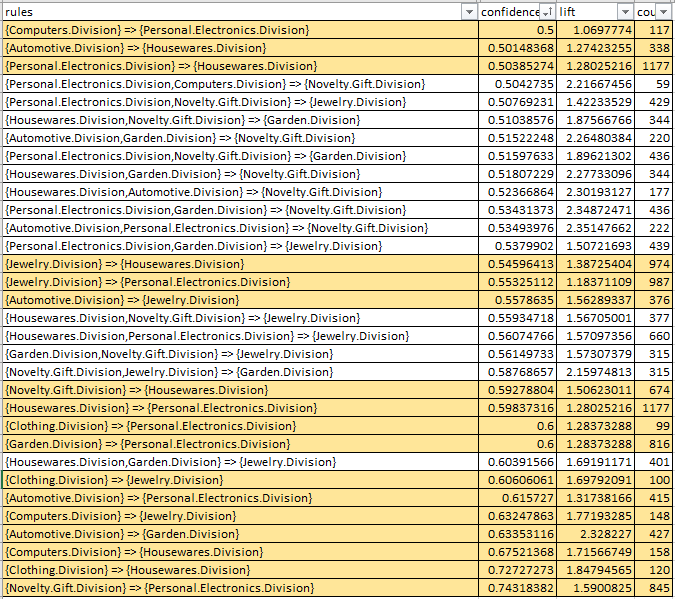
At this point, the use of subsets was beyond the scope of our textbook, and it rather described stripping excess rules by hand. It was because of this instruction that I chose to pull the file into Excel where I could continue to prune the rules. After removing rules that had a lift of 1, I first focused on the base rules (1 antecedent to 1 consequent), which would indicate a simple means of calculating which catalogs should be sent. There are 17 of these rules. I removed any rule that would be covered by these base rules (rules that had 2 antecedents, but were fully covered by the base rules). This left me with 32 rules with confidences between 0.5 to 0.75 and lifts between 1 and 2.35. For the purpose of focusing our marketing, these rules should cover all possibilities but limit what catalogs should go with what orders. If I was selling this to my superiors I would suggest that if more than 3 rules are triggered by a sale, you select the two best rules to determine what catalogs to send. My final list of rules is included in the figure below. Base rules are highlighted. 

Figure 1 Final Association Rules