Market Basket Analysis

Kiran Kumar Golanakonda

17/06/2022

Summary

Market basket analysis involves the data mining technique utilized in finding the association between item sets and those identifiable from the customer's behavior. The project has utilized the R programming tool to perform the market basket analysis process. The results for this project are essential as they will help provide information to the retailers to understand their potential buyers' purchase behavior, which is essential in the decision-making process. The current algorithms implemented in market basket analysis only work on static data and not on data that changes with time. Therefore, the project has implemented an associating rule data mining technique that has helped examine the customer's purchasing behavior. The project answered the research question of the likelihood of customers purchasing products along with a particular product category. The dataset for this project has been obtained from UCI Machine Learning Repository that includes transactions for UK-based online retail.

Introduction

Recently, customers have been provided with a wide range of independent options in almost any domain. In the past, customers had to choose from a catalog when purchasing anything from a given store. However, with technological advancement, there has been an exponential increase in the number of options for consumers where they can make choices from a wide variety of products. This has resulted in limitless possibilities for businesses. With this, many new competitors have entered the market, leading to increased competition. Therefore, the retail stores are seeking marketing strategies to attract new customers and maintain their current customers. There has been remarkable success with the Market basket analysis strategy. Therefore, to gain insights into market basket analysis, the project has used information concerning the customer purchasing what they buy and when they buy. The results from the project will help shop managers with the ability to determine the strong association between goods that are important in placing the products that co-occur close to one another. The shop managers will also be able to make a decision that involves store shelf arrangement, cross-selling, items to stock more, and up-selling.

Literature Review

Kamley et al. (2014), in their study, developed an association rule mining model to help gain insights into the interesting patterns in the stock market dataset. With the model, it was possible to predict the price of shares essential for investors and stock brokers to help them invest in the right many to understand market conditions. In a study by Solnet et al. (2016), the authors tried to understand the potential of market basket analysis for the revenue growth of hotels. From the research, the authors could derive the products and services that were satisfying and attractive to the guests. In their study, Kapadia & Kalyandurgmath (2015) analyzed consumer behavior patterns. The study provided insights related to the basket formation that helped manage stocks and product assortments for products sold, cross-selling, giving discounts to loyal customers, and making promotions on the likely products sold. There was a highlight by Mushtaq & Kanth (2015) on how helpful data mining is in marketing, especially on net profit, market analysis, and better marketing strategies.

Data

The dataset for the Online Retail was retrieved from <http://archive.ics.uci.edu/ml/machine-learning-databases/00352/Online> Retail.xlsx. This dataset includes:

## # A tibble: 6 × 8

## InvoiceNo StockCode Description Quantity InvoiceDate UnitPrice

## <chr> <chr> <chr> <dbl> <dttm> <dbl>

## 1 536365 85123A WHITE HANGING HEAR… 6 2010-12-01 08:26:00 2.55

## 2 536365 71053 WHITE METAL LANTERN 6 2010-12-01 08:26:00 3.39

## 3 536365 84406B CREAM CUPID HEARTS… 8 2010-12-01 08:26:00 2.75

## 4 536365 84029G KNITTED UNION FLAG… 6 2010-12-01 08:26:00 3.39

## 5 536365 84029E RED WOOLLY HOTTIE … 6 2010-12-01 08:26:00 3.39

## 6 536365 22752 SET 7 BABUSHKA NES… 2 2010-12-01 08:26:00 7.65

## # … with 2 more variables: CustomerID <dbl>, Country <chr>

To work with the data there needed to have clean rows with no missing values. To achieve this the complete.cases(data) approach has been used

retail\_data <- retail\_data[complete.cases(retail\_data), ]

head(retail\_data)

## # A tibble: 6 × 8

## InvoiceNo StockCode Description Quantity InvoiceDate UnitPrice

## <chr> <chr> <chr> <dbl> <dttm> <dbl>

## 1 536365 85123A WHITE HANGING HEAR… 6 2010-12-01 08:26:00 2.55

## 2 536365 71053 WHITE METAL LANTERN 6 2010-12-01 08:26:00 3.39

## 3 536365 84406B CREAM CUPID HEARTS… 8 2010-12-01 08:26:00 2.75

## 4 536365 84029G KNITTED UNION FLAG… 6 2010-12-01 08:26:00 3.39

## 5 536365 84029E RED WOOLLY HOTTIE … 6 2010-12-01 08:26:00 3.39

## 6 536365 22752 SET 7 BABUSHKA NES… 2 2010-12-01 08:26:00 7.65

## # … with 2 more variables: CustomerID <dbl>, Country <chr>

By use of the mutate function the Description column and Country column has to be converted to factor column.

Converts character data to date.Store InvoiceDate as date in new variable and Extract time from InvoiceDate and store in another variable, and lastly Convert and edit InvoiceNo into numeric

retail\_data$Date <- as.Date(retail\_data$InvoiceDate)

TransTime<- format(retail\_data$InvoiceDate,"%H:%M:%S")

InvoiceNo <- as.numeric(as.character(retail\_data$InvoiceNo))

## Warning: NAs introduced by coercion

Bind new columns TransTime and InvoiceNo into dataframe retail\_data

#get a glimpse of data

glimpse(retail\_data)

## Rows: 406,829

## Columns: 9

## $ InvoiceNo <chr> "536365", "536365", "536365", "536365", "536365", "536365"…

## $ StockCode <chr> "85123A", "71053", "84406B", "84029G", "84029E", "22752", …

## $ Description <chr> "WHITE HANGING HEART T-LIGHT HOLDER", "WHITE METAL LANTERN…

## $ Quantity <dbl> 6, 6, 8, 6, 6, 2, 6, 6, 6, 32, 6, 6, 8, 6, 6, 3, 2, 3, 3, …

## $ InvoiceDate <dttm> 2010-12-01 08:26:00, 2010-12-01 08:26:00, 2010-12-01 08:2…

## $ UnitPrice <dbl> 2.55, 3.39, 2.75, 3.39, 3.39, 7.65, 4.25, 1.85, 1.85, 1.69…

## $ CustomerID <dbl> 17850, 17850, 17850, 17850, 17850, 17850, 17850, 17850, 17…

## $ Country <chr> "United Kingdom", "United Kingdom", "United Kingdom", "Uni…

## $ Date <date> 2010-12-01, 2010-12-01, 2010-12-01, 2010-12-01, 2010-12-0…

Next, the dataframe is converted to transaction data to ensure items bought together in one invoice are placed in one row. The data is grouped by InvoiceNo and date which are placed as one row.

library(plyr)

transaction\_Retail\_Data <- ddply(retail\_data,c("InvoiceNo","Date"),

function(df1)paste(df1$Description,

collapse = ","))

head(transaction\_Retail\_Data)

## InvoiceNo Date

## 1 536365 2010-12-01

## 2 536366 2010-12-01

## 3 536367 2010-12-01

## 4 536368 2010-12-01

## 5 536369 2010-12-01

## 6 536370 2010-12-01

## V1

## 1 WHITE HANGING HEART T-LIGHT HOLDER,WHITE METAL LANTERN,CREAM CUPID HEARTS COAT HANGER,KNITTED UNION FLAG HOT WATER BOTTLE,RED WOOLLY HOTTIE WHITE HEART.,SET 7 BABUSHKA NESTING BOXES,GLASS STAR FROSTED T-LIGHT HOLDER

## 2 HAND WARMER UNION JACK,HAND WARMER RED POLKA DOT

## 3 ASSORTED COLOUR BIRD ORNAMENT,POPPY'S PLAYHOUSE BEDROOM,POPPY'S PLAYHOUSE KITCHEN,FELTCRAFT PRINCESS CHARLOTTE DOLL,IVORY KNITTED MUG COSY,BOX OF 6 ASSORTED COLOUR TEASPOONS,BOX OF VINTAGE JIGSAW BLOCKS,BOX OF VINTAGE ALPHABET BLOCKS,HOME BUILDING BLOCK WORD,LOVE BUILDING BLOCK WORD,RECIPE BOX WITH METAL HEART,DOORMAT NEW ENGLAND

## 4 JAM MAKING SET WITH JARS,RED COAT RACK PARIS FASHION,YELLOW COAT RACK PARIS FASHION,BLUE COAT RACK PARIS FASHION

## 5 BATH BUILDING BLOCK WORD

## 6 ALARM CLOCK BAKELIKE PINK,ALARM CLOCK BAKELIKE RED,ALARM CLOCK BAKELIKE GREEN,PANDA AND BUNNIES STICKER SHEET,STARS GIFT TAPE,INFLATABLE POLITICAL GLOBE,VINTAGE HEADS AND TAILS CARD GAME,SET/2 RED RETROSPOT TEA TOWELS,ROUND SNACK BOXES SET OF4 WOODLAND,SPACEBOY LUNCH BOX,LUNCH BOX I LOVE LONDON,CIRCUS PARADE LUNCH BOX,CHARLOTTE BAG DOLLY GIRL DESIGN,RED TOADSTOOL LED NIGHT LIGHT,SET 2 TEA TOWELS I LOVE LONDON,VINTAGE SEASIDE JIGSAW PUZZLES,MINI JIGSAW CIRCUS PARADE,MINI JIGSAW SPACEBOY,MINI PAINT SET VINTAGE,POSTAGE

Next, The InvoiceNo and Date are set to NULL as they were not needed in the rule mining.

#set column InvoiceNo

transaction\_Retail\_Data$InvoiceNo <- NULL

#set column Date

transaction\_Retail\_Data$Date <- NULL

#Rename column to items

colnames(transaction\_Retail\_Data) <- c("items")

#Show Dataframe transaction\_Retail\_Data

head(transaction\_Retail\_Data)

## items

## 1 WHITE HANGING HEART T-LIGHT HOLDER,WHITE METAL LANTERN,CREAM CUPID HEARTS COAT HANGER,KNITTED UNION FLAG HOT WATER BOTTLE,RED WOOLLY HOTTIE WHITE HEART.,SET 7 BABUSHKA NESTING BOXES,GLASS STAR FROSTED T-LIGHT HOLDER

## 2 HAND WARMER UNION JACK,HAND WARMER RED POLKA DOT

## 3 ASSORTED COLOUR BIRD ORNAMENT,POPPY'S PLAYHOUSE BEDROOM,POPPY'S PLAYHOUSE KITCHEN,FELTCRAFT PRINCESS CHARLOTTE DOLL,IVORY KNITTED MUG COSY,BOX OF 6 ASSORTED COLOUR TEASPOONS,BOX OF VINTAGE JIGSAW BLOCKS,BOX OF VINTAGE ALPHABET BLOCKS,HOME BUILDING BLOCK WORD,LOVE BUILDING BLOCK WORD,RECIPE BOX WITH METAL HEART,DOORMAT NEW ENGLAND

## 4 JAM MAKING SET WITH JARS,RED COAT RACK PARIS FASHION,YELLOW COAT RACK PARIS FASHION,BLUE COAT RACK PARIS FASHION

## 5 BATH BUILDING BLOCK WORD

## 6 ALARM CLOCK BAKELIKE PINK,ALARM CLOCK BAKELIKE RED,ALARM CLOCK BAKELIKE GREEN,PANDA AND BUNNIES STICKER SHEET,STARS GIFT TAPE,INFLATABLE POLITICAL GLOBE,VINTAGE HEADS AND TAILS CARD GAME,SET/2 RED RETROSPOT TEA TOWELS,ROUND SNACK BOXES SET OF4 WOODLAND,SPACEBOY LUNCH BOX,LUNCH BOX I LOVE LONDON,CIRCUS PARADE LUNCH BOX,CHARLOTTE BAG DOLLY GIRL DESIGN,RED TOADSTOOL LED NIGHT LIGHT,SET 2 TEA TOWELS I LOVE LONDON,VINTAGE SEASIDE JIGSAW PUZZLES,MINI JIGSAW CIRCUS PARADE,MINI JIGSAW SPACEBOY,MINI PAINT SET VINTAGE,POSTAGE

The transactions data had to be store into a .csv file.

write.csv(transaction\_Retail\_Data,"C:/Users/User/Downloads/transactions.csv", quote = FALSE, row.names = FALSE)

Methodology

Having the data converted to transactions data for rule mining, next was to load the transaction data into transaction class by use of arules package.

options(warn=-1)

trans\_class <- read.transactions('C:/Users/User/Downloads/transactions.csv', format = 'basket', sep=',')

The transactions object contains 22191 transactions and 7876 items. That is 7876 includes the product descriptions from the dataset and 22191 include the collection of items in the dataset. To view the distributiob of objects, an itemFrequencyPlot was created.

# Create an item frequency plot for the top 20 items

if (!require("RColorBrewer")) {

# install color package of R

install.packages("RColorBrewer")

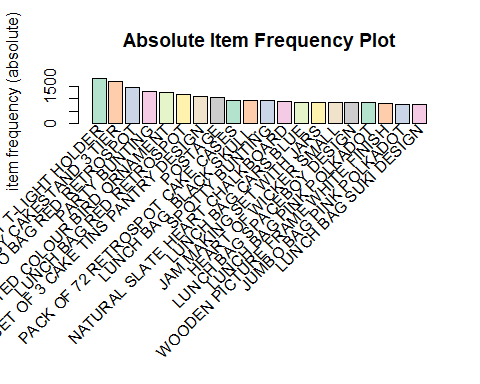
#include library RColorBrewer

library(RColorBrewer)

}

## Loading required package: RColorBrewer

itemFrequencyPlot(trans\_class,topN=20,type="absolute",col=brewer.pal(8,'Pastel2'), main="Absolute Item Frequency Plot")



Next step involve rule mining through the use of APRIORI algorithm.

# Min Support as 0.001, confidence as 0.8.

association.rules <- apriori(trans\_class, parameter = list(supp=0.001, conf=0.8,maxlen=10))

## Apriori

##

## Parameter specification:

## confidence minval smax arem aval originalSupport maxtime support minlen

## 0.8 0.1 1 none FALSE TRUE 5 0.001 1

## maxlen target ext

## 10 rules TRUE

##

## Algorithmic control:

## filter tree heap memopt load sort verbose

## 0.1 TRUE TRUE FALSE TRUE 2 TRUE

##

## Absolute minimum support count: 22

##

## set item appearances ...[0 item(s)] done [0.00s].

## set transactions ...[7876 item(s), 22191 transaction(s)] done [0.55s].

## sorting and recoding items ... [2324 item(s)] done [0.02s].

## creating transaction tree ... done [0.02s].

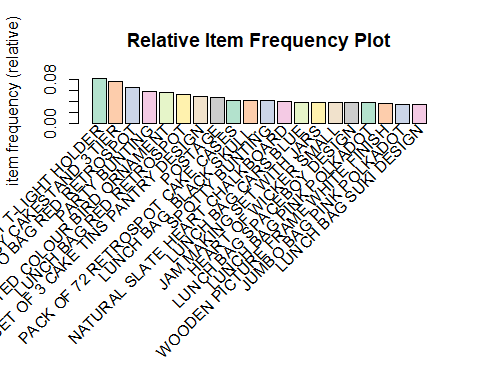
## checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [0.68s].

## writing ... [49122 rule(s)] done [0.07s].

## creating S4 object ... done [0.07s].

Results

Having the rule mining, it was clear that there were several association rules. The analysis indicates that ‘WHITE HANGING HEART T-LIGHT HOLDER’ and ‘REGENCY CAKESTAND 3 TIER’ have the most sales.

 100% of the customers who bought ‘WOBBLY CHICKEN’ also bought ‘METAL’.

inspect(association.rules[1:10])

## lhs rhs support confidence coverage lift count

## [1] {WOBBLY CHICKEN} => {DECORATION} 0.001261773 1.0000000 0.001261773 443.8200 28

## [2] {WOBBLY CHICKEN} => {METAL} 0.001261773 1.0000000 0.001261773 443.8200 28

## [3] {DECOUPAGE} => {GREETING CARD} 0.001036456 1.0000000 0.001036456 389.3158 23

## [4] {BILLBOARD FONTS DESIGN} => {WRAP} 0.001306836 1.0000000 0.001306836 715.8387 29

## [5] {WRAP} => {BILLBOARD FONTS DESIGN} 0.001306836 0.9354839 0.001396963 715.8387 29

## [6] {ENAMEL PINK TEA CONTAINER} => {ENAMEL PINK COFFEE CONTAINER} 0.001396963 0.8157895 0.001712406 385.1741 31

## [7] {WOBBLY RABBIT} => {DECORATION} 0.001532153 1.0000000 0.001532153 443.8200 34

## [8] {WOBBLY RABBIT} => {METAL} 0.001532153 1.0000000 0.001532153 443.8200 34

## [9] {ART LIGHTS} => {FUNK MONKEY} 0.001712406 1.0000000 0.001712406 583.9737 38

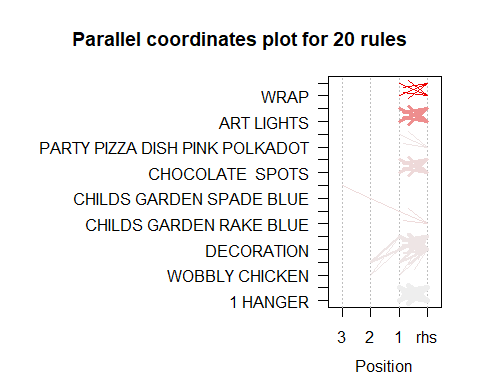
## [10] {FUNK MONKEY} => {ART LIGHTS} 0.001712406 1.0000000 0.001712406 583.9737 38

# Filter top 20 rules with highest lift

subRules<-association.rules[quality(association.rules)$confidence>0.4]

subRules2<-head(subRules, n=20, by="lift")

plot(subRules2, method="paracoord")



From the top most arrow, it is depicted that if a customer ‘CHILDS GARDEN SPADE PINK’ and ‘CHILDS GARDEN RAKE PINK’ in my shopping cart, they are likely to buy ‘CHILDS GARDEN RAKE BLUE’ along with these as well.

Conclusion

It is possible to perform market basket analysis and identify patterns through which items are bought together. The project has performed market basket analysis using the R and Apriori algorithms. From the analysis, it is clear that if a customer ‘CHILDS GARDEN SPADE PINK’ and ‘CHILDS GARDEN RAKE PINK’ in my shopping cart, they are likely to buy ‘CHILDS GARDEN RAKE BLUE’ along with these as well.

References

Kamley, S., Jaloree, S., & Thakur, R. S. (2014). An Association Rule Mining Model for Finding the Interesting Patterns in Stock Market Dataset. International Journal of Computer Applications, 93(9).

Mushtaq, A., & Kanth, H. (2015). Data mining for marketing. International Journal on Recent and Innovation Trends in Computing and Communication, 3(3), 985-991.

Kapadia, G., & Kalyandurgmath, K. (2015). Market basket analysis of consumer buying behaviour of a lifestyle store. In International Conference on Technology and Business Management (pp. 406-412).

Solnet, D., Boztug, Y., & Dolnicar, S. (2016). An untapped gold mine? Exploring the potential of market basket analysis to grow hotel revenue. International Journal of Hospitality Management, 56, 119-125.