“ONLINE MARKET BASKET ANALYSIS”

Submitted by
Jayesh Shelar (B194063)
Samruddhi Yadav (B194023)
Arjun Yachwad (B194032)
Prashant Walunj (B194158)

MIT Academy of Engineering

(An Autonomous Institute affiliated to Savitribai Phule Pune University)

School of Computer Engineering and Technology
MIT ACADEMY OF ENGINEERING
2019-2020

ACKNOWLEDGMENT

We take this opportunity to record our profound gratitude and indebtedness to out subject teacher for her inspiring guidance, valuable advices, constant encouragement and untiring supervision throughout our project work. We express our deep sense of gratitude to her for providing an opportunity to work on this project and also her continuous inspiration and encouragement.

Finally, we would like to acknowledge and express our gratitude towards our teammates, family, friends and classmates for their patience, encouragement and support without whom this project would not have been completed.

November 2019

Jayesh Shelar (B194063)
Samruddhi Yadav (B194023)
Arjun Yachwad (B194032)
Prashant Walunj (B194158)

CONTENTS

1. INTRODUCTION

Overview

The field of market basket analysis, the search for meaningful associations in customer purchase data, is one of the oldest areas of data mining. The typical solution involves the mining and analysis of association rules.

Market Basket Analysis is a modeling technique with which one could analyze patterns (generally shopping) from a given database. It works on the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items. With the trends predicted in shopping patterns, we could analyze the general shopping patterns, most and least bought items, shopping season etc. These inferences would help us to boost our sales and increase our profit incentives.

It is one of the key techniques used by retailers to uncover association between items. It works by looking at combinations of items that occur together frequently in transactions. Association rules are used to analyze the database and to identify strong association rules.

Apriori algorithm is used for frequent item set mining and predicting association rules. It works by identifying frequent individual items in the database and extends them to larger and larger item sets as long as there appear frequently (over minimum support and minimum confidence).

Now, those derived item sets determined by Apriori algorithm are used to define association rules which highlight the general trends in the database.

Background and Motivation

Market Basket Analysis is a key technique used to analyze market trends and shopping patterns.

These days, to boost the sales growth, one has to have a keen and sharp mind and must be in sync with the market trends. A good businessman realizes the need to go with the market, to know what the consumer needs, to observe the products which have a high demand and to identify seasonal items. One wouldn't buy a jacket in the hot month of June. Or rarely would one buy a swimming costume in the cold month of December. Of course, exceptions are there.

In order to increase sales growth and promote ones business, analyzing market trends in necessary. One must, after the analysis, apply the trends he/she observed for productive purposes.

This observation of trends and then realizing rules from them, motivated us to know and earn more about Market Basket Analysis and its related concepts like Association rules and Apriori Algorithm.

R provided us with a platform where this could become a reality. R helped us to actually analyze a dataset and predict shopping patterns and trends and helped us govern association rules.

Objective

- To identify the time at which most purchases are made
- To identify top 10 best sold products
- To identify number of items bought frequently
- To identify relationships between certain products using association rules
- To visualize the analysis done using graphs

Methodology

We have used Apriori Algorithm and Association Rules for an Analysis

Apriori Algorithm

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database.

Association rule

Consider the following transactions.

1	Flour	Chocolate
2	Fruits	Cream
3	Flour	Chocolate
4	Fruits	Cream
5	Fruits	Cream
6	Flour	Cream

Each row is a transaction. Each cell is an individual item of transaction.
These are transactions of 6 people who bought some items from a grocery store.

Support - Indication of how frequently an item is bought.
Confidence - Indication of how often a rule has been found true.
Lift - Correlation between 2 items.

The minimum support and minimum confidence is decided for an analysis, Support of items which exceeds minimum support will be considered for further analysis. Confidence of items which exceeds minimum confidence will be considered for further analysis.

Rule: All those who buy fruits, will also buy cream.
Support (Fruits, Cream) = 3/6
Confidence (Fruits, Cream) = 1

2. DATA SET

Description

The dataset contains transaction data from 01/12/2010 to 09/12/2011 for a UK-based registered non-store online retail.

Number of Rows: 541909
Number of Attributes: 08

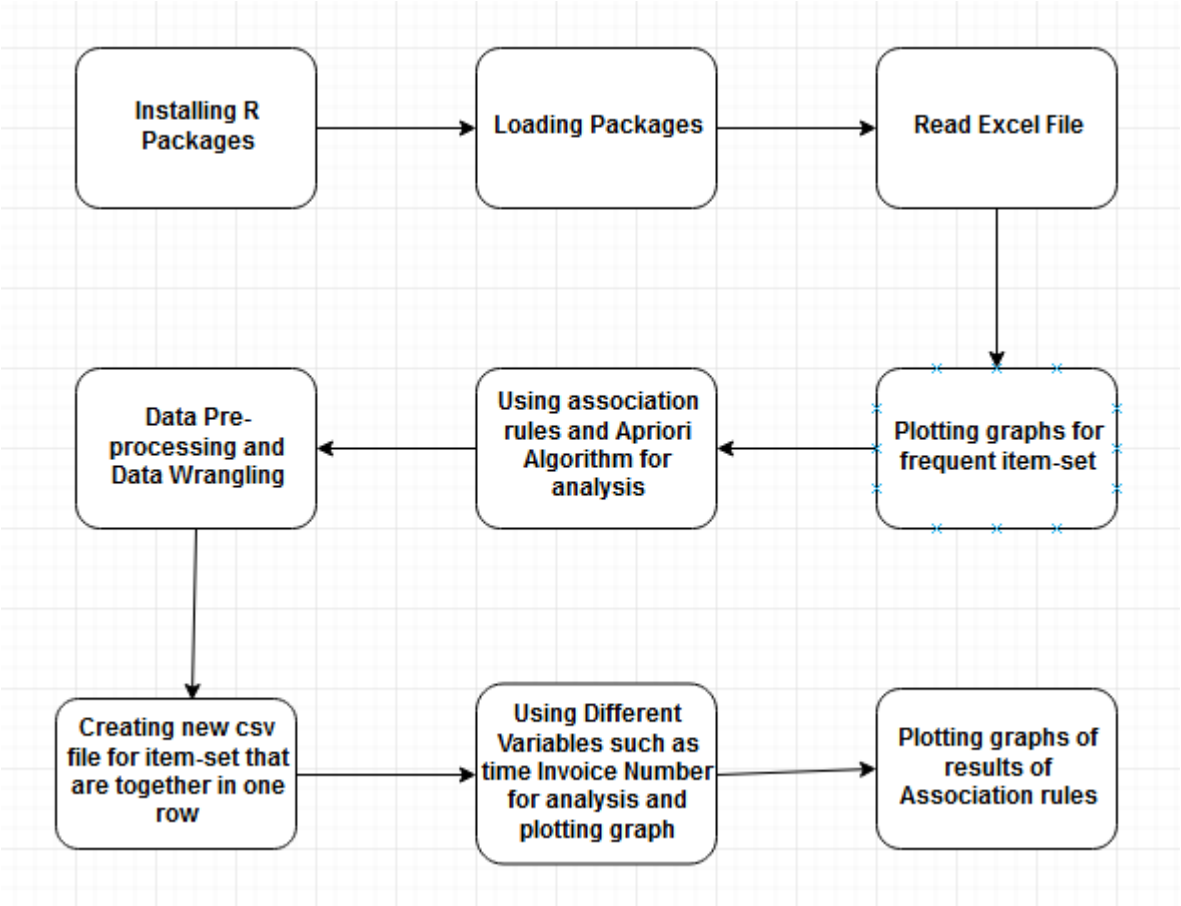
Number of attributes used

Attribute Information

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation. +StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- StockCode: Unique VarChar code for each product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated. Example from dataset: 12/1/2010 8:26
- UnitPrice: Unit price. Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

3. SYSTEM WORKING

Block Diagram



Code

```
#installing packages

install.packages("tidyverse")

install.packages("readxl")

install.packages("knitr")

install.packages("ggplot2")

install.packages("lubridate")

install.packages("arules")

install.packages("arulesViz")

install.packages("plyr")

install.packages("RColorBrewer")

#loading packages

library(tidyverse)     #designed for data science ->statistics

library(readxl)        # to read excel files

library(knitr)         # to identify default settings if not satisfied

library(ggplot2)       # for plotting graphs

library(lubridate)     # to work on date and time format

library(arules)        # for association rules and apriori algorithm functions

library(arulesViz)     # for visualization of association rules and apriori algorithm like plots

library(plyr)          # splitting and combining data
```

```
library(dplyr)        # for data manipulation

library(RColorBrewer) # for color palette

retail = read_excel("C:/rprogramming/Online Retail.xlsx")

retail = retail[complete.cases(retail),]

retail

#complete.cases(data) will return a logical vector indicating which rows have no missing values.

#Then use the vector to get only rows that are complete using retail[,]

retail %>% mutate(Description = as.factor(Description))

retail %>% mutate(Country = as.factor(Country))

retail$Date = as.Date(retail$InvoiceDate)

retail$Date

retail$Time = format(retail$InvoiceDate,"%H:%M:%S")

retail$Time

retail$InvoiceNo = as.numeric(as.character(retail$InvoiceNo))

retail$InvoiceNo

glimpse(retail)

#time at which people often purchase

retail$Time = as.factor(retail$Time)

retail$Time

retail
```

```
retail$Time1 = format(retail$InvoiceDate,"%H")

retail$Time1

ggplot(retail,aes(x=Time1)) + geom_histogram(stat = "count", fill = "indianred")



# how many items customer buy

detach("package:plyr", unload=TRUE)

retail %>%

  group_by(InvoiceNo) %>%

  summarize(n_items = mean(Quantity)) %>%

  ggplot(aes(x=n_items))+

  geom_histogram(fill="indianred", bins = 100000) +

  geom_rug()+

  coord_cartesian(xlim=c(0,80))



# top 10 best seller products

best_seller = retail %>%

  group_by(StockCode, Description) %>%

  summarize(count = n()) %>%

  arrange(desc(count))
```

```
best_seller = head(best_seller,10)

best_seller

best_seller %>%

  ggplot(aes(x=reorder(Description,count), y=count))+

  geom_bar(stat="identity",fill="indian red")+

  coord_flip()

library(plyr)



#to create list of items that are bought together

itemList = ddply(retail,c("InvoiceNo","Date"),

                 function(df1)paste(df1$Description,

                                    collapse = ","))

# paste() concatenates vectors to character and separated results using collapse

itemList$InvoiceNo = NULL

itemList$Date = NULL

#rename column

colnames(itemList) = c("items")

write.csv(itemList,"C:/rprogramming/market_basket.csv", quote = FALSE, row.names =
TRUE)
```

```
transactions =
  read.transactions('C:/rprogramming/market_basket.csv', format = 'basket', sep=',')
summary(transactions)


#graph for top 20 itemsets that are brought frequently
itemFrequencyPlot(transactions,topN=20,type="absolute",
                  col=brewer.pal(8,'Pastel2'),
                  main="Absolute Item Frequency Plot")


#graph for top 20 itemsets that are brought frequently relative to other items
itemFrequencyPlot(transactions,topN=20,
                  type="relative",
                  col=brewer.pal(8,'Pastel2'),
                  main="Relative Item Frequency Plot")
#relative frequency will help to know the items that are brought infrequently w.r.t to others
#Now using Association rules and apriori algo for analysis
association.rules =
  apriori(transactions, parameter = list(supp=0.001, conf=0.8,maxlen=10))
summary(association.rules)
```

```
inspect(association.rules[1:10])

subset.rules =

  which(colSums(is.subset(association.rules, association.rules)) > 1)

# get subset rules in vector

"which() returns the position of elements in the vector for which value is TRUE.

colSums() forms a row and column sums for dataframes and numeric arrays.

is.subset() Determines if elements of one vector contain all the elements of other

"

length(subset.rules)

# to delete the subset from superset as superset will have subset

subset.association.rules. = association.rules[-subset.rules]

length(subset.association.rules.)



#to identify what people buys before buying decoration

decoration.association.rules =

  apriori(transactions, parameter = list(supp=0.001, conf=0.8),appearance =
list(default="lhs",rhs="DECORATION"))

inspect(head(decoration.association.rules))
```

```
#to identify what people buys when they buy metal

metal.association.rules =

  apriori(transactions, parameter = list(supp=0.001, conf=0.8),appearance =
list(lhs="METAL",default="rhs"))

inspect(head(metal.association.rules))



# create vector of top 10 rules for Visualization

toprules = subset.association.rules.[1:10]

inspect(head(toprules))



#Visualization

plot(subset.association.rules.)

#items havinghigher lift have low support

plot(subset.association.rules.,method="two-key plot")

#order defines number of items



#graph will help to visualize rules better

plot(toprules, method="graph")

plot(toprules, method="graph", engine = "htmlwidget")
```

```
#to plot parallel coordinate plot to visualize sales pattern

toprules_lift = head(association.rules, n=20, by ="lift")

plot(toprules_lift, method="paracoord")
```

4. CONCLUSION

We have learned APRIORI algorithm, one of the most frequently used algorithms in data mining. We have learned all about Association Rule Mining, its applications, and its applications in retailing called as **Market Basket Analysis**. We are also now capable of implementing Market Basket Analysis in R and presenting our association rules with some great plots.

5. FUTURE WORK

- Develop software where retailers would upload their databases and from that rules would be generated along with that shopping patterns would also be shown with graphical representation.
- To analyze shopping pattern of a particular customer so as more recommendations and offers can be proposed to that customer

6. REFERENCES

1. https://datascienceplus.com/a-gentle-introduction-on-market-basket-analysis%E2%80%8A-%E2%80%8Aassociation-rules/
2. https://en.wikipedia.org/wiki/Sparse_matrix
3. https://cran.r-project.org/web/packages/arulesViz/vignettes/arulesViz.pdf
4. https://www.datacamp.com/community/tutorials/market-basket-analysis-r
5. https://data.world/aprasla0922/online-retail