What will you purchase next. A recommendar system for groceries

Prasanna Srinivasa Rao

Abstract

Grocery is a 675 brillion dollar annual business in the United States. Virtually every household shops for groceries and over 90% of households shop once a week or more (https://www.forbes.com/sites/richardkestenbaum/2017/01/16/why-online-grocers-are-so-unsuccessful-and-what-amazon-is-doing-about-it/#6c6d98a47f56). There is a huge potential to serve customers better by providing the right products to the right customer at the right time. In this project, I propose to use data science techniques to build such a system.

Specifically, I plan to build a recommendation engine to predict the next item that will be purchased. This will use three different techniques namely, 1. User based collaborative filtering 2. Item based collaborative filtering 3. Association analysis through apriori

This engine will enable the marketing teams to drive sales growth by suggesting products to purchase at the right time, planning campaigns for the right products, offer coupons and discount for the right customer. This can also be used to organize websites or stores and ultimately to provide a more superior shopping experience for the customer

Data

For this project, I will be using the data provided by Instacart (https://en.wikipedia.org/wiki/Instacart). Instacart is an US company that operates as a same-day grocery delivery service. Customers select groceries through a web application from various retailers and delivered by a personal shopper. Instacart has made a subset of its grocery shopping dataset available publicly at https://www.instacart.com/datasets/grocery-shopping-2017 (https://www.instacart.com/datasets/grocery-shopping-2017). This anonymized dataset is about 715MB and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, it provides between 4 and 100 of their orders, with the sequence of products purchased in each order. It also provide the week and hour of day the order was placed, and a relative measure of time between orders. ###Data citation. “The Instacart Online Grocery Shopping Dataset 2017”, Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 (https://www.instacart.com/datasets/grocery-shopping-2017) on 21-July-2017

Preprocessing

First the libraries will be loaded

```
library(knitr)
library(dplyr)
library(ggplot2)
library(arules)
library(arulesViz)
library(splitstackshape)
```

Then the data is loaded

```
ord_prod <- read.csv("order_products__prior.csv", header = TRUE)
aisles <- read.csv("aisles.csv", header = TRUE)
departments <- read.csv("departments.csv", header = TRUE)
orders <- read.csv("orders.csv", header = TRUE)
products <- read.csv("products.csv", header = TRUE)
```

Then preprocessing is done and data is cleaned up

```
#Removing the training and test sets
orders <- orders[orders$eval_set == "prior", ]

#Merge the order data and order product data to get a view of which user is buying which product
usr_prod <- merge(orders, ord_prod, by = "order_id")
usr_prod <- usr_prod %>% group_by(user_id, product_id) %>% summarise(cnt = n())
```

Exploratory Data Analysis

Now some exploratory data analysis is done

Average number of orders per user

Then orders per user is analyzed. This is shown in plot1

```
ord_cnt <- orders %>% group_by(user_id) %>% summarise(cnt = n())
ord_cnt %>% ggplot(aes(x=cnt)) + geom_histogram(stat="count",fill="blue")  +
           ggtitle("Average number of orders per user") + xlab("number of orders")
```

User and product heatmap

Next a heatmap of user and product is created. Since the data is huge, a sample 10000 rows is taken for visualization purpose. This is shown in plot2

```
usr_prod_s <- usr_prod[sample(nrow(usr_prod), 100000), ]
ggplot(usr_prod_s, aes(user_id, product_id)) + geom_tile(aes(fill = cnt), colour = "red") +
  scale_fill_gradient(low = "red", high = "steelblue")
```

Association analysis

Now association analysis is performed

Preprocessing

The data needs to be organized in a way that can be understood by the Arules package.

```
usr_prod1 <- usr_prod[ , c("user_id", "product_id")]
#split by transaction
agg <- split(usr_prod1$product_id, usr_prod1$user_id)
#aggregate by transactions
txn <- as(agg, "transactions")
```

Visualize top 20 product fequencies

Find top 20 product frequencies. This is shown in plot3

```
top20 <- head(sort(itemFrequency(txn), decreasing = TRUE),20)
product_id = as.numeric(names(top20))
top20df <- as.data.frame(product_id)
top20df$freq <- top20
top20df <- merge(top20df, products, by = "product_id")
top20df <- top20df[ , c("product_name", "freq")]
top20df <- top20df[order(top20df$freq, decreasing = TRUE), ]
ggplot(data=top20df, aes(x=reorder(product_name, freq), y=freq)) + xlab("Products") +
  geom_bar(stat="identity") + coord_flip() + ggtitle("Top 20 product frequency")
```

Create the rules and inspect top 10 rules

```
rules <- apriori(txn, parameter=list(support=0.02, confidence=0.8))
rules <- sort(rules, by="lift", decreasing=TRUE)
#top10 rules
inspect(rules[1:10])
```

```
 lhs                   rhs     support         confidence     lift
```

[1] {13176,38159} => {21137} 0.0204549752921 0.815072463768 2.85657700264 [2] {22035,27966} => {21137} 0.0211193497859 0.808727948004 2.83434143631 [3] {39275,39928} => {21137} 0.0225887327905 0.801859184025 2.81026854207 [4] {22825,27966} => {21137} 0.0239659762668 0.800323886640 2.80488779938 [5] {13176,27966,41950} => {21137} 0.0200136754458 0.821457006369 2.87895285065 [6] {8277,27966,47209} => {21137} 0.0217546275866 0.822968262704 2.88424933689 [7] {8277,13176,27966} => {21137} 0.0228700008244 0.812823164426 2.84869390382 [8] {8277,39275,47209} => {21137} 0.0202076534002 0.824985151455 2.89131790843 [9] {8277,13176,39275} => {21137} 0.0209011245872 0.816442508051 2.86137858429 [10] {27966,30391,47209} => {21137} 0.0214345639618 0.805246857351 2.82214128977

Vizualize the top 10 rules

```
plot(rules[1:10],method="graph",interactive=TRUE,shading=NA)
```

Build Queries

Then queries can be built.A sample query would be “What are customers likely to buy before they purchase”Banana (24852)?"

```
rules1<-apriori(data=txn, parameter=list(supp=0.02,conf = 0.8),
               appearance = list(default="lhs",rhs="24852"),
               control = list(verbose=F))
rules1<-sort(rules, decreasing=TRUE,by="confidence")
inspect(rules1[1:2])
```

The results are

lhs rhs support confidence lift
[1] {14927} => {24852} 0.00271084191282 0.849544072948 2.36875485070 [2] {34004} => {24852} 0.00169730710105 0.808314087760 2.25379468499

This shows that customer buys Blueberry (14927) and Mixed Berries Wildly Nutritious Signature Blends (34004) before buying bananas

User Based Colloborative filtering

To be completed

Item Based Colloborative filtering

To be completed

Comparison of three approaches

To be completed

Conclusions

To be completed