Project:

Effect of Relocating Products on Revenue of a Store

Course: Data Mining

Professor: Dr. Alice Smith

Student: Mohammad Maydanchi

Introduction

The shopping behavior of customers is a very interesting topic for managers of grocery stores. When customers go to grocery stores they want many things to buy. Mostly they buy specific products together. For example, if a person wants to buy something for her breakfast she may buy cheese, butter, bread, cereal, peanut butter and so on. If she is used to eating cereal for her breakfast she puts milk in her basket too. Or, if she is used to eating cheese she goes to buy bread too. Managers want to know what their customers buy and which products are bought together. Based on these kinds of information they can order products for the future and also locate the products in their stores.

In the past, managers predict the behavior of customers based on trial and error. They spent a lot of money and energy to figure it out. Also, it was time-consuming and they reached information lately. Also, for finding the relations and correlations between products they could only use their experiences.

After the dawn of computer age at 1936 collecting and processing large amounts of data got very easier. Computers helped the researcher to uncover the patterns in data. So, people got familiar with the concept of data mining. Data mining is the exploration and analysis of large data to discover meaningful patterns as rules [1]. Data miners extract information from data and transform them into a comprehensible structure [2].

One of the most popular methods of data mining is association rules. Association rules help to find patterns in a seemingly unrelated or related dataset. For example, if I buy a diaper, probably I will also buy beer. They seem unrelated but when you use data mining methods you can understand this relation. Association rule uses if/then patterns to analyze the data. An association rule has two parts, an antecedent(if) and a consequent(then). An item which is found in the data is antecedent and a consequent is an item that is found in combination with the antecedent [3].

Goal:

As mentioned before a customer buys some products together, it means if she buys A then she will buy B, so if the manager of the store decides to place these products far from each other, the customer should walk more in the store then by seeing more products he gets enticed to buy more products.

In this project, I want to see what is the effect of relocating products in a grocery store on the income of the store. For this **goal**, if I want to extract useful information I need a dataset with thousands or even millions of data. Finding relations between products without data mining approaches decreases the accuracy of results. Association rule is a data mining approach that is customized for finding these kinds of relations.

Association Rules Methods:

For finding association rules in the dataset there are 3 methods:

1- Apriori:

Apriori in Latin means “from the former”. Apriori knowledge uses the power of reasoning based on self-evident truths.
Apriori uses a “bottom-up” approach. The bottom-up approach extends the frequent subsets one item at a time(a step is known as candidate generation), and it tests a group of candidates against the data when no further successfully are found it terminates the algorithm [4].

2- Eclat:

The Equivalence Class Clustering and bottom-up Lattice Traversal (Eclat) algorithm are faster and more efficient than the Apriori algorithm. It works in a vertical manner just like but Apriori works horizontally [5].

3- FP-Growth:

FP stands for the frequent pattern. This algorithm uses an extended prefix-tree structure instead of candidates for storing crucial information about frequent patterns. [6]
For this project, I use the Apriori algorithm for analyzing the data.

Appriori parameters:

Appriori has 3 main parameters;

Support:

Support is a measure of how frequently the items(transactions) occur together.

$$P(A, B)= \frac{number\ of\ times\ A\ and\ B\ happen\ together}{total\ number\ of\ transactions}$$

Confidence:

Confidence is defined as the conditional probability of occurrence of consequent given the antecedent.

$$P(B|A)=\frac{number\ of\ times\ A\ and\ B\ happen\ together}{number\ of\ times\ A\ happens.}$$

Lift:

Lift is the ratio of the rule and the expected confidence of the rule [7].

$$Lift = \frac{P(A \cap B)}{P(A) * P(B)}$$

Data Mining Approach:

Steps of Apriori Algorithm:

1- First. We should read the data and see are they Boolean or not. If they are a boolean go to the next step but if they are not, transform it.
2- We should set minimum support. For setting minimum support, first, we have to set a minimum itemset frequency. Itemset frequency is a number of transactions containing itemset [8].
3- Third, in this step, we need to set minimum confidence. The default value in R is 0.8.
4- In this step, we should set the length of rules (we can specify the minimum length and the maximum length).
5- Now take all the subsets in transactions having higher support than minimum support.
6- Take all the rules of these subsets having higher confidence than minimum confidence.
7- Sort the rules by decreasing lift.
8- After sorting rules, we should check all rules and see whether rules are reasonable or not. If they are not reasonable, change the minimum confidence and go to step 4
9- last but not least, implement proper actions based on extracted rules.

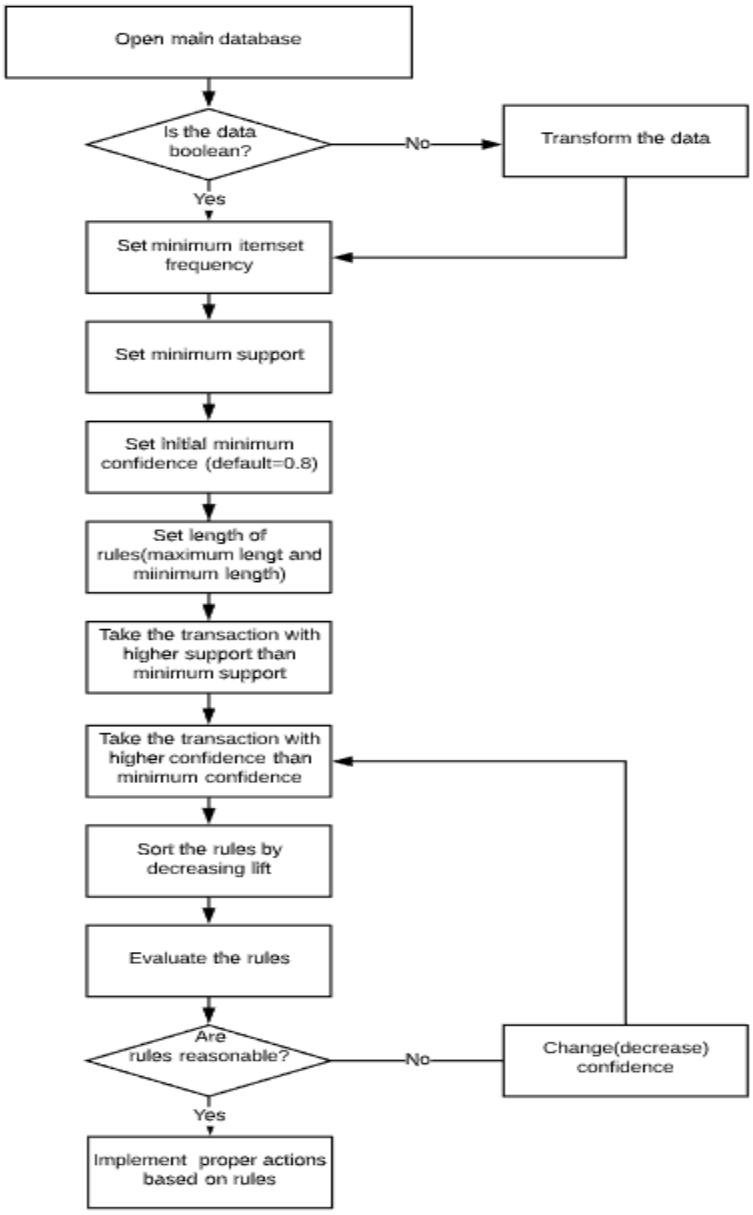

Figure 1, apriori algorithm flowchart

Datasets:

The datasets which I selected to work are for Ofogh Koorosh which is a grocery store in the south of the Caspian sea. There are 2 datasets. The first dataset is the transaction dataset. It has 7501 rows. Each row is one transaction and each transaction includes different products. It has 120 products in 21 categories. The second dataset includes categories and their corresponding aisle, also it has an average price of categories. Both datasets are formatted in CSV-file. Moreover, I could find some information about

the site plan of the store which shows the place of each aisle and how the categories are located in the aisles.

ground beef	energy bar	pet food	carrots	protein bar		
ground beef	tomato sauce	spaghetti	mineral water	almonds	eggs	
mineral water	olive oil	gums	cooking oil			
shrimp	pasta	mineral water	soup	avocado	milk	olive oil
shrimp	pasta	soup	cake	cooking oil	chicken	light mayo
spaghetti	mineral water	chocolate	french fries	champagne	escalope	mushroom cream sauce
shrimp	pasta	mineral water	eggs			
burgers	oil	tomato juice	fresh bread			

Table 1,dataset1,transactions

Category	Aisle	Average price
Frozen Food	A1	5
Nuts	A2	3
Clothes	B1	30
Books	B2	15
Patio&Garden	C1	7
Pet	C2	6
Sauces&Spices	D1	3
Dessert	D2	3
Health&Beauty	E1	8
Household Essential	E1	3
Beverage	E2	4

Table 2,dataset2,aisle

Use the Apriori Algorithm for the Dataset:

I use R-Studio for mining the data.

This transaction dataset is not Boolean so first, we should transform it to Boolean with R. I look for items that are bought at least 3 times during a day so since this data is collected in 7 days the minimum support is :

$$\textit{Minimum Support} = \frac{itemset\ frequency * period\ of\ collecting\ the\ data}{number\ of\ transactions} = \frac{3*7}{7501} = 0.0028$$

It is mentioned before that default value for confidence in R 0.8. I start with this value for minimum confidence and get all 3 items length rules with minimum support of 0.0028 and minimum confidence of 0.8. The algorithm cannot find any rules with these properties. So, I decrease the minimum confidence to 0.4 and sort the rules by decreasing the lift. In the below, you can see the part of R-code

```
#Training Apriori on the dataset
##minimum support: for minimum support
##we should set favor frequent number for buying a product in a day
p=7   #period of collecting the data
f=3   #favor frequent in a day for each product
mins=f*p/n #minimum support
###finding rules with length=3
rules=apriori(data=dataset,
              parameter = list(support=mins ,
                               confidence=0.4,target = "rules",
                               maxlen = 3, minlen =3))
```

R-studio could found 5 rules:

	lhs		rhs	support	confidence	lift	count
[1]	{cereals,ground beef}	=>	{spaghetti}	0.003066258	0.676470588	3.885303126	23
[2]	{olive oil,tomatoes}	=>	{spaghetti}	0.004399413	0.611111111	3.509911519	33
[3]	{frozen vegetables,soup}	=>	{mineral water}	0.005065991	0.633333333	2.656953766	38
[4]	{pancakes,soup}	=>	{mineral water}	0.004266098	0.62745098	2.632276177	32

Table 3, rules with a confidence equal to 0.4 and length of 3.

Most of these rules seem reasonable. But, the first one is interesting and unexpected. One of the plausible explanations for this rule is that when a person wants to cook food with spaghetti most of the time she uses ground beef too. This kind of food takes a long time (especially in a place which is close to the sea because of humidity), so if she has a kid they get hungry and the best thing for them can be cereals until the food gets ready. So, it means when a person buys aground beef and cereals she has a kid and she goes to buy spaghetti too.

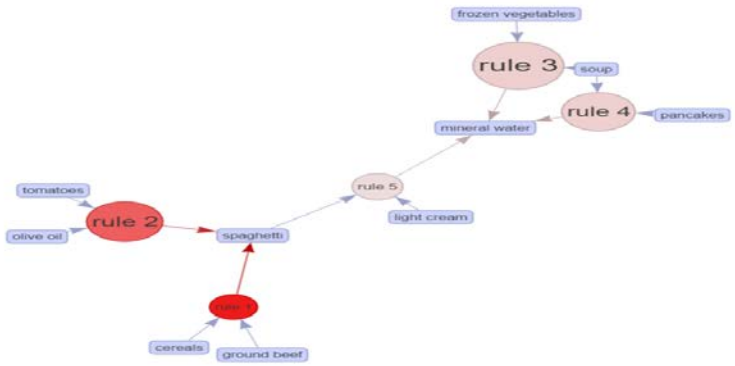

Figure 2, graph plot for rules with minimum confidence equal to 0.4 and length of 3

Comparison between the Current Site Plan and Optimum Site Plan:

This store has 18 aisles. Each aisle is divided into 3 blocks. So, totally there are 54 blocks.

Figure 3, current site plan

Now, based on rule1 and rule2 I am going to put items in rule1 and rule2 in maximum distance from each other to see how the path of a customer changes.

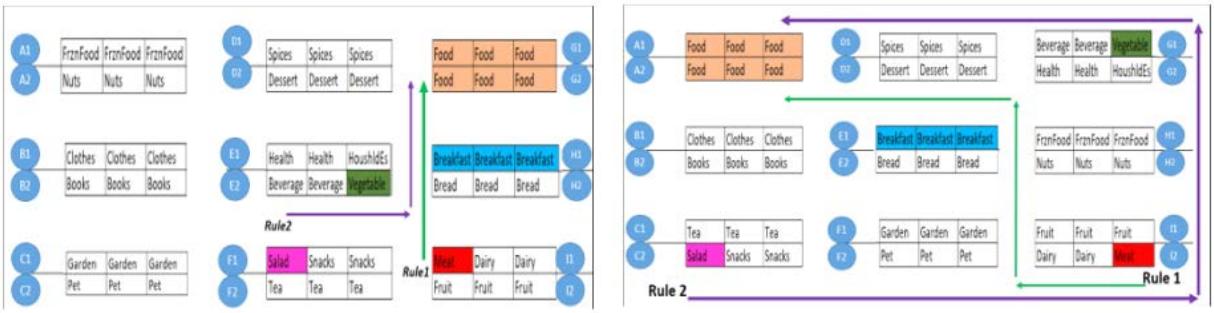

Figure 4, current site plan(left) vs. optimum site plan(right)

In both plans, if a customer wants to follow rule1 he should go through green arrow and for rule2 he should track purple arrow. As you can see in figure 4, in optimum site plan customer walks more in the store so he is able to see more products.

For getting a result from these observations we need **2 assumptions**:

1-First, we do not know if a customer passes a block how many products she bought from that block. I assume that she buys at least 1 product.

2- Second, when the customer is walking through the store she can see both sides (both aisles which are on the left-hand side and right-hand side) at the same time.

Based on these assumptions and also the average price of each block, When the customer walks through the path of **rule 1** and picks at least 1 item from each block he should pay **$32.5**. Also, if he walks through **rule2** and picks 1 item from each block he should pay **$42.5**.

Moreover, when a customer walks through path of **rule 1** and pick at least 1 item from each block he should pay **$97.5**($65 more than current place) Also, if he walks through **rule2** and pick 1 item from each block he should pay **$87**($54.5 more than current place).

Result:

The frequency of rule1 is 23 times during a week and the frequency of rule2 is 33 times during a week. So, if the manager changes the place of products like this project she is able to increase the income around $3277(23*65 +33*54=**$3277**) during a week and $157295 for a year.

Limitations:

This work has some limitations such as the period of collecting the data. This dataset is for one week, but if there is a dataset for an entire year we can calculate the total increasing with higher accuracy. Because customers have different shopping behaviors during a season and also a year. For example, definitely, their shopping behaviors are different in December(Chrismas) than their behaviors in May.

Future Work:

It can be very useful to investigate how many products a person can see when she passes an aisle. Also, I used 2 rules for relocating the products, for the future it should use more rules and also with different lengths of rules and different itemset frequency.

More, it should mention that it seems there is no similar work because of that I did not add literature review but there is an interesting paper which works on the opposite side of my topic The topic of this paper is “Data mining approach to optimize shelf space allocation in consideration of customer purchase and moving behaviors”. So, it can be very interesting if I can compare these two approaches to the income of a store.

References:

1. *C. D. Daniel T.Larose, Data Mining, and Predictive Analytics, Wiley, 2015.*

2. *"https://www.microstrategy.com/us/resources/introductory-guides/data-mining-explained," [Online].*
3. *"https://en.wikipedia.org/wiki/Data_mining," [Online].*
4. *"https://rpubs.com/Buczman/AssociationRules," [Online].*
5. *"wikipedia.org/wiki/Apriori_algorithm," [Online]. Available: https://en.wikipedia.org/wiki/Apriori_algorithm.*
6. *"geeksforgeeks.org/ml-eclat-algorithm/," [Online].*
7. *"StudyKorner," [Online]*
8. *"https://www.ibm.com/support/knowledgecenter/en/SSEPGG_11.1.0/com.ibm.im.model.doc/c_lift_in_an_association_rule.html," [Online].*