A picture containing black, indoor

Description generated with high confidence

Market Basket Analysis

ALY-6050 Week 6: Final Project

Data Mining applications

Submitted to:-

Mr. Sergiy Shevchenko

College of Professional Studies

Northeastern University, MA

Submitted by: -

Prashant Singh

Shrinath Patel

Academic term Winter-2018 Quarter II

Assignment Completion date: - 27th March 2018.

**Abstract**

Market basket analysis (MBA) or affinity analysis, is a data analysis and data mining technique that originated from marketing field and now has been used in other fields such as insurance, bioinformatics, immunology and geophysics. The purpose of market basket analysis is to determine what products customers purchase together. To identify the relationships between the purchase item. This analysis is very helpful to the retailer or ecommerce to increase their sales by knowing their customer behavior. A store could use this information to place products frequently sold together into the same area or putting one item at starting of the store and other at the end of the store in order to make customer pass through the store which might customer to make their mind to purchase other items, while a ecommerce will the recommend the products and give discount on two item purchasing together. Based on the sales, store should maintain their inventory to avoid extra inventory cost or overfilling of items which not needed. The reason for selecting this dataset as the topic for final project is to applied analytics concepts to construct a Recommender system using Market Basket Analysis (MBA) and using Apriori algorithm provide top relationships that have rules with the highest confidence.

**Introduction**

Market Basket Analysis is one of the main techniques used by large retailers to find association between items. It can be done by finding the combination of items which purchase together frequently in transactions. We can say it help the organization to know the relationship between the products people are buying. In most scenario, to analyze retail basket or transaction data association rules is being used.

To perform an Market Basket Analysis first we need a retail data set containing transaction. Each transaction contains a number of different items or products that have been bought together. For example, one itemset might be: {Pasta, vegetable, Sauce, soda} in which case all of these items have been bought in a single transaction.

In an MBA, every transaction is being analyzed to find rules of association. From above example, one rule could be: {Pasta, vegetable} => {Sauce}. It concludes, if a customer has a transaction contains a pasta and vegetable, then they are more likely to buy a Sauce also.

So before giving the rules, an analyst need to know whether there is sufficient evidence to support the given rules and the outcomes will be beneficial. In this market basket analysis, we used three concepts with which we will trying to find top relationship rules. These are

**Support:**  It is defined as number of transactions containing an item (e.g., pasta, vegetable and sauce) in the total number of transactions. Higher the support values the more frequently the itemset occurs in the transaction. Generally, in most of the case high support are preferred since they are more likely to be applicable to a large number of future transactions.

**Confidence:** It is define as a if a transaction contains the item (in our example, pasta and vegetable) then also contains the other item(in this case, Sauce). The higher the confidence, the greater the likelihood that the item on the right-hand side will be purchased or, in other words, the greater the return rate you can expect for a given rule.

**Lift:** is defined as the ratio of the observed support to that expected if X and Y were independent.  A typical and widely used example of association rules application is market basket analysis.

Here, we used data mining algorithm called the ‘[Apriori algorithm](https://en.wikipedia.org/wiki/Apriori_algorithm" \o "Apriori Algorithm" \t "_blank)’ for market basket analysis and recommendation of top relationship rules, which works in two steps:

First, identifying those itemsets that occur frequently in the data set with a support greater than a pre-specified threshold.

Second, calculate the confidence of all possible rules given the frequent itemsets and keep only those with a confidence greater than a pre-specified threshold. Threshold of the support and confidence are user-specified and are likely to vary between transaction data sets. R does have default values for support and confidence , so we have to check and try wit different values to see how they affect the number of rules returned.

**Description:** Recommender system using the concept of Market basket analysis. We have used Apriori Algorithm to predict top 20 most sold items and relevant items related to highest confidence. Expected growth in purchased rate is 14%.

A screenshot of a social media post

Description generated with very high confidence

In this Project we have 7500 overall transactions for the week. Library is used to use Apriori Algorithm to analyze the best fit. In this we are reading the file named as “Market Basket Optimization” and it contains 20 variables.

This dataset contains 20 variables with 7500 observations.7500 customers purchase history on weekly basis. But we are not going to use this dataset because Avril’s package doesn’t take dataset like this as input. It takes input as the sparse matrix.

It’s actually a matrix that contains a lot of zeroes in machinery and we will encounter a lot of times the word sparcity that corresponds to a large number of zeroes. So this matrix contains very few number of non-zero values. In this 120 different products are present and make 120 columns. Lines will be same as different transactions. So 0 and 1 in the new matrix.0 represent customer has not bought the product and 1 represent customer has bought the product. We need to use sep function because of read.transaction doesn’t understand comma separator rm.duplicates is to avoid duplicates.

A screenshot of a social media post

Description generated with very high confidence

We can observe that 7501 rows and 119 columns and a density of 0.03. Density is proportion of non-zero values is 0.03.3% non-zero and 97% zero. Most frequent item is mineral water. Eggs take 2nd place and so on. Length distribution defines item sets per transaction.1754 basket contains a single item.1358 basket contains two products. Mean is 3.9 and max are 20.

One more thing is to be analyzed here basket size like itemset per transaction or element distribution. We can observe here is Size of 1 is 1754 followed by 1358 ,1044 and so on. So most of the people purchases single item in as compared to others.

In this we can analyze from the statistical table that Min is 1 that means minimum one item is purchased and median and mean as 3 and 3.914 that means most of the customers purchases on an average 3 to 4 items as per stats.

A screenshot of a cell phone

Description generated with high confidence

Here is a list of top 20 most frequent purchased products. Its a graph between Item frequency vs Items. Mineral water frequency is even more than 0.20 that means mineral water is purchased more than 20% of times. Here we are interested in top 20 purchased items as they contribute more in the overall transaction and revenue turnover. Post calculation it has been observed that top 20 items out of 120 overall items they contribute the most.

A screenshot of a social media post

Description generated with very high confidence

Apriori algorithm is used for the dataset which is provided to find the rules. Rules are basically the most fit correlated items which is being purchased on the purchase of other items. Considering item to be bought 3 times a day that defines support as 0.003 and considering confidence 0.8 by default value.

“Support. This says how popular an itemset is, as measured by the proportion of transactions in which an itemset appears.

Confidence. This says how likely item Y is purchased when item X is purchased, expressed as {X -> Y}. This is measured by the proportion of transactions with item X, in which item Y also appears.”

So with the high level of confidence with 80% no rules has been generated. So this level of confidence is of no use as for now. We did some of more experiments on different level of confidence and lift values.

A screenshot of a social media post

Description generated with very high confidence

Considering item to be bought 3 times a day that defines support as 0.003 and considering confidence 0.4 value. We can observe 281 rules with 40% confidence.

A screenshot of a cell phone

Description generated with very high confidence

Here are the rules like if a person purchased mineral water, whole wheat pasta then chances of purchasing olive oil is 40% keeping in mind that it has been bought at least 3 times a week(Support).Lift states the ratio of confidence over support.

A close up of a map

Description generated with high confidence

This is a graph for 20 most frequent items and how they are correlated based on confidence and lift. As we can observe that mineral water is the most frequent item and playing the most significant role. Moreover, olive oil and ground beef also has the significant role in purchasing style of the customer.

Now, considering item to be bought 3 times a day that defines support as 0.003 and considering confidence 0.2 value. We can observe 1348 rules with 20% confidence.A screenshot of a cell phone

Description generated with very high confidence

Here we can observe different set of rules. With different confidence and lift values. In some of the cases this confidence level is making more sense. Like the person who has purchased light cream will also purchase chicken because in America most of the people has the habit of eating chicken with light cream.

Now, considering item to be bought 4 times a day that defines support as 0.004 and considering confidence 0.2 by default value. Support 4\*7/7500 ~ 0.004.

A screenshot of a cell phone

Description generated with very high confidence

In this case we can observe lower lift value as the support value is gone up to 0.004 from 0. 003.Here we can observe 811 rules which are less than when support war 0.003 and this is supposed to be like this. Here we can observe some more significant insight in the purchasing behavior of the customers. Here we will observe not only correlated items but also keeping in mind the minimum support of 0.004.

A close up of a map

Description generated with very high confidence

The plot uses the arulesViz package and plotly to generate an interactive plot. We can hover over each rule and see the Support, Confidence and Lift. As the interactive plot suggests, one rule that has a confidence of 0.38 is the one above. It has lift as well, at 1.59.

**Conclusion: -**

We can observe 811 rules with 20% confidence. With this confidence we are getting better and appropriate rules by visualizing these rules and plots, we can come up with a more detailed explanation of how to make business decisions in retail environments. we can make some specific aisles now in my store to help customers pick products easily from one place and also boost the store sales simultaneously.

Person who purchased light cream has also purchased chicken 30% times. Person who purchased pasta has also purchased escalope and shrimp 37 and 32% times. Person who purchased herb & pepper has also purchased spaghetti 57% times. Person who purchased cooking oil, ground beef has also purchased ground beef 39% times.

This analysis would help us improve our store sales and make calculated business decisions for people both in a hurry and the ones leisurely shopping.

**References: -**

Li, S. (2017, September 25). A Gentle Introduction on Market Basket Analysis - Association Rules. Retrieved March 27, 2018, from https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce

(n.d.). Retrieved March 27, 2018, from https://web.fhnw.ch/personenseiten/taoufik.nouri/Data Mining/Course/Case Study/PA-Tutorial/mba.html

Market Basket Analysis: Understanding Customer Behaviour. (2016, March 17). Retrieved March 27, 2018, from https://select-statistics.co.uk/blog/market-basket-analysis-understanding-customer-behaviour/