MARKET BASKET ANALYSIS (MBA)

Project Leader: Yashi Kesarwani, Reg no- 0000197

Project Faculty Coordinator: Dr. Anirban Lakshman

Table of Contents

Overview

- Association analysis is about discovering relationship among huge data sets. Just like the famous market basket analysis which gives a relationship between {Beer => Chips}. It says that whenever a person buys diapers he/she also buys beer.
- Data mining is commonly seen as a single step of a whole process called Knowledge Discovery in Databases. KDD is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.'

Introduction

- **Market Basket Analysis** is one of the key techniques used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions.
- Data Science Apriori algorithm is a data mining technique that is used for mining frequent item-sets and relevant association rules.
- **This application is basically used to find the buying pattern of consumer based on the associations of products in previous transactions.**

Objective

- The **main objective of the project is to see how different products in a grocery store interrelate and how to exploit these relations by marketing activities**.
- Mining association rules from transactional data will provide us with valuable information about co-occurrences and co-purchases of products.
- It is very important for retailers to get to know what their customers are buying. Some products have higher affinity to be sold together and hence the retailer can benefit from this affinity if special offers and promotions are developed for these products with the ultimate aim of increasing the revenue.

Why Association Analysis ?

- Association analysis is relatively light on the math concepts and easy to explain to non-technical people. In addition, it is an unsupervised learning tool that looks for hidden patterns so there is limited need for data prep and feature engineering.
- There are many complex ways to analyze data (clustering, regression, Neural Networks, Random Forests, SVM, etc.) but they can be difficult to tune, challenging to interpret and require quite a bit of data prep and feature engineering to get good results and hence require a lots of knowledge to work with.

Data Description

- **Dataset Description** : The specific data of Online Retails comes from the UCI Machine Learning Repository and represents transactional data from a UK retailer from 2010-2011.
- The training set consists of 5,41,911 rows and 8 columns having invoice no, stockCode, Description, Quantity, InvoiceData, UnitPrice, CustomerID, Country.

Apriori Algorithm

- Apriori algorithm is a classical algorithm in data mining that is used for mining frequent item-sets and association rule mining.
  The Apriori algorithm is used for the purpose of association rule mining.
- **Association rule mining is a technique to identify frequent patterns and associations among a set of items.**
- **The key concept in the Apriori algorithm is that it assumes all subsets of a frequent itemset to be frequent. Similarly, for any infrequent itemset, all its supersets must also be infrequent.**

Apriori Algorithm

It works on the various measuring terms such as :

- **Support**
- **Confidence**
- **List**
- **Conviction**

Apriori Algorithm

- **Support** of the item *x* is nothing but the ratio of the number of transactions in which the item *x* appears to the total number of transactions.
- **Confidence (*x* => *y*)** signifies the likelihood of the item *y* being purchased when the item *x* is purchased. This method takes into account the popularity of the item *x*.
- **Lift** is the ratio of the observed support to that expected if the two rules were independent. The basic rule of thumb is that a lift value close to 1 means the rules were completely independent.

Apriori Algorithm

- **Lift (*x* => *y*)** = 1 means that there is no correlation within the itemset.
- **Lift (*x* => *y*)** > 1 means that there is a positive correlation within the itemset
- **Lift (*x* => *y*)** < 1 means that there is a negative correlation within the itemset
- **Conviction** of a rule can be defined as follows:

  Conv (*x* => *y*) will be = (1-supp(y)) / (1-conf(x=>y))

Result

```
In [10]: rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
         rules.head()
```

Out[10]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE PINK)	0.096939	0.102041	0.073980	0.763158	7.478947	0.064088	3.791383
1	(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE GREEN)	0.102041	0.096939	0.073980	0.725000	7.478947	0.064088	3.283859
2	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.096939	0.094388	0.079082	0.815789	8.642959	0.069932	4.916181
3	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.094388	0.096939	0.079082	0.837838	8.642959	0.069932	5.568878
4	(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE RED)	0.102041	0.094388	0.073980	0.725000	7.681081	0.064348	3.293135

Result

```
In [11]: rules[ (rules['lift'] >= 6) &
                (rules['confidence'] >= 0.8) ]
```

Out[11]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
2	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.096939	0.094388	0.079082	0.815789	8.642959	0.069932	4.916181
3	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.094388	0.096939	0.079082	0.837838	8.642959	0.069932	5.568878
17	(SET/6 RED SPOTTY PAPER PLATES)	(SET/20 RED RETROSPOT PAPER NAPKINS)	0.127551	0.132653	0.102041	0.800000	6.030769	0.085121	4.336735
18	(SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES)	0.137755	0.127551	0.122449	0.888889	6.968889	0.104878	7.852041
19	(SET/6 RED SPOTTY PAPER PLATES)	(SET/6 RED SPOTTY PAPER CUPS)	0.127551	0.137755	0.122449	0.960000	6.968889	0.104878	21.556122
20	(SET/6 RED SPOTTY PAPER CUPS, SET/20 RED RETRO...	(SET/6 RED SPOTTY PAPER PLATES)	0.102041	0.127551	0.099490	0.975000	7.644000	0.086474	34.897959
21	(SET/6 RED SPOTTY PAPER CUPS, SET/6 RED SPOTTY...	(SET/20 RED RETROSPOT PAPER NAPKINS)	0.122449	0.132653	0.099490	0.812500	6.125000	0.083247	4.625850
22	(SET/20 RED RETROSPOT PAPER NAPKINS, SET/6 RED...	(SET/6 RED SPOTTY PAPER CUPS)	0.102041	0.137755	0.099490	0.975000	7.077778	0.085433	34.489796

Applications

- **Cross Selling:** Cross-selling is basically a sales technique in which seller suggests some related product to a customer after he buys a product. A seller influences the customer to spend more by purchasing more products related to the product that has already been purchased by him.
- **Fraud Detection:** Market basket analysis is also applied to fraud detection. It may be possible to identify purchase behaviour that can associate with fraud on the basis of market basket analysis data that contain credit card usage. Hence market basket analysis is also useful in fraud detection.

Application

- **Affinity Promotion:** Affinity promotion is a method of promotion that design promotional events based on associated products. Market basket analysis affinity promotion is a useful way to prepare and analyze questionnaire data.
- **Improve layout of the catalogue of e-commerce site.** Visual displays of products apply also to the catalogue of the firm online site. Ecommerce website interface plays significant part of customers' perceptions. A key success factor for profitable e-commerce site is the layout. In order to be able to determine an optimized layout for website it is important to know the interdependencies among different products.

Future Scope

- We will try to build a user-friendly web application where the user will be able to select as many items they wish and will be able to see the corresponding support count on a bar chart.
- From this they will be able to visualize and realize the relationship among particular items.
- The organizer of a super shop may use this application to arrange their products in the shop to boost their sell.

References

- [1] Saurabh Malgaonkar, Sakshi Surve and Tejas Hirave. “Use of Mining Techniques to Improve The Effectiveness of Marketing and Sales” IEEE Trans. Paper id-63.
- [2] Han, J., Kamber, M.: “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, 2006.
- [3] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining”
- [4] M.J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New algorithms for fast discovery of association rules," in Third International Conference on Knowledge Discovery and Data Mining, 1997.

Thank you !