Market-Based Analysis: Apriori approach to analyze purchase patterns

Bharadhwaj Reddy Lekireddy [1*], G. Michael[2], Naga Sai Ram Reddybathina[3], Sachi Nandan Mohanty[4]

[1,3,4] School of Computer Science and Engineering, VIT-AP University, Andhra Pradesh, India
[2]Department of Computational Intelligence, Saveetha School of Engineering, SIMATS, Chennai, India

[*1] Corresponding Author: B. R. Lekireddy, Email: reddy.bharadhwaj@gmail.com, [2] G. Michael, Email: micgeo270479@gmail.com, [3] N. S. R. Reddybathina, Email: lakshminagasairam.1717@gmail.com, S. N. Mohanty, Email: sachinandan09@gmail.com

Abstract

In this paper, a data mining approach has been developed to analyze customer behavioral patterns and to get a further idea of the scale of the products purchased. Using the approach we found the best and least performing products, the average number of the products that the customer is interested in buying, And using the apriori algorithm we can generate the frequent items which can be bought together which will help us in suggesting products for the customer to buy. Each pattern is supported by the support and confidence generated using the apriori algorithm.

Keywords: data mining, apriori algorithm

Received on 17 May 2023, accepted on 25 May 2023, published on 04 July 2023

doi: 10.4108/eetsis.3355

1. Introduction

Machine learning is helping retailers in many ways today. As you can imagine, there are many uses for machine learning (ML) in retail, from predicting sales performance to identifying shoppers. By analysing the past purchase behaviour of your customers, you can learn which products they often buy together. Behavioural analysis can be performed to understand customer buying behaviour patterns and is the foundation of marketing fundamentals. Market Based Analysis can be used anywhere from simple catalogue design to UI/UX and to Optimize store operations: An MBA not only helps you determine what's on the shelves, but also behind the store. Geographic patterns play an important role in determining the popularity and strength of a particular product, so MBA is increasingly being used to optimize inventory in each store or warehouse. MBA identifies customer buying habits by finding associations between the different items that customers purchase. This discovery of this kind of association will be helpful for retailers or marketers to develop marketing strategies by gaining insight into which items are frequently bought together.

MBA generally involves an algorithm such as Apriori to find the frequent patterns that are extracted from the data and EDA (Exploratory Data Analytics) which is used to find the nature and scope of the products in general and used as a primary source of information while designing marketing strategies.

Frequent itemset mining, Association rule mining and Apriori algorithm are used in MBA which are further discussed below under methodology.

2. Dataset description and Project Environment

In this project the dataset we used consists of 7500 transactions done by customers in a retail store. The data set is downloaded from Kaggle. Each column in the dataset contains the products that customers purchased at one go i.e., The dataset is a record of items purchased by the consumers at a time. The data set contained 7500 rows including some missing values and also contained words which are in incorrect format and require correction to do analysis.

Table 1: Sample of Dataset

Burgers	Meat balls	Eggs		
Chutney				
Turkey	Avacado			
Mineral Water	Milk	Energy Bar	Whole wheat	Green Tea
Yogurt				

We did the project on Jupyter Notebook and used several packages like pandas, profiling report, matplotlib, and sns to display the results.

3. Methodology

3.1 Frequent itemset mining

FIM is an attempt to extract all common itemsets in statistics (those whose frequency of occurrence does not fall below a certain threshold). This effort was proposed in the early 1990s to discover common items in the evaluation of market baskets (Agrawal et al., 1993) and became more commonly known as Large Itemset Mining.

Definition

"The frequency of a pattern (P) is the percentage or the number of data records, that contain the subset of items described by P" [1]

The Frequent Itemset Mining task is big trouble due to the patterns being very big that are required to be handled. Now by choosing a dataset that comprising 'n' items, the number of itemsets of size 'k' is equal to 'n! k! (n−k)!' for all $k \leq n$. The total number of likely itemsets is '$2n -1$'. In this type of dataset, the complexity of pattern finding are exponential. This intricacy is further enlarged when the frequency of each pattern is found, resulting in "O ($(2n - 1) \times m \times n$)".

In elegant, the overwhelming wide type of solutions is impractical for users who want actionable insights on this regard, exciting tremendous measures can be used to clear the output. these measurements can be divided into objective or records driven and subjective or customer driven. From the definition of FIM, the filtering of the quest place and the following set of solutions due to the anti-monotonic belongings of the helps were taken underneath attention. that is the foundation of FIM, and most existing pleasant measures are primarily based on such metrics.

3.2 Association Rule Mining

The association rule mining is one of the top important and well-studied technique in the field of statistical mining. It primarily aims at extracting exciting correlations, common associations, or informal systems between device units in transactional databases.

Let us consider a library in which if you consider taking a book, they will automatically suggest other books from similar sections, for example if you take a children's fairy tale book such as "100 fairy tales" they will suggest you other books such as Queen Fairy 35% and the Chaotic Prince from Ujjain 20% for further reading.

From the above example, the association rules are: when the book 100 fairy tales is brought, 35% of the time the book Queen Fairy is brought together, and 20% of the time the book Chaotic Prince from Ujjain is brought together. These rules extracted from the transaction database of the library can be used to structure the related books together further making those rules even strong. Additionally, the rules are used to help the library to make its own market strategies such as the promotion of the book 100 fairy tales, which can increase the sales/rent of the other books.

An association rule in the first place is the symbol $X \Rightarrow Y$, where X, Y are sets of items. The meaning of the association rule is that transactions in the DB which has X will tend to accommodate Y.

Given below is the formal statement containing the problem: Let $I=\{i1,i2,\ldots,im\}$ be the set of literals called 'items'. Now considering D to be a set of transactions, in which transaction 'T' is an itemset with implication '$T \subseteq I$'.A tuple 'T' of the database 'D' is represented by considering the attributes with value '1' that are associated with each transaction there is a unique identifier, called its 'TID'. And a set of items '$X \subseteq I$' is called an itemset. A transaction T contains an itemset X iff '$X \subseteq T$'.
An association rule is of form $X \Rightarrow Y$, where '$X \subset Y$', '$Y \subset I$', and '$X \cap Y= 0$'. The rule '$X \Rightarrow Y$' holds in the transaction set 'D' with confidence 'c'. If c% of transactions in 'D' contains X also accommodate Y. The rule $X \Rightarrow Y$ has support (s) in the transaction set 'D' if s% of transactions in 'D' contain '$X \cup Y$' missing items [3]

Association rules are broadly utilized in many diverse regions including telecommunications networks, market and risk management, inventory management, etc.

3.3 Apriori algorithm

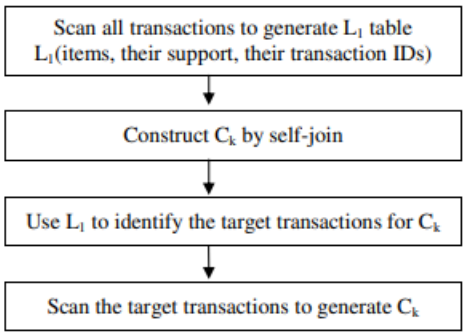
It mines all common itemsets in the database using the Apriori algorithm. It is simple and very easy to execute. The algorithm can perform many searches in the DB to find common itemsets which are used to generate 'k+1' itemsets.

So, every k-itemset must be more or equal to the MST to be the frequency, else it is called candidate itemsets. The algorithm to start with scans the DB to locate the frequency of all the '1-itemsets' which have exclusively one object by counting each item in the database.
The frequency of the '1-itemsets' is used to find the itemsets of '2- itemsets' and further it is used to find the '3-itemsets' and then it goes so on until there are no more 'k-itemsets'.If an itemset is not always frequent then any massive subset from it is also non-common, this type of circumstance prunes from search space in the database.

Definition 1: "
Suppose T={T1, T2, … , Tm},(m≥1) is a set of transactions, Ti= {I1, I2, … , In},(n≥1) is the set of items, and k-itemset = {i1, i2, … , ik},(k≥1) is also the set of k items, and k-itemset ⊆ I" [5]

Algorithm:

1. L1 = Find the frequent-1 itemsets (T)
2. for (j = 2; L(j-1) ≠Φ; j++) {
3. 'Ck' = candidates generated from L(k-1) ;
4. And x = Get item min support (Ck, L1);
5. T gt = get Transaction_ID (x);
6. For each transaction 'T' in Tgt. Do,
7. Now increase count of the items in Ck that are found



in Tgt by 1;
8. L(k) = items in Ck that are greater than min support;
9. }

Figure 1: Steps for Ck (Candidate Itemset) Generation

4. Result and Discussions

The figure-2 gives a basic overview of the dataset generated by the panda's profile report

Overview | Alerts 56 | Reproduction

Dataset statistics

Number of variables	20
Number of observations	7500
Missing cells	120657
Missing cells (%)	80.4%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.1 MiB
Average record size in memory	160.0 B

Variable types

Categorical	19
Unsupported	1

Figure 2: Statistics of the Dataset

From the figure, we can conclude that there are 20 different products in the dataset and there are 7500 transactions done by the customers.

Figure 3: Best-selling products

Figure-4: Least selling Products

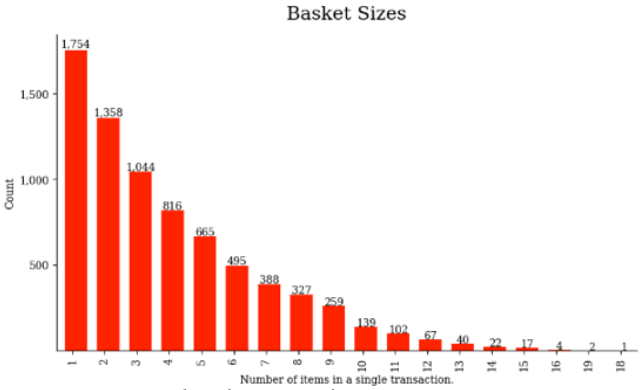


Figure 5: Basket size comparison

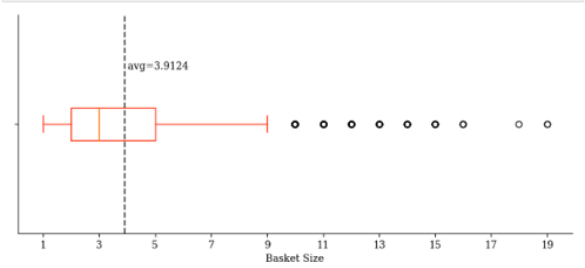


Figure 6: Average Basket size and Outliers

Through the Exploratory Data Analytics performed the above charts have been made

From the EDA performed We can observe that the products like mineral water, eggs, spaghetti, French fries, and chocolate are some of the best-selling products, and items such as water spray, napkins cream, and Bramble are sold considerably less we can also observe that the average customer buys 3 products every time he makes a transaction. this information can be used to find market strategists to increase the number of sales.

Furthermore, there is a figure to understand the products according to their scale.

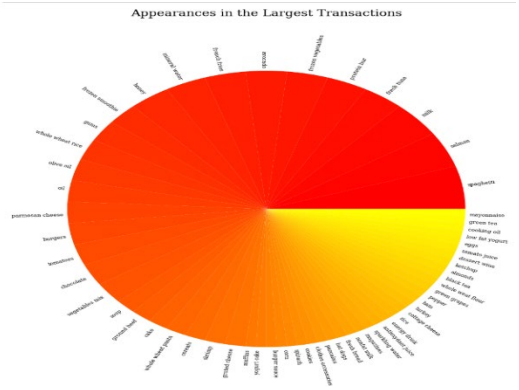


Figure 7: Pie chart showing products which appeared in large transactions.

Next by using the apriori algorithm we generated the values of confidence and support for the frequent patterns. The figure below shows the frequent patterns generated.

```
{eggs} -> {mineral water} (conf: 0.304, supp: 0.066, lift: 1.030, conv
{shrimp} -> {mineral water} (conf: 0.338, supp: 0.031, lift: 1.146, co
{low fat yogurt} -> {mineral water} (conf: 0.339, supp: 0.031, lift: 1
{chocolate} -> {mineral water} (conf: 0.342, supp: 0.069, lift: 1.159,
{cake} -> {mineral water} (conf: 0.356, supp: 0.036, lift: 1.207, conv
{spaghetti} -> {mineral water} (conf: 0.357, supp: 0.078, lift: 1.212,
{tomatoes} -> {mineral water} (conf: 0.370, supp: 0.032, lift: 1.257,
{pancakes} -> {mineral water} (conf: 0.375, supp: 0.044, lift: 1.273,
{milk} -> {mineral water} (conf: 0.383, supp: 0.063, lift: 1.300, conv
{frozen vegetables} -> {mineral water} (conf: 0.385, supp: 0.047, lift
{frozen vegetables} -> {spaghetti} (conf: 0.300, supp: 0.036, lift: 1.
{ground beef} -> {mineral water} (conf: 0.429, supp: 0.053, lift: 1.45
{olive oil} -> {mineral water} (conf: 0.438, supp: 0.036, lift: 1.487,
{burgers} -> {eggs} (conf: 0.341, supp: 0.038, lift: 1.556, conv: 1.18
{soup} -> {mineral water} (conf: 0.466, supp: 0.030, lift: 1.582, conv
{ground beef} -> {spaghetti} (conf: 0.411, supp: 0.051, lift: 1.881, c
```

Figure-8: List of frequent patterns

5. Conclusion

By the above results we can infer that the MBA (Market Based Analysis) really helps the marketer conclude on the products performance and their affinity towards each other in the market space when bought together. Also, the information generated by EDA is very essential for marketers to know their products and to come up with a plan to increase their sales in this research article Apriori algorithm is proposed to mine the frequent patterns that are generated to be used to suggest products to the customer based on their previous purchase and the cart. This application can be used widely not only in retail stores but also on e-commerce websites and other gaming applications.
Cross-selling and up-selling are the retail secrets that motivate consumers to buy. This has become a flourishing factor for an industry that uses patterns in market-based analytics to mine data to derive customer insights and improve brand performance.

6. Future Scope

In this project we found the frequent patterns of the items. Future scope of the project can be to find and evaluate the likeliness probability of the new mixed patterns which have the potential to attract the customers.

7. References.

1. Luna JM, Fournier-Viger P, Ventura S (2019) Frequent itemset mining: A 25 years review. WIREs Data Mining and Knowledge Discovery 9. doi: 10.1002/widm.1329

2. Zhao, Qiankun, and Sourav Bhowmick. “Association Rule Mining: A Survey.” Nanyang Technological University - NTU Singapore, Nanyang Technological University, Singapore, 2003, pp. 6–11, personal.ntu.edu.sg/assourav/Unpublished/UP-ARMSurvey.pdf. No. 2003116.

3. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast discovery of association rules. Advances in knowledge discovery and data mining, 12(1), 307-328.

4. Agrawal, Rakesh, et al. “Mining Association Rules between Sets of Items in Large Databases.” ACM SIGMOD Record, vol. 22, no. 2, 1993, pp. 207–216., https://doi.org/10.1145/170036.170072.

5. Al-Maolegi M, Arkok B (2014) An Improved Apriori Algorithm For Association Rules. International Journal on Natural Language Com DOIing 3:21–29. doi: 10.5121/ijnlc.2014.3103

6. Sinha, Anurag. “Implying Association Rule Mining and Market Basket Analysis for Knowing Consumer Behavior and Buying Pattern in Lockdown - a Data Mining Approach.” 2021, https://doi.org/10.20944/preprints202105.0102.v1.

7. Qisman, M, et al. “Market Basket Analysis Using Apriori Algorithm to Find Consumer Patterns in Buying Goods through Transaction Data (Case Study of Mizan Computer Retail Stores).” Journal of Physics: Conference Series, vol. 1722, no. 1, 2021, p. 012020., https://doi.org/10.1088/1742-6596/1722/1/012020.

8. Hermaliani, Eni Heni, et al. “Data Mining Technique to Determine the Pattern of Fruits Sales & Supplies Using Apriori Algorithm.” Journal of Physics: Conference Series, vol. 1641, no. 1, 2020, p. 012070., https://doi.org/10.1088/1742-6596/1641/1/012070.

9. Kurniawan, Fachrul, et al. “Market Basket Analysis to Identify Customer Behaviours by Way of Transaction Data.” Knowledge Engineering and Data Science, vol. 1, no. 1, 2017, p. 20., https://doi.org/10.17977/um018v1i12018p20-25.

10. Efrat, A R, et al. “Consumer Purchase Patterns Based on Market Basket Analysis Using Apriori Algorithms.” Journal of Physics: Conference Series, vol. 1524, no. 1, 2020, p. 012109., https://doi.org/10.1088/1742-6596/1524/1/012109.