Market Bsket Analysis using Datamining

S. Kishore Babu[1], Medindrao Kavya[2], Kokkira Rathna Kumari[3], Jorige Swarupa[4]

[1]Assistant Professor, Department of Information Technology, Andhra Loyola Institute of Engineering and Technology, Vijayawada, AP, INDIA
[2,3,4]UG Students, Department of Information Technology, Andhra Loyola Institute of Engineering and Technology, Vijayawada, AP, INDIA

Abstract: Market basket analysis is one of the key technique used by large retailers to uncover association between items. It works by looking for combination of items that occur together frequently in transactions. To put it another way, it allows retailers to identify relationships between the items that people buy. Association rules are widely used to analyse retail basket or transaction data, and are intended to identify strong rules discovered in transactions data using measures of interestingness. The results show that if top selling items are used, it is possible to get almost same frequent itemsets and association rules within a short time comparing with that outputs which are derived by computing all the items with Apriori algorithm.

Key words: Market Basket Analysis, Association Rule Mining,Apriori Algorithm.

DOI of the Article: https://doi.org/10.46501/IJMTST0707051

Available online at: http://www.ijmtst.com/vol7issue07.html

As per UGC guidelines an electronic bar code is provided to seure your paper

To Cite this Article: S. Kishore Babu; Medindrao Kavya; Kokkira Rathna Kumari and Jorige Swarupa. Market Bsket Analysis using Datamining. *International Journal for Modern Trends in Science and Technology* 2021, *7*, 0707124, pp. 291-296. https://doi.org/10.46501/IJMTST0707051

Article Info. Received: 14 June 2021; Accepted: 12 July 2021; Published: 25 July 2021

I. INTRODUCTION

Terabytes of commercial data are generated every second in today's digital world. Huge volumes of data are generated on a daily basis, and as a result, the volume of data is growing.significantly. Information can be gleaned from these rapid expansions.of data has become one of the most significant difficulties for data scientists.communities of management and mining Furthermore, the vast majority Several well-known organisations collect and retain large amounts of data.a large number of customer transaction records [1] Having said that,These vast amounts of data do not imply that the organisations were wealthy.[2] Commercial information The corporate world need it.to gain vital knowledge and information from this a huge amount of data Market basket analysis is the result of this.This method identifies a customer's buying habits by looking for trends in their purchases.Customers make connections between different goods they put in their carts.their grocery carts [3]. The goal of a market basket study is to find out how much money you have in your is to figure out which products are usually bought together.by the customers The word "frequent items" refers to groupings of goods.Must satisfy a specific percentage amount given by the user Customers who have purchased milk in a supermarket, for example,So, how many people have a good chance of buying bread.at the same time as milk [2]. This analysis is beneficial to the store.owners to make a number of crucial business decisions, as well as to identify increased product sales, catalogue design, and regular consumers There are many more. The primary purpose of market basket analysis is to determine the value of a basket of goods.find out what things people are buying and what they are buying. It's also beneficial. merchants to product placement on shelves by putting similar products next to each other goods that are near to each other For instance, if you have consumers who When you buy a computer, you should also acquire anti-virus software.at the same time, then putting the hardware display in close proximity to the software The use of a display may aid in the sale of both commodities [3].

For discovering new things, many algorithms have been proposed. knowledge gleaned from these massive databases Association of mining companies One of the most significant measurements is rules. a connection The rule is of the form X => Y, where X refers to something else.X is referred to as the antecedent, while Y is referred to as the consequent.Customers who buy X are more likely to buy Y, according to the rule.Y is more likely to be purchased [1]. The value of regulations is determined by how fascinating they are.by encouragement and trust The practicality and assurance They mirror the rules that have been identified. The organization The rules must meet the user-specified minimal support requirements.minimum trustworthiness Apriori and FP Growth are two of the most popular.basic techniques for locating and detecting common itemsets [3] There are correlations between products.

II. LITERATURE REVIEW

The field of association rule mining has seen a lot of research. Mr. Rakesh and colleagues [4], [5] presented an efficient technique for generating all relevant association rules and identifying frequent itemsets in a database. The authors used sales data from a large retailing organisation to try to uncover product connections with a minimum support of 1 percent and a minimum confidence of 50 percent. By assessing accuracy, they were able to ensure that the estimating and trimming approaches were effective.

For market basket analysis, authors Abdulsalam, Hambali, and others employed the Apriori algorithm. The authors attempted to portray a supermarket's sales pattern by representing six (6) different products across thirty (30) different transactions.The authors used the Apriori algorithm in the JAVA programming language to determine the itemsets that are frequent, assuming a minimum support of 50%.

In their paper [1], Dhanabhakyam and Punithavalli discussed Classification Dependent Predictive Association Rules (CPAR), Associative Classification, Classification Association Rule Mining (CARM), Distributed Apriori Association Rule, Six Sigma Technique, and the Apriori algorithm. They listed the benefits and drawbacks of each strategy and attempted to draw a judgement as to which way was superior.According to the authors, the Apriori algorithm is the best for association out of all the approaches, although it has a lot of drawbacks.

The authors recommended combining fuzzy logic with the Apriori algorithm to achieve a better outcome.In this study, the authors Liu and Guan employed FP Growth to address the shortcomings of Apriori [2].FP Growth, according to the authors, creates an FP tree with extremely compressed information. To discover the relationship between transactions, the authors created an FP tree using five transactions.

Authors Jiangtao Qiu and others attempted to develop a model of customer buy behaviour in the e-commerce setting, dubbed COREL (customer purchase prediction model) [7]. There are two steps to this model. Initially, a candidate production collection is created by identifying product linkages and predicting client motives. The second stage is used to determine which proposed products are most frequently purchased based on client preferences. Customers' information and product reviews were obtained from "Jingdong" by the authors. Customers' preferences play a large effect in purchase decisions, according to the findings of their article.

Authors Kaura and Kanga proposed employing association rule mining to find evolving trends in market data [8]. They began by describing several data mining approaches before attempting to explain why market basket analysis is significant. They attempted to locate outliers using extended bakery datasets. The authors also advised that this strategy be used to other fields.

III. PROPOSED APPROACH

This dataset has data which is taken from (https://www.kaggle.com/xvivancos/marketbasket-analysis/data)this data belongs to a bakery called "The Bread Basket", located in Edinburgh, Scotland. This bakery presents a refreshing offer of Argentine and Spanish products. This dataset contains over 21,293 records which consists of 95 unique items. Business analytics approach that mines visit segments from basket sales data. Association rules analysis is a technique to uncover the correlation between the items. The system will be helpful to predict the most purchased products, higher transactions by month-wise and year-wise. Determining the best-selling product in the most-selling time.

For Bakery Shop dataset there are four columns- Date, Time, Transaction, Item. We have first checked for 'NONE 'values in these four columns. In the Item column we found 'NONE' value which means no item was purchased and the number of such rows is 786. So these have no use in the dataset and we dropped such rows. Again, in Transactions column, the rows that share the same values belong to same transaction, for this the dataset has fewer transactions than observations. We have finally 9465 transactions in this dataset. Then we have computed the unique items and 94 items are found which means only these items are present in the Items column. After this, Transaction Encoder is used so that we can transform our data into a correct format for applying the mining algorithms

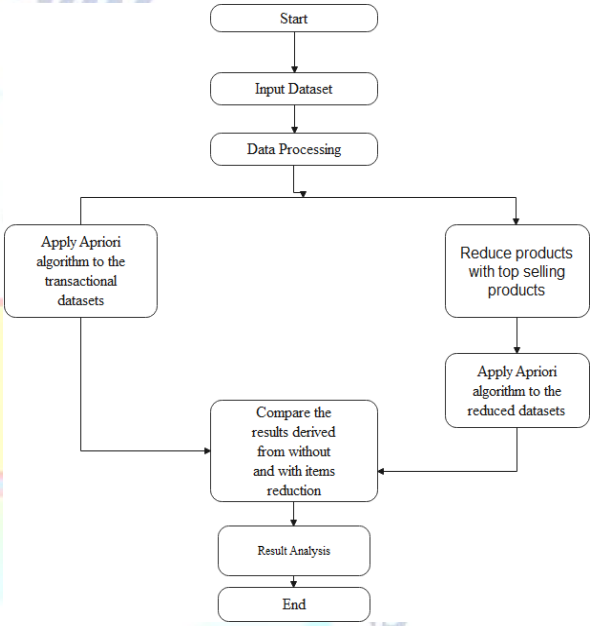

Fig 1: Proposed Approach

For finding frequent item sets and corresponding association rules in our datasets, we use

A.Apriori Algorithm - Apriori is the first and basic algorithm for finding frequent itemsets proposed by R. Agrawal and R. Srikant in 1994 [4]. Apriori involves an approach known as a level-wise search, where k-itemsets are used to explore (k +1)-itemsets. Here, at first frequent 1-itemsets are found by scanning the database which satisfy the minimum support. Again, frequent 2-itemsets are found by using frequent 1-itemsets. So this process continues until frequent k-item sets can be found [3]. Actually Apriori follows an antimonotonic property which states that every subset of a frequent item set must also be frequent and it uses a breadth-first search to count the candidate items frequently. This algorithm has two main steps-

Joining step: To find LK, a set of candidate k-item sets is generated by joining (Lk- 1) with itself [3].

Pruning step: Any (k -1)-item set that is not frequent cannot be a subset of frequent k-itemset [3].

The main goal of this study is to show performance evaluation of Apriori algorithms. First we discover all the frequent itemsets which satisfy a predefined minimum support and then find associations between frequent item sets which satisfy a predefined minimum confidence. Then we compare the execution time against transaction and mark out the results which are given in experimental analysis section. When the database is large Apriori generates a huge number of candidate sets. So if we can reduce our computation by some approach, it will be productive. Our proposed way is to reduce the items of datasets with top selling products. So we reshape the datasets by taking those products that bought most by the customers. But how much top selling products will be suitable for this proposed approach is a key question. For this, we have taken 30%, 40%, 50% and 55% top selling products and compared the results against frequent itemsets and association rules which are obtained by computing all products in the datasets. Figure 1 shows the flow-chart of our proposed approach.

IV. EXPERIMENTAL ANALYSIS

The overall experiment is performed on a PC with Intel(R) Core(TM) i5-4210U CPU 2.40 GHz processor, 4 GB main memory and running the Microsoft Windows 10 operating system.All the analysis are done by using python programming language.

A.Analysis over Bakery Shop Dataset

Here, minimum support=1% and minimum confidence =50%, Figure 5 displays that that the required time for Apriori algorithm.

B.Input Screens

```
In [3]: data = pd.read_csv("BreadBasket_DMS.csv")

In [4]: data.info()
```

Fig 2 Dataset Reading

```
In [5]: data.columns

Out[5]: Index(['Date', 'Time', 'Transaction', 'Item'], dtype='object')

In [7]: print('First Ten Rows of the DataFrame: \n')
        print(data.head(10))

First Ten Rows of the DataFrame:

         Date      Time  Transaction          Item
0  2016-10-30  09:58:11            1         Bread
1  2016-10-30  10:05:34            2   Scandinavian
2  2016-10-30  10:05:34            2   Scandinavian
3  2016-10-30  10:07:57            3  Hot chocolate
4  2016-10-30  10:07:57            3           Jam
5  2016-10-30  10:07:57            3       Cookies
6  2016-10-30  10:08:41            4        Muffin
7  2016-10-30  10:13:03            5        Coffee
8  2016-10-30  10:13:03            5        Pastry
9  2016-10-30  10:13:03            5         Bread
```

Fig 3: Attributes in the dataset

```
List of Items sold at Bakery:
Total Items:  95
---------------
Bread
Scandinavian
Hot chocolate
Jam
Cookies
Muffin
Coffee
Pastry
Medialuna
Tea
NONE
Tartine
Basket
Mineral water
Farm House
Fudge
Juice
Ella's Kitchen Pouches
Victorian Sponge
Frittata
Hearty & Seasonal
Soup
Pick and Mix Bowls
Smoothies
Cake
Mighty Protein
Chicken sand
Coke
Mv-5 Fruit Shoot
```

Fig 4: Items sold in bakery

C.Output Screens

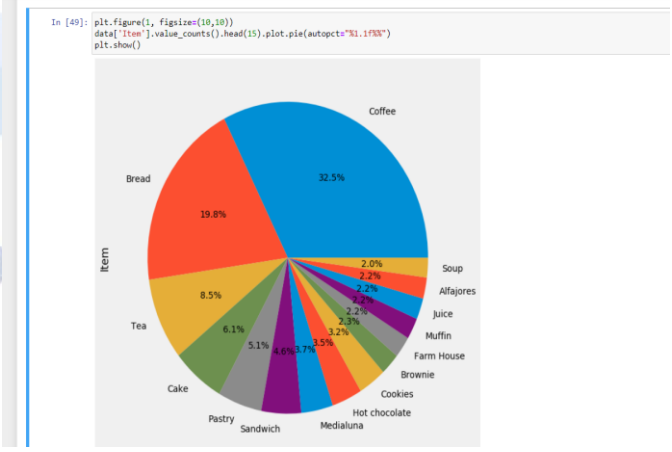

Fig 5: Top most selling products

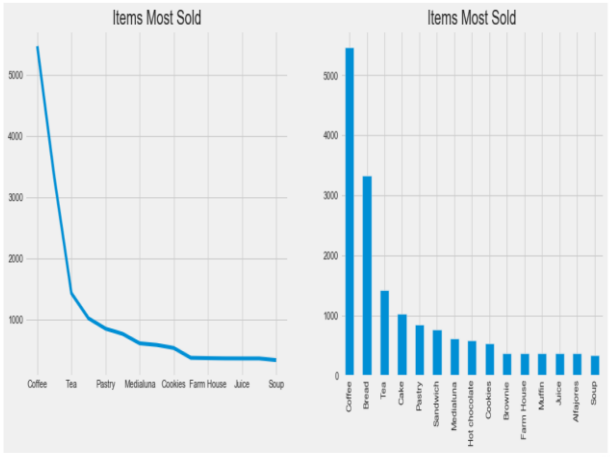

Fig 6: Most sold items

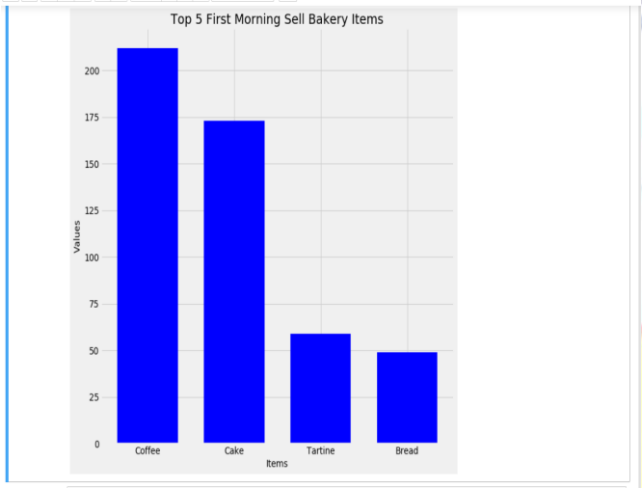

Fig 7 Top 4 First morning selling items

D.Rule Analysis Using Sampling Without Replacement:

At first we have computed 9465 transactions with all 94 products without product replacement while keeping minimum support=1% and minimum confidence=50%. The rules are-

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
31	(Toast)	(Coffee)	0.033597	0.478394	0.023666	0.704403	1.472431	0.007593	1.764582
28	(Spanish Brunch)	(Coffee)	0.018172	0.478394	0.010882	0.598837	1.251766	0.002189	1.300235
18	(Medialuna)	(Coffee)	0.061807	0.478394	0.035182	0.569231	1.189878	0.005614	1.210871
22	(Pastry)	(Coffee)	0.086107	0.478394	0.047544	0.552147	1.154168	0.006351	1.164682
0	(Alfajores)	(Coffee)	0.036344	0.478394	0.019651	0.540698	1.130235	0.002264	1.135648
16	(Juice)	(Coffee)	0.038563	0.478394	0.020602	0.534247	1.116750	0.002154	1.119919
24	(Sandwich)	(Coffee)	0.071844	0.478394	0.038246	0.532353	1.112792	0.003877	1.115384
6	(Cake)	(Coffee)	0.103856	0.478394	0.054728	0.526958	1.101515	0.005044	1.102664
26	(Scone)	(Coffee)	0.034548	0.478394	0.018067	0.522936	1.093107	0.001539	1.093366
12	(Cookies)	(Coffee)	0.054411	0.478394	0.028209	0.518447	1.083723	0.002179	1.083174
14	(Hot chocolate)	(Coffee)	0.058320	0.478394	0.029583	0.507246	1.060311	0.001683	1.058553
4	(Brownie)	(Coffee)	0.040042	0.478394	0.019651	0.490765	1.025860	0.000495	1.024293
20	(Muffin)	(Coffee)	0.038457	0.478394	0.018806	0.489011	1.022193	0.000408	1.020777
2	(Pastry)	(Bread)	0.086107	0.327205	0.029160	0.338650	1.034977	0.000985	1.017305
11	(Cake)	(Tea)	0.103856	0.142631	0.023772	0.228891	1.604781	0.008959	1.111865
38	(Tea, Coffee)	(Cake)	0.049868	0.103856	0.010037	0.201271	1.937977	0.004858	1.121962
32	(Sandwich)	(Tea)	0.071844	0.142631	0.014369	0.200000	1.402222	0.004122	1.071712
9	(Hot chocolate)	(Cake)	0.058320	0.103856	0.011410	0.195652	1.883874	0.005354	1.114125
39	(Cake, Coffee)	(Tea)	0.054728	0.142631	0.010037	0.183398	1.285822	0.002231	1.049923
10	(Tea)	(Cake)	0.142631	0.103856	0.023772	0.166667	1.604781	0.008959	1.075372
37	(Pastry)	(Bread, Coffee)	0.086107	0.090016	0.011199	0.130061	1.444872	0.003448	1.046033
36	(Bread, Coffee)	(Pastry)	0.090016	0.086107	0.011199	0.124413	1.444872	0.003448	1.043749

Fig 8: Association Rules

- [alfajores] => [coffee](1:96; 54:06)
- [cake] => [coffee](5:47; 52:69)
- [cookies] => [coffee](2:82; 51:84)
- [hotchocolate] => [coffee](2:95; 50:72)
- [juice] => [coffee](2:06; 53:42)
- [medialuna] => [coffee](3:51; 56:92)
- [pastry] => [coffee](4:75; 55:21)
- [sandwich] => [coffee](3:82; 53:25)
- [scone] => [coffee](1:80; 52:29)
- [spanish] => [coffee](1:08; 59:88)
- [toast] => [coffee](2:36; 70:44)

After 50% reduction, we have done sampling without replacement.As there are 9465 transactions, we have taken five samples, each has 1893 transactions in it. The results are-

Sample 1: Here, we get same two rules which we had chieved before by using without product replacement-
[alfajores] => [coffee](1:69; 56:14)
[spanishbrunch] => [coffee](1:0; 67:85)

Sample 2: Here, we get same six rules-
[hotchocolate] => [coffee](3:16; 53:09)
[medialuna] => [coffee](3:22; 58:65)
[sandwich] => [coffee](4:06; 62:09)
[scone] => [coffee](2:00; 61:29)
[spanishbrunch] => [coffee](1:21; 63:88)
[toast] => [coffee](3:27; 83:78)

Sample 3: We get same two rules here-
[hotchocolate] => [coffee](3:22; 55:45)
[pastry] => [coffee](5:01; 58:64)

Sample 4: Here, we get same six rules-
[alfajores] => [coffee](2:69; 67:10)
[cake] => [coffee](6:28; 57:48)
[hotchocolate] => [coffee](2:79; 58:24)
[pastry] => [coffee](5:44; 57:86)
[scone] => [coffee](1:58; 54:54)
[toast] => [coffee](2:58; 71:01)

Sample 5: Again, we get same six rules-
[cookies] => [coffee](3:38; 56:63)
[juice] => [coffee](2:27; 60:56)
[medialuna] => [coffee](3:38; 59:81)
[pastry] => [coffee](4:64; 56:41)
[scone] => [coffee](2:16; 56:16)
[spanishbrunch] => [coffee](1:47; 75:67)

So we can see that after using sampling without replacement,same rules are generated which are more accurate compared to those rules which were generated without product reduction.

V.CONCLUSION

Market basket analysis is a quick and easy first step towards uncovering hidden patterns from your customers. However, there are many more such interesting methods to delve into the world of analysis and data mining. It is a unsupervised machine technique that can be useful for finding patterns in transactional data. It can be a very powerful tool for analyzing the purchasing patterns of consumers. We have also done rule analysis by using sampling without replacement and results show that we get the same rules with higher confidence. So we can say the reduction of items is capable of identifying customers purchasing patterns which require less computation.In future, more transactional datasets can be used to determine the range of percentage for product reduction. Also analysis of individual rule with correlation analysis will be interesting.

REFERENCES

1. M. Dhanabhakyam and M. Punithavalli, "A survey on data mining algorithm for market basket analysis," Global Journal of Computer Science and Technology, 2011.
2. Y. Liu and Y. Guan, "Fp-growth algorithm for application in research of market basket analysis," in 2008 IEEE International Conference on Computational Cybernetics, pp. 269–272, IEEE, 2008.
3. J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques third edition," The Morgan Kaufmann Series in Data Management Systems, 2011.
4. R. Agrawal, T. Imieli´nski, and A. Swami, "Mining association rules between sets of items in large databases," in Acm sigmod record, vol. 22,pp. 207–216, ACM, 1993.
5. R. Agrawal, R. Srikant, et al., "Fast algorithms for mining association rules," in Proc. 20th int. conf. very large data bases, VLDB, vol. 1215,pp. 487–499, 1994.
6. S. Abdulsalam, K. Adewole, A. Akintola, and M. Hambali, "Data mining in market basket transaction: An association rule mining approach,"International Journal of Applied Information Systems, vol. 7, no. 10, pp. 15–20, 2014.
7. J. Qiu, Z. Lin, and Y. Li, "Predicting customer purchase behavior in the e-commerce context," Electronic commerce research, vol. 15, no. 4,pp. 427–452, 2015.
8. M. Kaur and S. Kang, "Market basket analysis: Identify the changing trends of market data using association rule mining," Procedia computer science, vol. 85, pp. 78–85, 2016.