SIMATS SCHOOL OF ENGINEERING

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES

CHENNAI-602105

A CAPSTONE PROJECT REPORT

CSA1601-DATA WAREHOUSE AND DATA MINING

"Optimizing Retail Strategies Through Market Basket Analysis: A Data-Driven Approach to Enhancing Customer Purchase Patterns and Maximizing Revenue"

Submitted in the partial fulfilment for the award of the degree of

BACHELOR OF ENGINEERING

IN

Computer Science Engineering

Submitted by

S. Chandana priya (192210091)

A V U .Pushyami(192211507)

Under the Supervision of

DR.S.ARUMUGAM

JANUARY 2025

ABSTRACT

Market Basket Analysis is an important part of the analytical system in retail organization to determine the placement of goods, designing sales promotion for different segments of customers to improve customer satisfaction and hence the profit of the supermarkets. MBA is a well known activity of ARMultimatelyusedfor business intelligent decisions. Mining frequent itemsets and hence deduce rules to build classifiers with good accuracy is essential for efficient algorithms. The issues for a leading supermarket are addressed here using frequent itemset mining. The project uses a file as a database. Here, the itemsets and transaction items are kept in a matrix form representing rows as list of items and columnas transactions. The frequent item sets are mined from the database using theApriorialgorithm and then the association rules are generated. The project is beneficial for supermarket managers to determine the relationship between the items purchased by their customers.

Market basket analysis is a powerful technique used by retailers to uncover associations between items purchased by customers. By analyzing transactional data, this method identifies frequently co-occurring items in baskets, enabling businesses to understand customer purchasing behavior and improve strategies such as product placement, cross-selling, and targeted marketing. This paper provides an overview of market basket analysis, including its algorithms (such as Apriori and FP-Growth), applications across various industries, and challenges faced in implementation. Furthermore, it discusses the significance of association rules generated from market basket analysis in enhancing business profitability and customer satisfaction. Through case studies and examples, the effectiveness of this approach in driving decision-making processes and fostering competitive advantage is demonstrated, highlighting its relevance in today's data-driven business environment.

Keywords: Market Basket Analysis, Association Rule Mining, AprioriAlgorithm

LIST OF ABBREVIATIONS

MBA- Market Basket Analysis

NFR- Non Functional Requirement

FR- Functional Requirement

LHS- Left Hand Side

RHS- Right Hand Side

INTRODUCTION

Market Basket Analysis is a data mining technique used to discover patterns and associations in customer purchases. By analyzing transaction data, businesses can understand the relationships between products and optimize marketing strategies.

MARKET BASKET ANALYSIS

Market basket is defined as an itemset bought together by a customer on a single visit to a store. In our visits to the supermarket we tend to buy a lot of products from different categories and put them all together in one single basket. Which Is Considered to be a single transaction. Market basket analysis is the analysis of those baskets all together.Market basket analysis encompasses a broad set of analytic techniques aimed at uncovering the associations and connections between specific objects, discovering customer behaviors and relation between items. In retail, it is used based on the following idea, if a customer buys a certain group of items, ismore(or less) likely to buy another group of items. For example, it is known that when a customer buys beer, in most cases, buys chips as well.The seller's/ 3 supermarkets are interested in analyzing which items are purchased together in order to create new marketing/sales strategies that can be helpful in improving the benefits of the company as well as customer experiences.

Market Basket Analysis (MBA) is a data mining technique that reveals associations and correlations between items in large datasets, commonly within the retail sector. It aims to identify item sets that frequently co-occur in transactions, such as bread and butter being often purchased together. This technique generates association rules, which help retailers understand the relationships between products. Key metrics in MBA include support, confidence, and lift, which measure the frequency, likelihood, and strength of these associations, respectively. Widely used for cross-selling, upselling, product placement, and targeted marketing, MBA informs decisions that enhance customer experience and boost sales. Various algorithms and tools, such as the Apriori algorithm and FP-Growth, facilitate MBA, making it a valuable asset for data-driven decision-making in retail and beyond.

FIG 1:Market Basket analysis

Market basket analysis is a powerful tool for the implementation of up-selling, cross-selling, inventory management strategies (Chen, Tang, Shen, &Hu, 2005). Market Basket Analysis is also known as association rule mining or affinity analysis, which have been used to understand consumer behavior regarding the types of the purchases they make. It is a Data Mining technique that originated in the field of marketing and was initially used to understand purchase patterns of the customers by extracting associations and co-occurrence from a transactional database (i.e. market basket data). For example, when shoppingina supermarket, consumers rarely buy one product, they are far more likely to purchase an entire basket of products, mostly from different product categories. This allows us to uncover nonobvious, usually hidden and counterintuitive associations between items, products, or categories. We are also able to extract products and product categories which are purchased together, and these associations can be represented in the form of association rules. These association rules enable managers to develop marketing strategies like developing interventions, promoting specific product categories, offering promotions, etc. 4 which eventually leads to customers spending more money based on two different principles. Upselling, which consists in buying a large quantity of the same product or adding new features and Cross-selling, which consists in adding more products from various categories. Market Basket Analysis is also very much useful in stock management and placement of items.

Key Concepts and Terminology in Market Basket Analysis

1. Itemset: A collection of one or more items. For example, {milk, bread} is an itemset comprising two items.

2. Transaction: A record in the dataset that contains a set of items bought together. For example, a shopping basket with milk, bread, and butter.

3. Support: The proportion of transactions in the dataset that contain a particular itemset. It measures the frequency of the itemset. Mathematically, it is defined as:

\[\text{Support}(A) = \frac{\text{Number of transactions containing } A}{\text{Total number of transactions}}\]

4. Confidence: The likelihood that a transaction containing itemset A also contains itemset B. It is calculated as:

\[\text{Confidence}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}\]

This metric evaluates the strength of the implication rule A → B.

5. Lift: The ratio of the observed support of itemsets A and B appearing together to the support expected if A and B were independent. It is calculated as:

\[\text{Lift}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A) \times \text{Support}(B)}\]

A lift greater than 1 indicates a positive association, meaning items A and B are more likely to be bought together than if they were independent.

6. Frequent Itemsets: Itemsets that have support greater than or equal to a specified minimum threshold. Identifying frequent itemsets is a crucial step in MBA as it helps to focus on the most relevant associations.

7. Association Rule: An implication expression of the form A → B, where A and B are disjoint itemsets. This rule suggests that if itemset A is purchased, itemset B is likely to be purchased as well.

8. Apriori Algorithm: A classic algorithm used to identify frequent itemsets and generate association rules. It operates on the principle that any subset of a frequent itemset must also be frequent.

9. FP-Growth Algorithm: An efficient alternative to the Apriori algorithm, which uses a frequent pattern tree (FP-tree) structure to mine frequent itemsets without candidate generation.

10. Support Count: The actual number of transactions that contain a particular itemset. It is the numerator in the support calculation.

Knowledge base - This is the domain knowledge that is used to guide the search or evaluate the interestingness of the result pattern. Such knowledge can include concept hierarchies, used to organize attribute or attribute values into different levels of abstraction. Knowledge such as user benefits, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata

Data mining engine- This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association, classification, cluster analysis, and evolution and deviation analysis.

Pattern evaluation module - This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search towards interesting patterns. It may use interestingness thresholds to filter out discovered patterns. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining methods used. For efficient data mining, it is highly recommended to push the evaluation of pattern interest as deep as possible into the mining process so as to confine the search to only the interesting patterns.

Graphical user interface- This module communicates between users and the data mining system, allowing user to interact with the system by specifying data mining query or task, providing information to help focus the search and performing exploratory data mining based on the intermediate data mining results. In addition this component allows the user to browse database and data warehouse schemes or data structures, evaluate mined patterns and visualize the patterns in different forms. From a data warehouse perspective, data mining can be viewed as an advanced stage of on-line analytical processing (OLAP).

Material and Methods

Data Collection Method: From the supermarket called Shetkari Bazar in kolhapur city in Maharashtra, India, the day today transactional data is gathered. The sale of products are the actual transactions made by the customer, which consist of various products. This various products transactions made by the customer are stored in the secondary storage medium i.e. hard disk or CD’s. It is stored in different categories i.e. department wise sale, counter wise sale, country wise sale, day wise sale, month wise sale. Depending on the necessary statistical analysis, required data is filtered through these databases.

Methodology:

In a supermarket, suppose as a manager, he may like to learn more about the buying habits of the customers. “Which groups or sets of items customers are likely to purchase on a given trip to the store? To answer this question, Market Basket Analysis from Association Rule mining may be performed on the retail data of customer transactions at stores. The result may be used to plan marketing or advertising strategies as well as catalog design different store layouts. In one of the strategies, items that are frequently purchased together can be placed in close proximity in order to further encourage the sale of such items together. Market Basket Analysis can help retailers to plan which items to put on sale at reduced prices.

If we think of the universe as the set of items available at the store, then each item has a Boolean variable represented by a Boolean vector of values assigned to these variables. The Boolean vector can be analyzed for buying patterns that reflect items that are frequently associated or purchased together. These patterns can be represented in the form of association rules. For example, the information that customers who purchase computers also tend to buy printers at the same time is represented in Association Rule below. Computer = Printer Support = 20%, Confidence = 80% Rule support and confidence are two measures of rule interest; they reflect the usefulness and certainty of discovered rules. A support of 20% means that 20% of all the transactions under analysis show that computer and printer are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the printer. Typically association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold that can be set by users or domain experts.

The Apriori Algorithm

The Apriori Algorithm is one of the most popular algorithms using association rule learning over relational databases. It identifies the items in a data set and further extends them to larger and larger itemsets. However, the Apriori Algorithm only extends if the itemsets are frequent, that is the probability of the itemset is beyond a certain predetermined threshold.

The Apriori Algorithm proposes that:

The probability of an itemset is not frequent if: P(I) < Minimum support threshold, where I is any non-empty itemset Any subset within the itemset has value less than minimum support. The second characteristic is defined as the Anti-monotone Property. Agoodexample would be if the probability of purchasing a burger is below the minimum support already, the probability of purchasing a burger and fries will definitely be below the minimum support as well.

Steps in the Apriori Algorithm

- The diagram below illustrates how the Apriori Algorithm Starts building from the smallest itemset and further extends forward.

- The algorithm starts by generating an itemset through the JoinStep, that is to generate (K+1) itemset from K-itemsets. For example, the algorithm generates Cookie, Chocolate and Cake in the first iteration.

- Immediately after that, the algorithm proceeds with the Prune Step, that is to remove any candidate item set that does not meet the minimum support requirement. For example, the algorithm will remove Cake itSupport(Cake) is below the predetermined minimumSupport.

- It iterates both of the steps until there are no possible further extensions left. Note That this diagram is not the complete version of the transactions table above. It Serves as an illustration to help paint the bigger picture of the flow.

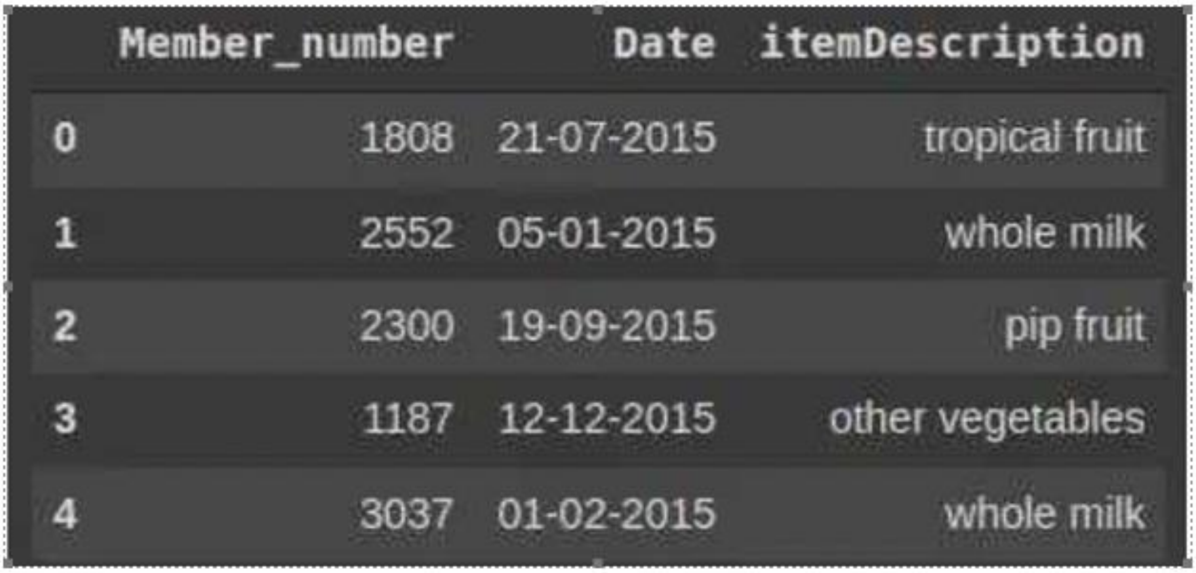
Code Implementation

To perform a Market Basket Analysis implementation with the Apriori Algorithm, we will be using the Groceries dataset from Kaggle. The dataset was publishedbyHeeral Dedhia on 2020 with a General Public License, version 2. The dataset has38765 rows of purchase orders from the grocery stores.

Program

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import re

from mlxtend.frequent_patternsimport apriori

from mlxtend.frequent_patterns import association_rules

from mlxtend.preprocessing

from mpl_toolkits.mplot3d import Axes3D

import network as  basket = pd.read_csv("Groceries_dataset.csv")

display(basket.head())
```

OUTPUT



	Member_number	Date	itemDescription
0	1808	21-07-2015	tropical fruit
1	2552	05-01-2015	whole milk
2	2300	19-09-2015	pip fruit
3	1187	12-12-2015	other vegetables
4	3037	01-02-2015	whole milk

Association Rule Mining: This technique uncovers relationships between items in a transaction and is the core of market basket analysis.

FP-Growth Algorithm: An efficient algorithm for mining frequent itemsets that significantly reduces the computational complexity of the process.

Conditional Pattern Base and Tree: These data structures play a crucial role in discovering frequent patterns in transaction databases.

Real-World Applications of Market Basket AnalysisRetail:

Retailers use market basket analysis to optimize product placement, create targeted promotions, and enhance the overall shopping experience

E-commerce: Online businesses utilize market basket analysis to provide personalized product recommendations and improve cross-selling and upselling strategies

Grocery Stores: Supermarkets leverage market basket analysis to understand customer behavior, manage inventory, and optimize sales and marketing efforts.

Benefits and Limitations of Market Basket Analysis

Benefits

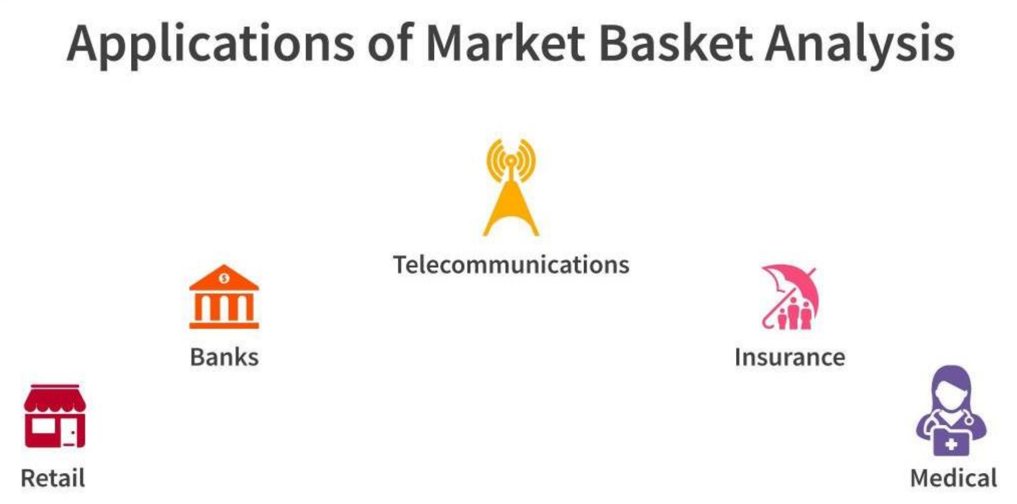
- Provides insights into customer behavior
- Enables targeted marketing strategies
- Supports inventory management
- Improved Product Placement
- Optimized Inventory Management

Limitations

- Does not reveal causality
- Dependent on transaction data quality
- Can be influenced by seasonal trends
- Data Quality and Quantity
- Complexity and Computational Resources

Conclusion

It is observed from the analysis that the data mining tools can be effectively used for optimizing the patterns associated with dynamic behaviors of the transactions which were made by the customers in purchasing some specific products. I have used the Market basket analysis algorithm, a widely and more pre dominantly used algorithm from association rule in Data Mining. Using this algorithm the frequent transactions made by the customers have been analyzed using the support and confidence of the customers in buying associated items. By using this methodology it is seen that there exists certain association between the products at the time of purchasing the products by the customers. Further it is observed that this analysis can best be used in managing the product placement on the shelves in the supermarket. This method can prove to fetch more profit to the seller. Thus the Data Mining tool can be used to improve the strategy in placement of the product on the shelf by using the Data mining tools.



References

1. M .Mohammed and B. Arkok. An Improved Apriori AlgorithmForAssociation Rules. (2014). International Journal on Natural LanguageComputing. 3. 10.5121/ijnlc.2014.3103.

2. D.H. Goh, R.P. Ang. An introduction to association rule mining: An Application in counseling and help-seeking behavior of adolescents. (2007). Behavior Research Methods 39, 259–266

3.S. Raschka. Machine Learning Extensions Documentation. (2021). Retrieved from:

4. A. Hagberg, D. Schult, P. Swart. NetworkX Reference Release 2.7.1. (2022). Retrieved from: https://networkx.org/

5. H. Dedhia. Groceries Dataset licensed under GPL 2. (2020).

6.Trnka, A. (2010, June). Market basket analysis with data mining methods. In *2010 International Conference on Networking and Information Technology* (pp. 446-450). IEEE.

7.Trnka, Andrej. "Market basket analysis with data mining methods." In *2010 International Conference on Networking and Information Technology*, pp. 446-450. IEEE, 2010.

8.Trnka, A., 2010, June. Market basket analysis with data mining methods. In *2010 International Conference on Networking and Information Technology* (pp. 446-450). IEEE.

9.Trnka A. Market basket analysis with data mining methods. In2010 International Conference on Networking and Information Technology 2010 Jun 11 (pp. 446-450). IEEE.

10. Vijaylakshmi S., Mohan V., Suresh Raja S., Mining of users access behavior for frequent sequential pattern from web logs, International Journal of Database Management System (IJDM), 2, (2010)

11. Yıldız B. and Ergenç B., (Turkey) in Comparison of Two Association Rule Mining Algorithms without Candidate Generation, International Journal of Computing and ICT Research, 674(131), 450-457 (2010)

12. Nan-chan Hsich, Kuo-Chang cha Enhancing consumer behavior analysis by data mining techniques (2009)

13. Peter P. Wakabi-Waiswa Venansius Baryamureeba, Extraction Of Interesting Association Rules Using Genetic Algorithms International Journal of Computing and ICT Research, 2(1), (2008)

14.Shrivastava A. and Sahu R., Efficient Association Rule Mining for Market Basket Analysis, Global Journal of e-Business and Knowledge Management, 3(1), (2007)

15.Junzo Watada and Kozo Yamashiro, A Data Mining Approach to consumer behavior-Proceedings of the