MARKET BASKET ANALYSIS

Sai Rajesh[CB.EN.U4ECE21241]

CH.K.VARDHAN [CB.EN.U4ECE21206]
T.KARTHIK [CB.EN.U4ECE21260]

Faculty Incharge : Dr. M. Venkateshkumar

Department of Electronics and Communication
Engineering, Amrita School of Engineering, Coimbatore,
Amrita Vishwa Vidyapeetham, India

Abstract - ***Market basket analysis, facilitated by the Apriori algorithm, provides a powerful means for retailers to glean actionable insights from customer purchasing behaviors. This guide presents a user-friendly approach to implementing this technique. By processing transactional data with Python, organisations can identify frequently co-purchased items and establish association rules. The algorithm's flexibility allows businesses to set thresholds, refining the analysis to suit their objectives. With a focus on support and confidence metrics, retailers can pinpoint robust correlations and enhance decision-making. Ultimately, this guide empowers businesses to fine-tune marketing strategies, optimise product placements, and boost customer satisfaction by aligning offerings with genuine consumer preferences. The accessible Python code ensures a straightforward application of Apriori, making advanced data mining techniques accessible for informed and strategic decision-making in the competitive landscape of retail. As businesses leverage Apriori insights, they gain a competitive edge by adapting offerings to align with genuine consumer preferences, fostering loyalty and sustained growth in the dynamic retail landscape.***

Keywords - ***Association Rules, Apriori Algorithm, Customer Preferences, Product Recommendations***

INTRODUCTION

The project is to create how enhance the customers are buying the products in Market. In the ever-evolving landscape of retail, understanding customer behaviour is paramount. Market Basket Analysis (MBA) stands out as a pivotal tool, offering retailers profound insights into the purchasing patterns of their customers. At the heart of MBA lies the Apriori algorithm, a robust data mining technique that empowers businesses to uncover associations between products, optimise marketing strategies, and enhance overall customer satisfaction. This introduction provides a comprehensive overview of market basket analysis, its significance in retail, and the foundational role played by the Apriori algorithm.

Understanding Market Basket Analysis (MBA):

Market basket analysis is a data mining technique that explores the relationships between products based on customer transaction data. It aims to answer a fundamental question: What items are frequently purchased together? By identifying these associations, businesses gain valuable intelligence that goes beyond individual product sales. Instead, MBA reveals hidden connections, enabling retailers to make informed decisions about product placements, promotions, and inventory management.

Significance in Retail:

In the competitive world of retail, understanding customer preferences is a game-changer. MBAs allow businesses to go beyond basic transactional data and decipher intricate patterns. For instance, if customers often buy cereal and milk together, a retailer can strategically place these items closer in the store or create bundled promotions, boosting sales and enhancing the shopping experience. The insights derived from MBA are not just about increasing revenue; they are about creating a personalised and satisfying customer journey.

Apriori Algorithm:

At the core of the MBA is the Apriori algorithm, a pioneering method for discovering frequent itemsets in transactional data. The algorithm, proposed by Rakesh Agrawal and Ramakrishnan Srikant in 1994, operates on the principle of "a priori property," which states that if an itemset is frequent, all of its subsets must also be frequent. This property significantly reduces the number of itemsets that need to be examined, making Apriori an efficient and widely adopted algorithm in the realm of market basket analysis.

Enhancing the Customer Experience:

Beyond its impact on sales and marketing, MBA contributes significantly to enhancing the overall customer experience. By understanding the natural affinities between products, retailers can make shopping more convenient and enjoyable for customers. This, in turn, fosters customer loyalty and satisfaction, key factors for long-term success in the retail industry. Retailers can uncover not only what products are frequently bought together but also the sequence in which they are purchased. This knowledge enables the optimisation of product placements, the creation of targeted promotions, and the development of personalised marketing strategies.

The Role of Apriori in Uncovering Associations Rule:

Aprio The Apriori algorithm serves as a cornerstone in market basket analysis, efficiently uncovering associations between products in transactional data. Through its systematic candidate generation and pruning, Apriori identifies frequent item sets based on adjustable support thresholds. The algorithm goes beyond generating association rules and expressing meaningful relationships between items. Metrics like confidence and lift provide quantitative measures of rule strength and significance. Apriori's adaptability allows analysts to tailor the analysis to specific datasets, making it a versatile and powerful tool for revealing hidden patterns in customer purchasing behavior. In real-world applications, businesses leverage Apriori's insights to optimise marketing strategies, enhance product placements, and

ultimately foster a more personalised and satisfying customer experience.

METHODOLOGY

A significant function called Apriori Property that decreases the search space helps increase the effectiveness of greater production of Periodic Arrays Apriori technique This rule indicates how well a package appears in a transaction. The market-based analysis is a classic result. We adopt an iterative or baseline search strategy where only k-frequently used itemset are identified. All frequent itemset non-empty subsets should be regular. The Apriori algorithm's core principle is its reference measure anti-monotonicity. Could well be a sluggish Apriori algorithm. The biggest downside is the time to keep a large number of candidates sets with common items, low minimum support, or vast item sets, i.e., for a large number of data that isn't an effective method. It will repeatedly scan the database for applicant items.
Apriori is very low and inefficient unless there are a big number of transactions with memory capacity restricted. The mining of associations reveals fascinating partnerships and correlations between vast sets of samples. A market- based analysis is one of the primary methods for showing correlations between things through wide relationships. It helps distributors to detect connections between the products customers commonly purchase together. A functional itemset is an item whose support exceeds or is equivalent to the threshold minus. The Association Rule is the statement of the X->Y form when X and Y are two collections of objects. The number of transactions with items in {X} and {Y} sections is a proportion of the overall number of transactions. It shows how often products are bought together as a percentage of the total transactions. In evaluating datasets, the association rules are highly beneficial. In supermarkets, the data are acquired via barcode scanners. These databases are a huge number of transacting records listed on a single buy all things purchased by the consumer. The administrator may however be knowledgeable whether particular item groupings are constantly acquired and used together for the revision of shop layouts, cross-selling, predictive marketing.

The given three components comprise the apriori algorithm.

- Support
- Confidence
- Lift

Support :-
Support is the proportion of transactions that contain a specific set of items.

$$\text{Support}(A) = \frac{\text{Transactions containg A}}{\text{Total Transactions}} \text{ ---------- Eq(1)}$$

For a rule A → B :

$$\text{Support}(A \cup B) = \frac{Transactions\ containing\ both\ A\ and\ B}{Total\ Tansactions}$$

Confidence :-
Confidence measures the likelihood that if item A is purchased, item B will also be purchased.

$$\text{Confidence}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)} \text{ ---------- Eq(2)}$$

Lift :-
Lift compares the observed support of a rule to the expected support under independence .

$$\text{Lift}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A) \times \text{Support}(B)} \text{ ---------- Eq(3)}$$

Alternatively, it can be expressed as the ratio of the confidence to the support of the consequent:

$$\text{Lift}\ (A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{Support(B)} \text{ ----------Eq(4)}$$

These formulas are fundamental in assessing the strength, significance, and impact of association rules generated by algorithms such as Apriori in Market Basket Analysis. Adjusting the thresholds for these metrics allows analysts to filter out less meaningful rules and focus on those that are more relevant to the specific objectives of their analysis.

The following are the main steps of the apriori algorithm in data mining:

- Set the minimum support threshold - min frequency required for an itemset to be "frequent".
- Identify frequent individual items - count the occurence of each individual item.
- Generate candidate itemsets of size 2 - create pairs of frequent items discovered.
- Prune infrequent itemsets - eliminate itemsets that do no meet the threshold levels.
- Generate itemsets of larger sizes - combine the frequent itemsets of size 3,4, and so on.
- Repeat the pruning process - keep eliminating the itemsets that do not meet the threshold levels.
- Iterate till no more frequent itemsets can be generated.
- Generate association rules that express the relationship between them - calculate measures to evaluate the strength & significance of these rules

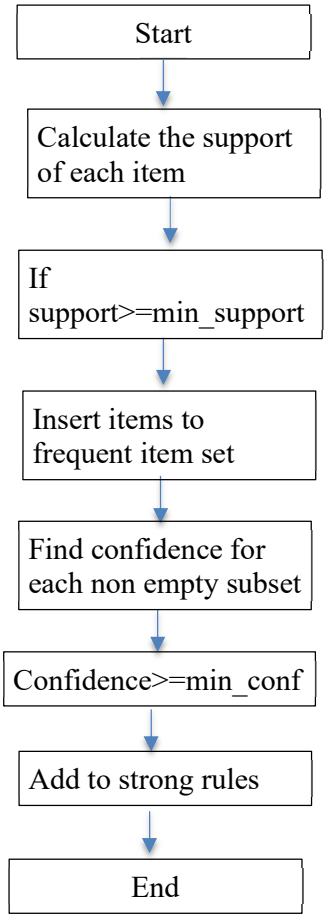

Figure 1 : Flowchart of Aprior Algorithm

RESULT & ANALYSIS:

The dataset represents transactions in a supermarket, where each row corresponds to a purchase, and columns represent items bought. The code utilizes association rule mining, specifically the Apriori algorithm, to uncover patterns in item co-occurrence. By generating association rules, the analysis identifies related items for a given product (e.g., 'yogurt') and calculates support percentages between pairs of items (e.g., 'other vegetables' and 'whole milk'). The code also visually presents the top frequent itemsets with two items. The objective is to derive insights into item relationships, aiding in optimizing product placement and marketing strategies based on customer purchasing behavior in the supermarket.
The goal is to discover relationships between items that frequently appear together in transactions. Let's break down the key components and provide a brief analysis of the results.
Data Loading and Preprocessing:
The code starts by importing necessary libraries, including NumPy, Pandas, CSV reader, and specific modules from the mlxtend library for association rule mining.
The read_and_prepare_data function reads a dataset from a CSV file, processes it, and converts it into a binary transaction format suitable for Apriori.
Association Rule Mining:
The Apriori algorithm is used to find frequent itemsets in the dataset. The apriori function is applied to the prepared dataset, specifying a minimum support threshold of 0.02.
The resulting frequent itemsets are then used to generate association rules, and these rules are stored in the variable rules.
Finding Related Items for a Given Item:
The find_related_items function takes an input item and the frequent itemsets as input, and it identifies rules where the input item is present either in the antecedents or consequents.
The output includes the antecedents, consequents, and the support percentage for each rule.
Plotting Top Frequent Itemsets:
The plot_top_frequent_itemsets function visualizes the top 10 frequent itemsets with two items. It creates a bar plot showing the support percentage for each itemset.
Finding Support Percentage Between Two Items:
The find_support_percentage_between_two_items function calculates the support percentage between two specified items based on the association rules.
Example:
The code provides an example usage where the dataset is loaded, frequent itemsets are mined, and then analysis is performed for a specific item ('yogurt') and a pair of items ('other vegetables' and 'whole milk').
The results include related items for 'yogurt,' support percentages between 'other vegetables' and 'whole milk,' and a plot of the top frequent itemsets with two items.
Analysis:
The analysis would involve interpreting the association rules, understanding which items are frequently bought together, and examining the support percentages.
For instance:
The output for 'yogurt' provides related items and their support percentages.
The support percentage between 'other vegetables' and 'whole milk' is calculated.
The bar plot visually represents the top 10 frequent itemsets with two items.
Users can analyze these results to make informed decisions, such as optimizing product placement in a store or creating targeted marketing strategies based on item associations.

OUTPUTS:

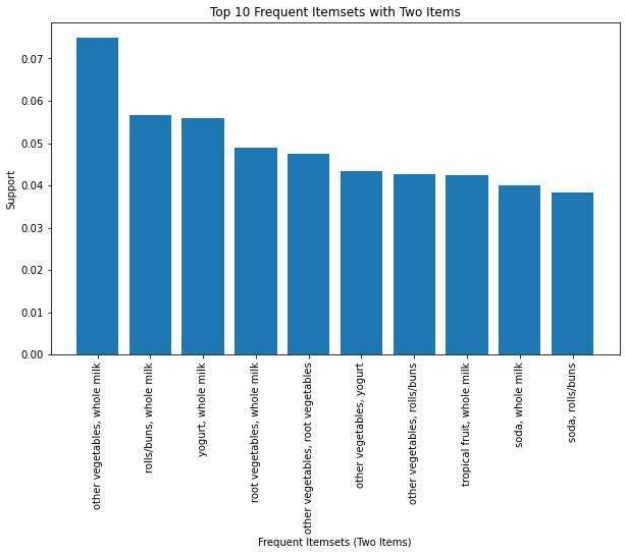

Fig:1 Frequent itemsets with Two items

Output for FPGrowth :

Antecedents: pork
Consequents: whole milk (2.22%)

Antecedents: pork
Consequents: other vegetables (2.17%)

The support percentage between 'other vegetables' and 'whole milk' is: 2.23%
Execution time: 0.54 seconds

Output for Apriori :

Antecedents: pork
Consequents: other vegetables (2.17%)

Antecedents: pork
Consequents: whole milk (2.22%)

The support percentage between 'other vegetables' and 'whole milk' is: 7.48%
Execution time: 0.49 seconds

CONCLUSION:

In summary, the Python script employs association rule mining techniques, specifically the Apriori algorithm and FP-Growth, catering to different dataset sizes. For smaller datasets, Apriori is utilized due to its ability to efficiently discover frequent itemsets. Apriori works well when the dataset fits into memory, and its step-by-step candidate generation approach is suitable for relatively modest transaction volumes.Conversely, FP-Growth is employed for larger datasets, offering improved scalability by constructing a compact tree structure known as the FP-Tree. This makes FP-Growth particularly advantageous when dealing with extensive transactional data, as it reduces the need for multiple database scans and optimizes memory usage. The script's flexibility in selecting the appropriate algorithm based

on dataset size enhances its applicability to diverse scenarios. By incorporating both Apriori and FP-Growth, the script provides a versatile solution for association rule mining in grocery datasets of varying scales. This approach acknowledges the strengths of each algorithm, contributing to the script's adaptability and effectiveness in extracting meaningful insights from transactional data in retail settings.

REFERENCES

- B. Ganguly Raich and M. Tota, "Machine Learning for Market Basket Analysis through", IOSR Journal of Engineering (IOSRJEN), pp. 22-23, 2019.
- M. Kaur and S. Kang, "Market Basket Analysis: Identify the changing trends of market data", International Conference on Computational Modeling and Security (CMS 2016), 2016.
- . R. Gangurde, D. B. Kumar and D. S. D. Gore, "Optimized Predictive Model using Artificial Neural for Market Basket Analysis" in Research Gate, Pune, Maharashtra, India, 2017.
- M. R. Wick and P. J. Wagner, "Using market basket analysis to integrate and motivate topics in discrete structures" in ACM SIGCSE Bulletin, Eau Claire, 2006.
- S. Mainali, "MARKET BASKET ANALYSIS" in GitHub, Kirtipur, 2016.
- M. A. Ulas, "MARKET BASKET ANALYSIS FOR DATA MINING" in Academia.edu, Istanbul, 2001.
- G. R. Grau, "Market Basket Analysis in Retail" in UPCommons. Global access to UPC knowledge, Barcelona, 2017.
- K. A. B. A. Kadir, "CLUSTERING ALGORITHM FOR MARKET-BASKET ANALYSIS: THE UNDERLYING CONCEPT OF DATA MINING TECHNOLOGY" in University Putra Malaysia Institutional Repository, Serdang, 2003.
- V. Gancheva, "Market Basket Analysis of Beauty Products" in SEMANTIC SCHOLAR, Rotterdam, 2013.
- F. Arasteh and F. Arbab, "Studying Changes in Corporative Society Retail Store Sales as a Result of Shelves' Rearrangement and Promotions Based on Market Basket Analysis" in Digitala Vetenskapliga Arkivet, Tehran, 2016.
- Berry, M. J. A. and Linoff, G. Data mining techniques for marketing, sales and customer support, USA: John Wiley and Sons,1997
- WANG Li-Zhen, ZHOU Li-Hua et al., D\warehouse and data mining principles and applications, Beijing:Science Press, 2005
- "Lcm ver.2: Efficient mining algorithms for frequent/closed/maximal itemsets", Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'04) volume 126 of CEUR Workshop Proceedings, 2004.
- "B.R'acz.nonordfp: An FP-growth variation without rebuilding the FPtree", Proceedings ofthe IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'04) volume 126 of CEUR Workshop Proceedings, 2004.
- A Adhikari and P R. Rao, "A framework for synthesizing arbitrary Boolean expressions induced by frequent itemsets[C]", Conference on Artificial Intelligence, 2007.
- Andre Bergmann, "Data Mining for Manufacturing: Preventive Maintenance, Failure Prediction, and Quality Control"
- Fayyad, U. M; Piatetsky-Shapiro, G. ; Smyth, P.; and Uthurusamy, R. 1 996. Advances in Knowledge Discovery and Data Mining. Menlo Park, Calif.: AAAI Press.
- Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.
- H. Mahgoub,"Mining association rules from unstructured documents" in Proc. 3rd Int. Conf. on Knowledge Mining, ICKM, Prague, Czech Republic, Aug. 25-27, 2006, pp. 1 67-1 72.
- M. Ashrafi, D. Taniar, and K. Smith "Redundant Association Rules Reduction Techniques". L ecture Notes in Computer Science, Volume 3S09, 2005, pp. 254-263 .