International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 10 | October - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

E-Commerce Market Basket Analysis using Apriori Algorithm

Khushi Gupta[1], Kashyapi Shah[2] , Ameya A. Kadam[3]

[1,2] Student, Dept. of Electronics and Telecommunication, Dwarkadas Jivanlal Sanghvi College of Engineering, Maharashtra, India

[3] Asst. Prof, Dept. of Electronics and Telecommunication, Dwarkadas Jivanlal Sanghvi College of Engineering, Maharashtra, India

---

Abstract - This paper presents the usage of the Apriori Algorithm to implement market basket analysis to identify purchase patterns for items that are frequently bought together by customers. The results of this analysis are primarily used to improve sales of multi-product stores by enhancing product placement based on consumers' shopping habit. In this particular scenario, we have used the data from an online E-commerce store that caters to customers across the world, but primarily focused to the United Kingdom.

Key Words: Market Basket, Apriori Algorithm, Association Rules, E-Commerce, Consumer Behaviour, Data Mining

1. INTRODUCTION

Consumer behavior refers to a consumer's decision to acquire, purchase and use purchased goods and services, as well as the elements that influence their decision. Every consumer has unique requirements and use diverse behaviors to meet their needs. However, they do share several characteristics, one of which is a desire to maximize their satisfaction when consuming an essential commodity or service. Research and analytical studies of consumption activity and consumer behavior beholds the potential to unfurl valuable information to store owners by predicting and forecasting consumer needs, preparing promotional marketing campaigns, managing inventory and ultimately boosting sales figure manifolds.

Market Basket Analysis is a widely used data mining method focused on examining customer buying habits to pinpoint items commonly bought in tandem. The objective is to discern connections between products found in a shopper's cart. By harnessing this information, retailers can enhance product positioning while aligning with customer preferences.

2. LITERATURE

The proposed paper is based on the transactional data of an online retail store, since as a of consequence of the Covid-19 pandemic, traditional offline marketplace shifted base to the internet, resulting into an upward hike of the E-Commerce sales worldwide. International Trade Association reported an active increase in online retail through the year 2019-2021. A predictive analysis forecasted online retail to acquire at least 22% of the total retail sales globally. Fig-1 illustrates the increase in the global retail ecommerce sales of the world from year 2012 to the year 2024[1].

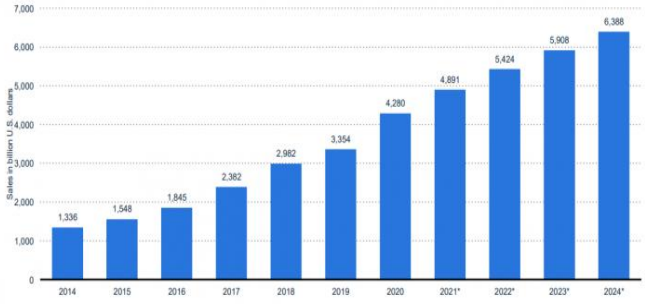

Fig -1: Global retail ecommerce sales worldwide from 2014 to 2024 (in billion USD)[1]

Dataset:

A special identification code given to each invoice produced is called an invoice number. Each product or item in the inventory is given a StockCode, which serves as an identification number. Description is a succinct statement that identifies the thing or thing's name. The quantity of a specific good or item are listed on the invoice. Date that the invoice was created or issued, according to invoice data. The cost of a single unit of the good or service is the price per unit. A special identification number given to every customer called as the Customer Id and Country Id depicts nation that the customer is from. This dataset can be helpful for a variety of analytical tasks, including comprehending sales patterns, consumer behavior, inventory management, and spotting global trends. With the use of this data, analysis of topics such as total revenue calculation, product identification, client country segmentation, analysis of purchase trends over time, and more, can be carried out.

The store data used was of the size: 541910 rows and 8 columns[2]. Each data in the set is a combination of products bought together by a customer. Figure below shows the head of the data.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom

Fig-2: First 5 rows of the Dataset [2]

Market Basket Analysis:
Market basket analysis comprehends the purchasing habits of clients by finding the points of correlation and connection between the things they put to their shopping carts.
In market basket analysis, association rules are used to predict the analogy of products being purchased together and its objective is to identify the goods that customers purchase concurrently. A firm will gain a lot from being able to leverage consumer purchasing trends to develop business strategies that can help retailers grow sales while improving customers' shopping experiences. Successful corporate plans can be developed with the help of such expertise.

Apriori Algorithm:
The Apriori algorithm is a cornerstone in relational database systems, primarily used for revealing patterns of co-purchased items through the lens of association rule learning. Its main goal is to identify commonly found items in the database and incrementally combine them to form larger item sets, provided these combinations consistently show up in the data. These frequent item sets, identified by Apriori, pave the way for generating association rules, which are instrumental in exposing dominant trends within the database. This becomes invaluable for tasks such as market basket analysis. The brainchild of Agrawal and Srikant in 1994, Apriori was crafted with transactional databases in mind. However, alternative techniques exist for detecting association rules in databases that lack transactions or timestamps. In the Apriori perspective, each transaction is treated as a distinct set of items. The technique utilizes a "ground-up" methodology, starting with commonly spotted individual items and methodically bundling them, using the data as a reference. The process halts when no further meaningful bundles can be identified.

Association Rule:
Within a dataset, the Apriori algorithm is used to create association rules. These rules suggest that if item A occurs, there's a particular likelihood of item B also occurring. One common application of the Apriori algorithm and its association rules is in market basket analysis. This study aims to pinpoint items that are frequently bought together, assisting businesses in improving sales and crafting impactful marketing strategies. By evaluating the buying habits through retailer transaction data, businesses can glean crucial insights about product affiliations and consumer preferences. Such insights aid companies in refining their product positioning, cross-promotion tactics, and tailored suggestions, all of which contribute to an improved shopping journey and an uptick in revenue.
When it comes to association rules, they categorize each customer's purchases into an array. This array, or item set, showcases the items purchased in the following manner:

$$I= \{i_1, i_2, i_3, i_4, \ldots\ldots\ldots, i_k\}$$

The transaction is represented by the following expression:

$$T= \{t_1, t_2, t_3, t_4, \ldots\ldots\ldots, t_n\}$$

Then, an association rule can be explained as an extension of the form:

$X \Rightarrow Y$, where $X \in I, Y \in I$ and $X \cap Y=0$

The 4 metrics used for the evaluation of this rule are:

i) Support:

Support is a metric of how often the item appears in the dataset.

$$supp(X \Rightarrow Y) = \frac{|X \cup Y|}{n}$$

In essence, support can be described as the ratio of transactions containing both X and Y to the total number of transactions. Rules tend to be less valuable when the support values are low.

ii) Confidence:
For association rules, confidence denotes the likelihood that the second item is purchased when the first one is

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

iii) Lift:
The lift of an association rule measures the observed support relative to what would be anticipated if the two items were independent. It can be formulated as:

$$lift(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)supp(Y)}$$

Strong association is indicated by greater lift values

iv) Conviction:
Conviction represents a measure used to determine the frequency of item X being present without item Y if they were assumed to be independent, in comparison to the actual frequency at which such predictions prove wrong. The formula for the Conviction of a rule can be described as:

$$conv(X \Rightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \Rightarrow Y)}$$

High value means that the consequent depends strongly on the antecedent.

Data Mining:
Data mining is the methodical process of finding hidden patterns and trends in sizable datasets with the aim of gaining insightful knowledge that can support data-driven decision-making. It enables the investigation and extraction of important knowledge and information from unstructured data, revealing intricate linkages and patterns that could otherwise go unnoticed. It does this by utilizing cutting-edge algorithms and approaches. By exposing previously unknown linkages, trends, and patterns, this method is crucial in enabling organizations to

make wise decisions based on thorough research of enormous data repositories. Data mining uses a variety of viewpoints to find important patterns and useful information in data. Its use improves productivity, competitive advantage, and decision-making. Data mining is a process that draws useful clues from vast amounts of data to support decision- and problem-solving.

3. Research Methodology:

Association rules are found to be useful if the minimum threshold for support and that for confidence complies the threshold set by the user or consultant. Market basket analysis rules can be written as Event A => Event B.

Data was first preprocessed by identifying the null values present within it. After a thorough identification, any row containing null values as well as rows from the invoice column that begin showcased cancelled orders were eliminated from the dataset.

The preprocessed data is visualized the count of purchase per country in the form of a tree map. Fig- 3 shows the country wise transactions [3]. Since United Kingdom contributes to the majority of the dataset, the analysis has been narrowed down to United Kingdom.

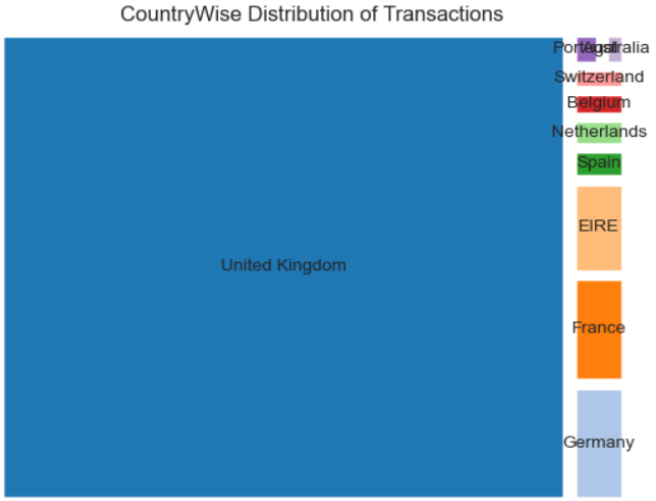

Fig-3: Country-wise Distribution of transactions[3]

In this research, a market basket analysis to uncover patterns of item co-occurrence in the United Kingdom retail market has been performed. Initially, filtering the dataset to focus solely on transactions from the United Kingdom has been done. The original data frame is filtered to include rows where the "Country" column is set to "United Kingdom". Subsequently, the data has been reshaped to organize it by individual invoices and their corresponding items. To facilitate analysis, The analysis required to encode the quantity values as binary indicators of item presence. Fig-4 shows the encoded values.[4]

Description	4 PURPLE FLOCK DINNER CANDLES	50'S CHRISTMAS GIFT BAG LARGE	DOLLY GIRL BEAKER	I LOVE LONDON MINI BACKPACK	NINE DRAWER OFFICE TIDY	OVAL WALL MIRROR DIAMANTE	RED SPOT GIFT BAG LARGE	SET 2 TEA TOWELS I LOVE LONDON	SPACEBOY BABY GIFT SET	TOADSTOOL BEDSIDE LIGHT
InvoiceNo										
536365	0	0	0	0	0	0	0	0	0	0
536366	0	0	0	0	0	0	0	0	0	0
536367	0	0	0	0	0	0	0	0	0	0
536368	0	0	0	0	0	0	0	0	0	0
536369	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
581582	0	0	0	0	0	0	0	0	0	0
581583	0	0	0	0	0	0	0	0	0	0
581584	0	0	0	0	0	0	0	0	0	0
581585	0	0	0	0	0	0	0	0	0	0
581586	0	0	0	0	0	0	0	0	0	0

16649 rows × 3844 columns

Fig -4: The encoded values[4]

Filtering for invoices containing at least two items results in a subset of the dataset that guarantees a sufficient quantity of items per transaction. Utilizing the Apriori algorithm, frequent item sets are identified based on a set minimum support threshold. Taking into account the support measure, item sets that often co-occur are prioritized. Fig- 5 displays items from the dataset commonly purchased by customers in the United Kingdom.[5]

Data entry 1 indicates the following:

	support	itemsets
99	0.121358	(WHITE HANGING HEART T-LIGHT HOLDER)
44	0.093197	(JUMBO BAG RED RETROSPOT)
80	0.090466	(REGENCY CAKESTAND 3 TIER)
6	0.084417	(ASSORTED COLOUR BIRD ORNAMENT)
71	0.082986	(PARTY BUNTING)
58	0.072841	(LUNCH BAG RED RETROSPOT)
86	0.064971	(SET OF 3 CAKE TINS PANTRY DESIGN)
52	0.064646	(LUNCH BAG BLACK SKULL.)
69	0.061004	(PAPER CHAIN KIT 50'S CHRISTMAS)
64	0.060939	(NATURAL SLATE HEART CHALKBOARD)

Fig -5: The most frequently bought items [5]

Next, association rule mining is applied to uncover relationships between items, assessing them using the lift metric. The derived association rules offer a glimpse into the co-purchasing tendencies of customers. These research outcomes illuminate the interrelationships among items and deliver essential data for marketing tactics, inventory oversight, and cross-selling

prospects.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.039802	0.043900	0.030957	0.777778	17.717202	0.029210	4.302452
1	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.043900	0.039802	0.030957	0.705185	17.717202	0.029210	3.256952
2	(LUNCH BAG RED RETROSPOT)	(LUNCH BAG PINK POLKADOT)	0.072841	0.055086	0.030632	0.420536	7.634188	0.026620	1.630668
3	(LUNCH BAG PINK POLKADOT)	(LUNCH BAG RED RETROSPOT)	0.055086	0.072841	0.030632	0.556080	7.634188	0.026620	2.088574
4	(JUMBO BAG RED RETROSPOT)	(JUMBO BAG PINK POLKADOT)	0.093197	0.052680	0.032908	0.353105	6.702899	0.027999	1.464412
5	(JUMBO BAG PINK POLKADOT)	(JUMBO BAG RED RETROSPOT)	0.052680	0.093197	0.032908	0.624691	6.702899	0.027999	2.416152
6	(LUNCH BAG BLACK SKULL)	(LUNCH BAG RED RETROSPOT)	0.064646	0.072841	0.031478	0.486922	6.684737	0.026769	1.807051
7	(LUNCH BAG RED RETROSPOT)	(LUNCH BAG BLACK SKULL)	0.072841	0.064646	0.031478	0.432143	6.684737	0.026769	1.647164

Fig -6: The group of items bought together

From our study with association rules, it's clear that the "Roses Regency Teacup and Saucer" and the "Green Regency Teacup and Saucer" have the highest "lift" value, suggesting a powerful relationship between these two products. They have a combined support of 0.0309, indicating that they were purchased together in 3.09% of all transactions. We visualized this data using a network graph with the help of Python's network module.

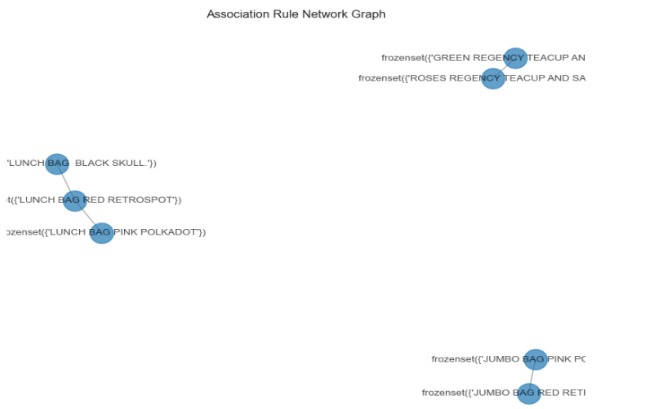

Fig -7: Network graph showing association between frequently bought together items

The generated graph displays the association rules with items as nodes, linked by edges. The density or prominence of these edges indicates the intensity of the association (lift) between these products. This graphical portrayal helps in understanding the intricate relationships and patterns discerned from the market basket analysis.

Interpreting the results, it's evident that customers often purchase the “Roses Regency Teacup and Saucer” alongside the “Green Regency Teacup and Saucer”. Next in line are the Lunch Bags in ‘Black Skull’, ‘Red Retrospot’, and ‘Pink Polkadot’ designs.

4. CONCLUSIONS

Market basket analysis, when applied to a dataset from the retail industry using the Apriori algorithm and association rules, offers a straightforward method for identifying patterns and relationships in transaction data, facilitating strategic marketing decision-making. The utility of market basket analysis isn't limited to the retail domain. In manufacturing, this technique aids in predictive analysis for machinery malfunctions, while in the pharmaceutical domain, it's utilized to find connections between diagnoses and active ingredients. Additionally, the financial industry can leverage it to spot fraudulent activities. The versatility and efficacy of this analysis emphasize its value in deriving actionable insights from diverse datasets across multiple industries.

REFERENCES

1. Online Retail. (2015). UCI Machine Learning Repository: https://doi.org/10.24432/C5BW33.
2. International Trade Association: https://www.trade.gov/impact-covid-pandemic-ecommerce
3. Aggarwal, C. C. (2015). Data Mining. New York: Springer. doi:10.1007/978-3-319-14142-8
4. Jirapatsil, Pattraporn, and Naragain Phumchusri. "Market basket analysis for fresh products location improvement: A case study of E-commerce business warehouse." In Proceedings of the 4th International Conference on Management Science and Industrial Engineering, pp. 23-28. 2022.
5. Raorane, A. A., Kulkami, R. V., & Jitkar, B. D. (2012). Association rule-extracing knowledge using market basket analysis. Research Journal of Recent Science(2277(2012):2502).
6. Zhao, Xian, and Pantea Keikhosrokiani. "Sales Prediction and Product Recommendation Model Through User Behavior Analytics." Computers, Materials & Continua 70.2 (2022).
7. Gupta Savi, Mamtora Roopal. A Survey on Association Rule Mining in Market Basket Analysis.International Journal of Information and Computationa Technology 2014:4(4):409-414.
8. He Zengyou, Xu Xiaofei, Huang Joshuaz, Deng Shengchun. FP-Outlier: Frequent Pattern Based Outlier Detection.ComSIS 2005:2(1):103- 118.