Market Basket Analysis - Apriori Algorithm

Team Kaizen

Atharva Ravi Puranik (JY77253)
Sayali Satish Dhavale (ZD62815)
Sujit Kandala (AI43610)
Sanjana Rajesh (IM71135)
Sanjay Sangaraju (TR02257)

Department of Information Systems,
University of Maryland, Baltimore County

IS 603: Decision Making Support System
Dr. Sanjay Purshottam

December 12, 2022

Table of Contents

ABSTRACT

We implemented Market Basket Analysis on E-commerce data using Apriori Algorithm which uses association rule mining. This was mainly done because Market basket analysis can potentially boost sales and customer satisfaction. Retailers employ market basket analysis, sometimes referred to as affinity analysis, as a significant data mining and statistical tool to comprehend consumer purchasing habits. It functions by examining consumer purchases that commonly occur in tandem. Retailers can use this to determine how certain products are related.

Association rules are used in market basket analysis to predict the likelihood of products being purchased together. Association rules count the frequency with which items occur together, looking for associations that occur far more frequently than expected.

Retailers can optimize product placement, offer special deals, and create new product bundles by using data to determine which products are frequently purchased together. These enhancements can potentially increase sales for the retailer while making the shopping experience more productive and valuable for customers. Customers may feel a stronger sentiment or brand loyalty toward the company if market basket analysis is used.

1. SIGNIFICANCE OF THE TOPIC

Retailers use market basket analysis, a data mining technique, to increase sales by better understanding customer purchasing patterns. It entails analyzing large data sets, such as purchase history, to identify product groups and products that are likely to be purchased together. The introduction of electronic point-of-sale (POS) systems aided the adoption of market basket analysis. In comparison to handwritten records maintained by store owners, digital records generated by POS systems made it easier for applications to process and analyze large volumes of purchase data.

Market Basket Analysis can be used to recommend a purchase based on the absence of a common pairing, such as when a customer orders only a small sandwich at a Quick Serve Restaurant (QSR). They are more likely to purchase a dessert or a second sandwich than someone who purchased a large sandwich. Staff who have been trained to recognize these situations can offer their customers the extra items, possibly at a discount to make the option more appealing.

When applied more deeply, Market Basket Analysis enables businesses to identify keystone products, those that distinguish them in the market and could potentially harm the business if they become unavailable or more expensive. Gourmet or other specialty items in a grocery store may have limited appeal, but the customers they attract (and the money they spend as a result) may justify prominent placement. Customers who order through the company's app may be interested in items or combinations that provide additional loyalty points. It works by looking for item combinations that occur frequently in transactions. To put it another way, it enables retailers to identify relationships between the items purchased by customers. Association Rules, which are based on the concept of strong rules, are widely used to analyze retail basket or transaction data. They are intended to identify strong rules discovered in transaction data using measures of interestingness.

2. PROBLEM STATEMENT

The goal of Market Basket Analysis models is to identify the next product that a customer might be interested in. As a result, marketing and sales teams will be able to devise more effective pricing, product placement, cross-sell, and up-sell strategies. It can improve shipping times and warehouse operations by predicting product sales in specific locations.

A typical Market Basket Analysis goal is to provide a set of association rules in the form:

IF [antecedent] THEN [as a result]

The "body" or "antecedent" of the rule is the first part of the rule, while the "head" or "consequent" is the second part of the rule. Furthermore, the antecedent and consequent can include multiple conditions, resulting in more complex rules. As an example:

IF ['Smartphone’] Then ['Case']

Problem statements that will be focused on throughout the paper are as below.

2.1 Product Association

To find the association between products purchased together:

With this problem statement, we will be able to relate two products. We would get an understanding of what changes are coming into the picture because of a particular relation between product A and product B.

2.2 Product Recommendation

To evaluate the lift caused in the sales of product B due to the recommendation shown while purchasing product A:

With the help of this problem statement, we will get an understanding of what is leading to the sale of product A because of product B and vice versa. With slight modifications, multiple use cases can be formed out of this objective or methodology.

2.3 Purchase Patterns

To understand customer purchase patterns for products:

With the help of this problem statement, we will be able to discover patterns between what customers buy by identifying product or menu item combinations that frequently co-occur in transactions. Retailers can use this information to create new products or pricing models that generate new revenue by identifying relationships between the products that people buy.

3. BACKGROUND OF THE TOPIC

Market Basket Analysis is a data processing technique that is employed to discover relationships between different items. The primary goal of market basket analysis in retail is to supply data to the distributor about a customer's purchasing behavior, which can aid the distributor in making the right selections (Kawale et al., 2018).

For several decades, data mining has played a significant role in marketing literature. Market basket analysis is among the oldest areas of data mining and the best illustration of mining association rules. Researchers have developed several algorithms for Association Rule Mining (ARM) to assist users in achieving their goals (Kawale et al., 2018).

One of the traditional algorithms for identifying patterns in data in Boolean association rules is the Apriori algorithm, which was first introduced by Ramakrishnan Srikant and Rakesh Agrawal. The idea of mining quantitative rules from large relational tables is explained in detail by the authors. Julander examined the proportion of customers who bought a specific product and the proportion of all sales that this product was responsible for. Making these associations makes it simple to identify the top products and their sales share. Since many customers interact with these particular product types on a daily basis, measuring which products are the leading products is crucial. It is essential to use this information for placing displays because the departments with popular products attract a lot of in-store traffic. The

process of generating association rules is another important area of study in the field of exploratory analysis (Kawale et al., 2018).

Berry and Linoff aimed to discover patterns by extracting correlations or co-occurrences from transactional data from a retail outlet. Customers who buy bread frequently also buy related products such as milk, butter, or jam. It makes perfect sense that these groups are units placed adjacent to one another in a retail center so customers can find them quickly. Such related product groups should also be placed side by side to notify customers of related items and to assist them through the center in a sensible manner (Kawale et al., 2018).

For this project, we are using the dataset from below kaggle source (Ostrowski, 2018). This is an E-commerce data set. This dataset contains transactions made by an online UK based store. It sells gift items. Transactions present are between 2010 and 2011. Many customers here are wholesalers. It consists of columns Invoice number, stock code, description, quantity, invoice data, Unit price, customer ID and country.

4. OVERALL OBJECTIVES AND GOALS

4.1 Technical Perspective

In a short period of time, new upcoming scientific technology, of which Market Basket Analysis is a component, could be integrated into existing decision support systems of online marketplaces. Product segmentation based on customer purchase patterns and network role assists marketing managers in improving marketing activities such as product recommendation, product placement, cross-selling, up-selling and customer retention (Kafkas et al., 2021).

Market Basket Analysis is a rule-based machine learning technique for identifying relevant relationships between variables in sizable databases. For instance, buyers of product A are more inclined to purchase product B as well. As a result, Market Basket Analysis would quantitatively prove that product A and product B are related. The same is true with products A, B, and C if they are related to each other. The model needs to be adjusted for three key metrics: Lift, Confidence, and Support (Provost, 2013).

So, from a technical perspective, major goal is to explore and identify these Key Metrics – Lift, Support and Confidence.

4.2 Business Perspective

A search is what Market Basket Analysis is. Every search turns up more information than it was looking for. In the course of looking for one document, one also comes across ten others. Obviously, a person stops looking for things once located. The last place one looks is where one finds what was looked for. That implies that things were searched in additional locations and discovered additional items. The Market Basket Analysis also results in this. There is a possibility of discovering other patterns and affinities that weren't expected. By doing so, an

analysis project like Market Basket Analysis will learn information about the company that goes beyond the solution to the immediate inquiry (Silvers, 2011).

We would like to Examine the client's internet browsing habits and previous sales data to help retailers gain insights into the sales patterns of any product. The process of market basket analysis is used to considerably boost up-sell and cross-sell chances while also improving marketing effectiveness.

With this, primary goal from a business perspective will be to get incremental sales and revenue and improve overall ROI.

5. IMPLEMENTATION DETAILS

We are using the dataset from below kaggle source (Ostrowski, 2018). This is an E-commerce data set. This dataset contains transactions made by an online UK based store. It sells gift items. Transactions present are between 2010 and 2011. Many customers here are wholesalers. It consists of columns Invoice number, stock code, description, quantity, invoice data, Unit price, customer ID and country.

Dataset Link (for Kaggle):
https://www.kaggle.com/datasets/yekahaaagayeham/online-retail-for-market-basket-analysis

Source Link:
https://archive.ics.uci.edu/dataset/352/online+retail

5.1. Data Preprocessing

5.1.1 Preprocessing

After the data source is collected, it is preprocessed to convert it into usable form for applying data mining techniques on it. Occurrence of null values in the data set is checked and removed. The data is additionally preprocessed by removing single transaction customers data, removing invalid entries etc. Hence these kinds of redundant rows are removed that may not contribute to the model.

5.1.2 Results after Preprocessing

We found 135,080 null values and 79 single-order products in the dataset consisting of 541909 values. Elimination of these data has been done so far. 244,050 i.e., 60% of the remaining values have been used as the training data for the model.

5.2. Methods used to Implement

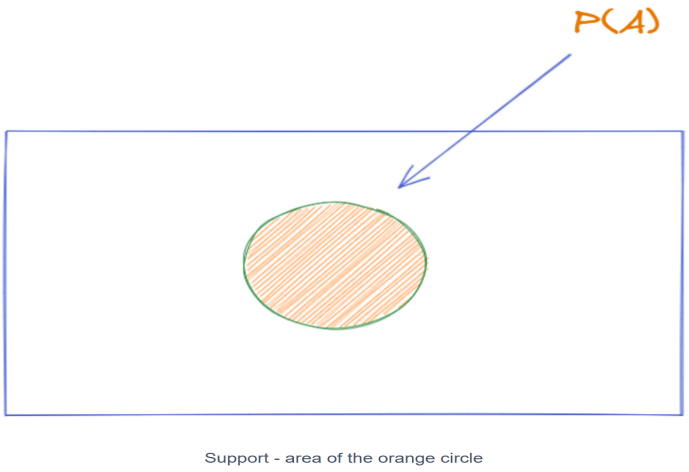
Inorder to achieve our goals, Market Basket Analysis. The classification and simplification of item sets which are purchased together by the customers will be done with the help of Apriori Algorithm. The algorithm majorly works around association between entities which are backed by 3 factors: Support, Confidence, Lift.

1. Support

This measurement reveals how frequently a given itemset appears in all transactions. Think of the item sets as "bread" and "shampoo," respectively.

Transactions involving bread will be much more prevalent than those involving shampoo. As a result, itemset1 will typically have more support than itemset2, as you correctly predicted. Now consider item sets 1 and 2, which are bread and shampoo respectively. Bread and butter will frequently be purchased together, but what about bread and shampoo? Not really. As a result, itemset1 in this instance will typically have more support than itemset2.Support is the percentage of all transactions in which the itemset appears, according to math.

support({A} ->{B} = Transactions containing both A and B / Total number of transactions.



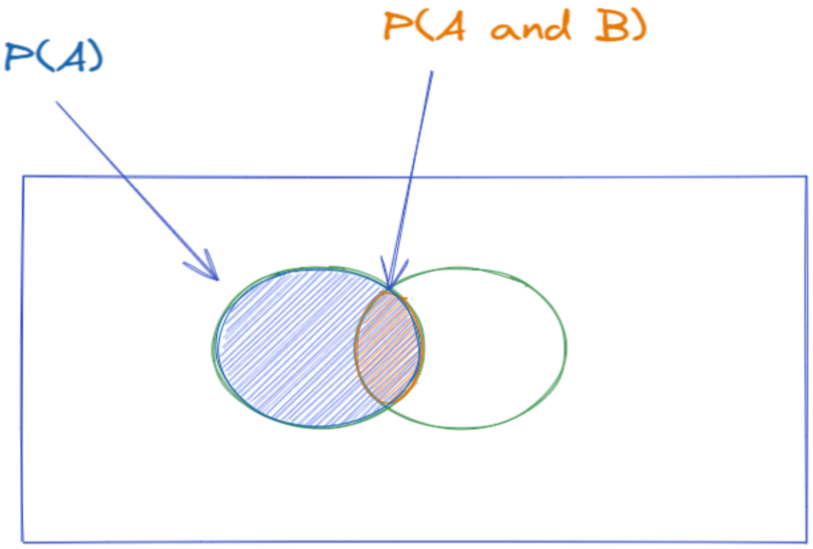
Venn Diagram for Support

The rules that merit further consideration for analysis using the value of support. For instance, if there were 10,000 transactions total, one might only wish to take into account the item sets that appeared at least 50 times, or 0.005 times. Since the lack sufficient knowledge of the link between an item set's items if it has a very low support, no inferences can be made from this rule.

2. Confidence

Given that the cart already possesses the antecedents, this measurement determines the likelihood of the consequence occurring on the cart. That is to say, how many of the transactions that contained, let's say, "Captain Crunch" also contained "Milk"? We can infer from common knowledge that the rule "Captain Crunch" and "Milk" should have a high degree of confidence. Confidence is defined technically as the conditional probability of the consequent occurring given the antecedent.

Confidence ({A}->{B}) = Transactions containing both A and B / Transactions containing A



Confidence - ratio of orange (overlap) area with respect to blue area

Venn Diagram for Confidence

Before continuing, let's think about a few additional instances. What do you think the confidence for "Butter" and "Bread" would be? What percentage of sales that included butter also included bread? Very high, or a value near 1? That' accurate. What about "milk" and "yogurt"? Once more feeling euphoric. Brush versus milk? Not certain? Due to the frequent occurrence of "Milk" and the fact that it will be included in every other transaction, confidence in this rule will likewise be strong.

Looks like a high value for confidence. However, a gut feeling that there is little correlation between these two items, and this high confidence number is deceptive. In order to overcome this obstacle, lift is introduced.

3. Lift

The conditional probability of occurrence of B given A, lift the restrictions for the support (frequency) of the consequent. This measure's name, lift, describes it in very precise terms. Consider it as the lift that having A on the cart gives our confidence. Rephrased, lift is the increase in likelihood of having B on the cart knowing that A is present over the likelihood of doing so without knowing that A is present.

Mathematically,
Lift({A}->{B}) = (Transactions containing both A and B) (Transactions containing A) / Fraction of Transactions containing B.

In probability terms lift is the ratio of confidence of A and Support B
P (A and B) / P(A)*P(B).

6. RESULTS OBTAINED

After executing the algorithm, it is discovered that "Roses Regency Teacup and Saucer" and "Green Regency Teacup and Saucer" have the greatest "lift" values, and as a result, the highest association of any two products, based on the outcomes of applying association rules. With a total support of 0.0309, this indicates that in 3.09% of all transactions, both items were bought simultaneously.

Below are the minimum and maximum results obtained for support, lift and confidence:

```
Minimum support is 0.00510
Maximum support is 1.00000
Minimum lift is 0.99942
Maximum lift is 195.55288
```

```
Minimum Support is between 'frozenset({'Description_WHITE HANGING HEART T-LIGHT HOLDER'})' and 'frozense
t({'Quantity'})'
Maximum Support is between 'frozenset({'InvoiceDate'})' and 'frozenset({'Quantity'})'
Minimum Lift is between 'frozenset({'Country_Netherlands'})' and 'frozenset({'UnitPrice'})'
Maximum Lift is between 'frozenset({'Description_WHITE HANGING HEART T-LIGHT HOLDER'})' and 'frozenset
({'StockCode_85123A'})'
Minimum Confidence is between 'frozenset({'Quantity'})' and 'frozenset({'Country_United Kingdom', 'UnitP
rice'})'
Maximum Confidence is between 'frozenset({'InvoiceDate'})' and 'frozenset({'Quantity'})'
```

7. CONCLUSION

The results obtained in terms of Support, Confidence and Lift are limited to the dataset obtained from Kaggle. This technique and methodology can be applied across businesses and firms belonging to numerous industries and domains. Slight modifications to the method can help firms obtain incremental results in terms of sales, revenue etc. and overall improve the whole funnel.

With Market Basket Analysis, there is a chance that further, unanticipated patterns and affinities will be found. By doing this, a project like Market Basket Analysis will acquire details about the business that go beyond answering the question at hand. Marketing managers can improve marketing operations such as product recommendation, product placement, cross-selling, up-selling, and customer retention by using product segmentation based on consumer purchasing patterns and network role.

8. TEAM MEMBER CONTRIBUTION

Member Name	Tasks Done
Atharva Puranik	Ideation of the Topic; Understanding significance of the Topic; Defining Problem Statement; Carried an end-to-end Project Regulation; Poster Making; Report Writing
Sayali Satish Dhavale	Define project, report, and research scope; Set overall objective and Goals; Suggest actionable basis the methodology presented; Poster Design and Ideation; Report Writing
Sujit Kandala	Background of the Topic – Market Basket Analysis, Preprocessing and its results, Report Writing
Sanjana Rajesh	Description; Experiments and Results of the algorithm; Report Writing
Sanjay Sangaraju	Implementation Details on Kaggle; Methods of Implementation; Report Writing

9. REFERENCES

Kafkas, K., Perdahçı, Z. N., & Aydın, M. N. (2021). Discovering customer purchase patterns in product communities: An empirical study on co-purchase behavior in an online marketplace. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(7), 2965–2980. https://doi.org/10.3390/jtaer16070162

Kawale, N. M., & Dahima, D. S. (2018). Market basket analysis using Apriori algorithm in R language. *International Journal of Trend in Scientific Research and Development*, *Volume-2*(Issue-4), 2628–2633. https://doi.org/10.31142/ijtsrd15677

Ostrowski. (2018, May 30). *Market basket analysis - exploring e-commerce data*. Kaggle. https://www.kaggle.com/code/ostrowski/market-basket-analysis-exploring-e-commerce-data/notebook

Provost, F. (2013). In T. Fawcett (Ed.), *Data Science for Business* (pp. 291–311). Essay, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Silvers, F. (2011). Data Warehouse Roi. *Data Warehouse Designs*, 1–12. https://doi.org/10.1201/b11692-1