SVKM'S NMIMS Deemed to be UNIVERSITY | NILKAMAL SCHOOL OF MATHEMATICS, APPLIED STATISTICS & ANALYTICS

UNSUPERVISED MACHINE LEARNING PROJECT

DOCUMENTATION

Implemented by

Charul Sankhe – A011

And

Anushka Avinash Kavli – A003

On the topic

CUSTOMER PURCHASING PATTERN USING MARKET BASKET ANALYSIS

INTRODUCTION:

Customer purchasing patterns refer to the recurring behaviors and tendencies exhibited by consumers when making purchases. These patterns encompass various aspects, including the types of products customers buy, the frequency of purchases, and the relationships between different items purchased. By analyzing transaction data through techniques like **Market Basket Analysis**, businesses can uncover valuable insights into these patterns. They discover that certain products are frequently bought together, indicating strong associations or dependencies. Understanding these patterns enables businesses to tailor their marketing strategies, optimize product offerings, and enhance the overall customer experience. Additionally, insights from purchasing patterns can inform inventory management decisions, promotional campaigns, and even store layout designs, ultimately driving sales and fostering long-term customer loyalty.

PROBLEM STATEMENT:

"Investigate and analyze customer purchasing behavior within the project through a market basket analysis to identify common product combinations. The aim is to gain insights that will inform sales strategies and improve overall customer satisfaction. The analysis will utilize a dataset comprising 38,765 transactions across 3 columns, seeking to understand patterns and relationships among purchased items."

ASSUMPTIONS:

Market basket analysis relies on several assumptions to identify customer purchasing patterns effectively:

1. Association Rule Mining Assumption: Market basket analysis assumes that purchasing behaviors can be represented as association rules, where certain products are frequently purchased together.

2. Frequent Itemset Assumption: It assumes that there are frequent itemsets within the dataset, meaning combinations of products that occur frequently enough to be considered meaningful.

3. Support and Confidence Metrics: The analysis assumes that support and confidence metrics accurately measure the significance and strength of associations between products.

4. Independence Assumption: It assumes that items are independent of each other, meaning that the purchase of one item does not significantly influence the purchase of another.

6. Complete Transaction Data: It assumes that the dataset contains complete and accurate transaction data, with each transaction representing a single purchase by a customer.

7. Homogeneous Customer Behavior: The analysis assumes that customers exhibit homogeneous behavior within the dataset, meaning that similar purchasing patterns are present across different customer segments.

8. Homogeneous Product Representation: It assumes that products are consistently represented within the dataset, with each product identifiable and distinguishable.

By acknowledging and working within these assumptions, market basket analysis can effectively uncover meaningful insights into customer purchasing behavior and inform business strategies.

MARKET BASKET ANALYSIS:

Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. It involves analyzing large data sets, such as purchase history, to reveal product groupings and products that are likely to be purchased together.

In market basket analysis, association rules are used to predict the likelihood of products being purchased together wherein they count the frequency of items that occur together, seeking to find associations that occur far more often than expected. Also, the Apriori algorithm is used to identify frequent items in the database, and then evaluate their frequency as the datasets are expanded to larger sizes.

DATASET LINK:

https://www.kaggle.com/datasets/heeraldedhia/groceries-dataset

DESCRIPTION:

The dataset contains 38,765 rows detailing purchase orders made by individuals at grocery stores. These records include columns for Member Number, Date, and Item Description.

Member Number - It is the unique ID for each customer.

Date – It includes the purchase date of the product purchased by the customers.

Item Description – It describes the product that has been purchased.

Data Analysis Tool: Python (using libraries like Pandas, etc.)

Data Visualization Tool: Matplotlib, Seaborn

Algorithm Applied: Market Basket Analysis

STEPS FOLLOWED:

Step 1: Importing the Libraries

Essential libraries such as `pandas`, `numpy` were used for data analysis, `seaborn` was used for visualization, and `apriori`, `association_rules` were used for association rule mining.

Step 2: Reading the dataset

By utilizing `data.info()`, one gains insights into the DataFrame's structure, including data types and non-null counts, while `data.describe()` offers a comprehensive overview of descriptive statistics such as mean, standard deviation, and quartiles for numerical columns.

Step 3: Pre-Processing the data

We transformed the date column into a datetime format and converted the member_number column into a string type to facilitate subsequent analysis. Subsequently, we conducted a null value check, revealing an absence of any null values within the dataset. After identifying 38,598 duplicate entries in the ItemDescription column, we proceeded to ascertain that there are approximately 167 unique items within the same column.

Step 4: Analysis and Visualization

Initially, we calculated the total number of items purchased by each member and then visualized the top 10 members with the highest purchase frequency. Additionally, we tallied the top 10 most frequently purchased items and presented them visually.

Step 5: Market Basket Analysis

In this process, we aggregated the items purchased by each customer daily, labeling this column as "unique transaction".

We utilized cross-tabulation on "uniqueTransaction" and "itemDescription" to generate a table named as “basket” illustrating the frequency of unique transactions associated with various item descriptions. Each cell in the table denotes how often a particular item description occurs within a unique transaction, facilitating the analysis of their relationship.

Step 6: Using Apriori Algorithm

Subsequently, we utilized the Apriori algorithm to identify frequently co-purchased items. Initially, we transformed the frequency-based basket table into a binary format to mitigate warnings, storing it as a variable named "apriori_data". Then, we applied the Apriori algorithm with a minimum support threshold of 0.01 to generate frequent itemsets.

Step 7: Applying Association rules Algorithm

Following that, we applied association rules to the frequent itemsets, using the lift metric with a minimum threshold of 1. The output, stored in a variable named "rules", comprises a table containing antecedents, consequents, antecedent support, consequent support, support, lift, confidence, leverage, conviction, and Zhang's metric.

Step 8: Making Recommendations

Moreover, we proceeded to formulate recommendations and constructed a table named "sub_list", encompassing the columns from the previously mentioned "rules" table.

Step 9: Evaluating Zhang Metric

To visualize the Zhang metric, we generated a pivot table with antecedents as the index, consequents as the columns, and the Zhang metric as the values. Subsequently, we constructed a heatmap to represent the Zhang metric values.

Step 10: Analysis by creating rules based on condition

In this context, we applied filters requiring the lift to exceed 1.18 and the confidence to surpass 0.1. Additionally, we imposed a condition where the Zhang metric value must be higher than 0.2.

PERFORMANCE METRICS:

1. Support
   Measures how frequently an itemset appears in the transaction dataset. It helps identify the most common itemset or item.
   Formula : $\text{Support}(X) = \frac{\text{Number of transaction containing X}}{\text{Total number of transactions}}$
   Summarizing the concept of 'Support' from our above dataset: Here, we observe that rolls/buns is the most common itemset being purchased. And it's support is 0.011964

2. Confidence
   Indicates the likelihood of item Y being purchased when item X is purchased. It's the measure of the predictive power or certainty of the association rule.
   Formula : $\text{Confidence } (X \rightarrow Y) = \frac{\text{Support } (X \cup Y)}{\text{Support}(X)}$

Summarizing the concept of ‘Confidence’ from our above dataset:
Here, we observe that if we buy sausages then there is 17.43% of buying whole milk.

3. Lift
Measures how much more often items X and Y occur together than if they were statistically independent. A lift value greater than 1 implies a positive association between X and Y.
Formula: $\text{Lift}(X\rightarrow Y) = \frac{\text{Confidence}(X \rightarrow Y)}{\text{Support}(Y)}$
Summarizing the concept of ‘Lift’ from our above dataset:
Here, we observe that sausage and whole milk are bought together more often.

INFERENCES:

1. A moderate association exists between 'Other Vegetables' and 'Rolls/Buns' (Zhang's metric value: 0.1324).
2. There's a moderate association between 'Soda' and 'Whole Milk' (Zhang's metric value: 0.028).
3. There's a relatively strong association between 'Whole Milk' and 'Sausage' (Zhang's metric value: 0.2381).
4. A moderate association exists between 'Whole Milk' and 'Yogurt' (Zhang's metric value: 0.179).
5. Empty cells (containing 0) indicate no association or very weak association between those particular pairs of items based on Zhang's metric.
6. The diagonal line represents the same item compared to itself, so the value is typically 1 (indicating a perfect association with itself).