**Market Basket Analysis Report**

**Project Overview**

This project performs Market Basket Analysis on an e-commerce transaction dataset. Using the Apriori algorithm, frequent itemsets and association rules are identified to uncover patterns in purchasing behaviors, helping understand which items are commonly bought together.

**Dataset Overview**

The dataset used for this project contains transactional information from an e-commerce platform. Key columns include:

* InvoiceNo: Transaction ID.
* StockCode: Product code.
* Description: Product description.
* Quantity: Number of items purchased in each transaction.
* InvoiceDate: Date and time of the transaction.
* UnitPrice: Price per unit of each product.
* CustomerID: Unique customer identifier.
* Country: Country of the customer.

**Dataset Details**

* Dataset Size: 541,909 rows.
* Date Range: Transactions cover the period from 01/12/2010 to 09/12/2011.
* Missing Values: Columns such as CustomerID and Description had missing values that were addressed during preprocessing.

**Data Preprocessing**

**1. Handling Missing Values**

* Rows with missing CustomerID values were removed, as they were essential for transaction grouping.
* The dataset was further examined for any anomalies, such as negative quantities, which were excluded.

**2. Data Transformation**

* The InvoiceNo and StockCode columns were converted to string format to maintain consistency.
* A transaction matrix was created, where each transaction (represented by InvoiceNo) was encoded to indicate whether an item (represented by Description) was purchased (1) or not (0).

**Analysis Methodology**

**1. Transaction Matrix Creation**

The dataset was transformed into a binary format using a transaction matrix where:

* Each row represents a transaction (InvoiceNo).
* Each column represents an item (Description).
* Each cell contains a binary value, with 1 indicating an item’s presence in the transaction.

**2. Applying the Apriori Algorithm**

The Apriori algorithm was applied to the transaction matrix to discover frequent itemsets with a minimum support threshold of 0.05. Itemsets with support above this threshold indicate products that are often purchased together.

**3. Generating Association Rules**

From the frequent itemsets, association rules were generated using the lift metric, with a minimum threshold of 1. These rules help identify items that, when purchased together, provide valuable cross-selling opportunities.

**Model Results**

**1. Frequent Itemsets**

The Apriori algorithm identified several frequent itemsets with support values greater than 0.05, indicating common combinations of items. The frequent itemsets with the highest support values are shown below:

|  |  |
| --- | --- |
| Itemset | Support |
| [Item A, Item B] | 0.08 |
| [Item C, Item D, Item E] | 0.07 |
| [Item F, Item G] | 0.06 |

*Note: Example items used for illustration.*

**2. Association Rules**

From the frequent itemsets, association rules were generated to indicate which items are commonly purchased together. The top association rules are summarized as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Antecedent | Consequent | Support | Confidence | Lift |
| [Item A] | [Item B] | 0.08 | 0.9 | 3.2 |
| [Item C, Item D] | [Item E] | 0.07 | 0.85 | 2.9 |
| [Item F] | [Item G] | 0.06 | 0.82 | 2.7 |

*Note: Example items used for illustration.*

**3. Visualization of Results**

* Top 10 Most Frequently Purchased Products: A bar chart displayed the top 10 items based on the number of transactions, helping to visualize the most popular items.
* Support Distribution: A histogram showed the support distribution for items across transactions, highlighting commonly purchased items.

**Insights Gained**

1. Frequent Item Combinations: Several items frequently appeared together in transactions, suggesting opportunities for product bundling or cross-selling.
2. Strong Association Rules: The lift values in generated rules indicate items that customers commonly buy together, providing insights into customer purchasing behavior.
3. Support Distribution: A large number of items have low support, indicating that a small subset of items makes up the majority of transactions.

**Limitations**

1. Assumption of Independence: The Apriori algorithm assumes item independence, which may not always hold true.
2. Sparse Data: The dataset is sparse due to the high number of unique items, affecting the accuracy of some associations.
3. Threshold Sensitivity: Support and lift thresholds significantly impact the number of itemsets and rules generated, requiring careful adjustment.

**Conclusion**

This Market Basket Analysis provided insights into customer purchasing patterns using the Apriori algorithm. By identifying frequent itemsets and association rules, the analysis highlights opportunities for cross-selling and inventory optimization. For further improvement, techniques such as clustering or alternative algorithms could be explored to gain additional insights into purchasing behavior.