PHASE – 5 MARKET BASKET INSIGHTS

Problem Statement: The problem statement for Market Basket Analysis (MBA) typically revolves around understanding the relationships between products purchased by customers in a retail environment. It seeks to answer questions such as:

- What products are often purchased together?
- Are there any patterns or associations between different items in the shopping carts?
- How can this information be used to improve sales and marketing strategies, optimize store layouts, or enhance the customer shopping experience?

Design Thinking Process for Market Basket Analysis:

Design Thinking is a human-centered problem-solving approach that can be applied to Market Basket Analysis:

1. **Empathize**: Understand the needs and pain points of both customers and the business. Gather data on customer behavior and shopping habits, and identify key challenges in optimizing sales and customer satisfaction.
2. **Define**: Clearly define the problem statement and objectives of the analysis. Determine what specific insights or actions are needed. For instance, you may want to increase cross-selling, reduce out-of-stock situations, or improve product placement.
3. **Ideate**: Brainstorm potential solutions or approaches to tackle the problem. Consider different data sources, algorithms, and tools for MBA. Engage cross-functional teams to contribute ideas and insights.
4. **Prototype**: Develop a preliminary MBA model or conduct an initial analysis to test your ideas. Use a small subset of data to create a proof of concept. This helps in refining your approach before full-scale implementation.
5. **Test**: Assess the prototype's performance, and gather feedback from stakeholders. Determine if the model is generating valuable insights and if it aligns with the defined objectives.
6. **Implement**: Once the MBA model has proven its value, deploy it on a larger scale. This may involve integrating it into the point-of-sale system, e-commerce platforms, or other relevant touchpoints in the retail environment.
7. **Measure and Learn**: Continuously monitor the performance of the MBA system and gather real-world data. Learn from customer behavior and adapt the system as needed to improve results.

Phases of Development in Market Basket Analysis:

1. **Data Collection**: Gather transaction data that includes details of products purchased by customers. This data should be clean, complete, and well-structured.
2. **Data Preprocessing**: Clean and preprocess the data, including tasks such as removing duplicates, handling missing values, and encoding categorical variables.
3. **Association Rule Mining**: Use algorithms like Apriori or FP-Growth to identify associations between products. This involves finding frequent itemsets and generating association rules that reveal which items are often bought together.
4. **Rule Evaluation**: Assess the generated rules based on metrics like support, confidence, and lift. These metrics help determine the strength and relevance of each association rule.
5. **Visualization and Interpretation**: Visualize the results to make them more understandable for non-technical stakeholders. This can include generating graphs, heatmaps, or tables to display product associations.
6. **Implementation**: Integrate the discovered associations into business processes. For example, update the point-of-sale system to suggest related products during checkout or reconfigure store layouts based on popular item groupings.
7. **Monitoring and Maintenance**: Continuously monitor the MBA system's performance and the changing preferences of customers. Update the analysis periodically and adapt to evolving customer behaviors and business goals.

Dataset Used:

The dataset is stored in the file Assignment-1_Data.xlsx located at /kaggle/input/marketbasket-analysis/. It contains information related to market transactions.

In[1]:

```
import numpy as np # linear algebra

import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

import os

for dirname, _, filenames in os.walk('/kaggle/input'):

    for filename in filenames:

        print(os.path.join(dirname, filename))
```

Transactional Data: The dataset is a collection of customer transactions, where each transaction is represented as a row and includes information like:

- Transaction ID: A unique identifier for each transaction.
- Date and Time: Timestamp of when the transaction occurred.
- Customer ID: A unique identifier for each customer.
- Items Purchased: A list of items or products bought in that transaction.

Data Preprocessing Steps:

Data preprocessing is a critical step in preparing the dataset for Market Basket Analysis:

1. **Data Cleaning**:

- Remove duplicate transactions to ensure data accuracy.
- Handle missing values if any (e.g., by removing transactions with missing items or using imputation techniques).

2. **Data Transformation**:
    - Convert the data into a suitable format for MBA algorithms. This often involves representing the items purchased as binary variables (1 if the item is in the basket, 0 if not) or using one-hot encoding for item names.
    - Aggregate data by customer if needed to understand customer behavior.
3. **Outlier Handling**:
    - Identify and handle outliers that might affect the analysis. For example, you might remove transactions with abnormally large numbers of items.
4. **Data Reduction**:
    - Reduce the dataset size if necessary. Large datasets can be computationally expensive, so you might consider working with a sample of the data for initial analysis.

Feature Extraction Techniques:

In MBA, the primary focus is on the relationships between items in shopping baskets. Feature extraction involves generating features that are relevant for discovering these relationships:

1. **Frequent Itemsets**:
    - Use Apriori or FP-Growth algorithms to identify frequent itemsets. These are combinations of items that appear together frequently in transactions. The size of the itemsets (number of items in a set) can vary.
2. **Association Rules**:
    - Generate association rules that describe the relationships between items. These rules consist of an antecedent (items in the basket) and a consequent (items that are likely to be bought if the antecedent is present).
    - Common metrics to assess the rules include support, confidence, and lift.
3. **Support**:
    - Support measures the frequency of occurrence of an itemset or rule in the dataset. It helps identify how popular a particular itemset is.
4. **Confidence**:
    - Confidence measures the likelihood that the consequent will be bought given the antecedent is in the basket. It quantifies the strength of the association.
5. **Lift**:
    - Lift measures how much more likely the consequent is to be bought when the antecedent is present compared to its independent purchase probability.
6. **Visualization**:

- Create visual representations of the extracted features, such as heatmaps, network graphs, or tables, to make the results more interpretable for business stakeholders.

Choice of Machine Learning Algorithm:

1. **Association Rule Mining Algorithms**: The most common algorithms for MBA are association rule mining algorithms, such as Apriori and FP-Growth. These algorithms are specifically designed to discover frequent itemsets and generate association rules. The choice between them depends on the dataset size and specific requirements.

    - **Apriori**: It's a widely used algorithm that generates frequent itemsets and association rules by iteratively pruning infrequent itemsets. It's effective for smaller datasets.

```
from mlxtend.frequent_patterns import apriori
rom mlxtend.frequent_patterns import association_rules
import pandas as pd
 # Sample transactional data
data = {
   'Transaction': [1, 2, 3, 4, 5],
   'Items': [['A', 'B', 'C'],
            ['A', 'C'],
            ['B', 'D'],
            ['A', 'B', 'C', 'D'],
            ['A', 'D']]
}
 df = pd.DataFrame(data)
 # Perform one-hot encoding
oht = df['Items'].str.join('|').str.get_dummies('|')
 # Apply Apriori algorithm to find frequent itemsets
min_support = 0.5  # Adjust this as needed
frequent_itemsets = apriori(oht, min_support=min_support, use_colnames=True)
 # Find association rules
min_confidence = 0.7  # Adjust this as needed
rules = association_rules(frequent_itemsets, metric='confidence',
min_threshold=min_confidence)
 # Display frequent itemsets and association rules
print("Frequent Itemsets:")
print(frequent_itemsets)
```

```
print("\nAssociation Rules:")
print(rules)
```

- **FP-Growth**: This algorithm constructs a compact data structure (the FP-tree) to efficiently discover frequent itemsets and association rules. It can be more efficient for large datasets compared to Apriori.

```
from mlxtend.frequent_patterns import fpgrowth
from mlxtend.frequent_patterns import association_rules
import pandas as pd
 # Sample transactional data
data = {
   'Transaction': [1, 2, 3, 4, 5],
   'Items': [['A', 'B', 'C'],
            ['A', 'C'],
            ['B', 'D'],
            ['A', 'B', 'C', 'D'],
            ['A', 'D']]
  }
 df = pd.DataFrame(data)
# Perform one-hot encoding
oht = df['Items'].str.join('|').str.get_dummies
min_support = 0.5  # Adjust this as needed
frequent_itemsets = fpgrowth(oht, min_support=min_support, use_colnames=True)

# Find association rules
min_confidence = 0.7  # Adjust this as needed
rules = association_rules(frequent_itemsets,
metric='confidence',min_threshold=min_confidence
# Display frequent itemsets and association rules
print("Frequent Itemsets:")
print(frequent_itemsets)

print("\nAssociation Rules:")
print(rules)
```

2. **Deep Learning**: For more complex and extensive datasets, deep learning techniques, such as neural networks, can be applied to capture intricate patterns in market basket data. However, deep learning is typically not the first choice for MBA due to the simplicity and interpretability of association rule mining.

Model Training:

- Model training in MBA primarily involves setting the algorithm's parameters (e.g., minimum support, minimum confidence) and running the chosen algorithm on the preprocessed dataset.
- The model should be trained using historical transactional data to discover frequent itemsets and association rules.
- Depending on the chosen algorithm, you may need to fine-tune parameters to achieve the desired level of detail and comprehensibility in the rules generated.

Evaluation Metrics:

The evaluation of Market Basket Analysis models is essential to ensure that the generated association rules are valuable and meaningful for business decision-making. Common evaluation metrics include:

1. **Support**: It measures the percentage of transactions in which the itemset appears. High support indicates that the itemset is popular among customers.
2. **Confidence**: Confidence measures the likelihood that the consequent item(s) will be bought when the antecedent item(s) are present in the basket. High confidence indicates a strong association between items.
3. **Lift**: Lift measures how much more likely the consequent is to be bought when the antecedent is present compared to its independent purchase probability. A lift greater than 1 indicates a positive association.
4. **Leverage**: Leverage measures the difference between the observed support of the itemset and what would be expected if the items were statistically independent.
5. **Interest**: Interest measures the interestingness of a rule by comparing its actual support with what would be expected if the items were independent.