PERFORMINTG ASSOCIATION ANALYSIS GREATING INSIGHTS

Phase 4

BASIC INTRODUCTION

- **Introduction**
- Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions. To put it another way, it allows retailers to identify relationships between the items that people buy.
- Association Rules are widely used to analyze retail basket or transaction data, and are intended to identify strong rules discovered in transaction data using measures of interestingness, based on the concept of strong rules.
- **An example of Association Rules**

ASSOCIATION RULE ANALYSIS

- **What Is Association Rule Analysis?**
- Association rule analysis is a **data mining technique** used to discover relationships between items or events in large datasets. It identifies patterns or co-occurrences that **frequently appear** together in a transactional database.
- Association rule analysis is commonly used for **market basket analysis**, product recommendation, fraud detection, and other applications in various domains.
- In other words, it helps to find the association between different events or items in a dataset

IMPORTANT OD ASSOCIATION ANALYSIS

- **Importance of Association Rule Analysis In Data Mining**
- Association rule analysis plays a vital role in data mining by providing insights into complex data relationships that would be difficult to identify manually. It is an important tool for businesses to understand **customer behaviour,** preferences, and trends.
- For example, retail businesses use association rule analysis to determine which **products are frequently purchased together** and to improve product placement and promotion strategies.
- Association rule analysis can also be used in medical research to identify potential drug interactions or adverse effects.

BASIC CONCEPTS

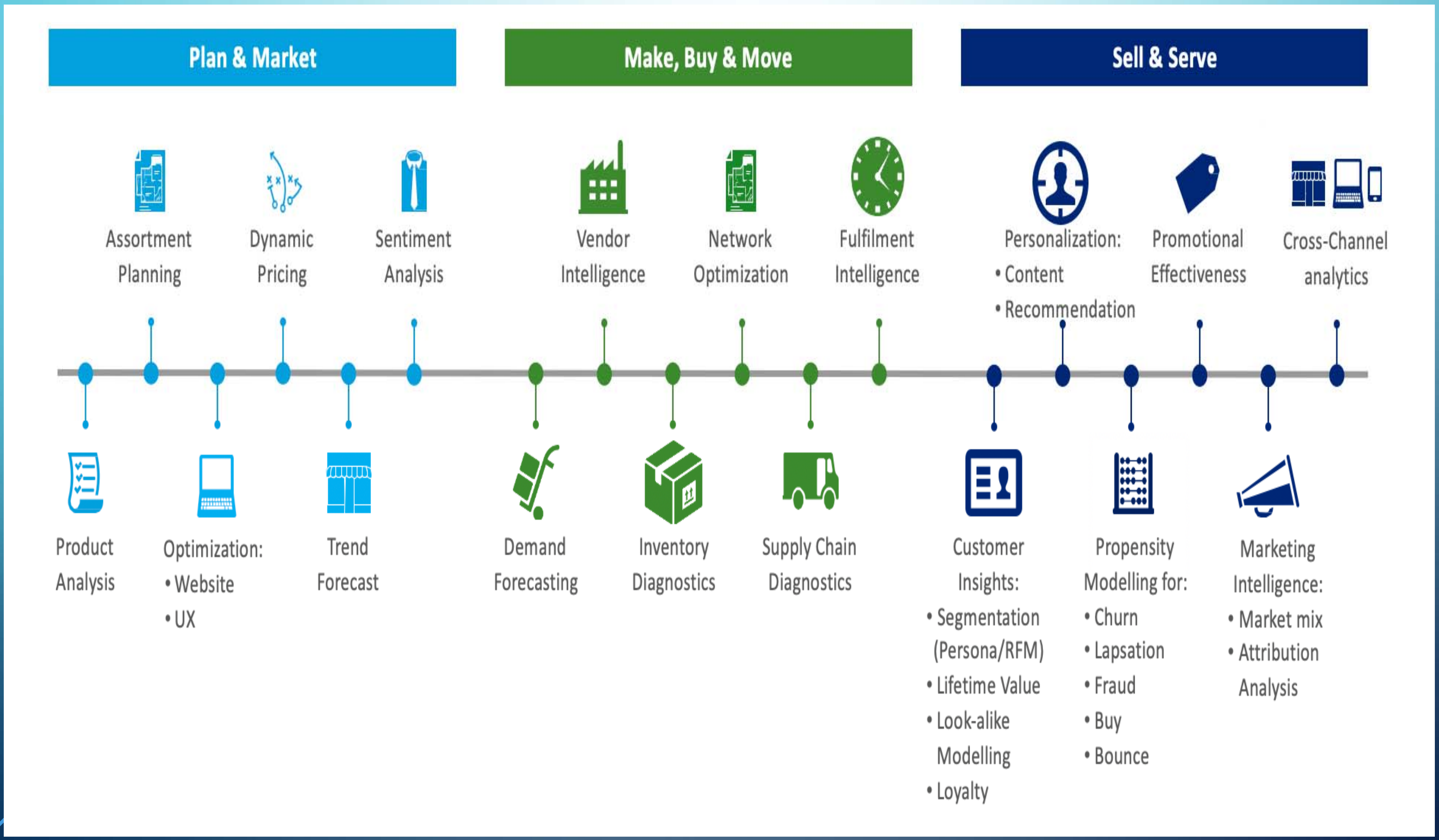
- **Basic Concepts and Terminology**
- The following terms are commonly used in association rule analysis:
- **Item:** An element or attribute of interest in the dataset
- **Transaction:** A collection of items that occur together
- **Support:** The frequency with which an item or itemset appears in the dataset.
  - **(Item A + Item B) / (Entire dataset)**
- **Confidence:** The likelihood that a rule is correct or true, given the occurrence of the antecedent and consequent in the dataset.
  - **(Item A + Item B)/ (Item A)**
- **Lift:** A measure of how often the antecedent and consequent occur together than expected by chance.

MEASUREMENT OF ASSOCIATION RULES

- **Measures For Evaluating Association Rules**
- Association rule analysis generates a large number of **potential rules,** and it is important to evaluate and select the most relevant rules.
- The following measures are commonly used to evaluate association rules:
- **Support:**
  - Rules with high support are more significant as they occur more frequently in the dataset
- **Confidence:**
  - Rules with high confidence are more reliable, as they have a higher probability of being true
- **Lift:**
  - Rules with high lift indicate a strong association between the antecedent and consequent, as they occur together more frequently than expected by chance

5 STEPS FOR ASSOCIATION ANALYSIS

BASIC THREE CONCEPTS OD ASSOCIATION ANALYSIS



APRIORI ALGORITHM

- **Apriori Algorithm**
- One of the most popular association rule mining algorithms is the Apriori algorithm. The Apriori algorithm is based on the concept of **frequent itemsets,** which are sets of items that occur together frequently in a dataset.
- The algorithm works by first **identifying all the frequent itemsets** in a dataset, and then generating **association rules** from those itemsets.
- These association rules can then be used to make predictions or **recommendations** based on the patterns and **relationships discovered**.

APRIORI ALGORITHM

```
from mlxtend.frequent_patterns import apriori
from mlxtend.preprocessing import TransactionEncoder
import pandas as pd

# define a sample dataset
dataset = [['apple', 'bread', 'milk'],
['apple', 'bread', 'diaper', 'milk'],
['apple', 'diaper', 'milk'],
['bread', 'diaper', 'milk']]

# create a transaction encoder
te = TransactionEncoder()
te_ary = te.fit(dataset).transform(dataset)
df = pd.DataFrame(te_ary, columns=te.columns_)

# apply the Apriori algorithm
frequent_itemsets = apriori(df, min_support=0.5, use_colnames=True)

# print the frequent itemsets
print(frequent_itemsets)
```

FP GROWTH ALGORITHM

- **FP-Growth Algorithm**
- In **large datasets**, FP-growth is a popular method for mining frequent item sets.
- It generates frequent itemsets efficiently without generating candidate itemsets using a tree-based data structure called the **FP-tree**. As a result, it is faster and more memory efficient than the **Apriori algorithm** when dealing with large datasets.
- First, the algorithm constructs an FP-tree from the input dataset, then **recursively** generates frequent itemsets from it.

FP GROWTH ALGORITHM

```
from mlxtend.frequent_patterns import fpmax
from mlxtend.preprocessing import TransactionEncoder
import pandas as pd

# Example dataset
dataset = [['beer', 'nuts', 'diapers'],
['beer', 'cheese', 'diapers'],
['beer', 'cheese', 'nuts'],
['cheese', 'nuts']]

# Convert dataset to one-hot encoded DataFrame
te = TransactionEncoder()
te_ary = te.fit(dataset).transform(dataset)
df = pd.DataFrame(te_ary, columns=te.columns_)

# Apply FP-max algorithm with min_support = 0.5
frequent_itemsets = fpmax(df, min_support=0.5, use_colnames=True)

# Print the frequent itemsets
print(frequent_itemsets)
```

ECLAT ALGORITHM

- **Eclat Algorithm**
- Equivalence Class Transformation, or Eclat is another popular algorithm for Association Rule Mining.
- Compared to Apriori, Eclat is designed to be more **efficient at mining frequent itemsets**. There are a few key differences between the Eclat algorithm and the Apriori algorithm.
- To mine the frequent itemsets Eclat uses a **depth-first search** strategy instead of candidate generation. Eclat is also designed to use less memory than the Apriori algorithm, which can be important when working with large datasets.

ECLAT ALGORITHM

```
import pandas as pd
from pyECLAT import ECLAT

transactions = [
['bread', 'milk', 'eggs'],
['bread', 'milk'],
['milk', 'eggs'],
['bread', 'butter'],
['butter', 'jam']
]

# convert the list to a Pandas DataFrame
df = pd.DataFrame(transactions)

# instantiate an ECLAT object with minimum support 0.4
eclat = ECLAT(df, 0.4)

# find frequent itemsets
frequent_itemsets = eclat.fit()

# print the frequent itemsets
```

ADVANCED TECHNIQUES

- **Advanced Techniques in Association Rule Analysis**
- While traditional association rules mining techniques, such as Apriori, FP-growth, and Eclat, are effective in discovering frequent itemsets and association rules, they are limited in terms of their ability to handle complex relationships and patterns in large and diverse datasets.

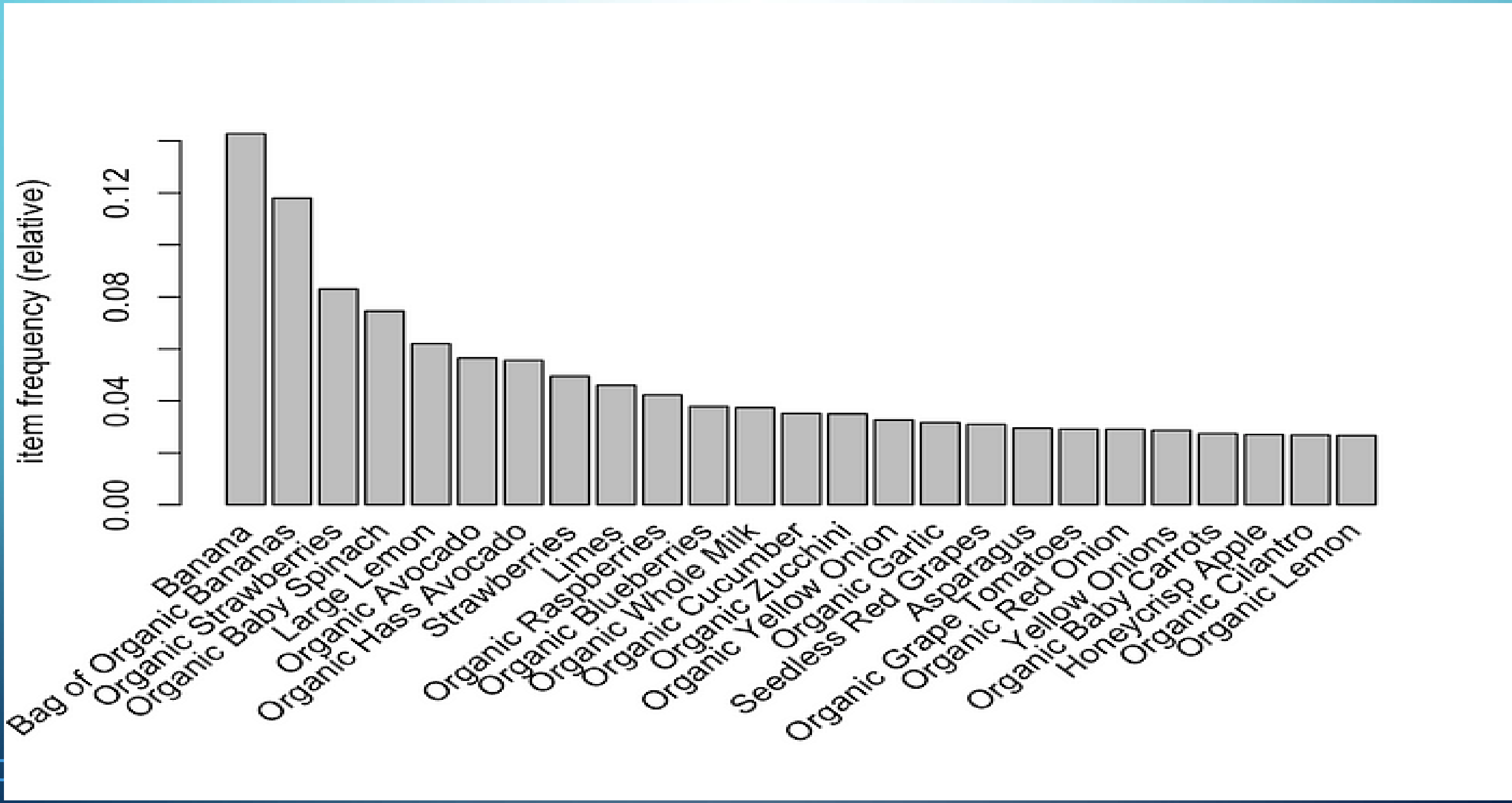
CONSTRAINT AND SEQUENTIAL MINING

- **Constraint Based Mining**
- One of the advanced techniques in association rule analysis is constraint-based mining.
- Constraint-based mining is a method of mining association rules that incorporates prior knowledge, domain constraints, and background knowledge into the mining process.
- This approach can improve the accuracy and relevance of the mined rules by reducing the search space and avoiding mining irrelevant or redundant rules.
- Constraint-based mining is particularly useful in domains with complex relationships and patterns, such as bioinformatics, where prior knowledge about the domain can be incorporated into the mining process.
- **Sequential Pattern Mining**
- Mining patterns in sequential data, such as time series data or online clickstreams, is known as **sequential pattern mining** .
- This method can aid in discovering patterns in data that occur in a specified order or with a temporal lag between them. Several applications exist for sequential pattern mining, such as anticipating **consumer behaviour** or finding abnormalities in **time-series** data.

OTHER INDUSTRIES PROCESS

- **Other Industries**
- Telecommunications, insurance, and e-commerce are some of the other areas that might profit from association rule analysis.
- Association rule analysis can be used in the **telecommunications** sector to find trends in call data to enhance network efficiency and improve customer service.
- It may be used in the insurance business to detect **risk variables** and develop more accurate risk models, allowing for more effective and efficient insurance policies.
- It may be used in e-commerce to improve product suggestions and generate focused marketing campaigns based on clients' purchase behaviour.

FLOW CHAT FOR ANALYSIS



VISUALIZE THE RULES

- **visualize the rules**
- The plot.graph function is used to visualize the rules that we have shortlisted based on their leverage values. It internally uses a package called igraph to create a graph representation of the rules:

PRESENTED BY P.SARAVANAN MRK IT