Association Rules

Name: ULLI VENKATA SAI KUMAR **Batch ID:** 04072024HYDAM
Topic: Association Rules

Problem Statement: -
Q1. Kitabi Duniya, a famous bookstore in India, was established before Independence, the growth of the company was incremental year by year, but due to the online selling of books and widespread Internet access, its annual growth started to collapse. As a Data Scientist, you must help this heritage bookstore gain its popularity back and increase the footfall of customers and provide ways to improve the business exponentially to an expected value at a 25% improvement of the current rate. Apply the pattern mining techniques (Association Rules Algorithm) to identify ways to improve sales. Explain the rules (patterns) identified, and visually represent the rules in graphs for a clear understanding of the solution.

1.) Data: Books.csv

	ChildBks	YouthBks	CookBks	DoItYBks	RefBks	ArtBks	GeogBks	ItalCook	ItalAtlas	ItalArt	Florence
1	0	1	0	1	0	0	1	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	0	1	0	1	0	0	0	0
5	0	0	1	0	0	0	1	0	0	0	0
6	1	0	0	0	0	1	0	0	0	0	1
7	0	1	0	0	0	0	0	0	0	0	0
8	0	1	0	0	1	0	0	0	0	0	0
9	1	0	0	1	0	0	0	0	0	0	0
10	1	1	1	0	0	0	1	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0
12	0	0	1	0	0	0	1	0	0	0	0
13	1	0	0	0	0	1	0	0	0	0	1

```
# Let's first load and inspect the contents of the uploaded file to understand its structure and data.
import pandas as pd
# Load the CSV file
book_data = pd.read_csv('book.csv')
# Display the first few rows of the dataset to understand its structure
book_data.head()
# Importing necessary libraries for association rule mining
from mlxtend.frequent_patterns import apriori, association_rules
# Applying the apriori algorithm to find frequent itemsets with a minimum support of 0.05 (5%)
frequent_itemsets = apriori(book_data, min_support=0.05, use_colnames=True)
# Generating association rules from the frequent itemsets with a minimum confidence threshold of 0.2
```

```
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.2)
# Display the top rules for insights
rules.head()
# We can implement a simple form of frequent itemset mining without the apriori module.
# To identify pairs of books that are frequently bought together, we will calculate the support for itemsets.
# Calculate the correlation matrix to see which book types are commonly bought together
correlation_matrix = book_data.corr()
# Let's visualize the correlation matrix using a heatmap for better understanding
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(10,8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Matrix of Book Purchases')
plt.show()
# To further explore, let's manually compute some pairwise itemset supports (support is the fraction of transactions containing the itemset)
# Create a function to calculate the support for a pair of book categories
def support_pair(book_data, book1, book2):
    return (book_data[book1] & book_data[book2]).mean()
# Example: Calculate support for a few pairs of highly correlated book categories
support_pairs = {
    ("ChildBks", "YouthBks"): support_pair(book_data, "ChildBks", "YouthBks"),
    ("GeogBks", "ChildBks"): support_pair(book_data, "GeogBks", "ChildBks"),
    ("DoItYBks", "YouthBks"): support_pair(book_data, "DoItYBks", "YouthBks"),
    ("CookBks", "ArtBks"): support_pair(book_data, "CookBks", "ArtBks"),
}
# Display the support values for these pairs
support_pairs
```

Questions to Ignite your Thinking process:

Q1. Which library/package is used for the Association rules algorithm?

- The mlxtend library is commonly used for implementing association rule mining in Python.

Q2. Which functions are used in the Association rules algorithm?
The main functions used are:

- apriori(): To find frequent itemsets.
- association_rules(): To generate association rules from the frequent itemsets.

Q3. What is the keyword used to import any package to the Python session's memory?

- The keyword is import.

Q4. What type of data is usually worked in Association rules?

- Transactional data, typically in the form of a binary matrix (indicating the presence or absence of items in transactions).

Q5. Association rules are also named as

- Market basket analysis.

Q6. What is the IF part called in an Association rule?

- The **antecedent**.

Q7. What is the THEN part called in an Association rule?

- The **consequent**.

Q8. In which sector is the Association rules algorithm mainly used?

- Retail and e-commerce, for tasks such as market basket analysis and recommendation systems.

Q9. What do slotting fees mean?

- Slotting fees refer to the fees that manufacturers pay to retailers to have their products placed on shelves in prime locations.

Q10. How is Support calculated in the Association rules algorithm?

- **Support** is calculated as the fraction of transactions that contain the itemset

$$Support(X) = \frac{Transactions\ containing\ X}{Total\ Transactions}$$

Q11. What is the drawback of Support in Association rules algorithm?

- It does not take into account how commonly items occur individually, which can result in overly frequent but uninteresting associations.

Q12. To remove the infrequent items from data which algorithm is used?

- The **Apriori algorithm** is used to eliminate infrequent items.

Q13. The drawback of Support is captured by

- **Confidence** and **Lift** measures.

Q14. How is confidence calculated in the Association rules algorithm?

- **Confidence** is the ratio of the number of transactions containing both the antecedent and consequent to the number of transactions containing only the antecedent:

$$Confidence(X \rightarrow Y) = \frac{Support(X\ U\ Y\ )}{Support(X)}$$

Q15. What is the drawback of Confidence in Association rules algorithm?

- Confidence can be misleading when the consequent is frequent on its own, as it doesn't account for the baseline probability of the consequent.

Q16. How is the Lift ratio calculated?

- **Lift** is calculated as the ratio of the observed support to the expected support under independence:

$$Lift(X \rightarrow Y) = \frac{Confidence(X \rightarrow\ Y\ )}{Support(Y)}$$

Q17. What is the threshold value of the Lift ratio of a rule to declare it as a good Association rule?

- A lift value greater than **1** indicates a good association rule, meaning the occurrence of the antecedent increases the likelihood of the consequent.