A yellow and blue logo

AI-generated content may be incorrect.

Lab 8

TEB3123: Machine Learning

Online Retail Market Basket Analysis

|  |  |  |
| --- | --- | --- |
| **Name** | **ID** | **Course** |
| Cheng Pin-Jie | 21000548 | Computer Science |

*Mr Abdul Muiz Fayyaz*

1.0 Data Understanding

The dataset is in excel format which is downloaded from:

<https://archive.ics.uci.edu/ml/machine-learning-databases/00352/Online%20Retail.xlsx>

The dataset is loaded by using:

# Import libraries

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from mlxtend.frequent\_patterns import apriori

from mlxtend.frequent\_patterns import association\_rules

# Load the dataset

df = pd.read\_excel("Online Retail.xlsx")

# Display the first few rows of the dataset

print(df.head())

A screenshot of a computer screen

AI-generated content may be incorrect.

2.0 Data Preparation

The dataset needs to be further process for market basket analysis. Thus, data cleaning, including removing extra spaces in product descriptions, dropping missing value row and filtering datasets.

# Remove spaces in the descriptions and drop rows with missing invoice numbers

df['Description'] = df['Description'].str.strip()

df.dropna(axis=0, subset=['InvoiceNo'], inplace=True)

# Filter out records for transactions in France

df = df[df['Country'] == 'France']

Before passing the data to the algorithm, encoding is needed to convert values into binary values.

def encode\_units(x):

    if x <= 0:

        return 0

    if x >= 1:

        return 1

A matrix is then created to represent ‘Trasaction vs Product”.

basket = (df.groupby(['InvoiceNo', 'Description'])['Quantity']

          .sum().unstack().reset\_index().fillna(0)

          .set\_index('InvoiceNo'))

basket\_sets = basket.applymap(encode\_units)

basket\_sets

A screenshot of a video game

AI-generated content may be incorrect.

3.0 Modeling

The data is now ready to implement association rules. First, *apriori* is used to find the frequent itemsets. Then, association rules are created in the format of *“If customer buys A, they also buy B”*

frequent\_itemsets = apriori(basket\_sets, min\_support=0.07, use\_colnames=True)

rules = association\_rules(frequent\_itemsets, metric="lift", min\_threshold=1)

A screenshot of a computer screen

AI-generated content may be incorrect.

The rules generated are then further filtered for:

* Life ≥ 6
* Confidence ≥ 0.8

It Is because we want rules that has a strong positive association (lift) and itemsets that appear at least 80% of the time (confidence).

filtered\_rules = rules[(rules['lift'] >= 6) & (rules['confidence'] >= 0.8)]

print(filtered\_rules)

A screen shot of a computer

AI-generated content may be incorrect.

Lastly, a scatterplot is visualized to view the rules:

sns.scatterplot(x='support', y='confidence', data=rules, size='lift', sizes=(10, 200))

plt.title('Association Rules - Support vs Confidence')

plt.xlabel('Support')

plt.ylabel('Confidence')

plt.show()

A chart with blue dots

AI-generated content may be incorrect.

4.0 Findings & Conclusion

Based on the *filtered\_rules*, I can spot that there is a strong relationship among red spotty tableware in France. For example:

|  |  |  |
| --- | --- | --- |
| Transaction | Lift | Confidence |
| Red spotty paper plates -- Red spotty paper cups | 8.2 | 0.96 |
| Red spotty paper cups -- Red spotty paper plates | 8.2 | 0.89 |
| Red spotty paper cups -- Red spotty paper plates -- Red retrospot paper napkins | 7.2 | 0.81 |

These are the top transaction in this online retail business located in France, which suggest that these items can be displayed together in online store layout. A “Party Package” can be created with promotion price to further boost the France market. Besides, business owner can maintain the stocks for all these items for future inventory management.

In conclusion, these association rules let us understand the customer behaviour and their preference when come into our online retail store.