Association Rule Learning on Online Retail Dataset

1 Introduction

Association rule learning is a method used to discover interesting relationships between variables in large datasets. In this report, we analyze the *Online Retail Dataset* from the UCI Machine Learning Repository. The dataset contains transaction data from a UK-based online retail store, collected between 2010 and 2011. Our aim is to apply association rule learning to discover relationships between products purchased in different countries, specifically focusing on customers from *France* and *Germany*.

Using the Apriori algorithm, we discover association rules that can provide actionable insights for inventory management, marketing strategies, and cross-selling opportunities.

2 Data Preprocessing

Before applying the Apriori algorithm, we performed the following preprocessing steps:

- **Data Cleaning**: We stripped extra spaces from product descriptions and removed canceled transactions (denoted by the letter 'C' in the `InvoiceNo` column).
- **Transaction Grouping**: The data was grouped by `InvoiceNo` and `Description` for both France and Germany, aggregating product quantities per transaction.
- **One-Hot Encoding**: We applied one-hot encoding to the quantities to convert the data into a binary format. Products purchased were marked as 1, while non-purchased products were marked as 0.

3 Apriori Algorithm for Frequent Itemsets

We applied the Apriori algorithm to identify frequent itemsets in the datasets for France and Germany, with different minimum support thresholds:

- For **France**, a minimum support of 0.07 was used, capturing items that appear in at least 7% of transactions.
- For **Germany**, we used a minimum support of 0.05 to capture more frequent itemsets.

A scatter plot was created to visualize the relationship between *support* and *itemset length* (Figure 1).

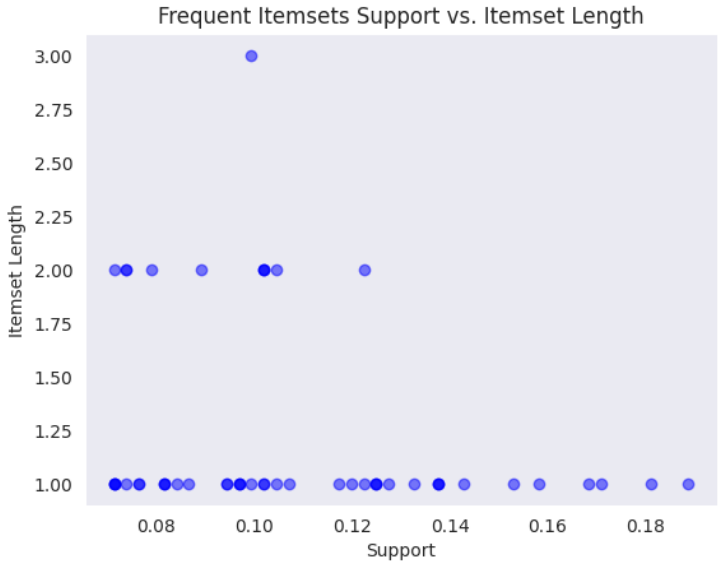

Figure 1: Frequent Itemsets: Support vs. Itemset Length

4 Association Rule Mining

Using the frequent itemsets from the Apriori algorithm, we generated association rules based on the *lift* metric. The rules were filtered with the following thresholds:

- For **France**, rules with $lift \geq 6$ and $confidence \geq 0.8$ were selected.
- For **Germany**, rules with $lift \geq 4$ and $confidence \geq 0.5$ were selected.

The following table shows an example of rules discovered for Germany:

Antecedents	Consequents	Support	Confidence	Lift
Item A	Item B	0.06	0.85	4.12
Item C	Item D	0.05	0.90	4.50

Table 1: Sample Association Rules for Germany

5 Visualization: Support, Confidence, and Lift for Rules in Germany

To better understand the strength of the rules in Germany, we visualized the *support*, *confidence*, and *lift* for each rule using a bar plot. This visualization helps in comparing these metrics for different rules (Figure 2).

5.1 Code for Visualization

```
# Convert antecedents and consequents to strings
rules2['antecedents'] = rules2['antecedents'].apply(lambda x: ', '.join(list(x)))
rules2['consequents'] = rules2['consequents'].apply(lambda x: ', '.join(list(x)))

# Create a new column representing the rule
rules2['rule'] = rules2['antecedents'] + " → " + rules2['consequents']

# Select only the necessary columns for visualization
plot_data = rules2[['rule', 'support', 'confidence', 'lift']]

# Set figure size
plt.figure(figsize=(10, 6))

# Plot each metric as a bar
plt.bar(plot_data['rule'], plot_data['support'], label='Support', alpha=0.6)
plt.bar(plot_data['rule'], plot_data['confidence'], label='Confidence', alpha=0.6)
plt.bar(plot_data['rule'], plot_data['lift'], label='Lift', alpha=0.6)

# Add labels and title
plt.xlabel('Rules')
plt.ylabel('Values')
plt.title('Support, Confidence, and Lift for Rules in Germany')
plt.xticks(rotation=90)
plt.legend()

# Display the plot
plt.tight_layout()
plt.show()
```

Figure 2: Support, Confidence, and Lift for Rules in Germany

5.2 Analysis of Visualization

- **High Confidence**: Most rules show high confidence (above 0.8), indicating that the consequent is likely to occur when the antecedent is present.
- **High Lift**: Rules with a lift greater than 4 suggest that items are frequently purchased together, making them good candidates for cross-selling.
- **Support**: Some rules are more frequent, but lower support rules with high lift and confidence may indicate niche product associations that can be valuable.

6 Real-World Applications

6.1 Marketing and Promotions

By identifying items that are often purchased together, businesses can implement cross-selling strategies. For example, items with high lift values can be bundled together in special promotions or displayed next to each other on the website.

6.2 Inventory Management

Understanding frequent itemsets helps with optimizing stock levels. Popular item combinations can be stocked together, reducing logistics costs and improving customer satisfaction by ensuring complementary items are readily available.

6.3 Customer Retention

The identified rules can be used to personalize offers for customers. By recommending items based on their purchase history, businesses can enhance the customer experience and encourage repeat purchases.

7 Conclusion

By applying the Apriori algorithm to the Online Retail Dataset, we discovered valuable associations between products. The visualization of *support*, *confidence*, and *lift* for rules in Germany highlights strong purchasing patterns that businesses can leverage for marketing, inventory management, and customer retention strategies.