Online Retail Datamining Project Proposal

redacted
redacted
redacted, redacted
redacted

Abstract

This report outlines a data mining project focused on the UCI Online Retail dataset. I aim to investigate product associations, customer spending distributions, and seasonal sales patterns to provide business insights.

1 Dataset Description

1.1 Overview and Source

The study utilizes the **Online Retail Dataset** from the UCI Machine Learning Repository. This dataset contains a large number of transactions from a UK-based online retail store.
Data Source: https://archive.ics.uci.edu/dataset/352/online+retail

1.2 Data Characteristics

- **Size:** 541,909 rows, 8 columns.
- **File Size:** 22.6 MB.
- **Representation:** A single row represents a single transaction.

1.3 Key Features and Schema

The primary attributes utilized for data-mining include timestamps, customer identifiers, and product codes as detailed in Table 1.

Table 1: Dataset Schema

Feature Name	Type	Description
InvoiceNo	String	6-digit unique ID. 'c' prefix indicates cancellation.
StockCode	String	5-digit unique product ID.
Description	String	Nominal product name.
Quantity	Integer	Units per transaction.
InvoiceDate	Date	Timestamp of transaction.
UnitPrice	Float	Product price per unit (GBP).
CustomerID	Integer	5-digit unique customer ID.
Country	String	Name of the customer's country.

1.4 Data Quality Issues

Some pre-processing is required due to:

- Missing `CustomerID` values for approximately 25% of records.
- Occasional negative `Quantity` and `UnitPrice` values (cancellations/adjustments).
- Some redundant entries and inconsistent product descriptions.

2 Discovery Questions

(1) **1:** What products are frequently purchased together?
- *Value:* This is a classic discovery question for retail datasets. Shops can use this information to push associated products together to drive sales.
- *Techniques:* Market Basket Analysis (Apriori)

(2) **2:** What is the distribution of spending across the customer base?
- *Value:* Understanding the customer base is also valuable for shops. If certain segments of the customer base dominate spending, then tailoring the shop to their preferences can increase revenue.
- *Techniques:* Customer Segmentation (K-Means Clustering)

(3) **3:** Which products have seasonal sales patterns?
- *Value:* Certain products have a seasonality to their sales patterns. If shops exploit this by pushing the products more aggressively or offering discounts, they can move more product. This also applies to products that are popular for a time, but fade away afterwards. This type of analysis is also extremely useful for inventory management, since the shop can predict when changes in demand will occur.
- *Techniques:* Time Series Analysis (Time-Series Clustering)

3 Planned Techniques

3.1 Analysis Pipeline

The project will follow a standard Knowledge Discovery in Databases (KDD) process as illustrated in Figure 1.

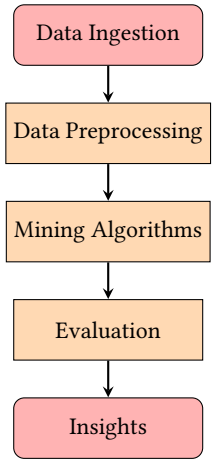

Figure 1: Planned Overall Analysis Pipeline

3.2 Mining Techniques

(1) **1:** Market Basket Analysis

- *Algorithm(s):* The Apriori algorithm will be used to perform a market basket analysis of the retail data. The data will be transformed into a matrix that groups purchased items invoice number and description. The algorithm will be executed on this matrix to generate association rules. This will require tuning of the Apriori "hyperparameters", namely support, confidence, and lift. Iteration will likely be required to output strong, insightful association rules.

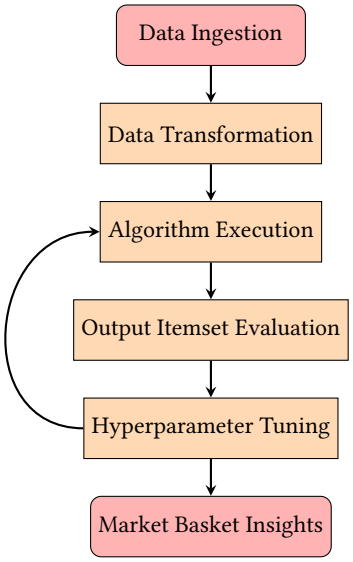

Figure 2: Apriori Data Processing Pipeline

(2) **2:** Customer Segmentation

- *Algorithm(s):* A K-Means clustering algorithm will be used to perform retail segmentation. To perform customer segmentation, the transactional data is first transformed into an RFM (Recency, Frequency, Monetary) matrix that aggregates behavior at a customer level. This data will then be scaled or normalized to ensure consistent clustering. The K-Means algorithm is then executed on the resulting feature space. The resulting clustering outputs will then be analyzed and clustering algorithm parameters tuned to produce useful customer segments that can provide data insights.

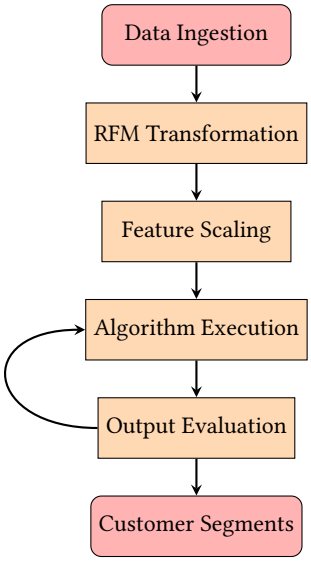

Figure 3: K-Means Clustering Analysis Pipeline

(3) **3:** Time Series Analysis/Seasonality Assessment

- *Algorithm(s):* To discover seasonal patterns, individual sales items will be aggregated into product categories based on description keywords. These categories will then be analyzed using Seasonal-Trend Decomposition (STL) to isolate their recurring cyclical components. By applying Time-Series Clustering with Dynamic Time Warping, clusters of seasonally associated products can be identified. This provides insights on the seasonality of various item categories and hidden seasonal associations.

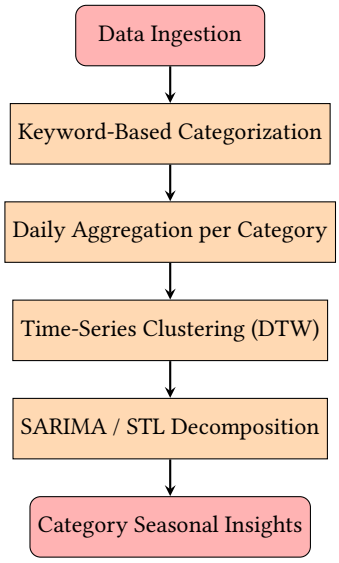

Figure 4: Categorical Seasonality Discovery Pipeline

4 Preliminary Timeline

- **Milestone 2 (M2):** Data preprocessing and Apriori algorithm implementation. Some limited evaluation of results.
- **Milestone 3 (M3):** Implementation and evaluation of other 2 algorithms.
- **Milestone 4 (M4):** Complete evaluation and final report.

4.1 Anticipated Challenges

This dataset is fairly large so there might be a large computational overhead when processing it, especially when performing the time-series decomposition. Furthermore, the multi-dimensional nature of the seasonality analysis make it potentially the most challenging mining technique applied here. There are also a very large number of stock codes (items) which will make Apriori basket analysis potentially just as long. However, this means there is plenty of data to extract patterns and insight from.