HW2: Association Rule Mining (FP-Growth & Apriori) and Time Series Forecasting

Supervisor: Dr. Fakhrahmad

Deadline:Monday, 1404/09/15

---

1. Introduction

Understanding customer purchasing behavior and predicting future sales are two critical components of modern retail analytics. This project combines two powerful data mining techniques to analyze retail transaction data: **Association Rule Mining** to discover product purchasing patterns, and **Time Series Analysis** to forecast future sales.

In the first part, association rule mining reveals which products are frequently pur chased together, enabling retailers to optimize product placement and create targeted promotions. In the second part, time series forecasting predicts future purchase amounts, helping businesses plan inventory and make informed strategic decisions.

Through systematic data preprocessing, exploratory analysis, and evaluation, this assignment demonstrates how data mining techniques can provide actionable insights for retail business operations.

2. Dataset Description

This project uses the **Online Retail Dataset** from the UCI Machine Learning Repository. The dataset contains all transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based online retail company that specializes in unique all-occasion gifts.

Dataset attributes:

- InvoiceNo: Invoice number (unique per transaction)
- StockCode: Product code
- Description: Product name/description
- Quantity: Number of each product per transaction
- InvoiceDate: Invoice date and time
- UnitPrice: Product price per unit
- CustomerID: Unique customer identifier
- Country: Customer’s country

The dataset contains approximately 540,000 transactions with around 4,000 unique products and serves customers from multiple countries.

Dataset download link: UCI Online Retail Dataset

Include proper citation of the UCI Machine Learning Repository in your report and provide detailed information about the dataset structure.

3. Data Preprocessing

Data preprocessing ensures data quality and prepares the dataset for both association rule mining and time series analysis.

3.1 Loading and Initial Cleaning

1. Load the dataset: Use Pandas to load the dataset (typically available as CSV or Excel format).

2. Examine data structure: Display the first few rows, check data types, and iden tify the dimensions of the dataset.

3. Handle missing values:

- Identify columns with missing values (especially Description and CustomerID)
- Remove rows with missing Description (critical for association rules)
- For CustomerID, you may keep or remove based on your analysis needs

4. Remove invalid records:

- Filter out transactions with negative or zero quantities
- Remove records with negative or zero prices

5. Data type conversion: Convert InvoiceDate to datetime format.

6. Feature engineering: Create TotalAmount by multiplying Quantity × UnitPrice.

3.2 Preprocessing for Association Rule Mining

Prepare the data in transaction format:

- Group items by InvoiceNo to create individual shopping baskets
- Each basket should contain a list of unique product descriptions
- Remove duplicate products within the same transaction
- Filter out products that appear less than 10 times

3.3 Preprocessing for Time Series Analysis

Prepare the data for temporal analysis:

- Aggregate purchase amounts by date (daily totals)
- Calculate daily total sales: sum of TotalAmount for each day
- Sort data based on time

Document all preprocessing steps in your report, showing statistics before and after cleaning (e.g., number of records removed, missing value percentages).

4. Exploratory Data Analysis (EDA)

Perform exploratory analysis to understand the dataset characteristics.

4.1 General Dataset Exploration

Create the following visualizations:

- **Summary statistics:** Descriptive statistics for numerical columns
- **Distribution plots:** Histograms for quantity and unit price
- **Top products:** Bar chart of top 10-15 most frequently purchased products
- **Geographic distribution:** Sales by country (bar chart)

4.2 EDA for Association Rules

- **Basket size distribution:** Histogram showing number of products per transaction
- **Item frequency:** Bar chart of most frequently purchased items

4.3 EDA for Time Series

- **Time series plot:** Line plot showing daily total sales over time
- **Trend analysis:** Identify overall trends in sales
- **Moving average:** Plot a 7-day moving average

Use Matplotlib and Seaborn for visualization. Include all plots in your report with clear labels and brief interpretations.

5. Part 1: Association Rule Mining

Apply association rule mining to discover product purchasing patterns.

5.1 Methodology

1. Transform data to transaction format: Convert the preprocessed data into a list of transactions where each transaction contains the products purchased together.

2. Apply the Apriori algorithm: Use the Apriori algorithm to discover frequent itemsets. Start with min support = 0.01 (1%) and adjust if needed.

3. Generate association rules: From the frequent itemsets, generate association rules showing relationships: ”If customers buy product A, they are likely to buy product B.”

Use min confidence = 0.3 (30%) as starting point.

4. Calculate evaluation metrics: For each rule A ⇒ B, compute:

- **Support:** $\text{support}(A \Rightarrow B) = P(A \cap B)$
- **Confidence:** $\text{confidence}(A \Rightarrow B) = \frac{P(A \cap B)}{P(A)}$
- **Lift:** $\text{lift}(A \Rightarrow B) = \frac{P(A \cap B)}{P(A) \times P(B)}$

5. Filter and rank rules:

- Keep only rules with lift > 1
- Sort rules by lift value
- Select the top 10 most interesting rules

5.2 Implementation Requirements

- Explain your choice of support and confidence thresholds
- Present the top 10 discovered rules in a table showing:
  - – Antecedent (product A)
  - – Consequent (product B)
  - – Support, Confidence, Lift
- Create ONE visualization:
  - – Scatter plot: support vs. confidence (colored by lift)

5.3 Analysis and Interpretation

- Interpret the top 5 discovered rules in business terms
- Explain what the lift values indicate
- Suggest 2-3 practical business applications (e.g., product placement, bundle promotions)

6. Part 2: Time Series Forecasting

Apply time series analysis to predict future daily sales.

6.1 Methodology

1. Prepare the time series: Use the daily aggregated sales data. Ensure it is sorted chronologically.

2. Split the data:

- Training set: First 80% of the time series
- Testing set: Last 20% of the time series

3. Implement ONE forecasting method:

Choose ONE of the following:

Option 1: Moving Average

- Compute simple moving average with window size = 7 days
- Predict the next day's sales as the average of the previous 7 days

Option 2: Exponential Smoothing

- Apply Simple Exponential Smoothing
- Experiment with different alpha values (e.g., 0.1, 0.3, 0.5)

Option 3: ARIMA

- Use auto ARIMA to find optimal parameters
- Or manually set simple parameters like ARIMA(1,1,1)

4. Generate predictions: Make predictions on the test set.

5. Evaluate model performance:

Calculate TWO of the following metrics:

- Mean Absolute Error (MAE): $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\mathrm{yi} - \hat{}\mathrm{yi}|$

- Root Mean Squared Error (RMSE): RMSE = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \hat{}yi)^2}$
- Mean Absolute Percentage Error (MAPE): MAPE = $\frac{100\%}{n}\sum_{i+1}^{n}\left|\frac{yi - \hat{}yi}{yi}\right|$

6.2 Implementation Requirements

- Clearly document your choice of model and parameters
- Present model performance metrics
- Create TWO visualizations:
  - Plot of actual vs. predicted sales on test data
  - Forecast plot showing predictions for next 14 days

6.3 Analysis and Interpretation

- Interpret the forecasting results
- Comment on the reliability of the forecasts
- Suggest 2-3 business applications (e.g., inventory planning, staffing decisions)

7. Bonus Points (Optional)

Implement up to TWO of the following for additional points (maximum 10% total):

1. Compare multiple time series models: Implement two different forecasting methods and compare their performance.

2. Advanced association rule visualization: Create a network diagram showing product relationships.

3. Interactive visualizations: Use Plotly to create interactive plots.

4. Seasonal decomposition: Decompose the time series into trend, seasonal, and residual components.

Notes:

- Allowed programming language: **Python**
- Required libraries:
  - Core: NumPy, Pandas
  - Visualization: Matplotlib, Seaborn
  - Time series: Statsmodels (for ARIMA if chosen)
  - Association rules: mlxtend
- If you wish to use any other libraries ,please tell us for approval before doing so.
- Any sign of cheating will result in a zero grade for this assignment.
- **Submission:**
  All submissions must be uploaded to Quera as a **ZIP file** named in the following format.
  FirstNameFamilyName-StudentNumber.zip
  Example: RezaZahedi-40432506.zip
  The ZIP file should include both your implementation code and the project report.
- Your report should be in PDF format and include a summary of the implementation steps, final results, and answers to any assignment questions within their sections.