Level-2-Data Analytics and Predictive Modeling Project Documentation

Project Title:

Market Basket Analysis

Team Members:

Soumajit Mandal

Project Overview:

The primary goal of this project is to perform market basket analysis and customer segmentation using the FP-Growth algorithm. The objectives include identifying associations between products purchased together, optimizing product placement, enhancing marketing strategies, and segmenting customers based on their purchasing behavior. The problem statement addresses the need for insights into customer purchasing patterns to drive effective marketing and product placement decisions.

1. Introduction

- **Purpose**: The purpose of this project is to analyze transaction data to uncover patterns in customer purchasing behavior, enabling better decision-making for product placement, cross-selling, and marketing strategies. By understanding these patterns, retailers can enhance customer satisfaction and increase sales.
- **Scope**: Transaction data analysis, association rule mining, customer segmentation, and recommendation generation.

2. Data Collection

- **Data Sources**: UCI Online Retail Dataset ‘This dataset contains transactional data from a UK-based online retailer from December 2010 to December 2011’.
- **Data Acquisition**: The data was downloaded from the UCI Machine Learning Repository.

3. Data Preprocessing

Data Cleaning:

Removed rows with missing values in critical columns (InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country).

Filtered out transactions with non-positive quantities and unit prices.

- **Data Transformation**: Stripped whitespace from product descriptions.
- **Data Integration**:Integrated transactional data into a unified dataset suitable for market basket analysis by grouping transactions and converting them into a transaction-product matrix.

4. Exploratory Data Analysis (EDA)

- **Descriptive Statistics**:Summary statistics for key variables such as Quantity and UnitPrice.
- **Visualization**: Histograms for Quantity and UnitPrice distributions.
- **Correlations**: Analyzed the correlation matrix for different variables.

5. Feature Engineering

- **Feature Selection**: Selected key features for market basket analysis: InvoiceNo, Description, Quantity..

6. Model Development

- **Model Selection**:

  Association Rule Mining: Selected the FP-Growth algorithm for its efficiency in finding frequent item sets.

  Customer Segmentation: Chose K-Means clustering for its simplicity and effectiveness in segmenting customers based on RFM (Recency, Frequency, Monetary) values.

- **Model Training**:

  · **FP-Growth:** Applied the algorithm with a minimum support threshold of 0.02.

  · **K-Means:** Standardized RFM values and trained the model with 4 clusters.

- **Model Evaluation**: Evaluated association rules using support, confidence, and lift metrics.

7. Model Interpretation

- **Feature Importance**: Identified important features for association rules based on their support, confidence, and lift values.
- **Model Insights**:Provided actionable insights for product placement and cross-selling strategies based on identified association rules.

8. Model Deployment

- **Deployment Plan**: The models can be deployed as part of a retail analytics platform using tools like Flask for the web interface.
- **Monitoring and Maintenance**:Regularly monitor model performance and update rules and clusters based on new data.

9. Conclusion

- **Summary**: Successfully identified significant product associations and customer segments, providing valuable insights for retail strategies.
- **Challenges**: Dealing with missing values and ensuring data consistency were significant challenges, addressed through robust data cleaning procedures.

Future Work:

Explore more advanced models for predictive analytics.

Integrate real-time data processing to update insights continuously.

10. Appendices:

- **Additional Visualizations**: Included additional visualizations such as the bar charts for product counts in the EDA section.
- **References**:

Dataset: UCI Machine Learning Repository - Online Retail Dataset.

Libraries Used: Pandas, NumPy, Matplotlib, Seaborn, mlxtend, scikit-learn.