Data Mining Final Project - Real World E-Commerce Data Mining

Tanishk Singh, Kritvirya Singh, Kanit Mann, Umesh Kumar Siyak

University of Arizona

INFO-523 SP25 001

Dr. Kunal Arekar

April 2, 2025

Dataset

For this project, our team will be using the Brazilian E-commerce Public Dataset by Olist, available on Kaggle. This dataset contains information on 100,000 orders from 2016 to 2018 made at multiple marketplaces in Brazil. It's a rich, real-world e-commerce dataset that provides an opportunity to explore various data mining techniques in a commercial context.

Project Goal and Questions:

Question 1: Delivery Performance

- What factors most significantly influence delivery performance, and how can delivery performance be improved?

Question 2: Customer Segmentation and Purchase Pattern Analysis

- How can we segment customers based on their purchasing behaviors, and what distinctive patterns exist across these segments?

Question 3: Time Series Analysis for Sales Forecasting

- Can we develop a reliable time series model to forecast future sales volumes and identify seasonal patterns in the Brazilian e-commerce market?

Tentative Plan of Analysis:

- Feature engineering (create features like total spend, average order value, purchase frequency, average delivery time based on locations, etc.)
- Aggregate orders data to appropriate time intervals (daily/weekly/monthly).

- Identify initial patterns through various correlation analysis techniques like the chi-square test.
- Apply K-means clustering to segment customers.
- Apply the Apriori algorithm to discover product associations within segments.
- Calculate support, confidence, and lift metrics.
- Implement ARIMA/SARIMA models to predict sales volumes.
- Create interactive visualizations of customer segments.
- Perform cross-validation with time series data and calculate error metrics like MAE, RMSE, etc.

These techniques may be altered/changed/dropped (if deemed unfeasible) as further data exploration takes place.

Expected Outcomes:

- Identification of 3-5 distinct customer segments with clear behavioral patterns.
- Geographic visualization showing the distribution of different customer types across Brazil.
- Identification of clear seasonal patterns in Brazilian e-commerce (e.g., holiday effects, monthly patterns).
- Identifying factors that impact delivery performance the most.

References

Kaggle. (2018). *Brazilian E-Commerce Public Dataset by Olist*. Kaggle. https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/data