Ghulam Ishaq Khan Institute of Engineering Sciences and Technology

Customer Behavior Analytics & Predictive Insights from E-Commerce Data

Submitted by:

Abdullah Mustafa (2022327)
Muhammad Bilal (2022360)
Hamza Motiwala (2022380)
Muhammad Umer Sami (2022684)

Spring 2025

Abstract

This project presents a comprehensive analysis of customer behavior using a real-world e-commerce dataset. Our objective was to derive actionable insights and predictive patterns through data mining techniques including preprocessing, exploratory analysis, association rule mining, classification, and clustering. We also developed a dark-themed Tkinter-based UI for interactive exploration, offering users a more intuitive way to analyze trends and patterns in the data. The resulting insights can be used to drive business decisions such as bundling strategies, targeted campaigns, and segmentation-based marketing.

1. Technical Report

1.1. Introduction

Customer behavior analytics has become a cornerstone of modern e-commerce strategies. Understanding how customers interact with products, and identifying purchasing patterns, can empower businesses to make informed decisions. In this project, we explore the "Online Retail" dataset using a wide array of data mining techniques. Our methodology integrates a thorough data preprocessing pipeline, exploratory data analysis, association rule mining, classification of high-value customers, and customer segmentation through clustering.

In addition to a complete Jupyter-based implementation, we created a graphical user interface (GUI) using Python's `tkinter` and `ttkbootstrap` libraries. The GUI is styled in a Breeze Dark theme, aligning with modern aesthetic and accessibility preferences. It allows users to load the dataset, view analytics, and interact with machine learning outputs in a user-friendly environment. This dual-deliverable approach enhances both technical depth and usability.

1.2. Dataset Overview

The analysis was conducted on the **Online Retail dataset**, sourced from the UCI Machine Learning Repository / Kaggle. This dataset contains **541,909 rows** and **8 columns**, detailing transactions for a UK-based retailer. The features include `InvoiceNo`, `StockCode`, `Description`, `Quantity`, `InvoiceDate`, `UnitPrice`, `CustomerID`, and `Country`. The primary problem statement driving this project was to analyze customer purchasing behavior to extract frequent buying patterns, segment customers, and predict high spenders using classification models.

1.3. Data Preprocessing

A crucial initial step involved cleaning and preparing the raw data. This included removing approximately ˜135,000 rows with missing `CustomerID`, which are essential for customer-level analysis. We also filtered out transactions with negative quantities, as these likely represent returns and do not reflect typical purchasing behavior. Finally, a new column, `TotalPrice`, was computed for each transaction as the product of `Quantity` and `UnitPrice`. These steps ensured the dataset was clean and properly structured for downstream data mining tasks.

1.4. Exploratory Data Analysis (EDA)

We performed a comprehensive visual analysis of the dataset using the Seaborn and Matplotlib libraries, styled with a dark theme for clarity.

1.4.1. Key Visualizations

Our analysis included several key visualizations. The chart showing **Top Countries by Transactions** (excluding the UK) highlighted international market activity. The **Monthly Sales Trend** visualization revealed significant seasonal peaks. An analysis of **Top Products** by quantity sold showed which items were exceptionally popular. Lastly, the **Transaction Distribution** plot illustrated the range of transaction values.

1.5. Association Rule Mining

Using the Apriori algorithm from the `mlxtend` library, we mined frequent itemsets to discover associations between products commonly purchased together. Rules were filtered based on a minimum support threshold of 5% and a lift greater than 1.0 to identify non-trivial and interesting relationships.

1.5.1. Methodology

The process involved transforming the transaction data into a one-hot encoded format suitable for the Apriori algorithm. We then generated frequent itemsets and subsequently derived association rules from these sets. Filtering was applied to focus on rules with sufficient support and lift, indicating strong and potentially useful associations.

1.6. Classification Models

To predict high-value customers, we engineered relevant features from the transaction data. These features include the `TotalSpend` per customer, their purchase `Frequency` (number of unique invoices), and the `AvgQuantity` of items per transaction. Customers were then labeled as `HighSpenders` if their total spend exceeded the dataset's median total spend, creating a binary classification target.

1.6.1. Models Used and Evaluation

We trained and evaluated several standard classification models: Decision Tree, Gaussian Naive Bayes, and K-Nearest Neighbors (KNN). Each model's performance in predicting high spenders was assessed using key evaluation metrics including accuracy, precision, recall, and F1 score. Cross-validation techniques were employed to ensure the robustness of the model evaluations.

1.7. Clustering: Customer Segmentation

Customer segmentation was performed using the K-Means clustering algorithm on the standardized features (`TotalSpend`, `Frequency`, `AvgQuantity`). Principal Component Analysis (PCA) was employed to reduce the dimensionality for visualization purposes, allowing us to visualize the clusters in a 2D space. We utilized the elbow method and validated the result with the silhouette score (approximately ˜0.58), which suggested that **3 clusters** provided a reasonable and interpretable segmentation.

1.7.1. Methodology

The clustering process involved scaling the selected features to ensure equal contribution, applying the K-Means algorithm to partition customers into a predefined number of clusters, and using PCA to project the data into a lower-dimensional space for visual inspection of the resulting segments.

1.8. Challenges and Reflections

Throughout the project, we encountered and addressed several technical and implementation challenges:

1.8.1. Key Challenges

Memory Usage in ARM: Performing association rule mining on the full dataset initially led to memory overflow issues due to the large number of transactions and unique items. This was successfully resolved by implementing data downsampling techniques for the ARM task, focusing on a representative subset of the data.

Handling Sparse Data: The e-commerce dataset is inherently sparse, with many products being purchased only once or by a very small number of customers. This required careful consideration and dimensionality reduction techniques before applying methods like one-hot encoding for ARM to manage computational complexity and improve rule quality.

Visualization Consistency: Ensuring that visualizations were aesthetically pleasing and easily readable within a dark theme environment, while also supporting potential high-DPI displays, required careful configuration of Matplotlib and Seaborn parameters and testing across different display settings.

Tkinter UI Development: Building a user-friendly interactive GUI with Tkinter presented challenges related to managing dynamic layouts, seamlessly integrating Matplotlib plots within the interface, handling user inputs for parameters (like K in clustering), and styling elements consistently, which was significantly aided by the `ttkbootstrap` library.

1.9. Tools and Libraries

The project leveraged the following tools and Python libraries:

Languages: Python 3.11

Libraries: pandas, seaborn, matplotlib, scikit-learn, mlxtend, networkx, tkinter, ttkbootstrap

Platform: Jupyter Notebook, VSCode

1.10. Team Contributions

The successful completion of this project is the result of collaborative effort, with key contributions from each team member:

Abdullah Mustafa (2022327): Focused on feature engineering, developing the classification and clustering pipelines, and analyzing their results.

Muhammad Bilal (2022360): Primarily responsible for creating the EDA visualizations, implementing the association rule mining, and contributing significantly to the development of the Tkinter user interface.

Hamza Motiwala (2022380): Contributed to the UI development, assisted with data visualization aesthetics, and worked on integrating different components of the project.

Muhammad Umer Sami (2022684): Handled the initial data preprocessing steps, focused on performance tuning for computationally intensive tasks, and managed the LaTeX documentation for the report.

1.11. Conclusion

This technical report details the methodology and implementation of a data mining project focused on analyzing customer behavior in an e-commerce context. We successfully applied preprocessing techniques to clean the dataset, performed exploratory data analysis to understand key trends, utilized association rule mining to discover product relationships, built classification models to predict high spenders, and segmented customers using clustering. The development of a dark-themed Tkinter GUI provides an interactive platform for exploring these technical outputs. This project demonstrates the technical feasibility and application of standard data mining techniques to derive valuable insights from complex transactional data.

2. Stakeholder Report: Business Insights and Recommendations

This report summarizes the key business insights derived from the customer behavior analysis of the Online Retail dataset and provides actionable recommendations aimed at improving sales performance, enhancing customer engagement, and optimizing operational strategies. The technical details of the analysis can be found in the accompanying Technical Report.

2.1. Executive Summary

Our analysis reveals significant patterns in customer purchasing behavior, identifies distinct customer segments, and provides a method for predicting high-value customers. Key findings include the identification of top international markets, seasonal sales trends, popular product combinations, and the distribution of transaction values. These insights form the basis for targeted strategies to increase revenue, improve customer loyalty, and make more efficient use of resources.

2.2. Key Business Insights

2.2.1. Market and Sales Trends

International Opportunities: Beyond the primary UK market, countries like Germany, France, and EIRE represent significant transaction volumes. Understanding the specific preferences and behaviors within these regions is crucial for targeted international growth strategies.

Seasonal Peaks: The e-commerce sales exhibit strong seasonality, with a pronounced peak towards the end of the calendar year (November-December). This surge in activity presents both a major revenue opportunity and a logistical challenge.

Popular Products: Certain product categories, particularly small decorative items and craft-related kits, consistently appear among the top-selling items by quantity. These products drive significant volume and can influence overall sales patterns.

Transaction Value Distribution: The majority of individual transactions are of relatively low value, while a smaller proportion of customers contribute through high-value purchases. This indicates a diverse customer base with varying spending habits.

2.2.2. Customer Behavior Patterns

Product Associations: Customers frequently purchase certain items together. For example, specific craft kits are often bought alongside ribbons and packaging materials. Identifying these associations reveals natural product bundles and cross-selling opportunities.

Customer Segments: The customer base can be segmented into distinct groups based on their spending habits (Total Spend, Purchase Frequency, Average Quantity). These segments include high-value loyal customers, infrequent big spenders, and lower-value but consistent buyers. Each segment has unique characteristics and potential.

High Spender Predictability: It is possible to predict which customers are likely to be high spenders based on their early purchase behaviors. Features like initial total spend and frequency are strong indicators.

2.3. Actionable Recommendations for Sales Improvement

Based on the insights gained, we recommend the following strategies to enhance sales and customer engagement:

2.3.1. Targeted Marketing and Personalization

Segment-Specific Campaigns: Develop tailored marketing campaigns for each identified customer segment. For High Value Loyalists, focus on loyalty programs, early access to new products, or premium support. For Potential Loyalists, encourage increased engagement through personalized recommendations and tiered discounts. For Lower Value Customers, explore strategies to increase their average order value or purchase frequency through targeted promotions on popular items.

High Spender Nurturing: Implement the high spender classification model to proactively identify customers with the potential to become high spenders. Engage these customers with exclusive offers, personalized product suggestions based on their browsing/purchase history, and potentially dedicated customer service.

Personalized Product Recommendations: Utilize the insights from association rule mining and individual customer purchase history to provide highly relevant product recommendations on the website, in emails, and during the checkout process. Suggesting frequently bought-together items can significantly increase basket size.

2.3.2. Product and Inventory Management

Strategic Product Bundling: Create attractive product bundles based on the discovered association rules. Offer these bundles at a slight discount compared to purchasing items individually to incentivize customers and increase average transaction value. Promote these bundles prominently on product pages and category listings.

Optimize Inventory for Popular Items: Ensure robust stock levels for the consistently top-selling products, especially leading up to and during peak seasons. Running out of stock on high-demand items directly impacts sales and customer satisfaction.

Cross-Selling Initiatives: Train sales or customer service staff (if applicable) to suggest complementary items based on customer purchases. On the e-commerce platform, enhance the ”Customers who bought this also bought...” or ”Frequently bought together” sections using the association rule data.

2.3.3. Operational and Strategic Planning

Seasonal Readiness: Use the detailed monthly sales trend data to inform operational planning. This includes forecasting demand more accurately, optimizing inventory levels throughout the year, scheduling adequate staffing for warehousing and customer service during peak periods, and planning marketing spend to align with seasonal opportunities.

International Market Focus: Develop tailored strategies for the identified top non-UK markets (Germany, France, EIRE). This could involve localizing the website and marketing content, offering region-specific promotions, adapting product assortments to local preferences, and potentially exploring localized warehousing or shipping solutions to improve delivery times and costs.

Increase Average Transaction Value (ATV): Implement strategies specifically designed to encourage customers to spend more per order. This could include offering free shipping above a certain order value, providing tiered discounts (e.g., save 10% on orders over £X), or promoting product bundles as mentioned earlier.

Customer Loyalty Programs: Design and promote loyalty programs that reward frequent purchases and high spending. Tailor the rewards to appeal to the characteristics of the high-value segments identified through clustering.

2.4. Conclusion for Stakeholders

The data mining analysis provides a solid foundation for understanding our customer base and their purchasing behaviors. By implementing the recommended strategies derived from this analysis – focusing on targeted marketing, leveraging product associations, optimizing inventory based on popularity and seasonality, and nurturing high-value customers – the business can expect to see improvements in sales performance, customer loyalty, and overall profitability. The interactive GUI developed as part of this project offers a user-friendly tool to explore these insights further and monitor key metrics.

Note: For detailed technical methodology and implementation, please refer to the accompanying Technical Report. The interactive GUI `projectui.py` is available for exploring the analysis outputs.