Analysis of Online Shopping Patterns and Consumer Behavior in E-commerce

Tesfamariam Tsegay. Ghezehey

Northwest Missouri State University, Maryville MO 64468, USA
`s568695@nwmissouri.edu`

Abstract. The growth of e-commerce has led to vast consumer data, offering opportunities to enhance customer experience and retention. This project, "Analysis of Online Shopping Patterns and Consumer Behavior in E-commerce," explores key purchasing trends, product preferences, and churn factors using a public dataset. [To be completed at the end]

Keywords: e-commerce · consumer behavior · retail analytics · shopping patterns · data mining

1 Introduction

1.1 Understanding Online Shopping Patterns and Customer Behavior

E-commerce is a rapidly growing industry that generates vast amounts of data from customer interactions. Analyzing online shopping patterns and consumer behavior helps businesses better understand their customers, optimize marketing strategies, enhance the shopping experience, and reduce customer churn.

1.2 Data Sources

primary Dataset: - E-commerce Customer Behavior (Kaggle) - Contains customer demographics, purchase frequency and churn data
Secondary Dataset: - E-Commerce Order and Sales (Kaggle) - Includes order details, customer location (State Code), products, categories, order date, status, brand, cost, and sales revenue. I also referred to a collection of project examples and ideas for inspiration from UpGrad [4]. Some project guidance and clarification were supported using AI-powered tools such as ChatGPT [3].

1.3 Data Problem and Importance

Problem Statement: This project aims to analyze customer shopping patterns, churn behavior, and sales performance. By integrating order-level data, we can **identify key factors that influence customer retention, purchasing decisions, and revenue generation.**
Why is this important:
 – Businesses can predict customer churn and improve engagement strategies.
 – Helps in analyzing which **products and brands** drive the most revenue.

- Optimizes marketing efforts through customer segmentation and sales trend analysis.
- Helps in pricing strategies by analyzing **cost vs. sales profitability.**

1.4   steps

Planning - Implementation Plan

- **Data Collection:** Import customer behavior and order details datasets from Kaggle.
- **Data Cleaning and Preprocessing:** Handle missing values, normalize category names, and merge datasets.
- **Exploratory Data Analysis (EDA):**
    - Analyze customer demographics and purchase behavior.
    - Identify the best-selling products, brands, and revenue trends.
    - Detect patterns in customer retention and churn.
- **Sales and Churn Prediction:** Apply machine learning models to identify high-value customers and predict churn risks.
- **Results and Insights:** Summarize findings and provide recommendations for customer retention and sales improvement.

1.5   key component and Limitations

- **key component:**
- **Customer Segmentation:** Identifying high-value customers.
- **Churn Analysis:** Analyzing why customers leave.
- **Sales and Product Analysis:** Understanding purchasing trends.
- **Limitations:**
- **Data Bias:** Dataset might not represent all global e-commerce businesses.
- **Feature Availability**: Limited variables may affect the accuracy of the churn prediction.
- **Assumptions**: Missing customer feedback in the dataset.

2   Methodology

2.1   Data Collection

The datasets used in this, project were sourced from **Kaggle**, Specifically, two datasets were utilized: Although the datasets contain similar customer-related information, they originate from different sources with no shared customer identifiers. Therefore, they are analyzed separately, each offering complementary insights into the behavior and patterns of online commerce.

- **E-commerce Customer Behavior Dataset** [1] - Contains customer demographics, purchase patterns, and churn indicators.
- **Online E-commerce Orders Dataset** [2] - Includes order details, product categories, brands, and sales revenue.

Both datasets were obtained in **CSV format**. No additional web scraping was required, as the data was readily available for download. To prepare the datasets for analysis, I focused on the following major attributes.

– **Customer Demographics:** Age, location (state code), and anonymized customer name.
– **Purchase Behavior:** Order number, order date, status (completed, pending, canceled).
– **Product Information:** Product name, category, brand, cost, and total sales.
– **Churn Prediction Indicators:** Purchase frequency, average order value, and payment method.

2.2 Data Cleaning

The cleaning and preparation steps, as shown in Figure 1, ensured that the data was structured and suitable for further analysis, including exploratory visualizations and predictive modeling.
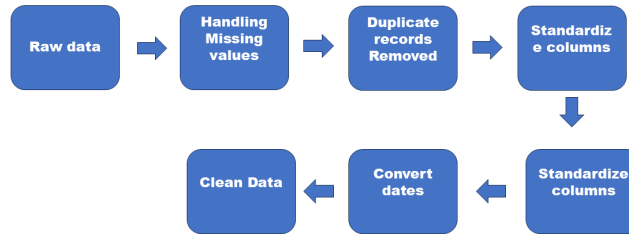

Fig. 1. Overview of the Data Cleaning Workflow

Two public datasets from Kaggle were curated for this project: one containing customer behavior and churn information, and the other consisting of detailed order-level records. The data curation process began with identifying relevant features, inspecting the structure of the raw CSV files, and aligning column names and formats for compatibility. Since the datasets came from different sources and lacked common identifiers, they were cleaned and analyzed separately.

Data preprocessing was conducted using Python in a Jupyter Notebook environment. The main tools and techniques included the `pandas` library for data manipulation and exploratory inspection using functions such as `.isnull()`, `.dropna()`, `.fillna()`, and `.duplicated()`. Column headers were standardized to lowercase and stripped of spaces for consistency across datasets.

To cleanse missing values, rows with missing values in critical fields like `customer_name`

and `order_number` were removed to ensure data integrity. For non-critical fields, missing values were filled with default values such as zero to preserve dataset structure. Duplicate records were dropped, and all date fields were converted into a standard datetime format to support time-based feature engineering.

After cleaning, `cleaned_orders.csv` contained 13 attributes and 5,095 records (from an original 5,110), while `cleaned_customers.csv` retained 13 attributes and 250,000 records with no rows dropped.

Dependent variables: `churn`, `customer_value_segmentation`, and `product_category_preference`.

Independent variables: `customer_id`, `age`, `gender`, `state_code`, `purchase_date`, `order_date`, `product_price`, `quantity`, `total_purchase_amount`, `payment_method`, `returns`, `brand`, `status`.

Engineered features: `days_since_last_purchase`, `purchase_frequency`, and `value_segment`.

These engineered features were created to support both exploratory data analysis and downstream modeling tasks such as customer segmentation and churn prediction.

2.3   Exploratory Data Analysis (EDA)

Customer Demographics

Product Trends

Churn Insights

3   Results and Analysis

4   Limitations

5   Conclusion and Future Work

6   References

References

1. Kaggle: E-commerce customer behavior analysis dataset (2025), https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis/data
2. Kaggle: Online e-commerce orders dataset (2025), https://www.kaggle.com/datasets/ayushparwal2026/online-ecommerce
3. OpenAI: Chatgpt - ai-powered conversational assistant (2025), https://openai.com/chatgpt
4. UpGrad: Top data science project ideas & topics for beginners (2025), https://www.upgrad.com/blog/data-science-project-ideas-topics-beginners/