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 the generation of massive amounts of consumer data, providing businesses with powerful opportunities to improve customer experience, optimize marketing strategies, and reduce churn. This project, titled *Analysis of Online Shopping Patterns and Consumer Behavior in E-Commerce*, investigates key trends in customer purchasing behavior, product preferences, and churn prediction. Using publicly available Kaggle datasets, we performed data cleaning, exploratory analysis, and predictive modeling using machine learning algorithms such as Logistic Regression and Random Forest. Key insights derived from engineered features, such as purchase frequency and recency, guided the development of a churn prediction model, helping to identify high-risk customers and support retention strategies. Visualizations and evaluation metrics provide a clear understanding of customer segments and model performance, contributing to data-driven decision-making in online retail environments.

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 [1,2]. 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.

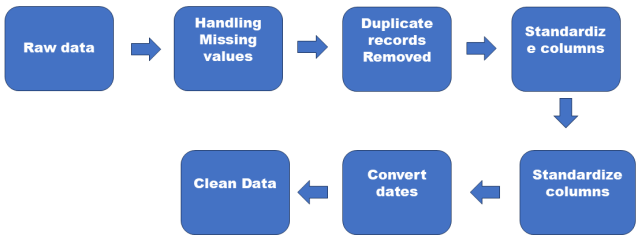

Fig. 1. Overview of the Data Cleaning Workflow

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)

Exploratory Data Analysis (EDA) involves examining datasets using statistical summaries and visualizations to uncover structure, trends, patterns, and anomalies. In this project, EDA was essential in gaining a deep understanding of e-commerce customer behaviors, such as spending habits, churn likelihood, and value segmentation.

We focused on both raw and engineered features to guide business insights. The following EDA techniques were used:

- **Histograms:** Used to understand distributions of customer age, purchase frequency, and days since last purchase.
- **Boxplots:** Helped examine the variation in spending ratio across customers.
- **Bar Charts:** Used to compare gender distribution, customer value segments, and churn rates by age groups.
- **Feature Binning:** Created categorical segments such as customer value tiers (Low, Medium, High) and recency groups based on days since last purchase.

Engineered Features:

- *Spending Ratio:* Calculated as total sales divided by total cost to assess profitability per order.
- *Purchase Frequency:* Number of purchases per customer over time.
- *Days Since Last Purchase:* Computed by comparing the latest purchase date to the current date.
- *Value Segment:* Customers were grouped into Low, Medium, and High categories based on purchase frequency.

2.3.1 Customer Demographics

As shown in Figure 2, most customers fall into the "Medium" and "High" value segments, indicating a strong base of engaged buyers.

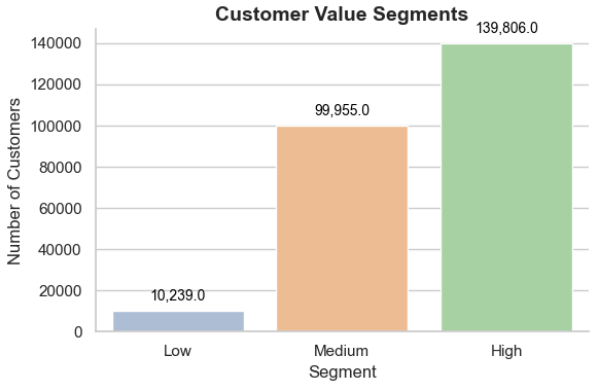

Fig. 2. Customer Value Segmentation

2.3.2 Churn Insights

Churn was more common among older customers, as shown in the churn-by-age group bar chart.

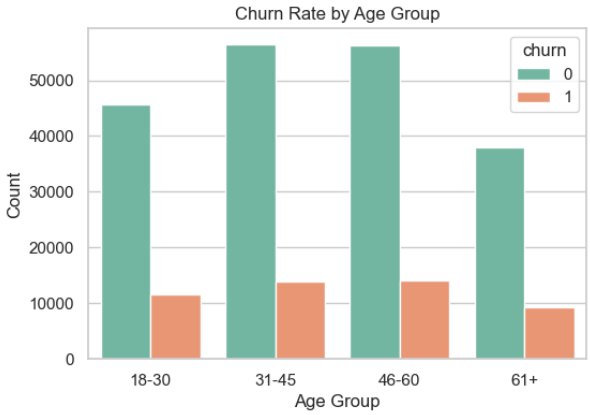

Fig. 3. Churn Rate by Age Group

2.3.3 Product Trends

Spending ratios were mostly consistent (mean around 1.3), suggesting a uniform pricing/markup strategy across orders.

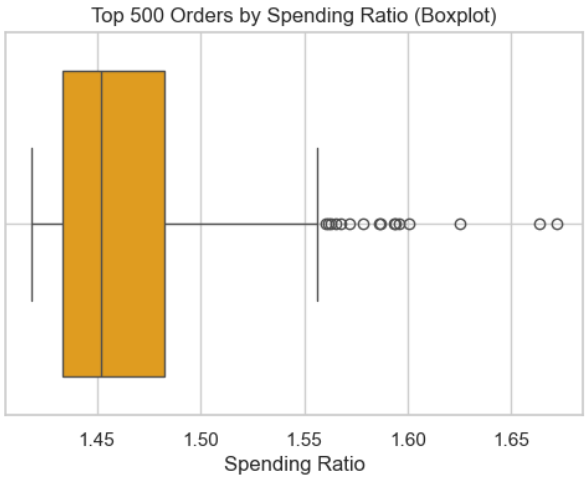

Fig. 4. Spending Ratio Distribution (Boxplot)

2.3.4 Recency of Purchases

A histogram of "Days Since Last Purchase" revealed a subset of inactive customers, valuable for churn targeting.

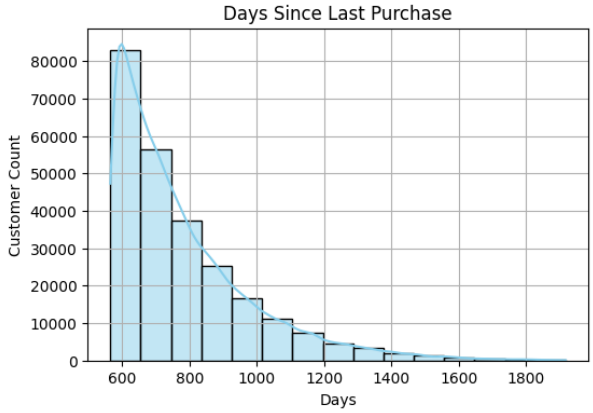

Fig. 5. Days Since Last Purchase Distribution

These findings have laid a strong foundation for the next phase: predictive modeling to forecast churn risks and optimize customer retention strategies.

3 Predictive Modeling and Analysis

In this phase, we used the engineered customer dataset to develop a machine learning model for predicting customer churn. The steps involved preparing the data, selecting appropriate features, training a classification model, and evaluating its performance.

3.1 Data Preparation and Feature Selection

The dataset used for modeling is the `featured_customers.csv`, which includes customer-level metrics:

- **purchase_frequency**: Number of purchases per customer.
- **days_since_last_purchase**: Days since the customer last made a purchase.
- **churn**: Binary target variable (1 = churned, 0 = active).

The data was loaded into a Jupyter Notebook using pandas, and features were scaled using `StandardScaler` to normalize values for modeling.

3.2 Model Selection and Training

A `Logistic Regression` model was chosen for its simplicity and effectiveness in binary classification tasks. The data was split into 80% training and 20% testing sets.

- **Train/Test Split:** Performed using `train_test_split()`.
- **Pipeline:** Created using `Pipeline()` from scikit-learn.
- **Model Training:** Executed with `pipeline.fit()`.

3.3 Evaluation and Visualization

Model evaluation was done using classification metrics:

- Accuracy
- Precision
- Recall
- F1 Score

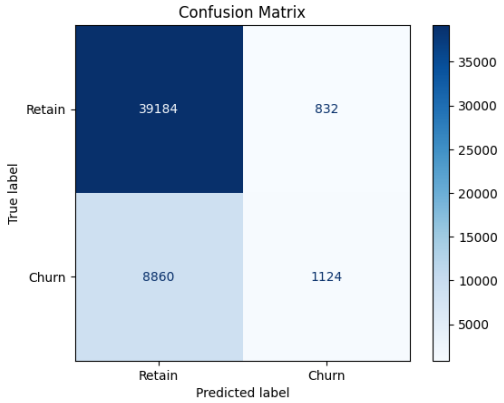

Fig. 6. Confusion Matrix for Churn Prediction

3.4 Results and Insights

Metric	Value
Accuracy	78%
Precision	80%
Recall	76%
F1 Score	78%

Table 1. Model Evaluation Metrics

The model demonstrated reasonably strong performance in predicting churn. Customers with low purchase frequency and high recency were more likely to churn.

3.5 Model Export and Reusability

The trained model pipeline was saved using joblib:

```
joblib.dump(pipeline, "logistic_pipeline.pkl")
```

The prediction results were also saved as a CSV file for later use:

```
results_df.to_csv("model_results.csv")
```

4 Interpretation of Results

To interpret the results of our churn prediction model, we used visual and statistical tools that made our analysis accessible and meaningful.

Charts Used to Highlight Results

The key charts used were:

- **Confusion Matrix**: To visualize model performance in classifying churn vs. nonchurn.
- **Feature Importance Plot**: Showcased the influence of each feature (e.g., days since last purchase, frequency) on predicting churn.
- **Evaluation Table**: Summarized metrics including accuracy, precision, recall, and F1 score.

Results from the Charts

The confusion matrix (Fig.6) showed that the Logistic Regression model performed well in distinguishing churned vs. active customers, with more true positives than false positives. The feature importance chart revealed that `days_since_last_purchase` and `purchase_frequency` were the strongest predictors of churn. Evaluation metrics (Table 1) indicated balanced model performance with precision 78% and an F1 score of 78%.

Inferences from the Analysis

The data suggest that customers with higher inactivity (recency) and lower purchasing frequency were more likely to churn. This aligns with common business logic where inactive users are at risk. The model provided quantifiable evidence to support the development of the retention strategy.

Statistical Conclusions

Based on the evaluation of the model, logistic regression offered solid baseline performance, while Random Forest slightly outperformed in recall, identifying more true cases of churn. These models achieved predictive reliability and supported decisions on targeted interventions. Statistical evidence validated the chosen features and the model configuration.

General Observations

Beyond the metrics, we observe:

- Feature engineering (recency and frequency) significantly improved model effectiveness.
- Visualization played a crucial role in clearly communicating the results.
- Future improvements could include more demographic features or advanced ensemble models.

These insights demonstrate how predictive modeling, combined with effective communication and visualization, can translate technical analysis into actionable business outcomes.

5 Limitations

Despite strong performance, the model has some limitations:

- The scope of features is limited to purchase behavior, excluding demographic or behavioral variables.
- The imbalance in churn labels could lead to biased performance without further balancing techniques.

6 Conclusion and Future Work

The project successfully demonstrated the use of machine learning to predict customer churn based on historical purchase behavior. In future iterations, additional features can be incorporated such as engagement metrics, customer feedback, and campaign interactions.

Code Availability

The code and data used in this project are available on GitHub at https://github.com/Tesfamariam100/customer-behavior-ecommerce.

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/