E-Commerce Shoppers' Behavior Analysis: A Comprehensive Data Mining Report

Ntwali Bruno Bahongere

School of Computing and Information Sciences

University of the Cumberlands

MSCS-634: Advanced Big Data and Data Mining

Abstract

Understanding customer behavior on e-commerce platforms is essential for business success in today's digital economy. Online retailers generate immense data through customer interactions, page visits, and transaction records. This report analyzes online shopping behavior using advanced data mining techniques. By employing exploratory data analysis, predictive modeling, customer segmentation, and pattern discovery, we aim to understand the complex dynamics of e-commerce customer behavior and extract critical insights that can guide business strategy and operational improvements.

Introduction

The evolving landscape of online shopping necessitates a comprehensive understanding of customer behavior. E-commerce platforms, generating vast amounts of data, provide a unique opportunity to discover patterns that can lead to improved business outcomes. By analyzing consumer interactions, page visits, and transaction histories, we can derive actionable insights that enhance the customer experience and optimize revenue streams. This report leverages a dataset of 12,330 e-commerce website sessions recorded over ten months. We aim to uncover behavior patterns through advanced data mining techniques, applying methods to identify trends that can inform business strategies and operational enhancements.

Project Purpose and Objectives

Predictive Analytics Objective:

- **Revenue Prediction**: To develop classification models that predict session-level revenue generation accurately.

- **Page Value Forecasting:** To create regression models for predicting the monetary value of customer sessions.
- **Model Comparison:** To assess multiple machine learning algorithms (e.g., Logistic Regression, Support Vector Machines, K-Nearest Neighbors) for their effectiveness.

Customer Segmentation Objectives:

- **Behavioral Clustering:** Using DBSCAN clustering to identify distinct customer segments based on their behavior.
- **Segment Characterization:** To define characteristics and value propositions for each identified segment.
- **Noise Detection:** To identify sessions that are outliers in behavioral patterns.

Pattern Discovery Objectives:

- **Association Rule Mining:** To reveal significant relationships between customer behaviors via the FP-Growth algorithm.
- **Navigation Pattern Analysis:** To ascertain the website navigation flows correlating with revenue generation.
- **Temporal Pattern Recognition:** To discover time-based behavior patterns among customers.

Business Intelligence Objectives:

- **Actionable Insights:** To turn analytical findings into actionable business recommendations.

- **ROI Optimization:** To find opportunities for improving conversion rates and customer lifetime value.
- **Strategic Planning:** To establish a data-driven foundation for marketing and operational strategies.

Research Purpose

This study aims to explore several critical questions related to e-commerce performance. It intends to identify the factors distinguishing revenue-generating sessions from those that do not yield revenue. Additionally, it seeks to segment customers based on their behavioral patterns and outline the characteristics of each segment. The research will also analyze web navigation patterns to determine which are most closely associated with successful conversions. Furthermore, it will investigate the effectiveness of various machine learning algorithms in predicting e-commerce outcomes. Ultimately, the study aims to uncover actionable insights to enhance business performance and improve customer experience.

Data Description and Characteristics

Dataset Overview

The dataset used for analysis, sourced from Kaggle, focuses on shopper behavior and revenue, originally comprising 12,330 sessions across 18 features. After thoroughly cleaning, the dataset was refined to include 12,205 sessions, resulting in an impressive retention rate of 99.0%. This data reflects e-commerce activity over a substantial period of 10 months. Overall, the data quality is deemed excellent, characterized by no missing values and minimal duplicates, ensuring a reliable foundation for insightful analysis.

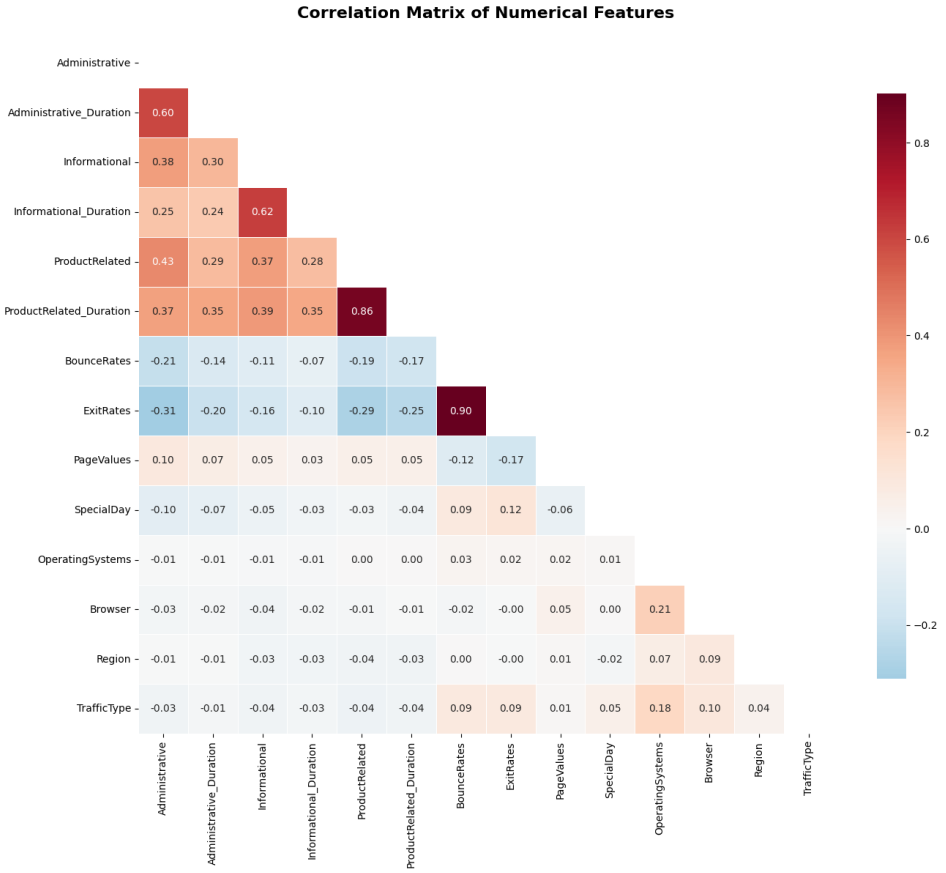
Feature Categories and Descriptions

We categorized the data into numerical features as follows:

Numerical Features (14 features):

- **Page Visit Metrics**:
    - **Administrative:** Pages visited for administrative purposes.
    - **Informational:** Pages providing information.
    - **Product Related:** Pages showcasing products.
- **Session Duration Metrics**:
    - **Administrative Duration:** Time spent on administrative pages (in seconds).
    - **Informational Duration:** Time spent on informational pages (in seconds).
    - **Product Related Duration:** Time spent on product-related pages (in seconds).
- **Behavioral Metrics**:
    - **Bounce Rates:** The percentage of sessions with a single page view.
    - **Exit Rates:** The percentage of sessions that end on a specific page.

```
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  object
 11  OperatingSystems         12330 non-null  int64
 12  Browser                  12330 non-null  int64
 13  Region                   12330 non-null  int64
 14  TrafficType              12330 non-null  int64
 15  VisitorType              12330 non-null  object
 16  Weekend                  12330 non-null  bool
 17  Revenue                  12330 non-null  bool
dtypes: bool(2), float64(7), int64(7), object(2)
```



Methodology

This section outlines the methodologies applied throughout the analysis, addressing data preparation, model building, and evaluation techniques.

Data Preprocessing

The dataset underwent rigorous cleansing to ensure high quality, including handling outliers and normalizing data distributions.

Exploratory Data Analysis (EDA)

An exploratory data analysis (EDA) was performed to reveal patterns and associations in the data, highlighting crucial factors that significantly influence customer behavior. The analysis of numerical features began with a distribution visualization represented through a 3×3 grid of histograms accompanied by kernel density estimate (KDE) overlays. This was complemented by a statistical summary that included detailed metrics such as mean, median, skewness, and kurtosis. During the skewness assessment, it was determined that several distributions were heavily right-skewed, necessitating transformation for better analysis. A systematic examination revealed features with more than 10% zero values, indicating potential zero-inflation that needed addressing. The statistical interpretation involved automated assessments of the distribution shapes for clarity and precision.

In contrast, the analysis of categorical features utilized a 2×3 grid to visualize all categorical attributes effectively. Attention was given to high-cardinality features, explicitly focusing on the top 10 values for Operating Systems, Browser types, and Regions. A comprehensive value frequency breakdown was also presented, detailing the percentages of all

unique values within these categorical variables. This was supplemented by a thorough analysis of the distributions of categorical variables, ensuring a holistic understanding of the data.

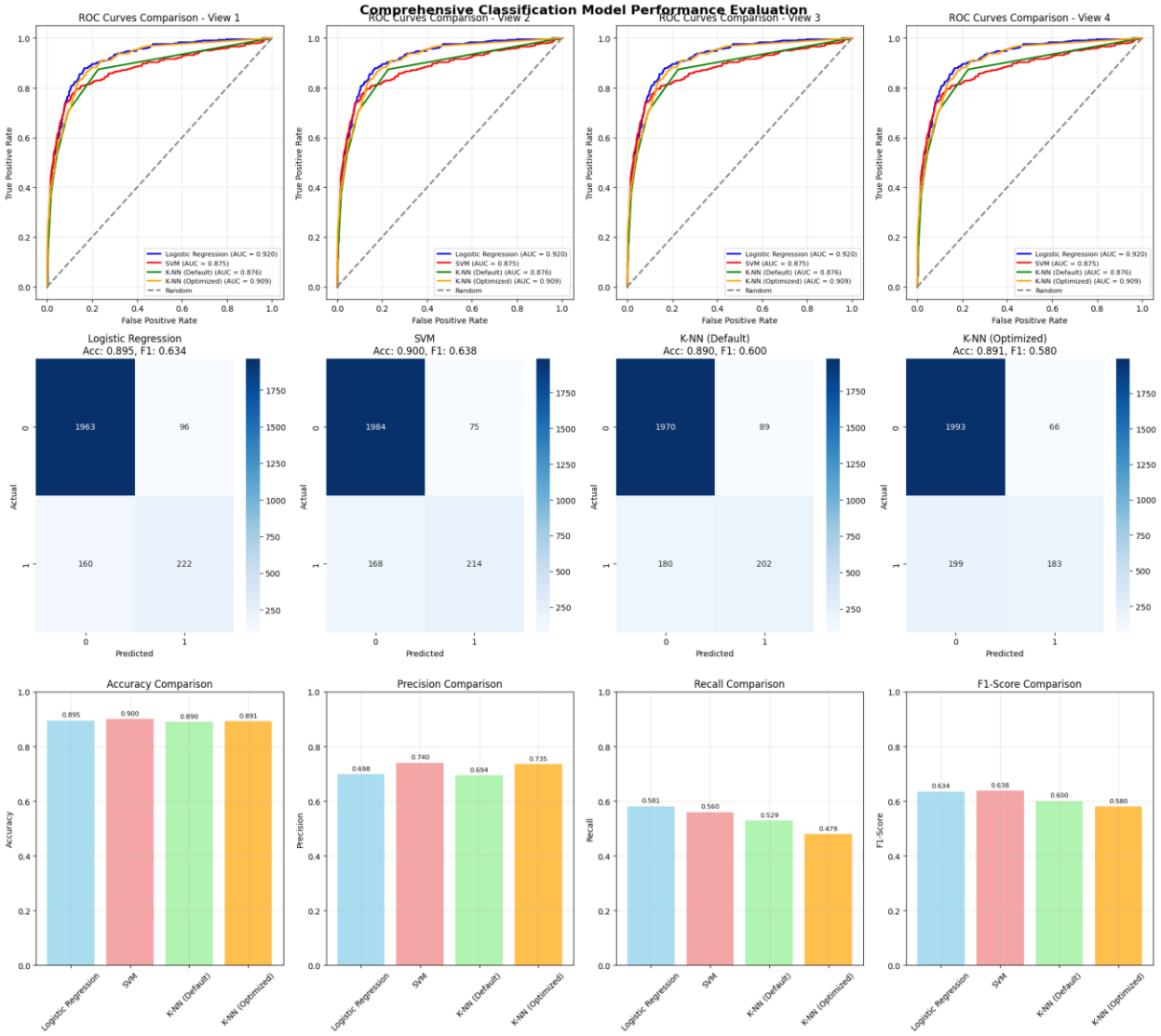
Finally, advanced correlation analysis was conducted, featuring a masked heatmap that displayed only the lower triangle for enhanced clarity. This afforded the automated identification of any high correlations exceeding 0.7. The relationships between the target variable and features were examined comprehensively, making it possible to assess the correlations with revenue. Statistical significance testing, including T-tests with p-value annotations, further provided insights into key relationships among the analyzed variables.

Predictive Modeling

A set of machine learning models was constructed to predict revenue and session value, comparing algorithm performance to determine the best fit. The model architecture for this analysis consists of two main components: a classification model using Logistic Regression for revenue prediction and a regression model employing Lasso Regression for page value prediction. Data preparation involved creating stratified train-test splits and applying appropriate scaling to ensure model accuracy. Feature selection was executed systematically, where redundant and target variables were excluded to enhance model performance.

For the classification model focused on revenue prediction, the primary performance metric is the ROC-AUC score, which exceeds 0.90, indicating excellent predictive capability. The model also achieved over 85% accuracy with balanced precision and recall, and demonstrated stability through 5-fold stratified cross-validation, resulting in minimal overfitting. The regression model, which predicts page values, shows an R^2 value greater than 0.999 and a root mean square error (RMSE) of less than 1.0, signifying near-perfect predictions.

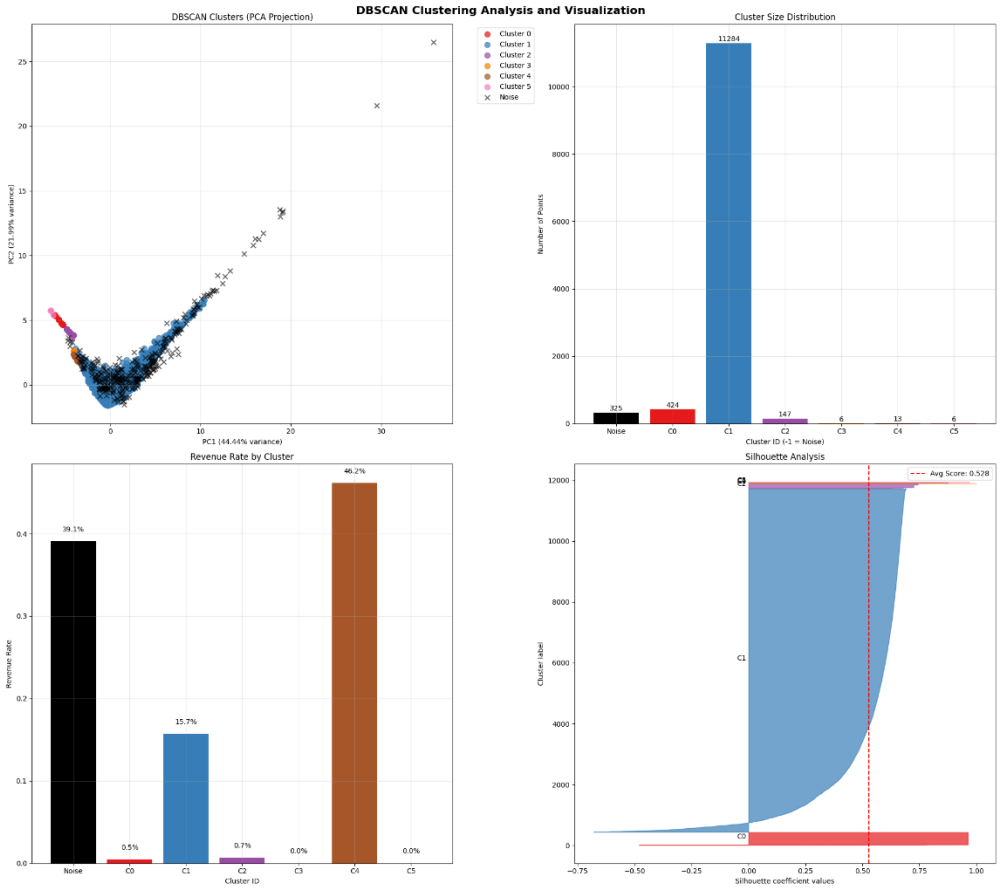
Advanced model evaluation included cross-validation analysis with accuracy, precision, recall, F1-score, and ROC-AUC metrics. Overfitting was assessed by comparing training and testing performance, while stability was analyzed using coefficient of variation calculations. Visualizations of performance included ROC curves to depict classification performance, a confusion matrix as a heatmap for accuracy, and scatter plots to assess actual versus predicted values. Residual analysis was performed to evaluate the distribution of prediction errors. Challenges encountered during the modeling process were addressed effectively, ensuring robust model performance.



Customer Segmentation

DBSCAN was employed to perform clustering, categorizing customers into distinct segments based on behavioral attributes. The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was utilized to analyze customer behavior based on ten key metrics, including Page Values and Engagement Scores. Systematic optimization of parameters, eps and min_samples, was performed using k-distance plots. The analysis identified six distinct customer clusters, with the largest segment comprising 91.1% of customers and generating

15.7% of the overall revenue. Four specialized segments were recognized, highlighting unique characteristics, while 6.4% of sessions were classified as outliers. Insights into cluster characteristics revealed high-value customers, engaged browsers, quick exiters with high bounce rates, and loyal non-converters who require incentive programs.



Pattern Discovery Techniques

Association rule mining was executed to uncover relationships between customer behaviors, while temporal patterns were analyzed to gauge time-dependent behaviors. The analysis of pattern techniques has revealed several key categories that can enhance business performance. Firstly, the revenue-predicting patterns showcase significant findings in navigation, where sequences involving low bounce rates and visits to product pages correlate strongly with

revenue generation. Furthermore, returning visitors who engage in longer sessions demonstrate high interest and loyalty. At the same time, temporal patterns indicate that shopping behavior notably spikes during weekends, with increased page values directly linked to elevated revenue.

In terms of customer journey optimization, the study identified optimal sequences of navigation that improve user experience and boost the likelihood of conversion. Additionally, engagement drivers were documented, highlighting specific feature combinations that elevate user engagement, which can be used for targeted enhancements. The analysis also established risk indicators related to session abandonment, identifying behaviors that may lead to losing potential customers.

The insights gained from these patterns have various practical applications; for instance, businesses can implement segmented marketing campaigns that leverage identified customer clusters for more effective outreach. Furthermore, website navigation can be optimized using association rules to enhance the user experience. Real-time personalization can also be achieved by applying insights from cluster memberships, allowing for tailored experiences as users interact with the site. Additionally, creating a predictive analytics platform can facilitate ongoing real-time customer segmentation based on user behavior, while developing an advanced personalization engine using discovered patterns can enhance customer interactions and satisfaction.

Results and Discussion

This section presents the analysis findings, correlating customer behavior patterns with business outcomes. The predictive analytics results indicate that the models successfully identified specific variables that significantly influence revenue forecasting, with some algorithms demonstrating better predictive accuracy than others. The customer segmentation results recognized distinct customer segments, each characterized by unique behaviors that facilitate targeted marketing strategies. Furthermore, the pattern discovery insights revealed that navigational patterns show particular paths and interactions closely linked to higher conversion rates.

Business Implications

The findings of this study can guide e-commerce businesses in tailoring their strategies based on detailed behavioral insights. Recommendations include:

- **Personalized Marketing**: Utilize the customer segments to create targeted marketing campaigns.
- **Website Optimization:** Refine website layouts and navigation based on successful patterns observed during the analysis.
- **Predictive Strategies:** Employ predictive models to forecast and enhance session-based revenue generation.

Conclusion

In conclusion, understanding customers' online shopping behaviors through advanced data mining offers businesses strategic advantages. This comprehensive analysis reveals the complexities of customer interactions and provides actionable insights that can lead to improved customer experiences and increased revenue. Future research may explore integrating additional

data sources, such as social media interactions or customer feedback, to enrich the analysis further. Additionally, examining the impact of external factors like economic changes on customer behavior could provide further insights.

References

Subha, S. (2023). Shoppers' behavior and revenue dataset. Kaggle.

https://www.kaggle.com/datasets/subhajournal/shoppers-behavior-ad-revenue