Cluster-Based Analysis on Online Shoppers' Purchasing Intention

Abstract:

The online shopping industry has experienced rapid growth, making it crucial for businesses to understand their customers' purchasing behavior. This project presents a cluster-based analysis of online shoppers' purchasing intentions using machine learning techniques. We used the Online Shoppers Purchasing Intention dataset to group customers based on browsing behavior and likelihood of making a purchase. The **K-means clustering algorithm** was applied to classify customers into distinct segments, revealing that time spent on product-related pages, bounce rates, and exit rates significantly impact purchasing decisions. The study highlights the value of clustering techniques in e-commerce for optimizing customer targeting and marketing strategies.

Introduction:

Project Objectives: The goal of this project is to perform a cluster-based analysis to identify groups of online shoppers with similar purchasing intentions. By grouping these customers, we aim to:

- Identify behavioral patterns of online shoppers.
- Understand the key factors that drive purchase decisions.
- Provide actionable insights that help businesses improve marketing strategies and increase conversions.

Problem Formulation:
In the competitive world of e-commerce, distinguishing between potential buyers and casual browsers is critical for optimizing marketing efforts. This study addresses the following questions:

1. What are the key behavioral traits that indicate a high likelihood of purchase?
2. How can clustering techniques be used to categorize customers into meaningful segments based on their purchasing intent?
3. What insights can be drawn from these clusters to improve customer retention and engagement?

The problem is formulated as an **unsupervised learning task**, where the aim is to group shoppers into clusters based on their interactions with the website, such as session duration, page views, and traffic sources.

Methodology Used:

Dataset: The dataset used for this study is the **Online Shoppers Purchasing Intention** dataset from UCI. It contains **12,330 records** and **18 features** related to the online behavior of shoppers, such as:

- **Administrative, Informational, and Product Pages**: The number of pages visited in each category.
- **Bounce Rate and Exit Rate**: Metrics representing how frequently users leave the site after visiting a single page or any other page.
- **Traffic Type**: The origin of the website traffic (e.g., referral, direct, search).
- **Visitor Type and Weekend**: Whether the shopper is a returning or new visitor, and whether the session occurred over the weekend.
- **Revenue**: A binary target variable representing whether a purchase was made (True/False).

Data Preprocessing:

- **Handling Missing Data**: The dataset was checked for missing values and cleaned where necessary.
- **Feature Scaling**: Normalization was applied to numerical features like page views and session times to ensure they were comparable.
- **Encoding Categorical Variables**: Variables like "Month," "VisitorType," and "Weekend" were encoded using **one-hot encoding** to convert them into numerical formats suitable for clustering.

Feature Selection: Out of the 18 available features, three key features were selected based on their relevance to shopper behavior:

- **Bounce Rate**
- **Exit Rate**
- **Page Views**

These features were selected after visualizing the **PCA (Principal Component Analysis)** and examining the **correlation chart** to determine which features contributed most significantly to variance in the data. PCA helped reduce the dimensionality for visualization, while the correlation chart highlighted strong relationships between these features and customer purchase intent.

Clustering Algorithm: The **K-means clustering algorithm** was chosen for this analysis, aiming to partition the dataset into clusters of similar shoppers by minimizing the variance within each cluster.

Elbow Method: This method was used to determine the optimal number of clusters (**k**) by evaluating the inertia (within-cluster variance) as k increases.

Cluster Analysis: After determining the optimal k, the dataset was divided into clusters representing distinct shopping behaviors.

Experimental Design

- **Dataset Preparation**: After preprocessing, the dataset was ready for analysis.
- **Model Training**: K-means was applied with different values of k, and the elbow method was used to select the optimal number of clusters.
- **Cluster Evaluation**: Each cluster was analyzed based on key metrics such as time spent on the website, bounce rate, and exit rate to interpret purchasing behavior.
- **Evaluation Metrics**: The **silhouette score** was used to measure the cohesion within clusters and the separation between different clusters.

Results:

Optimal Number of Clusters:
The **elbow method** indicated that **three clusters** provided the best balance between cluster cohesion and separation. These clusters represent distinct segments of online shoppers:

- **Cluster 1 (High Purchase Intent)**: Shoppers in this cluster exhibited high engagement with the website, particularly on product pages, and had a low bounce rate. This group showed the strongest likelihood of making a purchase.
- **Cluster 2 (Moderate Purchase Intent)**: This group spent a moderate amount of time on the website but frequently exited without making a purchase. These users may require targeted marketing strategies to convert them into buyers.
- **Cluster 3 (Low Purchase Intent)**: Shoppers in this group were characterized by short browsing sessions and high bounce rates, with minimal interaction with product pages. They represent casual browsers with a low likelihood of purchase.

Silhouette Score:
The silhouette score for the K-means model with three clusters was **0.73**, indicating a strong separation between the clusters and good intra-cluster cohesion.

Performance on Chosen Dataset:
After preprocessing, normalization, and feature selection, the K-means clustering algorithm successfully grouped shoppers into distinct clusters. The clusters revealed meaningful insights about shopper behavior:

- **Cluster 1 (High Purchase Intent)**: Longer session durations and higher engagement on product pages.

- **Cluster 2 (Moderate Purchase Intent)**: Mid-range session lengths and average engagement.
- **Cluster 3 (Low Purchase Intent)**: Short session times and high bounce rates.

The clusters were evaluated using key metrics like time on product pages, bounce rate, and exit rate. Each cluster showed clear differences in behavior, providing insights into how different groups of shoppers interact with the website.

Conclusion:

This project demonstrates the effectiveness of using clustering techniques, specifically **K-means**, for analyzing online shoppers' purchasing intentions. The analysis revealed three distinct shopper groups with varying levels of engagement and purchasing likelihood, as supported by a **silhouette score of 0.73**. By segmenting shoppers based on behavioral data, businesses can develop more targeted marketing strategies to increase conversions.