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 utilized the "Online Shoppers Purchasing Intention" dataset to group customers based on browsing behavior and the likelihood of making a purchase. Three clustering algorithms—K-means, DBSCAN, and Hierarchical clustering—were applied. The analysis revealed that hierarchical clustering provided the most distinct segmentation with a silhouette score of 0.7614. 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:

1. Identify behavioral patterns of online shoppers.
2. Understand the key factors that drive purchase decisions.
3. Provide actionable insights to 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:

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

Methodology

Dataset

The dataset used for this study is the "Online Shoppers Purchasing Intention" dataset from UCI, which contains 12,330 records and 18 features related to the online behavior of shoppers. The key features include:

- **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 other pages.
- **Traffic Type**: The origin of the website traffic (referral, direct, search, etc.).
- **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.

Data Preprocessing

- **Handling Missing Data**: The dataset was cleaned to remove any missing values.
- **Feature Scaling**: Normalization was applied to numerical features like page views and session times to make them comparable.
- **Encoding Categorical Variables**: Variables such as "Month," "VisitorType," and "Weekend" were one-hot encoded to convert them into numerical formats suitable for clustering.

Clustering Algorithms

Three clustering algorithms were applied for this analysis:

1. **K-means**: A partition-based method that groups shoppers into clusters by minimizing variance within each cluster.
2. **DBSCAN**: A density-based algorithm that identifies clusters of varying shapes and handles noise points effectively.
3. **Hierarchical Clustering**: This technique creates a hierarchy of clusters and allows the exploration of different levels of granularity.

- **Elbow Method**: Used to determine the optimal number of clusters for K-means.
- **DBSCAN Parameters**: The parameters were optimized using `eps = 0.1` and `min_samples = 7`.
- **Dendrogram**: Visualized for hierarchical clustering to identify meaningful clusters.

Feature Selection and PCA

All 18 features were used in the clustering process. Additionally, PCA was employed for dimensionality reduction and visualization, helping to interpret the clusters formed by each algorithm.

Results and Analysis:

Results Comparison:

Clustering Algorithm	Best Parameters/Clusters	Silhouette Score
DBSCAN	eps = 0.1, min_samples = 7	0.5405
K-means	Optimal number of clusters = 8	0.6048
Hierarchical	Dendrogram visualized	0.7614

Results and Analysis:

The analysis involved applying three clustering algorithms—**DBSCAN**, **K-means**, and **Hierarchical Clustering**—to the "Online Shoppers' Purchasing Intention" dataset, which consists of 18 features that capture various aspects of online shopper behavior. Each algorithm's performance was evaluated based on the **silhouette score**, a metric that measures how well each data point fits within its cluster (cohesion) compared to other clusters (separation).

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- **Best Parameters**: `eps = 0.1`, `min_samples = 7`
- **Silhouette Score**: 0.5405

DBSCAN is a density-based clustering algorithm that is particularly useful for identifying clusters with irregular shapes and handling noise points. In this analysis, DBSCAN identified some meaningful clusters but produced a moderate silhouette score of 0.5405, which suggests the following:

- DBSCAN was able to detect clusters that were moderate evident using other algorithms, especially clusters with varied densities. It successfully managed to identify small groups of shoppers who exhibited highly specific browsing behaviors.

Insights:

- DBSCAN highlighted certain niche behaviors in the dataset, such as customers who might be looking for very specific products or information but did not follow general browsing patterns.
- However, the noise and overlapping areas show that this method might not be the most suitable for a dataset with such varied customer behavior, where not all clusters follow a dense structure.

K-means (K-means Clustering)

- **Best Number of Clusters**: 8
- **Silhouette Score**: 0.6048

K-means clustering works by partitioning the dataset into **k clusters** by minimizing the variance within each cluster. After applying the elbow method to determine the optimal number of clusters, **8 clusters** were selected, leading to a silhouette score of 0.6048, which is significantly higher than DBSCAN's score. This suggests the following:

- K-means was able to segment online shoppers into **well-separated clusters**, indicating a relatively clear distinction between shopper behaviors. The algorithm effectively identified groups with similar browsing patterns, such as shoppers who spent a considerable amount of time on product pages or those with frequent exits (high bounce rates). The clustering captured diverse shopper intentions, including potential buyers, casual browsers, and highly engaged users.

Insights:

- The **8 clusters** provide a solid foundation for segmenting shoppers. These clusters revealed a variety of purchasing intentions, ranging from casual browsers to potential buyers with strong engagement on product-related pages.
- K-means performed well in distinguishing major shopper groups but might have missed some more nuanced shopper behaviors, which could be better captured by other methods.

Hierarchical Clustering

- **Best Structure**: Dendrogram Visualized
- **Silhouette Score**: 0.7614

Hierarchical clustering outperformed both DBSCAN and K-means, achieving the highest silhouette score of **0.7614**, indicating strong cluster separation and cohesion. This method creates a tree-like structure of clusters, enabling the analysis of clusters at different levels of granularity. The high silhouette score suggests the following:

- Hierarchical clustering captured the underlying structure of the dataset more effectively than K-means and DBSCAN. The dendrogram visualization revealed clear relationships between clusters, providing insights into how certain shopper behaviors evolved and grouped together. Hierarchical clustering was particularly effective in capturing both large, well-defined clusters and smaller, more specific clusters of shoppers.

Insights:

- **Hierarchical clustering** uncovered well-defined clusters with distinct behavioral patterns. For example, it distinguished between shoppers who only browsed briefly and those who engaged deeply with product pages, indicating strong purchase intent.
- The dendrogram allowed for flexibility in analyzing shopper behavior at different levels of granularity, revealing subtle differences between shoppers with moderate engagement and those showing a higher likelihood of making a purchase.
- The method's superior silhouette score underscores its ability to offer the most accurate segmentation, making it the best choice for understanding detailed shopper behavior in this dataset.

Comparative Insights:

By comparing the silhouette scores and the clustering outputs from each algorithm, it's clear that **Hierarchical clustering** provided the best results, offering well-separated and cohesive clusters that accurately represented the underlying shopper behaviors. The dendrogram provided rich insights into how different groups of shoppers interacted with the website, making it a powerful tool for customer segmentation.

- **K-means** performed well in identifying distinct clusters, but the lower silhouette score indicates that some groups were less distinct than those found by hierarchical clustering.
- **DBSCAN** was able to detect clusters with unusual patterns and noise points, but the presence of noise and overlapping clusters limited its effectiveness compared to the other two methods.

Overall, the **comparative analysis** shows that while each algorithm has its strengths, **hierarchical clustering** provides the most precise segmentation for the "Online Shoppers' Intention" dataset, making it the most suitable for businesses looking to understand and target distinct shopper groups based on behavior.

Conclusion:

In conclusion, hierarchical clustering proved to be the most effective method for analyzing online shoppers' purchasing intentions, achieving the highest silhouette score of 0.7614. It revealed distinct shopper segments, providing actionable insights into customer behavior. K-means also offered valuable insights with a good balance between computational efficiency and cluster quality, while DBSCAN highlighted more complex, niche patterns but struggled with noise. By utilizing all 18 features and employing PCA for visualization, this project demonstrates how machine learning can aid e-commerce businesses in targeting specific shopper groups to improve marketing strategies and conversion rates.