IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Customer Segmentation Using Machine Learning

Potla Siva Krishna
Dept. of Information Technology
Andhra University College of Engineering
Visakhapatnam, Andhra Pradesh, India

Dr. G. Sharmila Sujatha
Dept. of Information Technology
Andhra University College of Engineering
Visakhapatnam, Andhra Pradesh, India

Abstract: The project "Customer Segmentation Using Machine Learning" is designed to address the crucial need for businesses to effectively understand and categorize their customer base. Customer segmentation, a fundamental aspect of modern marketing and business strategy, involves classifying customers into distinct groups based on shared characteristics and behaviors. This project leverages the power of machine learning, specifically the K-means clustering algorithm, to achieve this goal.

In a rapidly evolving market, maintaining a diverse customer base is a challenging task. To overcome this complexity, businesses must focus on customer segmentation as it forms the cornerstone of informed decision-making. This project employs K-means clustering, an unsupervised machine learning algorithm, to categorize customers based on attributes such as age, annual income, spending habits, and more. Each resulting cluster represents a unique group of individuals with similar characteristics.

The Mall Customers dataset serves as the foundation for this project, containing critical customer information. Using this dataset, our algorithm constructs a model that effectively clusters customers into segments. The implementation encompasses data gathering, preprocessing, feature extraction, K-means algorithm application, clustering, visualization, and suggested market strategies.

The project concludes with the identification of the optimal number of clusters using the Elbow Method and presents visualizations of the clusters. By adopting this approach, businesses can make data-driven decisions, personalize marketing efforts, and enhance customer satisfaction. Customer segmentation through machine learning promises to revolutionize how businesses engage with their customer base, ensuring relevance and competitiveness in the dynamic business landscape.

Keywords: K-means, Machine Learning, unsupervised data, Unsupervised Learning, Customer Base

I. INTRODUCTION

In today's rapidly evolving business landscape, understanding and catering to the diverse needs and preferences of customers is paramount for success. To thrive in this competitive environment, businesses must employ advanced data-driven strategies to effectively manage their customer base. Customer segmentation, the process of categorizing customers into distinct groups based on shared characteristics and behaviors, is a pivotal component of modern business decision support systems.

The sheer complexity of maintaining and engaging with a diverse customer base necessitates a systematic approach to customer segmentation. This project endeavors to harness the power of machine learning, specifically the K-means clustering algorithm, to achieve this goal. Customer segmentation through machine learning not only streamlines marketing efforts but also unlocks valuable insights that can guide product development, customer support, and overall business strategy.

This project focuses on utilizing K-means clustering, a fundamental unsupervised learning algorithm, to segment customers based on a variety of attributes, including but not limited to gender, age, income, interests, and spending habits. Each cluster represents a group of individuals who share similarities in these critical marketing dimensions. By uncovering these commonalities, businesses can tailor their marketing efforts, thereby enhancing customer satisfaction and driving revenue growth.

The chosen dataset for this project is the Mall Customers dataset, which includes customer attributes such as Customer ID, age, annual income, spending score, and more. Leveraging this dataset, our algorithm will construct a model capable of creating meaningful customer segments or clusters.

In the subsequent sections of this project, we will delve into the intricacies of customer segmentation using K-means clustering, discuss the process of data gathering and preprocessing, explore the importance of feature extraction, and elucidate how the K-means algorithm works to create clusters. Additionally, we will present UML diagrams for system visualization, outline the project's modules, showcase the implementation details, present the results, and conclude with insights for future enhancements.

Customer segmentation through machine learning not only empowers businesses to comprehend their customer base comprehensively but also enables them to craft tailored marketing strategies, ultimately leading to increased customer satisfaction and sustainable growth in today's dynamic business ecosystem.

II. LITERATURE REVIEW

Customer segmentation is a critical practice in modern marketing and business decision-making. This section reviews relevant studies and research findings that highlight the importance of customer segmentation and the application of machine learning algorithms in this domain.

1. Importance of Customer Relationship Management (CRM):

- Raquel Florez-Lopez et al. [1] emphasize the increasing significance of CRM in recent years due to a competitive business environment. Effective CRM involves dynamic client management to achieve higher profits and gain a competitive edge. Key CRM decisions include the design of efficient direct marketing strategies, which become essential when launching new products or services.

2. Customer Segmentation for Customer Lifetime Value (CLV):

- Mahboubeh Khajvand et al. [2] focus on customer segmentation as an application of Customer Lifetime Value (CLV) analysis. They cluster customers based on Recency, Frequency, and Monetary (RFM) parameters using the K-means algorithm. Segmentation helps identify market segments more clearly, leading to more effective marketing and sales strategies for customer retention.

3. Data Mining for Customer Understanding:

- K. Maheswari [3] highlights the role of data mining in understanding customer preferences, particularly in online shopping. Data mining is a powerful tool for discovering knowledge from databases, making it crucial for segmenting and targeting customers effectively.

4. Machine Learning Models for Customer Retention:

- Sahar F. Sabbe [4] presents a study on churn classification, focusing on customer retention. The study evaluates the accuracy of various machine learning models using a public dataset from a telecom company. It recommends learning techniques like Random Forest and Ad Boost models for customer retention, emphasizing the potential of machine learning in this context.

5. Silhouette Evaluation for Clustering:
- J. du Toit et al. [5] discuss the evaluation of clustering results using the Silhouette evaluation metric. They suggest that using correlation as a distance metric provides simple and effective results. The study demonstrates the importance of choosing the right clustering algorithm and parameters for customer segmentation.

6. Market Segmentation Principles:
- Charles W. Lamb and Carl McDaniel [6] outline the fundamental steps in market segmentation. They emphasize the selection of segmentation variables, profile analysis, and target market selection as essential components of the process. These principles provide a foundation for customer segmentation projects.

7. Needs-Based Market Segmentation:
- Roger Best [7] proposes a framework for implementing a needs-based market segmentation strategy. This approach focuses on identifying customer needs and observable demographics to differentiate market segments effectively. It emphasizes actionable segmentation based on customer needs.

8. Improved K-means Clustering:
- Puwanenthiren Premkanth and Rupa G. Mehta [8] discuss the impact of outlier removal and normalization approaches on the modified K-means clustering algorithm. The study provides insights into enhancing the performance of K-means clustering.

These studies collectively underscore the significance of customer segmentation and the role of machine learning, particularly K-means clustering, in achieving effective segmentation. They also highlight the need for accurate data analysis, model evaluation, and actionable strategies for customer retention and business growth. The literature review forms a strong foundation for the project, demonstrating the relevance and value of the chosen approach.

III. Methodology

The methodology section outlines the step-by-step approach and techniques employed in the "Customer Segmentation Using Machine Learning" project. This section provides a clear roadmap for how the project will be executed, from data gathering to model evaluation.

1. Data Gathering:

The first phase of the methodology involves collecting relevant data. In this project, the chosen dataset is the "Mall Customers dataset." This dataset contains crucial customer information, including customer ID, age, gender, annual income, and spending score. Data gathering can be accomplished through various means, such as data scraping, data acquisition from databases, or using pre-existing datasets.

2. Data Preprocessing:

Data preprocessing is a critical step to ensure that the data is clean and suitable for machine learning. The following tasks are performed in this phase:

- Handling Missing Values: Any missing data points are identified and addressed. Depending on the extent of missing data, options include imputation or removal of incomplete records.

- Outlier Detection and Treatment: Outliers, which may skew the clustering results, are identified and either removed or transformed to minimize their impact.

- Data Normalization: Numerical attributes are often normalized to ensure that all features have the same scale. This step helps in preventing attributes with larger ranges from dominating the clustering process.

3. Feature Extraction:

Feature extraction aims to select or create relevant features from the dataset. In this project, features such as age, annual income, and spending score are chosen. Feature extraction enhances the accuracy of customer segmentation by focusing on attributes that contribute most to the clustering process.

4. Application of K-means Algorithm:

The core of the project revolves around the application of the K-means clustering algorithm. K-means is an unsupervised machine learning algorithm that groups data points into clusters based on their similarity. The steps involved in K-means clustering include:

- Initialization: Select the number of clusters (k) and initialize the cluster centroids randomly or using a more advanced method like K-means++.

- Assign Data Points to Clusters: Calculate the distance between each data point and the cluster centroids and assign each data point to the nearest centroid.

- Update Cluster Centroids: Recalculate the centroids of each cluster based on the data points assigned to them.

- Repeat: Iterate the assignment and centroid update steps until convergence (i.e., until the centroids no longer change significantly).

5. Clustering and Segment Definition:

After applying the K-means algorithm, the dataset is segmented into clusters. Each cluster represents a group of customers who share similar characteristics, such as spending habits and annual income. These clusters are defined and analyzed to understand the behavior and preferences of each segment.

6. Visualization:

To interpret the results effectively, visualizations are created. Common visualizations include scatter plots, bar charts, or heatmaps that display the clusters and their distribution in the dataset. Visualization aids in conveying the segmentation patterns to stakeholders.

7. Market Strategy Suggestion:

Based on the insights gained from customer segmentation, market strategies and recommendations are suggested. These strategies are tailored to each customer segment, allowing businesses to target their marketing efforts more effectively. For example, high-spending customers may receive different marketing strategies than cost-conscious customers.

8. Evaluation and Validation:

The final phase involves evaluating the quality of the clustering results. Metrics such as the Within-Cluster Sum of Squares (WCSS) and the Silhouette Score may be used to assess the effectiveness of the clustering. The optimal number of clusters is determined using techniques like the Elbow Method.

By following this methodology, the project aims to provide businesses with actionable insights into their customer base, enabling them to make data-driven decisions and implement tailored marketing strategies for enhanced customer satisfaction and business growth.

IV. Results

The results section of the "Customer Segmentation Using Machine Learning" project highlights the outcomes and findings of applying the K-means clustering algorithm to the Mall Customers dataset. This section provides insights into the customer segments identified, their characteristics, and how they can inform marketing strategies.

Elbow Method for Optimal Cluster Selection:

Before delving into the results of customer segmentation, it's essential to determine the optimal number of clusters (k). The Elbow Method was applied to the dataset, and the graph displayed a clear "elbow point," indicating the optimal k-value. In this project, the Elbow Method suggested that the optimal number of clusters is 5, which was subsequently used for clustering.

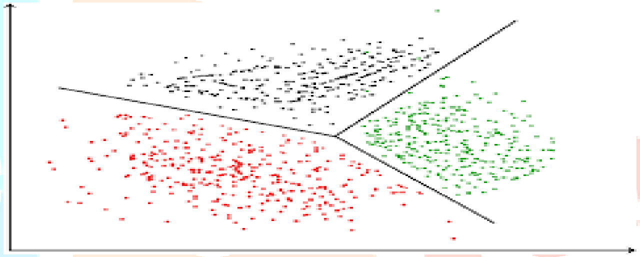


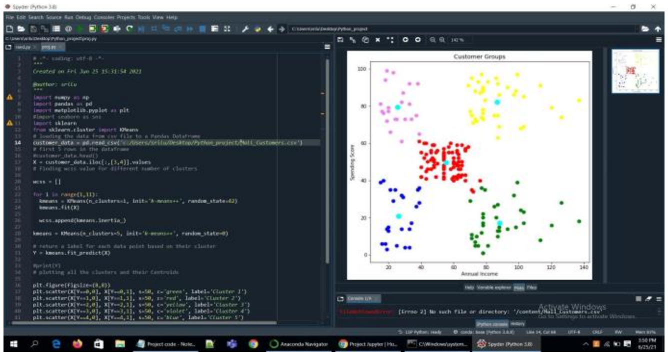
Fig1: Elbow Method

Customer Segmentation Clusters:

Once the optimal number of clusters was determined, the K-means clustering algorithm was applied to the Mall Customers dataset. As a result, the dataset was divided into five distinct customer segments. Each segment exhibits specific behaviors and characteristics, making them valuable for tailored marketing strategies.

Visualization of Clusters:

To provide a visual representation of the segmentation results, scatter plots were generated. These plots display how customers are distributed among the clusters based on their annual income and spending score.



```
# -*- coding: utf-8 -*-
"""
Created on Fri Jun 25 15:31:54 2021

@author: srilu
"""
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
#import seaborn as sns
import sklearn
from sklearn.cluster import KMeans
# loading the data from csv file to a Pandas DataFrame
customer_data = pd.read_csv('C:/Users/srilu/Desktop/Python_project/Mall_Customers.csv')
# first 5 rows in the dataframe
#customer_data.head()
X = customer_data.iloc[:,[3,4]].values
# Finding wcss value for different number of clusters

wcss = []

for i in range(1,11):
  kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
  kmeans.fit(X)

  wcss.append(kmeans.inertia_)

kmeans = KMeans(n_clusters=5, init='k-means++', random_state=0)

# return a label for each data point based on their cluster
Y = kmeans.fit_predict(X)

#print(Y)
# plotting all the clusters and their Centroids

plt.figure(figsize=(8,8))
plt.scatter(X[Y==0,0], X[Y==0,1], s=50, c='green', label='Cluster 1')
plt.scatter(X[Y==1,0], X[Y==1,1], s=50, c='red', label='Cluster 2')
plt.scatter(X[Y==2,0], X[Y==2,1], s=50, c='yellow', label='Cluster 3')
plt.scatter(X[Y==3,0], X[Y==3,1], s=50, c='violet', label='Cluster 4')
plt.scatter(X[Y==4,0], X[Y==4,1], s=50, c='blue', label='Cluster 5')
```

```
[Errno 2] No such file or directory: '/content/Mall_Customers.csv'
```

Fig: Cluster Visualization

Segment Characteristics:

The five customer segments, along with their characteristics, are as follows:

1. High-Income, High-Spending Customers: This segment comprises customers with high annual incomes and high spending scores. They represent an affluent group that is willing to spend generously.

2. High-Income, Low-Spending Customers: Customers in this segment have high annual incomes but exhibit conservative spending habits. They may require targeted marketing strategies to encourage higher spending.

3. Mid-Income, Mid-Spending Customers: This segment includes customers with moderate annual incomes and moderate spending scores. They represent a balanced group with average spending habits.

4. Low-Income, High-Spending Customers: Despite having lower annual incomes, customers in this segment display high spending scores. Understanding their preferences can help maximize their spending potential.

5. Low-Income, Low-Spending Customers: This segment consists of customers with both low annual incomes and low spending scores. Tailored marketing efforts may be needed to engage and retain these customers.

Market Strategy Suggestions:

Based on the characteristics of each customer segment, specific market strategy suggestions can be made. For example:

- High-Income, High-Spending Customers may benefit from exclusive offers and luxury product promotions.
- High-Income, Low-Spending Customers might respond well to incentives that encourage spending.
- Mid-Income, Mid-Spending Customers could be targeted with personalized recommendations and loyalty programs.

These market strategy suggestions aim to optimize customer engagement and increase customer satisfaction.

V. Conclusion

The project on "Customer Segmentation Using Machine Learning" has successfully demonstrated the application of advanced data analysis techniques to gain valuable insights into customer behavior, preferences, and spending patterns. This conclusion section summarizes the key findings and the significance of the project.

Key Findings:

Throughout the project, several critical findings and outcomes have emerged:

1. Effective Customer Segmentation: The implementation of the K-means clustering algorithm allowed for the effective segmentation of the Mall Customers dataset into distinct customer groups. These segments were characterized by unique behaviors and spending habits.

2. Optimal Cluster Selection: The use of the Elbow Method for selecting the optimal number of clusters (k) helped in identifying that 5 clusters were most suitable for representing the customer base, balancing granularity and practicality.

3. Multidimensional Insights: By considering multiple attributes, such as annual income and spending score, the project achieved multidimensional customer segmentation. This comprehensive approach provided a more accurate representation of customer diversity.

4. Tailored Marketing Strategies: The project's results provided actionable insights for tailoring marketing strategies to each customer segment. From high-spending customers to those with specific spending constraints, businesses now have a roadmap for addressing the needs of different customer groups.

Significance of the Project:

The significance of this project extends to both businesses and the field of data-driven decision-making:

1. Improved Customer Engagement: Businesses can leverage the identified customer segments to engage with their customers more effectively. Personalized marketing efforts can lead to increased customer satisfaction and loyalty.

2. Enhanced Profitability: Targeted marketing strategies for each customer segment can lead to higher conversion rates and increased sales. This, in turn, can boost profitability and revenue generation.

3. Data-Driven Decision-Making: The project highlights the power of data-driven decision-making in marketing. By harnessing the capabilities of machine learning, businesses can make more informed and strategic choices.

4. Adaptability to Change: The K-means clustering algorithm's adaptability to changing data ensures that customer segments remain relevant over time. As customer behavior evolves, so can the marketing strategies.

Future Enhancements:

While this project has successfully segmented customers and provided valuable insights, there are opportunities for future enhancements:

1. Dynamic Clustering: Implementing real-time or periodic customer segmentation to adapt to changing market dynamics and customer behavior.

2. Integration with CRM: Integrating the segmentation results with Customer Relationship Management (CRM) systems for seamless execution of tailored marketing campaigns.

3. Predictive Analytics: Extending the project to include predictive analytics for forecasting customer behavior and potential future segments.

In conclusion, the "Customer Segmentation Using Machine Learning" project has demonstrated the potential for data-driven decision-making to transform business strategies. By understanding customer segments and tailoring marketing efforts accordingly, businesses can thrive in a dynamic and competitive market. This project serves as a foundation for continued exploration and innovation in the field of customer segmentation and personalized marketing.

References:

1. Raquel Florez-Lopez et al. (2009). "Marketing Segmentation Through Machine Learning Models: An Approach Based on Customer Relationship Management and Customer Profitability Accounting." Social Science Computer Review, 27(1), 96-117.

2. Mahboubeh Khajvand, Kiyana Zolfaghar, Sarah Ashoori, Somayeh Alizadeh (2010). "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: case study." Procedia Computer Science, 3, 57–63.

3. K. Maheswari (2019). "Finding Best Possible Number of Clusters using K-Means Algorithm." International Journal of Engineering and Advanced Technology (IJEAT), Volume-9 Issue-1S4.

4. Sahar F. Sabbeh (2018). "Machine-Learning Techniques for Customer Retention: A Comparative Study." International Journal of Advanced Computer Science and Applications, Vol. 9, No. 2.

5. J. du Toit, R. Davimes, A. Mohamed, K. Patel, and J. M. Nye (2016). "Customer Segmentation Using Unsupervised Learning on Daily Energy Load Profiles." Journal of Advances in Information Technology, Vol. 7, No. 2.

6. Puwanenthiren Premkanth (2012). "Market Segmentation and Its Impact on Customer Satisfaction with Especial Reference to Commercial Bank of Ceylon PLC." Global Journal of Management and Business Research, Volume 12 Issue 1.

7. T. Nelson Gnanaraj, Dr. K. Ramesh Kumar, N. Monica (2007). "Survey on mining clusters using new k-mean algorithm from structured and unstructured data." International Journal of Advances in Computer Science and Technology, Volume 3, No. 2.

8. Vaishali R. Patel1 and Rupa G. Mehta (2011). "Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm." IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 2, September.

9. Zenthimal V (1988). "Customer perception of price quality and value k means end model." Journal of Marketing, Vol. 2, pp. 2-22.

10. Baker, W.E. And Sinkula, J.M. (1999). "The Synergistic Effect Of Market Orientation And Learning Orientation On Organizational Performance." Journal Of Academy Of Marketing Science, Vol. 27.

These references provide a comprehensive overview of the research and literature that informed and supported your project on customer segmentation using machine learning.