Some possibilities for the utilization of machine learning methods for customer segmentation based on consumer habits

Dragana Radojicic[1][0000-0001-7850-2623] and Bojana Milunovic[1]

[1] Faculty of Economics and Business, University of Belgrade
Belgrade, Serbia

dragana.radojicic@ekof.bg.ac.rs,bojanamilunovic12@gmail.com

Abstract. Customer segmentation is the marketing practice of grouping customers according to certain characteristics. This paper presents a thorough exploration of customer segmentation using machine learning techniques, Logistic Regression, and Support Vector Machine (SVM), applied to data obtained from a mall customers database. By labeling the customer groups and analyzing their characteristics to gain deeper insights into their shopping behavior and preferences, the goal is to develop targeted marketing strategies and allocate resources efficiently to meet the specific needs of each customer segment. Applying statistical analyses and data visualization techniques, the study seeks to derive valuable insights from the data and identify discernible patterns and trends. Utilizing logistic regression yields a remarkable model accuracy of 98%. Subsequently, we employ another machine learning technique for data classification, namely the Support Vector Machine, which achieves an equally notable accuracy of 96%. Using these classification models, potential customers can be effectively converted into loyal ones and enhance the satisfaction of existing customers through tailored marketing strategies for each segment. The research offers insights into effective strategies for distinct customer groups. Applying these methods in a business setting can yield valuable information, forming a basis for informed decision-making and improving customer relationships through customer relationship management strategies.

Keywords: Customer segmentation, cluster analysis, classification.

1 Introduction

1.1 Motivation

Customer segmentation is an essential marketing practice that entails categorizing customers into separate groups based on specific characteristics. The main goal is to tailor strategies for each group to maximize the value extracted from all types of customers, including both highly profitable and less profitable segments. The segmentation process typically includes collecting and analyzing customer data,

determining relevant criteria for division, selecting the most suitable segments, and developing unique marketing approaches for each segment.
Customer segmentation advantages include offering better-tailored products and services, increasing customer satisfaction, and meeting their needs and expectations more effectively. Retaining existing customers is more cost-effective than acquiring new ones, making customer segmentation a valuable retention strategy. By aligning prices with customers' financial capacity, businesses can optimize revenue through price optimization. Moreover, targeted marketing campaigns focused on segmented groups can yield a significantly higher Return on Investment (ROI), with up to 77% of ROI attributed to such initiatives. The utilization of Machine Learning has sparked significant interest from researchers across various fields, see [1], [2], [3], [4]. In recent times, the application of machine learning for customer segmentation has gained popularity as a method to identify customer segments and improve business (see [5], [6], [7], [8], [9]). In order to identify a group of customers with unsatisfied needs, authors in [10] introduce a novel approach that uses machine learning techniques to detect the importance of product features. However, our approach differs from the aforementioned works. A novel approach "Probabilistic Linguistic Group Decision (PLGD)-FlowSort methodology" for modeling customer satisfaction based on online customer reviews is proposed in [11]. The COVID-19 pandemic has raised the use of mobile payment services, and in paper [12] probabilistic linguistic indifference threshold-based attribute ratio analysis (PL-ITARA) is introduced for determining attributes' importance in mobile payment services.

In this paper, we utilize classification models, namely Logistic Regression and Support Vector Machine to train our model to divide customers into different groups. In the first section, we outline the structure of our database and provide a statistical analysis of its contents. Further, in section 2, we give a brief explanation of the machine learning techniques implemented within this research to analyze and classify customers. Section 3 presents the implementation of machine learning algorithms, including a cluster analysis to determine the optimal number of clusters. Within section 4 we propose marketing strategies for each group. Finally, in section 5 we give a conclusion and some possibilities for future research based on the given results.

1.2 **Data in brief**

We conducted an analysis using the Mall customers database[1], which contains information about customers in the shopping center. The dataset consists of 200 observations and includes 5 attributes that characterize the customers as follows:

- CustomerID: A unique identifier for each customer.
- Gender: The gender of the customer (male, female).
- Age: The age of the customer.

[1] https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python, (16.07.2023)

- Annual Income (k$): The annual income of the customer in thousands of dollars.
- Spending Score (1-100): The consumer's spending score is derived from their spending behavior and specific purchase patterns.

The purpose of this analysis is to gain insights into the shopping behavior and preferences of customers in the mall. By examining these attributes, we aim to identify potential patterns and trends that can aid in formulating effective marketing strategies and enhancing the overall shopping experience. Additional statistical analyses and data visualization techniques are utilized to extract valuable insights from the data.

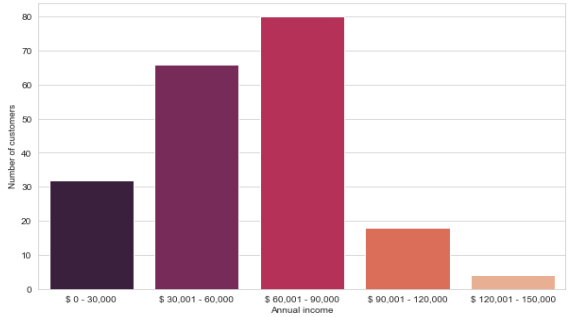

Fig. 1. Structure of customers by annual income

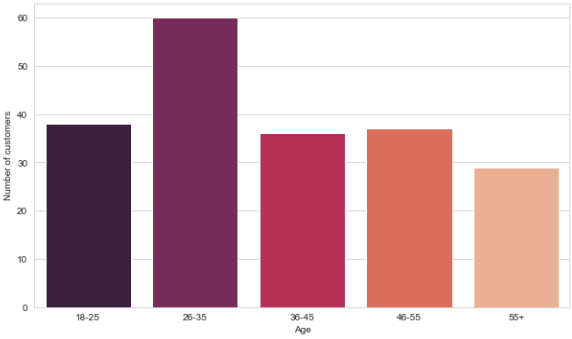

Fig. 2. Structure of customers by age

Notably, most customers belong to a younger demographic with lower annual incomes (Fig.2 and Fig.1, respectively), which could be valuable in devising sustainable business strategies.

The goal is to group customers based on similar characteristics to understand their behavior better and tailor marketing strategies accordingly.

1.3 **Data analysis, cluster analysis**

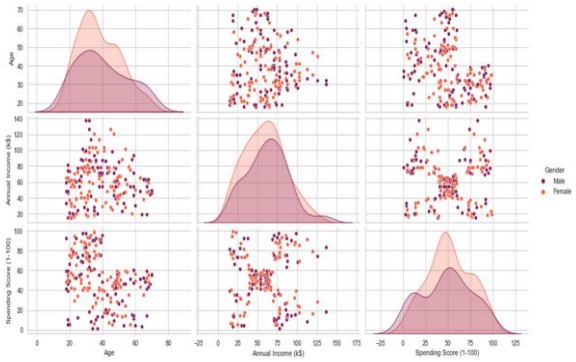

Fig. 3. Plot pairwise relationships of age, annual income, and spending score

Given that we have discarded gender as a segmentation criterion, we attempt to visualize the remaining three characteristics. As a result, gender will not be considered as a differentiating characteristic in our customer segmentation efforts. However, a visual examination of the dataset indicates the potential existence of distinct clusters based on annual income and spending scores (Fig.3). Additionally, though less significant, there are indications of two to three potential clusters when considering customers' ages.

From the provided display (Fig.4), we cannot observe any clear division among our customers, necessitating cluster analysis to determine desired groups.

Indeed, we will create graphs to depict the remaining attributes (customer age, annual income, and spending score) and proceed with cluster analysis to identify distinct customer groups. Through this analysis, we aim to group customers based on similar characteristics and behaviors, gaining insights into their preferences and needs.

Once the cluster analysis is complete, we can label the customer groups and further examine their characteristics to gain a better understanding of their behaviors. This will serve as a foundation for devising targeted marketing strategies and allocating resources effectively to the appropriate customer segments, thereby enhancing business efficiency and customer satisfaction.

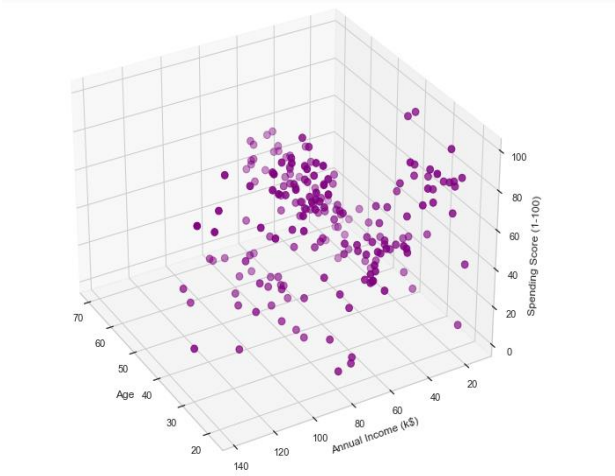

Fig. 4. 3D plot of age, annual income, and spending score

2 Machine learning algorithms

Classification techniques leverage training data to predict the probability of new data points belonging to specific categories, using insights gained from the training process. Particularly for this task, we employ statistical models for classification, namely Logistic regression and Support Vector Machine. The main purpose of the logistic regression method is to calculate the likelihood of an input point being associated with a particular class, considering its set of features. A support vector machine aims to find the best hyperplane that separates different classes within the feature space.

2.1 Logistic Regression

Logistic Regression is a machine learning method used to predict a binary outcome. This technique is based on a logistic model and it has two classes as possible outcome variables. Using the logistic function (sigmoid function) the algorithm produces a probability score between 0 and 1, which represents the probability that a particular data point belongs to one of the two predefined classes considering 0.5 as the threshold. Logistic Regression has successfully been used in different areas. For a further description of the Logistic regression concept see [13], [14], [15], [16], [17].

2.2 **Support Vector Machine**

A support vector machine (SVM) is a supervised machine learning method for classification and regression purposes, its role is to identify the optimal hyperplane that most effectively divides distinct classes within the feature space.

SVM is particularly effective for tasks involving binary classification, but in cases when the dataset is unbalanced, SVM may have suboptimal performance. For a further description of the SPV concept see [13], [14], [15].

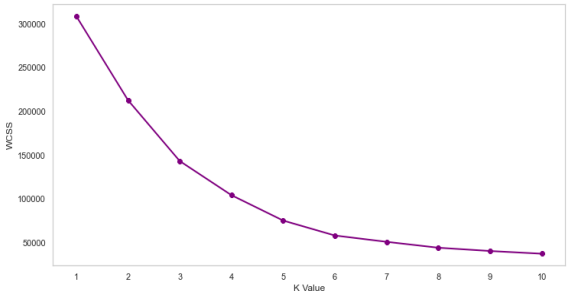

Fig. 5. Within-Cluster Sum of Square for different numbers of clusters

3 **Measurements, machine learning algorithm implementations**

Within this section, we first conduct a cluster analysis within which we determine the optimal number of clusters. After that, we apply classification algorithms and measure their performance.

3.1 **Cluster Analysis**

Before moving on to the cluster analysis, it's essential to standardize and prepare the data. Firstly, we determine the optimal number of clusters for our dataset. Using the Elbow method (Fig.5), we decide on 5 clusters and continue with the analysis. In particular, we apply the K-means algorithm to group similar data points together. Plotting the results (Fig.6), we can see that this method gave us five spatially separated groups of customers.

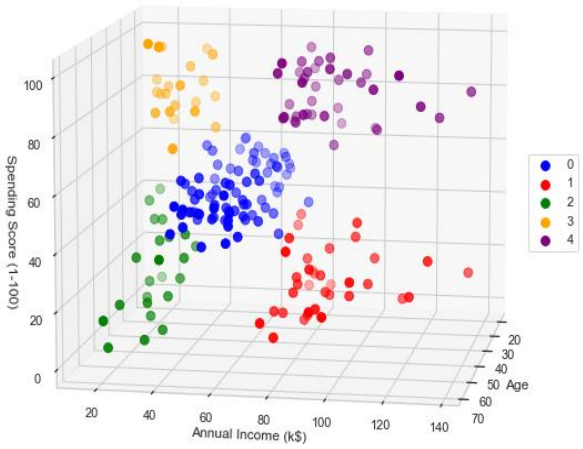

Fig. 6. Plot obtained by cluster analysis

3.2 **Training and testing the model**

Initially, we divide the dataset into training data and testing data (75:25). We apply multiclass logistic regression, resulting in a remarkable model accuracy of 98%, indicating high performance. Next, we explore another machine learning approach, Support Vector Machine (SVM), for data classification. Once again, we achieve a notably high accuracy of 96%.

3.3 **PCA method**

Principal component analysis (PCA) is a method that is often used to reduce the dimensionality of large datasets, by transforming a large set of variables into a smaller one, while preserving the most significant information and variations in the data. Principal components are new variables that are constructed as linear combinations of the initial variables. They are ranked based on their eigenvalues, with the first component capturing the most variance and the subsequent components in decreasing order of variance. The total variance explained by all components should be between 70% to 80% variance. To assess the significance of the "Age" attribute, we apply PCA. If the significance proves to be low, we can exclude the attribute from consideration, simplifying the overall analysis.

PCA results confirm our assumption that the "Age" attribute has a relatively lower impact on customer segmentation (10.76%) compared to the other two attributes (around 45% each), see Fig.7. As a result, we can simplify the overall picture of a customer segmentation problem using two remaining variables.

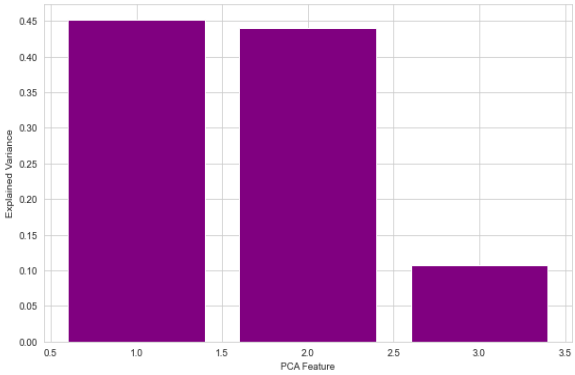

Fig. 7. Feature explained variance

This dimensionality reduction technique aids in gaining a clearer understanding of customer groupings while maintaining a meaningful representation of the data. By visualizing the clusters in a 2D space, we can further comprehend the distinct patterns and relationships among customers.

This method acts as a preliminary step before visualizing data. It helps to spot patterns and groups within the data, which makes it easier to grasp the underlying structure. This understanding then paves the way for moving on to the clustering process.

When it comes to the clustering process, it is important that PCA can reduce the dimensionality of the data by transforming it into a lower-dimensional space while retaining most of the variability, thus clustering algorithms can perform in the most efficient way.

Our aim of a dimensionality reduction technique is to gain better insights in the underlying structure of data so that we can derive better strategies that are closely aligned with our customers.

Customer segmentation plays a vital role in the field of e-commerce, enabling businesses to approach each consumer uniquely and incentivize their purchasing behavior. By obtaining voluntary consent from consumers through online surveys, we can collect personal data, and annual income range information and through specified questions determine their buying behavior. Leveraging the gathered data, we can classify them into one of five pre-defined groups. Based on this classification, we can target each potential customer individually with a customized marketing campaign that is given in the next section, aiming to engage and convert them from potential prospects into loyal customers.

4 Derived Strategies

As evident from the previous visualization (Fig.6), we have successfully segmented customers into five distinct groups. Leveraging these segments, we can implement targeted marketing strategies to optimize customer engagement and sales. The identified segments and suggested business strategies are as follows:

- Segment 0 - Customers with moderate income and spending scores: The goal is to uplift them to a higher segment by promoting additional products/services. As this segment constitutes the largest customer base, they are a significant target audience.
- Segment 1 - Customers with high income and low spending scores: Special offers and quantity discounts should be offered to incentivize purchases.
- Segment 2 - Customers with low income and low spending scores: This segment should have the lowest marketing priority. Investments should focus on the other four segments until all marketing techniques have been exhausted.
- Segment 3 - Customers with low income and high spending scores: These customers should be treated similarly to those with high incomes. They exhibit a stable cash flow and can be offered similar benefits. Regular contact, especially during discounts, along with better credit payment terms, can enhance their loyalty.
- Segment 4 - Customers with high income and high spending scores: This segment should be the top priority target group. They possess the highest potential for high-value purchases and frequent buying. Loyalty programs, exclusive VIP benefits, and priority access to products are suitable for this category. Regular communication via email and messages, keeping them informed about new and exclusive products, will enhance their engagement.

5 Conclusion and future research

In conclusion, customer segmentation based on the given dataset and subsequent analysis enables businesses to better understand their customer base and tailor their marketing efforts accordingly. By implementing the proposed strategies for each segment, companies can significantly improve customer satisfaction, loyalty, and overall business performance.

After implementing customer segmentation, this research offers valuable insights into consumer behavior and preferences. The characteristics that we excluded from our research, like gender and age, can be used for demographic market segmentation with the intention of making more direct paths to our customers.

References

1. Li Y, Jiang W, Yang L, Wu T. On neural networks and learning systems for business computing. Neurocomputing. 2018 Jan 31;275:1150-9.
2. Sharifani K, Amini M. Machine Learning and Deep Learning: A Review of Methods and Applications. World Information Technology and Engineering Journal. 2023;10(07):3897-904.

3. Nti IK, Quarcoo JA, Aning J, Fosu GK. A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. Big Data Mining and Analytics. 2022 Jan 25;5(2):81-97.
4. Milakovic A, Draskovic D, Nikolic B. Visual Simulator for Mastering Fundamental Concepts of Machine Learning. Applied Sciences. 2022 Dec 17;12(24):12974.
5. Smeureanu I, Ruxanda G, Badea LM. Customer segmentation in private banking sector using machine learning techniques. Journal of Business Economics and Management. 2013 Nov 1;14(5):923-39.
6. Monil P, Darshan P, Jecky R, Vimarsh C, Bhatt BR. Customer segmentation using machine learning. International Journal for Research in Applied Science and Engineering Technology (IJRASET). 2020 Jun;8(6):2104-8.
7. Narayana VL, Sirisha S, Divya G, Pooja NL, Nouf SA. Mall customer segmentation using machine learning. In2022 International Conference on Electronics and Renewable Systems (ICEARS) 2022 Mar 16 (pp. 1280-1288). IEEE.
8. Othayoth SP, Muthalagu R. Customer segmentation using various machine learning techniques. International Journal of Business Intelligence and Data Mining. 2022;20(4):480-96.
9. Dileep PS, Seshashayee M. Customer segmentation using machine learning. International Research Journal of Modernization in Engineering Technology and Science. 2022 May;4(5):3484-7.
10. Joung J, Kim H. Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. International Journal of Information Management. 2023 Jun 1;70:102641.
11. Darko AP, Liang D. Modeling customer satisfaction through online reviews: A FlowSort group decision model under probabilistic linguistic settings. Expert Systems with Applications. 2022 Jun 1;195:116649.
12. Darko AP, Liang D, Xu Z, Agbodah K, Obiora S. A novel multi-attribute decision-making for ranking mobile payment services using online consumer reviews. Expert Systems with Applications. 2023 Mar 1;213:119262.
13. Robert C. Machine learning, a probabilistic perspective.
14. Witten D, James G. An introduction to statistical learning with applications in R. springer publication; 2013.
15. LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015 May 28;521(7553):436-44.
16. Hilbe JM. Logistic regression models. CRC press; 2009 May 11.
17. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. John Wiley & Sons; 2013 Apr 1.
18. Ma Y, Guo G, editors. Support vector machines applications. New York: Springer; 2014 Mar 3.
19. Christmann A, Steinwart I. Support vector machines.