ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/387907097

Harnessing Machine Learning for Effective Customer Segmentation

Conference Paper · January 2025

DOI: 10.1109/SPARC61891.2024.10829373

CITATIONS
0

READS
40

4 authors, including:

Shikha Singh
Amity School of Engineering & TechnologyAmity University
54 PUBLICATIONS **62** CITATIONS
SEE PROFILE

Vineet Singh
IFTM University
24 PUBLICATIONS **79** CITATIONS
SEE PROFILE

All content following this page was uploaded by Shikha Singh on 13 January 2025.

The user has requested enhancement of the downloaded file.

Harnessing Machine Learning for Effective Customer Segmentation

Siddharth Agarwal
Department of CSE,
Amity School of Engineering and Technology, Amity University (U.P),
Lucknow Campus, India
siddharth.agarwal@s.amity.edu

Shikha Singh
Department of CSE,
Amity School of Engineering and Technology, Amity University (U.P),
Lucknow Campus, India
ssingh8@lko.amity.edu

Bramah Hazela
Department of CSE,
Amity School of Engineering and Technology, Amity University (U.P),
Lucknow Campus, India
bhazela@lko.amity.edu

Vineet Singh
Department of CSE,
Amity School of Engineering and Technology, Amity University (U.P),
Lucknow Campus, India
vsingh@lko.amity.edu

2024 International Conference on Signal Processing and Advance Research in Computing (SPARC) | 979-8-3503-8520-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/SPARC61891.2024.10829373

***Abstract*— The paper's objective is to utilize machine learning techniques for customer segmentation. We aim to identify significant patterns and segments within a customer base to enhance marketing strategies, improve customer satisfaction, and optimize business operations. By mastering Python programming alongside essential libraries like Pandas, NumPy, and data visualization tools, we focus on data collection, training, and testing to construct a robust customer segmentation model. Such a model will enable businesses to enhance their marketing strategies, customer satisfaction, and streamline operations. This paper delves into customer segmentation using k-means clustering, analysing the Mall Customers dataset to demonstrate how machine learning can effectively segment customers created on their purchasing behaviour and demographic information. The study includes an extensive review of literature, detailed research methodology, implementation of the clustering algorithm, evaluation of models, and discussion of the results. The model's accuracy is verified using two metrics: the Silhouette score, which is 0.69777, and the Davies-Bouldin Index (DBI), which is 0.42110.**

Keywords*— *Machine learning, Customer segmentation, Elbow method, K-means clustering, Silhouette score, Davies-Bouldin Index (DBI)

I. Introduction

Customer segmentation, the manner of dividing a company's customers into groups with similar characteristics, has become essential in modern business. Evaluating the important needs and inclinations of different customer segments helps businesses create targeted marketing strategies, enhance customer satisfaction, and boost overall profitability. The rise of machine learning has transformed customer segmentation, providing more advanced and precise methods for analysing customer data.

Machine learning techniques, particularly clustering algorithms, are widely used for customer segmentation. Clustering is an unsupervised learning method that assembles data points based on similarities. Among various clustering techniques, K-means clustering is popular due to its clarity and efficiency. It subsets data into K clusters, with each data point belonging to the cluster with the closest mean. This algorithm has been extensively applied in marketing, retail, and finance to identify distinct customer segments and generate actionable insights [1].

The importance of customer segmentation in business can't be overstated. Recent advancements in clustering algorithms and their applications in customer segmentation emphasize the growing role of machine learning in understanding customer behaviour and improving business strategies [5]. Segmentation allows businesses to allocate resources more effectively, create tailored marketing campaigns, and improve customer retention rates. Effective segmentation leads to a better understanding of customer needs, helping develop products and services that cater to specific segments [7]. Additionally, segmentation enables businesses to identify prominent-value customers and focus on retaining them [8].

The use of K-means clustering in customer segmentation has been widely researched. Studies have shown its usefulness in segmenting customers based on purchasing behaviour [9] and highlighted the importance of selecting appropriate features and preprocessing data to improve clustering performance.

In this paper, we explore customer segmentation using the K-means clustering algorithm applied to the Mall Customers dataset. This dataset includes information about customers' annual income, spending scores, and demographic details such as age and gender. We aim to identify distinct customer segments and provide insights into each segment's characteristics. The verdicts from this study will add to the awareness on customer segmentation and offer practical perceptions for businesses looking to leverage machine learning for better customer understanding and targeted marketing.

The paper commences with an Introduction providing a concise overview of customer segmentation utilizing machine learning. This is followed by a review of literature presenting comprehensive analyses of studies conducted by various authors in section 2. The Research Methodology then walks

979-8-3503-8520-5/24/$31.00 ©2024 IEEE
2024 International Conference on Signal Processing and Advance Research in Computing (SPARC)

Authorized licensed use limited to: AMITY University. Downloaded on January 11,2025 at 06:22:40 UTC from IEEE Xplore. Restrictions apply.

through the key steps taken to develop the model in section 3. In the Implementation of Work in section 4, we explain exactly how we built and trained our model. Subsequently, the Evaluation of Model in section 5 addresses the final testing outcomes and the accuracy of the model. Then in the Results and Discussions, we analyse the data in graphical representations and gain insights. Lastly, in the Conclusion and Future Scope in section 6, we share the valuable insights gained throughout the process and discuss future possibilities for customer segmentation.

II. Literature Review

The application of machine learning in customer segmentation has been widely studied. Early work by Jain and Dubes (1988) [1] introduced the basic concepts of clustering algorithms. More recent studies have demonstrated the effectiveness of K-means clustering in various domains, including retail and marketing (Wu et al., 2008)[2].

Xu and Wunsch (2005)[3] provided a comprehensive review of clustering techniques, highlighting the strengths and limitations of different algorithms. They emphasized the importance of selecting appropriate features and preprocessing data to improve clustering performance. Tibshirani et al. (2001)[4] introduced the Gap Statistic, a method for determining the optimal number of clusters, which has become a standard approach in evaluating clustering models. Recent advancements in clustering algorithms and their applications in customer segmentation have been discussed by Berkhin (2006)[5] and Aggarwal and Reddy (2013)[6]. These studies highlight the growing importance of machine learning in understanding customer behaviour and improving business strategies. Kotler (1997)[7] and Wedel and Kamakura (2000)[8] provided foundational insights into the importance of customer segmentation in marketing. McLachlan et al. (2004)[9] explored the application of K-means clustering in segmenting customers based on purchasing behaviour, while Punj and Stewart (1983)[10] discussed the importance of feature selection and data preprocessing in achieving meaningful segmentation results. Also grocery stores in Kenya can enhance their competitive ever-changing retail market, build stronger connections with their customers, and boost their business growth [11].

Recent studies have extended these concepts into more specialized applications of customer segmentation using machine learning. For instance, Sukru Ozan (2018) [12] investigated customer segmentation using various machine learning methods. Patankar et al. (2021) [13] provided insights into customer segmentation through machine learning within parallel computing frameworks. Pyla and Seshashayee (2022) [14] presented a detailed analysis of customer segmentation using machine learning in the International Research Journal of Modernization in Engineering Technology and Science.

Further contributions include Regmi et al. (2022) [15], who focused on customer market segmentation using machine learning algorithms, and Abidar et al. (2020) [16], who proposed a new strategy for targeted actions in customer segmentation. Lewaa (2023) [17] applied RFM analysis in customer segmentation, offering a contemporary perspective on segmentation strategies. Papetti and Thompson (2019) [18] explored customer segmentation in the cannabis retail sector using machine learning, while Parab and Dave (2023) [19] provided a comprehensive study on customer segmentation using machine learning in IJNRD. Sathyanarayana et al. (2023) [20] implemented clustering methodologies to categorize mall customers, yielding valuable insights into consumer behaviour patterns and supporting more informed business decision-making processes.

III. Research Methodology

Our research methodology involves several key steps as shown in the flow chart in Figure 1:

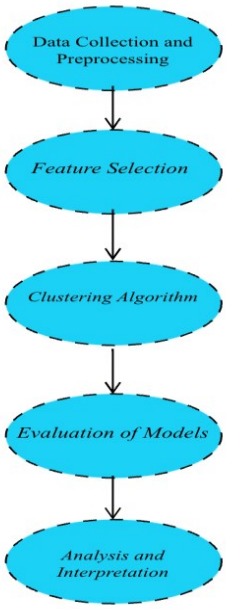

Fig 1 Reasearch Methodology Flowchart

A. Data Collection and Preprocessing

We use the Mall Customers dataset, which includes 2000 entries of customer information. The dataset is well pre-processed to handle the missing values and easily normalize numerical features.

- Check for Missing Values: Makes sure the dataset does not contain any missing values.
- Gender Encoding: Convert the 'Gender' column from categorical to numerical values (e.g., Male = 0, Female = 1) using label encoding or one-hot encoding.
- Standardization: was applied to 'Annual Income (k$)' and 'Spending Score (1-100)' using Scikit-learn's StandardScaler. This ensured the features had a mean of zero and a standard deviation of one, ensuring comparability in scale. The Standard Scaler was fitted and transformed, resulting in the standardized feature matrix X_scaled, which improved the performance of the K-means clustering algorithm.
- Drop 'CustomerID': This column is not needed for clustering because it is simply a unique identifier and does not provide any meaningful information for grouping customers. Including it could introduce noise and distort the clustering process. Therefore, it is dropped during preprocessing to focus on relevant features like 'Annual Income (k$)' and 'Spending Score (1-100)'.

Authorized licensed use limited to: AMITY University. Downloaded on January 11,2025 at 06:22:40 UTC from IEEE Xplore. Restrictions apply.

B. Feature Selection

Key features for clustering are chosen based on their importance for customer segmentation. These features include age, annual income, and spending score.

C. Clustering Algorithm

We apply the K-means clustering algorithm to segment customers into distinct groups. The algorithm is implemented using Python and popular machine learning libraries such as Scikit-learn.

D. Evaluation of Models

The performance of the K-means clustering model is assessed using metrics like the silhouette score and the Davies-Bouldin Index. These metrics help identify the optimal number of clusters and evaluate the quality of the segmentation..

E. Analysis and Interpretation

The resulting clusters are analyzed to understand the characteristics of each segment. Insights are drawn regarding the demographic and purchasing behaviour of customers in each cluster. To finalize the optimal number of clusters, the elbow method was used, plotting the Within-Cluster Sum of Squares (WCSS) against different cluster numbers. The optimal number of clusters was identified where the WCSS curve flattened, indicating a balance between minimizing intra-cluster variance and avoiding overfitting. The performance of the clustering model was then evaluated using the Silhouette Score and the Davies-Bouldin Index (DBI). A higher Silhouette Score and a lower DBI suggested well-defined and distinct clusters. Visualization through scatter plots of Annual Income and Spending Score, color-coded by cluster assignments, provided insights into the distinct characteristics of each cluster. For example, one cluster might include high-income, high-spending customers, while another might feature younger individuals with lower incomes.

IV. Implementation of Work

The initial step in our implementation involves data collection and preprocessing. The Mall Customers dataset contains 2000 entries, each representing a customer with following attributes: CustomerID, Gender, Age, Annual Income (in thousands of dollars), and Spending Score (ranging from 1 to 100). Data preprocessing is critical to make sure the dataset is clean and ready for analysis. This process includes encoding categorical variables, addressing missing values, and normalizing numerical features.

First, the dataset is inspected for any missing or inconsistent data. Any missing values are handled appropriately to ensure they do not affect the clustering algorithm. The 'Gender' column, a categorical variable, is converted into numerical format using label encoding. This transformation is necessary because the K-means algorithm requires numerical input to compute distances between data points. Following this, numerical features such as 'Age', 'Annual Income (k$)', and 'Spending Score (1-100)' are normalized using standard scaling. Normalization is essential to ensure that all features contribute equally to the distance calculations, as K-means is sensitive to the scale of input data. The scaling process transforms these features to have a mean of zero and a standard deviation of one, making them comparable.

For effective clustering, it is crucial to select features that provide meaningful insights into customer behaviour. In this study, we choose 'Age', 'Annual Income (k$)', and 'Spending Score (1-100)' as the primary features for clustering. These features offer a balanced view of the customers' demographic and purchasing behaviour, which are key indicators for segmentation. By focusing on these attributes, we aim to uncover distinct patterns and groupings within the customer base.

With the pre-processed data, we apply the K-means clustering algorithm. The first step is to determine the optimal number of clusters (K). To do this, we use the elbow method, which involves running the K-means algorithm for a range of cluster numbers (typically from 1 to 10) and calculating the Within-Cluster Sum of Squares (WCSS) for each K. WCSS measures the compactness of the clusters and decreases as the number of clusters increases.

As illustrated in Figure 2, the WCSS values decrease significantly as the number of clusters increases from 1 to 4. Specifically, the WCSS drops from approximately 4000 for 1 cluster to around 1000 for 4 clusters. After this point, the rate of decrease slows down, indicating diminishing returns from adding more clusters. The "elbow" point in the graph, observed at K=4, suggests that four clusters offer an optimal balance between minimizing intra-cluster variance and avoiding overfitting.

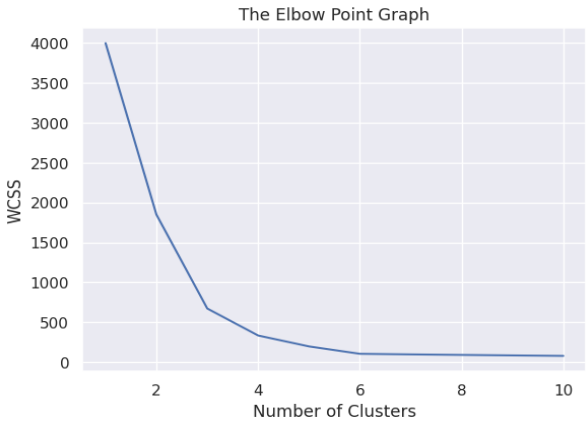

Fig 2 WCSS plot graph showing Elbow point

Once the optimal number of clusters is determined, the K-means algorithm is executed with this specific K value. The algorithm partitions the dataset into K clusters by minimizing the total squared distances between each data point and the centroid of its assigned cluster. Each data point is assigned to the cluster with the nearest centroid, and the centroids are iteratively recalculated until convergence is achieved.

V. Evaluation Of Models

We evaluated the clustering model using two key metrics-the silhouette score and the Davies-Bouldin Index (DBI) which are given in Figure 3:

Authorized licensed use limited to: AMITY University. Downloaded on January 11,2025 at 06:22:40 UTC from IEEE Xplore. Restrictions apply.

A. Silhouette Score: This evaluates the effectiveness of data point clustering. A higher score signifies better-defined clusters. The score is calculated by comparing the average distance of a point to other points within its cluster to the average distance to points in the nearest different cluster.

B. Davies-Bouldin Index (DBI): This metric evaluates how similar each cluster is to the cluster that is most similar to it, on average. A lower DBI signifies higher clustering quality, indicating that the clusters are tightly packed and distinctly separated.

SILHOUETTE SCORE

```
[105] silhouette_avg = silhouette_score(X_scaled, kmeans.labels_)
      print(f'Silhouette Score: {silhouette_avg}')

Silhouette Score: 0.697775129966753
```

DAVIES-BOULDIN SCORE

```
[106] db_index = davies_bouldin_score(X_scaled, kmeans.labels_)
      print(f'Davies-Bouldin Index: {db_index}')

Davies-Bouldin Index: 0.42110159608316533
```

Fig 3 Accuracy scores of the model

After evaluating the optimal number of clusters using the elbow method, we applied K-means and calculated the silhouette score and DBI. The silhouette score indicated well-defined clusters, and the low DBI suggested compact and distinct clusters. These metrics confirmed the effectiveness of K-means clustering in segmenting the Mall Customers dataset into meaningful groups, providing important observations for targeted marketing strategies and improved customer satisfaction.

VI. Results and Discussions

The results from the K-means clustering are analysed to understand the characteristics of each identified customer segment. The clusters are visualized using scatter plots, which display the division of customers based on their annual income and spending score, color-coded by their cluster assignments as shown in Figure 4. These visualizations help in interpreting the distinct groups formed by the algorithm.

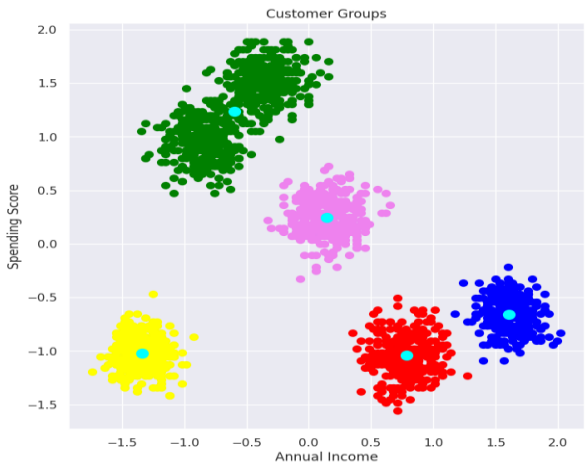

Fig 4 Visualizing all the Clusters by using scatter plots

A. Clustering Results Overview

Applying K-means clustering to the Mall Customers dataset resulted in the creation of distinct customer segments. The clustering process was visualized with scatter plots, showing the distribution of customers based on their annual income and spending scores. Each data point was color-coded according to its cluster assignment, providing a clear visual representation of the customer groupings.

B. Visual Analysis of Clusters

Cluster Visualization: Scatter plots revealed how clusters were distributed in terms of annual income and spending score. For example:

- Cluster 1: This cluster might represent high-income, high-spending customers. The plot would show these customers positioned in the upper right quadrant, indicating high values in both annual income and spending score.
- Cluster 2: This could include low-income, low-spending customers, appearing in the lower left quadrant.
- Cluster 3: A cluster of moderate-income, high-spending customers might be located in the upper middle of the plot.
- Cluster 4: Another cluster might include younger customers with varying income levels but consistent spending patterns, appearing in a different region of the plot.

The scatter plots help in understanding the separation between different customer segments and provide insights into the characteristics that define each cluster.

Cluster Characteristics:

- High-Income, High-Spending Customers: This segment includes affluent customers who spend significantly on their purchases. Marketing strategies for this group could include exclusive offers, high-end product promotions, and loyalty programs to maintain engagement.
- Low-Income, Low-Spending Customers: This group includes customers with lower disposable income and spending patterns. Targeted promotions, budget-friendly products, or discounts might be effective in engaging this segment.
- Moderate-Income, High-Spending Customers: Customers in this segment have moderate income but show high spending behaviour. They might be interested in products offering good value for money or premium options within their budget.
- Young Customers with Varied Income: This segment includes younger customers whose spending patterns differ from other groups. Tailoring marketing campaigns to their preferences, such as digital promotions or social media campaigns, could be beneficial.

Authorized licensed use limited to: AMITY University. Downloaded on January 11,2025 at 06:22:40 UTC from IEEE Xplore. Restrictions apply.

C. Cluster Insights and Implications

1) *Behavioral Patterns:* Each cluster's behavioral patterns reveal significant insights into customer preferences and spending habits. For instance, high-spending clusters might prioritize luxury or premium products, while low-spending clusters might seek value deals or discounts.

2) *Marketing Strategy Recommendations:*

- Personalized Campaigns: Developing personalized marketing strategies for each segment can increase effectiveness. For example, high-income customers might respond better to exclusive, high-value offers, while low-income customers might be more attracted to discounts or sales.
- Product Development: Understanding the distinct needs of each cluster can guide product development. For instance, products tailored for high-spending segments could focus on quality and luxury, while products for low-spending segments might emphasize affordability and value.

3) *Customer Retention and Engagement:*

- High-Income Segments: Implementing loyalty programs and premium services can enhance customer retention for high-income segments.
- Low-Income Segments: Engaging these customers with cost-effective solutions and regular promotions can improve their loyalty and spending.
- Young Customers: Engaging younger customers through digital platforms, social media, and interactive campaigns can boost their involvement and satisfaction.

D. Comparison and Validation

1. Silhouette Score and Davies-Bouldin Index: The evaluation metrics, with a Silhouette score of 0.69777 and a Davies-Bouldin Index of 0.42110, confirm the effectiveness of the clustering. The high Silhouette score suggests that the clusters are well-defined, with data points closely related to their own clusters and distinct from other clusters. The low DBI value specifies that the clusters are well-separated and compact, supporting the validity of the segmentation.
2. Benchmarking Against Previous Studies: The results align with findings from previous studies on K-means clustering in customer segmentation. The effectiveness of the clustering model can be compared with benchmarks from other research, providing context for the current study's results.

E. Practical Applications

1. Strategic Planning: Businesses can use these insights to develop strategic plans tailored to each customer segment. This includes targeted marketing strategies, product offerings, and customer service enhancements.
2. Resource Allocation: The segmentation helps in efficient resource allocation by focusing efforts on high-value customer segments and optimizing marketing expenditures.
3. Customer Experience: By comprehending the distinct needs and behaviors of each segment, businesses can enhance the overall customer experience, resulting in increased satisfaction and loyalty.

VII. Conclusion

This study successfully demonstrated the application of the K-means clustering algorithm for customer segmentation using the Mall Customers dataset. By focusing on key features such as age, annual income, and spending score, we effectively categorized customers into distinct groups, each characterized by specific traits. The analysis provided valuable insights into customer demographics and purchasing behaviors, offering a solid foundation for businesses to develop targeted marketing strategies, enhance customer satisfaction, and optimize resource allocation.

Key findings include:

- The K-means clustering approach was effective in creating meaningful customer segments, as confirmed by metrics like the silhouette score and Davies-Bouldin Index, which indicated well-defined and distinct clusters.
- Proper data preprocessing, including handling missing values, encoding categorical variables, and normalizing numerical features, was crucial for achieving accurate and meaningful clustering results.
- Feature selection played a significant role in ensuring that the clustering outcomes were both interpretable and relevant, capturing essential aspects of customer behavior.

In conclusion, customer segmentation using K-means clustering equips businesses with actionable insights into their customer base, allowing them to tailor strategies more effectively. This approach empowers organizations to make informed, data-driven decisions that drive growth and improve overall customer engagement.

References

[1] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. London, England: Prentice-Hall, 1988.

[2] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.

[3] R. Xu and D. Wunsch 2nd, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, 2005.

[4] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 63, no. 2, pp. 411–423, 2001.

[5] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping Multidimensional Data*, Berlin/Heidelberg: Springer-Verlag, 2006, pp. 25–71.

[6] C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications*. CRC Press, 2013.

[7] "Kotler, P. (1997) Marketing Management Analysis, Planning, Implementation, and Control (9th ed.). Upper Saddle River, NJ Prentice Hall. - references - scientific research publishing," *Scirp.org*. [Online]. Available:

Authorized licensed use limited to: AMITY University. Downloaded on January 11,2025 at 06:22:40 UTC from IEEE Xplore. Restrictions apply.

https://www.scirp.org/reference/referencespapers?referenceid=3000227. [Accessed: 12-Aug-2024].

[8] M. Wedel and W. A. Kamakura, *Market Segmentation: Conceptual and Methodological Foundations*. Springer Science & Business Media, 2000.

[9] G. J. Mclachlan, K.-A. Do, and C. Ambroise, *Analyzing Microarray Gene Expression Data*. Wiley, 2004.

[10] G. Punj and D. W. Stewart, "Cluster analysis in marketing research: Review and suggestions for application," *J. Mark. Res.*, vol. 20, no. 2, p. 134, 1983.

[11] E. Omol, O. Onyango, J. Wachira, and E. Njeru, "Application of K-Means Clustering for Customer Segmentation in Grocery Stores in Kenya," *Int J Sci Technol Manag*, vol. 5, no. 4, pp. 112–128, 2020.

[12] S. Ozan, "A case study on customer segmentation by using machine learning methods," in *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, 2018.

[13] N. Patankar, S. Dixit, A. Bhamare, A. Darpel, and R. Raina, "Customer segmentation using machine learning," in *Recent Trends in Intensive Computing*, IOS Press, 2021.

[14] S. D. Pyla and M. Seshashayee, "CUSTOMER SEGMENTATION USING MACHINE LEARNING," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 04, pp. 3484–3485, 2022.

[15] S. R. Regmi, J. Meena, U. Kanojia, and V. Kant, "Customer market segmentation using machine learning algorithm," in *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2022.

[16] L. Abidar, D. Zaidouni, and A. Ennouaary, "Customer segmentation with machine learning: New strategy for targeted actions," in *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications*, 2020, pp. 1–6.

[17] I. Lewaaelhamd, "Customer segmentation using machine learning model: An application of RFM analysis," *Journal of Data Science and Intelligent Systems*, vol. 2, no. 1, pp. 29–36, 2023.

[18] R. H. Papetti and R. H. Thompson, *CUSTOMER SEGMENTATION ANALYSIS OF CANNABIS RETAIL DATA: A MACHINE LEARNING APPROACH (By The Honors College & University of Arizona)*. 2019.

[19] Y. Parab and J. Dave, "Customer Segmentation Using Machine Learning: A Comprehensive Research study," *IJNRD*, vol. 8, pp. 718–720, 2023.

[20] M. M. Sathyanarayana, S. Dhanish, P. S. Kumar, and A. N. Reddy, "Mall customer segmentation using clustering algorithm," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 1, pp. 587–593, 2023.

Authorized licensed use limited to: AMITY University. Downloaded on January 11,2025 at 06:22:40 UTC from IEEE Xplore. Restrictions apply.

View publication stats