International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Customer personality analysis and clustering for targeted marketing

Ijegwa David Acheme [1,*] and Esosa Enoyoze [2]

[1] *Department of Computer Science, Edo State University, Uzairue, Nigeria.*
[2] *Department of Mathematics Science, Edo State University, Uzairue, Nigeria.*

International Journal of Science and Research Archive, 2024, 12(01), 3048–3057

Publication history: Received on 23 April 2024; revised on 21 June 2024; accepted on 24 June 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.12.1.1003

Abstract

A customer's personality has been stated to be a key determinant in his or her purchasing behavior. Therefore, in today's highly competitive market, understanding customer behavior and preferences is very important for businesses aiming to stay ahead. This research presents a customer personality analysis and clustering of consumers using machine learning into different consumer groups, this kind of clustering is key for targeted marketing strategies. The machine learning technique used in this work is the k-means machine learning clustering algorithm which divides a set of n observations into k clusters with the aim of designating each observation as a representative of a cluster. The model revealed insightful patterns and correlations between customer characteristics and their purchasing behavior. Through comprehensive data analysis and feature engineering, the model identifies key personality traits such as extroversion, openness, conscientiousness, agreeableness, and neuroticism, based on customer interactions, demographics, and psychographic data. The dataset that was used for this work was obtained from kaggle, it originated from different sources such as customer relationship management (CRM) systems, e-commerce platforms, social media platforms, surveys. The dataset comprises information on "people," "products," and "promotion," aiming to understand customer behavior, preferences, and responses to various marketing efforts. The output of this work provides businesses with actionable insights to tailor marketing campaigns and product offerings to individual customer preferences, thereby enhancing customer satisfaction and maximizing profitability.

Keywords: Customer Personality Analysis; Clustering; Targeted Marketing; Machine Learning

1. Introduction

Targeted marketing in e-commerce refers to the practice of tailoring marketing efforts to specific individuals or groups based on their characteristics, preferences, behaviors, or past interactions within an online platform [1]. This approach aims to deliver personalized content, product recommendations, promotions, and advertisements to enhance the overall shopping experience of customers. The growth and advancements in machine learning has led to an increase in recommendation systems that also work on this same principle and electronic commerce stores because of their competitive nature are increasingly embracing this technique of marketing which is considered to be more effective.

Various computing methods and technologies have been reported in literature for the analysis of customer data and identifying relevant segments in order to deliver personalized targeted marketing products. Some of these computing methods include; Data mining techniques such as clustering, classification, and association rule mining, which are used to extract valuable insights from large datasets [2]. Machine learning algorithms have also been reported, particularly decision trees, neural networks, and collaborative filtering, which are applied to predict customer behavior, segment customers, and make personalized recommendations. Also popular are clustering algorithms (e.g., K-means, hierarchical clustering) and dimensionality reduction techniques (e.g., PCA) used to identify meaningful segments that can be targeted with specific marketing campaigns [3], and predictive analytics which are used to anticipate customer

* Corresponding author: Ijegwa David Acheme

behavior, such as purchasing patterns, churn likelihood, or product preferences, enabling businesses to tailor marketing strategies accordingly. Collaborative filtering, content-based filtering, and hybrid approaches are also common computing methods used in recommendation systems to personalize product recommendations and enhance the shopping experience [4].

While these techniques offer solutions, the success of marketing efforts in contemporary business landscape, relies heavily on the ability to understand and connect with customers on a personal level. As such traditional demographic segmentation methods are insufficient in capturing the nuanced preferences and behaviors of modern online consumers. Consequently, there is a growing interest in employing advanced data-driven techniques, to delve deeper into customer insights. This research seeks to address this need by developing a ma-chine learning model capable of predicting customer personality traits and utilizing these predictions for targeted marketing strategies.

2. Literature Review

In this section, related works that provide a diverse range of perspectives and methodologies related to ma-chine learning, customer personality analysis, customer clustering and targeted marketing, are presented. These review highlights the contributions, strengths, and potential areas for improvement in each related work.

The study in [5] proposed a machine learning approach to predict consumer personality traits using datasets obtained from Facebook, twitter and Instagram social media sites. The approach presented in the paper showed the application of machine learning in predicting personality traits from social media, offering insights into consumer behavior. However, the methodology and model performance should be thoroughly evaluated for robustness.

A systematic review on the impact of personality on online shopping behavior was presented in [6]. The review synthesized existing literature on the impact of personality on online shopping behavior. The paper provided a comprehensive overview of research in the domain of online marketing and targeted advertising, offering valuable insights which are useful for understanding the role of personality in online consumer behavior.

A survey of various machine learning algorithms used for predicting consumer behavior and targeted marketing was presented by [7]. The work covered the different areas of the application of machine learning in targeted marketing. It offered a thorough examination of machine learning techniques used in targeted marketing, specifically the authors highlighted the strengths, limitations, and potential applications of the various machine learning algorithms. This works provides a solid foundation for research on the application of machine learning to consumer behavior and analytics.

Facebook digital marketing was studied in [8]. With the growing interest in digital marketing especially on social media sites, the study explored the effectiveness of personality-based targeted advertising using Facebook advertising as a case study. The paper presented a real world application of personality-based targeting that offered insights into the effectiveness of digital marketing campaigns. However, further research may be needed to generalize findings beyond the specific platform examined.

Deep Learning Models for Customer Personality Prediction from Online Activity Data was reported in [9]. The paper investigated deep learning models for predicting customer personality traits from online activity data. The study introduces deep learning techniques for personality prediction, which may offer improved accuracy and generalization compared to traditional machine learning approaches. However, the feasibility and interpret-ability of such models warrant careful consideration.

Ringbeck, D. et al. [10] examined the Role of Personality Traits in Predicting Consumer Response to Personalized Recommendations. The study examined the influence of personality traits on consumer response to personalized recommendations. The paper contributes to understanding how personality traits impact consumer responses to personalized recommendations, offering implications for personalized marketing strategies. How-ever, further research could explore the interplay between personality and other factors influencing consumer behavior

Several other studies have also reported the application of machine learning to sales in online environment. Some of these studies are found in[11-13].

3. Methodology

This research adopts a multi-stage methodology comprising data collection, pre-processing, model development, and evaluation. Data collection involves gathering a diverse range of customer-related data, including demographic information, psychographic attributes, social media activity, and past purchasing behavior. Pre-processing steps focuses on cleaning the data, handling missing values, and encoding categorical variables. Feature engineering techniques will be employed to extract relevant features and construct input variables for the machine learning model. The K-means machine learning algorithm is then applied for customer segmentation.

3.1. Data Collection

The dataset contains a mix of structured data (e.g., numerical data, categorical data) and unstructured data (e.g., text data, images) that was obtained from kaggle.com, it originated from different sources such as CRM systems, e-commerce platforms, social media platforms, surveys. The dataset comprises information on "people," "products," and "promotion," aiming to understand customer behavior, preferences, and responses to various marketing efforts. Table 1, shows of each category and its description. Analyzing and modeling this dataset can help businesses personalize marketing strategies, optimize product offerings, improve customer satisfaction, and increase sales and profitability. The snapshot of the dataset is also shown in figure 1.

Table 1 Features and Description

People (Customers)		**Products**		**Promotion**	
Feature	**Description**	**Feature**	**Description**	**Feature**	**Description**
ID	Customer's unique identifier.	MntWines:	Amount spent on wine in last 2 years.	NumDealsPurchases	Number of purchases made with a discount.
Year_Birth	Customer's birth year.	MntFruits:	Amount spent on fruits in last 2 years.	AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise.
Education:	Customer's education level.	MntMeatProducts:	Amount spent on meat in last 2 years.	AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise.
Marital_Status	Customer's marital status.	MntFishProducts	Amount spent on fish in last 2 years.	AcceptedCmp3	1 if customer accepted the offer in the 3rd campaign, 0 otherwise.
Income	Customer's yearly household income.	MntSweet Products	Amount spent on sweets in last 2 years.	AcceptedCmp4	1 if customer accepted the offer in the 4th campaign, 0 otherwise.
Kidhome:	Number of children in customer's household.	MntGoldProds	Amount spent on gold in last 2 years.	AcceptedCmp5	1 if customer accepted the offer in the 5th campaign, 0 otherwise.
Teenhome	Number of teenagers in	MntWines:	Amount spent on wine in last 2 years.	Response	1 if customer accepted the offer in the last

	customer's household.				campaign, 0 otherwise.
Dt_Customer	Date of customer's enrollment with the company.	MntFruits	Amount spent on fruits in last 2 years.	NumDealsPurchases	Number of purchases made with a discount.
Recency	Number of days since customer's last purchase.			AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise.
Complain	1 if customer complained in the last 2 years, 0 otherwise				

The variables of the dataset are shown in table. In figure 1 below, a snapshot of the dataset is presented.

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	88	546	172
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38	11	1	6	2
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26	426	49	127	111
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26	11	4	20	10
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	43	118	46
5	7446	1967	Master	Together	62513.0	0	1	09-09-2013	16	520	42	98	0
6	965	1971	Graduation	Divorced	55635.0	0	1	13-11-2012	34	235	65	164	50
7	6177	1985	PhD	Married	33454.0	1	0	08-05-2013	32	76	10	56	3
8	4855	1974	PhD	Together	30351.0	1	0	06-06-2013	19	14	0	24	3
9	5899	1950	PhD	Together	5648.0	1	1	13-03-2014	68	28	0	6	1
10	1994	1983	Graduation	Married	NaN	1	0	15-11-2013	11	5	5	6	0
11	387	1976	Basic	Married	7500.0	0	0	13-11-2012	59	6	16	11	11

Figure 1 Snapshot of the dataset

3.2. Data Preprocessing

Data pre-processing is an important step in the machine learning pipeline, this is the point where the raw dataset is transformed into more usable format for analysis and modeling [14]. It involves a variety of techniques to clean, transform, and organize data before feeding it into machine learning algorithms. Figure 2 Shows the various activities that were carried out in this phase.

As part of the data preprocessing activities, basic statistical analysis were performed in order to observe the shape and distribution of the entire dataset. These statistics include, the count of each variable, the mean, standard deviation, the min and max etc. A snapshot of this is shown in figure 3 for the first ten features of the dataset.

The next step in the pre-processing stage was detection of missing values and duplicates. The 'income' variable/feature had the highest number of missing values from the analysis, this was handled by filling all the NaN values with the median of the variable. Duplicate values were also found in all the rows for the 'contact' and 'revenue' variables, this were dropped because those values will eventually have no significant effect on the performance of the model.

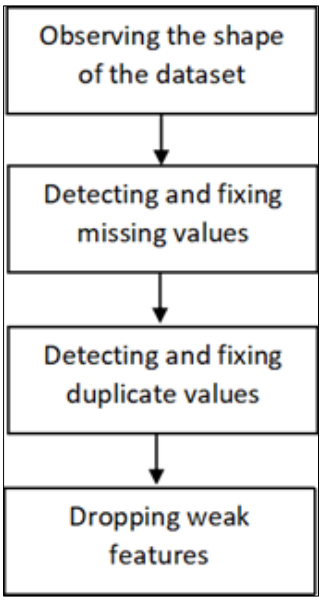

Figure 2 Flow of data pre-processing activities

3.3. Univariate Analysis on Selected Features

Univariate analysis is a statistical method used to analyze data that involves examining the distribution, summary statistics, and patterns within a single variable [15]. It is looking at the distribution of one variable at a time. For example, in a dataset with information about people's ages. A univariate analysis on this dataset would focus solely on the age variable. calculating the mean (average) age, median age, mode (most common age), standard de-viation (how spread out the ages are from the mean), and perhaps create a histogram or box plot to visualize the distribution of ages.

Univariate analysis helps to reveal the characteristics and behavior of individual variables in isolation, which can be useful for detecting outliers, understanding the central tendency and variability of the data, and identifying any patterns or trends within that specific variable. However, it doesn't consider relationships between variables. Univariate analysis was performed on the following select features; Educational Level, marital status, income, number of kids at home and Duration of being a custom-er.

	ID	Year_Birth	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts
count	2240.000000	2240.000000	2216.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000
mean	5592.159821	1968.805804	52247.281250	0.444196	0.506250	49.109375	303.935714	26.302232	166.950000	37.525446
std	3246.662198	11.984069	25173.074219	0.538398	0.544538	28.962453	336.597393	39.773434	225.715373	54.628979
min	0.000000	1893.000000	1730.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2828.250000	1959.000000	35303.000000	0.000000	0.000000	24.000000	23.750000	1.000000	16.000000	3.000000
50%	5458.500000	1970.000000	51381.500000	0.000000	0.000000	49.000000	173.500000	8.000000	67.000000	12.000000
75%	8427.750000	1977.000000	68522.000000	1.000000	1.000000	74.000000	504.250000	33.000000	232.000000	50.000000
max	11191.000000	1996.000000	666666.000000	2.000000	2.000000	99.000000	1493.000000	199.000000	1725.000000	259.000000

Figure 3 Basic statistics of the dataset

3.3.1. Univariate Analysis of the Customer

Educational Level Variable

The dataset contained various categories for the education variable. These categories include graduate, MSc., PhD, Basic etc. For the purpose of the analysis, they were converted to only two categories, these are: Undergraduate and postgraduate. As shown in figure 4, 97.88% of the customers have a postgraduate qualification while 2.4% have an undergraduate qualification.

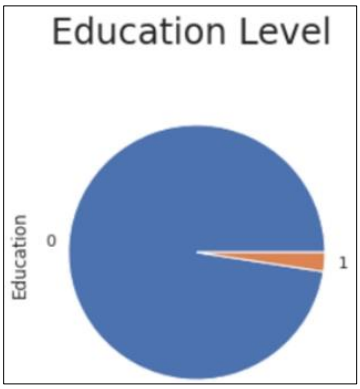

Figure 4 Analysis of the 'Education Variable'

3.3.2. Univariate Analysis of the 'Marital_Status" Variable

The categories in the marital status variable include the following, married, divorced, together, widow, and single. The count of each of the categories is shown in table 2.

Table 2 Analysis of the 'Marital Status Variable'

Category	Count
Married	864
Divorced	232
Together	580
Widow	77
Single	3

3.3.3. Univariate Analysis of the 'Income' Variable

The maximum annual income of the customers quoted in the USD is 666666, while the minimum is 1730. The mean annual income is 52238.004. figure 5 shows the density plot for the income variable.

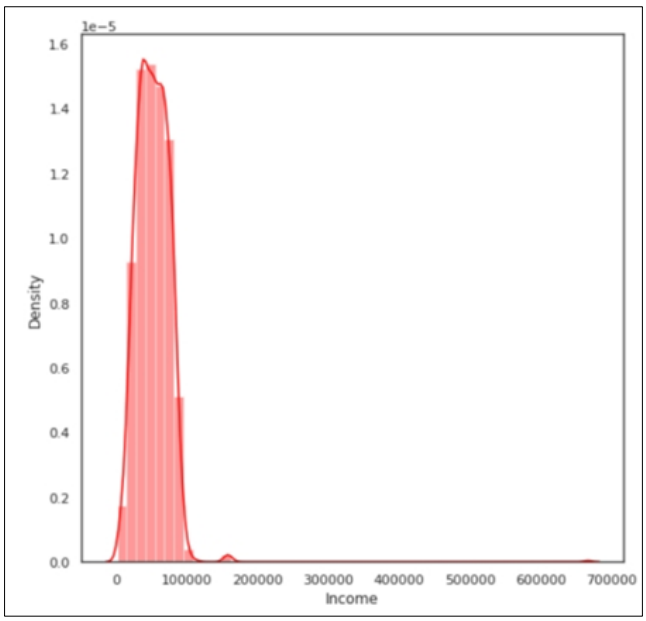

Figure 5 Density plot for the income variable

3.3.4. Univariate Analysis on the 'Number of Kids' Variable

Analysis of the number of children variable revealed 28.4% of customers had no children, 50.35% had one child only, 18.79% had two children while 2.35% had three or more children. Figure 6 shows this distribution.

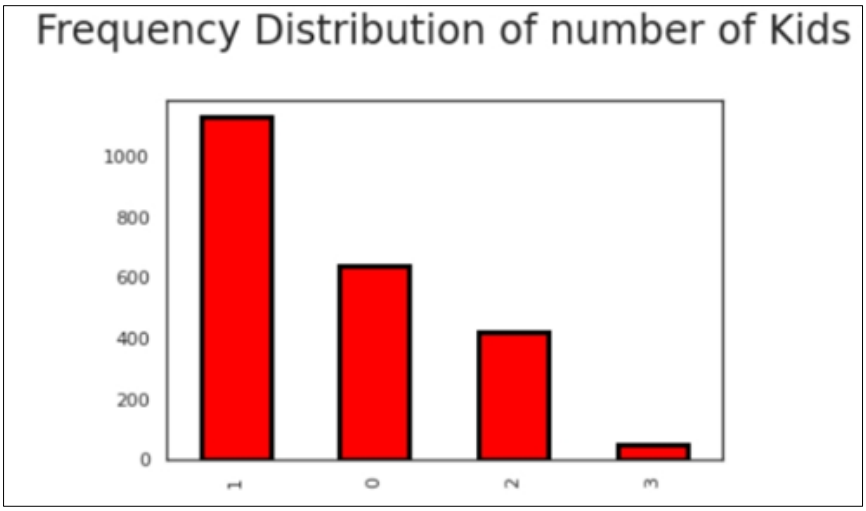

Figure 6 Distribution of the number of kids variable

3.3.5. Univariate Analysis on the 'Duration of Being a Customer' Variable

The dataset also revealed the age of the customers. This is the total number of years they have been registered as customers.

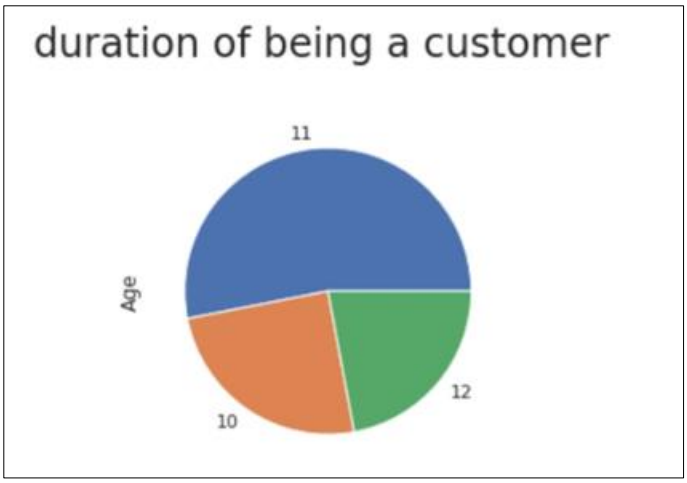

Figure 7 Customer age

4. Implementation

In order to achieve the main objective of this work, the K-means algorithm is applied for segmentation of the customers. K-means clustering is a vector quantization approach that comes from signal processing. A set of n observations is divided into k clusters with the aim of designating each observation as a representative of a cluster by assigning it to the cluster whose mean (also known as the cluster center or centroid) is closest to it.The number of clusters in the dataset is first of all established using the the elbow method. The "elbow" method is a heuristic approach which is used to find the optimal number of clusters (k) in k-means clustering. The following are the steps followed to obtain establish the number of clusters using the elbow method:

- STEP 1: The K-means is run for a Range of K, Typically, a value between 1 and 10 is chosen.

- STEP 2: The Within-Cluster Sum of Squares (WCSS) is computed. For each value of k, we calculate the with-in-cluster sum of squares (WCSS). WCSS represents the sum of squared distances between each data point and its assigned centroid within a cluster. In other words, it measures the compactness of the clusters. Lower WCSS indicates tighter clusters, as data points are closer to their centroids.
- STEP 3: Plot WCSS against K: Next, plot the WCSS values against the corresponding k values. This creates an elbow-shaped curve (figure 8)
- STEP 4: Identification of the "Elbow" Point: The "elbow" point on the plot is the point where the rate of decrease in WCSS starts to slow down. Visually, it appears as the point where the curve begins to bend or flatten out, resembling an elbow.
- STEP 5: Choose the Optimal K: The optimal number of clusters is typically chosen as the value of k at the "elbow" point. This point represents the trade-off between maximizing the number of clusters (which reduces WCSS) and minimizing the complexity of the model (which in-creases with the number of clusters).

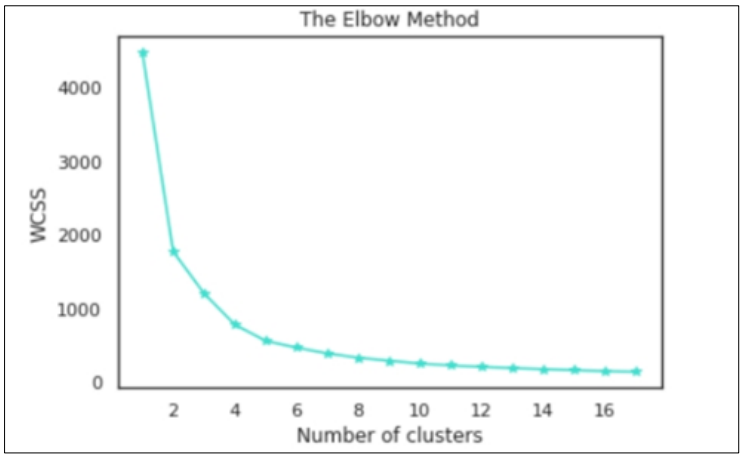

Figure 8 The elbow method showing optimum number of clusters

4.1. K-Means Clustering

Applying the K-means clustering shows two distinct customer clusters as shown in figure 9. These are the high spending and moderate spending customers. The results also includes the cluster centers otherwise called the centroid for each cluster, These help in understanding the structure of the data and can be useful for a business owner who can better understand these two segments of his customers in order to carryout targeted marking

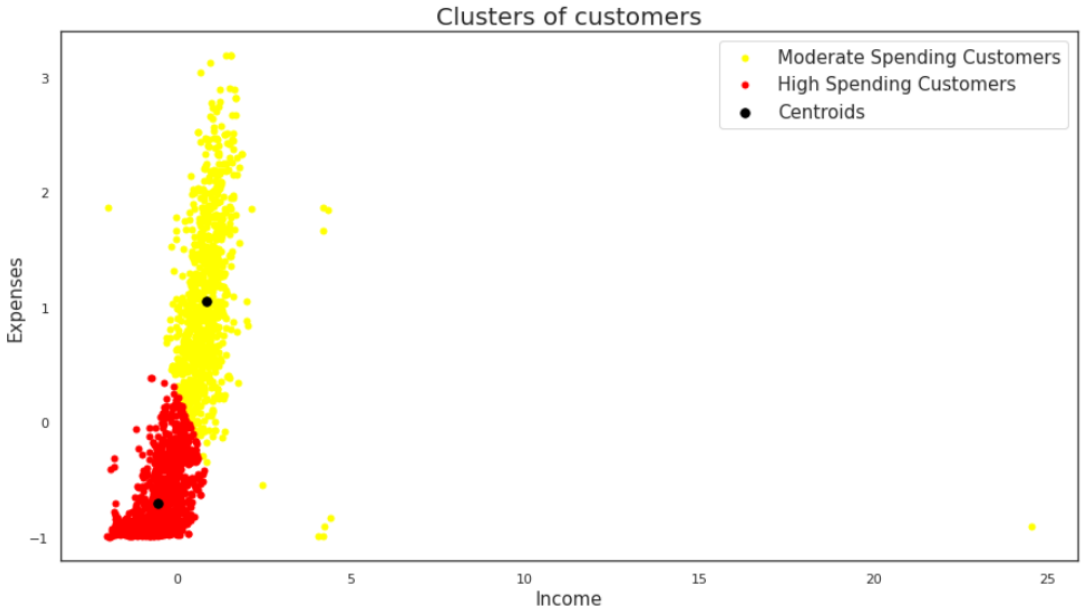

Figure 9: k-means clustering of customers

5. Conclusions

Business managers and owners are continually striving to better satisfy their customers in order to maintain their loyalty. It has been noted that a consumers' personality greatly influences his or her purchasing behavior, hence this research presented an analysis of customer dataset and applied K-means clustering algorithm for segmentation these customers. Based on their personalities and previous purchase behaviors, with the ultimate goal of enhancing targeted marketing strategies, which are more effective and less expensive. The analysis presented in this work will also provide valuable insights into customers' preferences and behaviors.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare no conflicts of interest.

Data Availability Statement

The data is available from the corresponding author upon reasonable request.

Authors Contributions

- **Ijegwa David Acheme:** Conceptualization, Data Curation and methodology, review and editing
- **Esosa Enoyoze:** Modelling and Formal Analysis Methodology, Writing original draft

References

[1] Saluja, A., Soth, R. K. D., Pawar, B. R., Pasunoori, V., Mungi, A., & Faldu, R. (2023). Precision Marketing Strategy for E-Commerce By Using Big Data Technology. Journal of Informatics Education and Research, 3(2).

[2] Ziafat, H., & Shakeri, M. (2014). Using data mining techniques in customer segmentation. Journal of Engineering Research and Applications, 4(9), 70-79.

[3] Abdulhafedh, A. (2021). Incorporating k-means, hierarchical clustering and pca in customer segmentation. Journal of City and Development, 3(1), 12-30.

[4] Parthasarathy, G., & Sathiya Devi, S. (2023). Hybrid Recommendation System Based on Collaborative and Content-Based Filtering. Cybernetics and Systems, 54(4), 432-453

[5] Sümer, S. I., & Parilti, N. (Eds.). (2023). Social Media Analytics in Predicting Consumer Behavior. CRC Press.-137.

[6] Hermes, A., & Riedl, R. (2021). Influence of personality traits on choice of retail purchasing channel: literature review and research agenda. Journal of Theoretical and Applied Electronic Commerce Research, 16(7), 3299

[7] Baderiya, S. H., & Chawan, P. M. (2018). Customer buying Prediction Using Machine-Learning Techniques: A Survey. International Research Journal of Engineering and Technology (IRJET), 5(10), 931-935.

[8] Curran, K., Graham, S., & Temple, C. (2011). Advertising on facebook. International Journal of E-business development, 1(1), 26-33. e12345.

[9] Boustani, N., Emrouznejad, A., Gholami, R., Despic, O., & Ioannou, A. (2023). Improving the predictive accuracy of the cross-selling of consumer loans using deep learning networks. Annals of Operations Research, 1-18.-3320.

[10] Ringbeck, D., Seeberger, D., & Huchzermeier, A. (2019). Toward personalized online shopping: Predicting personality traits based on online shopping behavior. Available at SSRN 3406297.

[11] Ijegwa, A. D., Olufunke, V. R., Folorunso, O., & Richard, J. B. (2019). A Bayesian based system for evaluating customer satisfaction in an online store. In Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 2 (pp. 1047-1061). Springer International Publishing.

[12] Acheme, I. D., Osemengbe, U., Makinde, A. S., & Vincent, O. R. (2021). Online Stores: Analysis of user Experience with Multiple Linear Regression Model. In 4th International Conference on Information Technology in Education and Development.

[13] Acheme, I. D., Olayinka, A., Jegede, A., Uddin, O. O., Nwanwo, W., & Vincent, O. (2023, November). Investigating the Most Influential Factors for Customer Satisfaction in Online Stores. In 2023 2nd International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS) (Vol. 1, pp. 1-4). IEEE.

[14] Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., & Herrera, F. (2020). Big data preprocessing. Cham: Springer.

[15] Mertler, C. A., Vannatta, R. A., & LaVenia, K. N. (2021). Advanced and multivariate statistical methods: Practical application and interpretation. Routledge.