Customer Segmentation using Machine Learning with a Coupon Generator GUI

Asif Iqbal, Rajeev Ranjan Pandey, Subhraneel Bagchi, Saikat Ruj, Sujata Dawn
Computer Science and Engineering Department and Information Technology Department of
Durgapur Institute Of Advanced Technology and Management, Durgapur

Abstract:- The rise of numerous competitors and entrepreneurs which has led to a great deal of competition among businesses, compelling them to seek out new customers while retaining their existing ones. Consequently, the importance of delivering exceptional customer service has become crucial, regardless of a business's size or scale [2]. Moreover, understanding the unique needs of each customer is paramount to providing targeted support and developing personalized customer service strategies. This level of comprehension can be achieved through the implementation of a well-structured customer service framework, as different customer segments often share similar market characteristics [5].

To tackle the challenges posed by a large customer base, the integration machine learning has gained traction, surpassing traditional market analytics methods that tend to falter under such circumstances. This paper adopts the k-means clustering algorithm to address this issue [8]. The implementation of the k-Means algorithm, facilitated by the Sklearn library (refer to the Appendix), involves training a program using a dataset comprising 100 patterns and two factors

Keywords:- *data mining; machine learning; customer segment; k-Mean algorithm; sklearn; extrapolation.*

I. INTRODUCTION

Over time, the intensifying competition among businesses and the wealth of historical data available have led to the widespread utilization of data mining techniques for uncovering crucial and strategic insights hidden within organizations' information [1]. Data mining involves the extraction of meaningful information from datasets and its presentation in a format that is easily understandable and can support decision-making. This field encompasses various disciplines, including statistics, artificial intelligence, machine learning, and data systems. The applications of data mining are vast and diverse, ranging from bioinformatics and weather forecasting to fraud detection, financial analysis, and customer segmentation.

This paper focuses on employing machine learning to identify customer segments within a commercial business. Customer segmentation is a crucial task for businesses aiming to understand their customer base and deliver personalized experiences.

Traditional approaches to segmentation often rely on manual categorization or rule-based methods, which can be time-consuming and subjective. With the advent of machine learning, specifically the K-means clustering algorithm, businesses have gained a powerful tool for automating the segmentation process.

In this journal, we will delve deeper into the implementation of K-means clustering for customer segmentation. We will explore the theoretical foundations of the K-means algorithm, provide a step-by-step guide for its application in customer segmentation, discuss the challenges that may arise during implementation, and evaluate the effectiveness of K-means clustering compared to other popular segmentation techniques. Buried in a database of integrated data proved to be effective for detecting subtle but subtle patterns or relationships. This mode of learning is classified under supervised learning. Integration algorithms include the K means algorithm, K-nearest algorithm, sorting map (SOM), and more.[4] These algorithms, without prior knowledge of the data, are able to identify groups in them by repeatedly comparing input patterns, as long as static aptitude in training examples is achieved based on subject matter or process. Each set has data points that have very close similarities but differ greatly from the data points of other groups. Integration has great applications in pattern recognition, image analysis, and bioinformatics and so on.

In this paper the k-means clustering algorithm was implemented in the customer segment. standard silhouette -score with two feature sets of 100 training patterns found in the retail trade. After several indications, five stable intervals or customer segments were identified. Two factors are considered annual income and spending score. And the clusters are names as cluster 1, cluster 2, cluster 3, cluster 4 and cluster 5.

II. LITERATURE SURVEY

A. Customer Clustering

In the ever-evolving business landscape, organizations face increasing competition and the challenge of meeting the diverse needs and desires of their customers while also attracting new ones to enhance their operations. However, catering to the individual requirements of every customer can be a daunting task. Customers vary in terms of their needs, preferences, demographics, size, tastes, and characteristics. Treating all customers equally is considered an ineffective practice in business. To address this challenge, the concept of customer segmentation, or market segmentation, has emerged. This approach involves dividing consumers into distinct subgroups or segments based on their similar market behaviors or characteristics. Customer segmentation aims to categorize the market into indigenous groups, enabling businesses to tailor their strategies accordingly.

B. Data repository

Data collection entails gathering and measuring information to assess targeted modifications within a predetermined system. It serves as a vital component of research across various disciplines, including the physical and social sciences, humanities, and business [12]. The primary objective of data collection is to obtain high-quality evidence that allows for the analysis and generation of accurate and insightful responses to the research questions at hand. In this particular study, the data was obtained from the UCI machine learning repository, which serves as a valuable source for datasets used in research and analysis.

C. Clustering data

Clustering refers to the procedure of organizing data within a dataset into distinct groups based on shared characteristics or similarities. Numerous algorithms have been developed to handle clustering tasks, each suited for specific conditions or requirements. However, since there is no universally applicable clustering algorithm, it becomes crucial to select the appropriate technique for a given dataset.

In this paper, we have applied the k-means algorithm to perform clustering analysis. The k-means algorithm, implemented using the Python scalar library, allows us to group information within the dataset based on similarities. It is a widely used and effective clustering technique for various applications. By utilizing the k-means algorithm, we can gain valuable insights and uncover meaningful patterns in the dataset under investigation.

D. K-means

The k-means algorithm is widely recognized as one of the most popular clustering algorithms used for classification. This algorithm operates by assigning each data point to one of the predefined clusters, which are formed based on centroids. The clusters represent underlying patterns within the data, offering valuable insights to aid decision-making processes.

To determine the optimal number of clusters, we will employ the elbow method, which is a commonly used approach in k-means clustering. This method helps identify the appropriate number of clusters by analyzing the variance explained as more clusters are added. By applying the elbow method, we can make informed decisions regarding the configuration of the k-means algorithm for our analysis.

III. METHODOLOGY

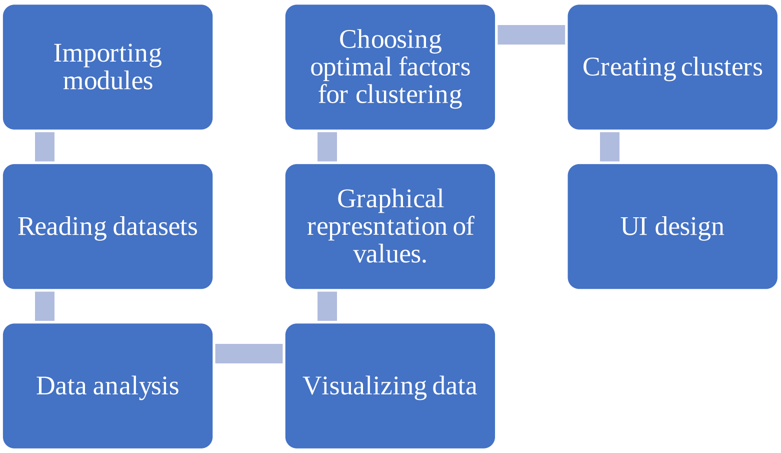
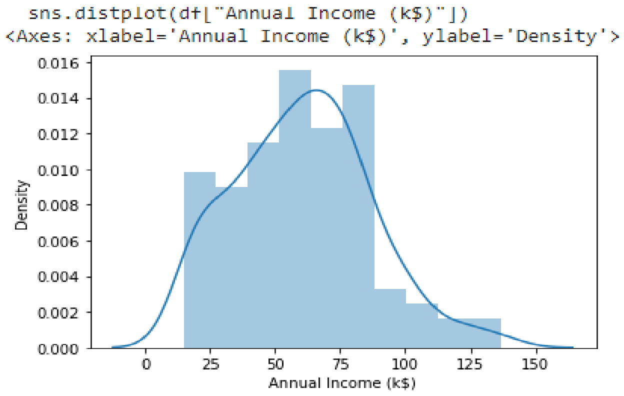


Fig. 1: Flow chart

The dataset used for implementing the clustering analysis and K-means algorithm was gathered from a retail store within a shopping mall. It consists of 200 tuples, representing the data of 200 customers. The dataset encompasses 5 attributes, including CustomerId, gender, age, annual income (in thousands of dollars), and spending score (ranging from 1 to 100).

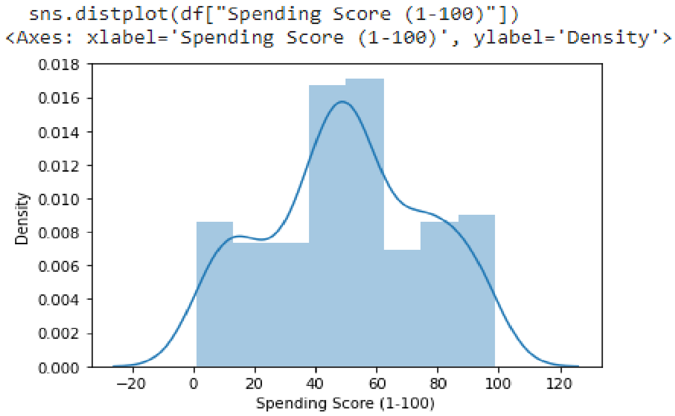
A. Visualize our data:

To gain a better understanding of our dataset, we will visualize it using the matplotlib and seaborn libraries. This visualization will help us analyze the relationships between different columns and gain insights into our customers' categorizations based on their annual income, location, and spending score. By visualizing the data, we can uncover patterns and trends that may assist in making informed decisions and formulating effective strategies.



```
sns.distplot(df["Annual Income (k$)"])
<Axes: xlabel='Annual Income (k$)', ylabel='Density'>
```

Fig. 2: Graph of Annual Income



```
sns.distplot(df["Spending Score (1-100)"])
<Axes: xlabel='Spending Score (1-100)', ylabel='Density'>
```

Fig. 3: Graph of spending score

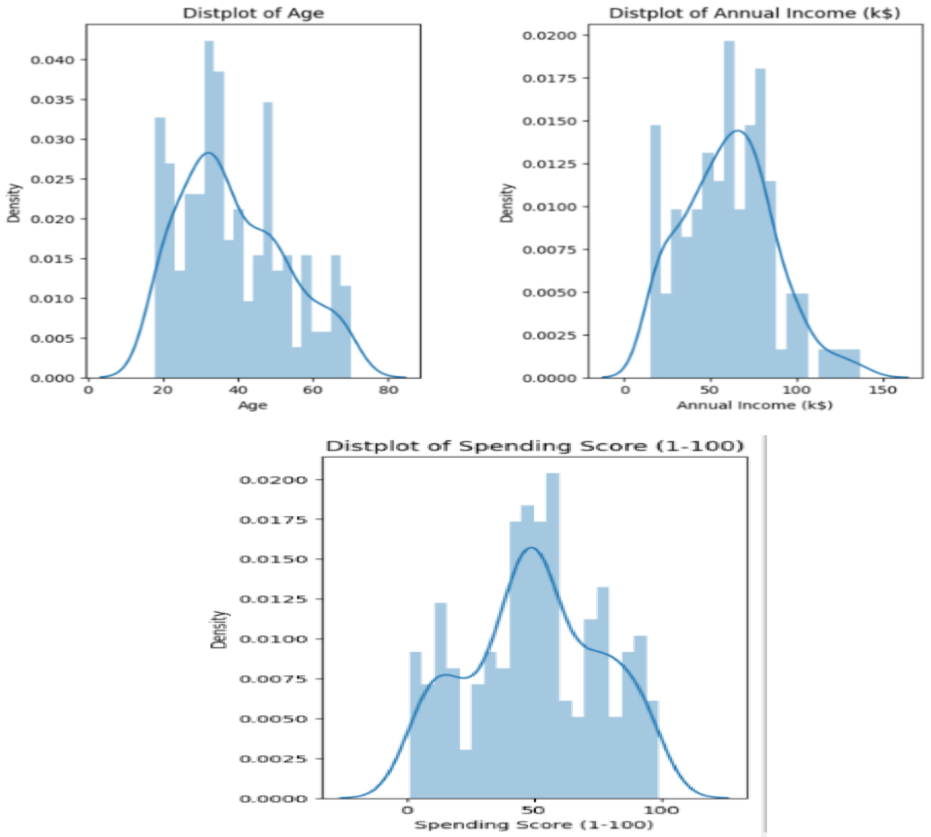


Fig. 4: Graphical representation of customers age, annual income and spending score

IV. ELBOW METHOD

The elbow method is a technique used to determine the optimal number of clusters in a dataset. It is based on the observation that increasing the number of clusters can lead to a decrease in the sum of within-cluster variance. This means that by having more clusters, we can capture finer groups of data objects that exhibit greater similarities.

To apply the elbow method, we start by running the clustering algorithm for different values of k, ranging from 1 to 10 clusters. For each value of k, we calculate the total intra-cluster sum of squares. This metric represents the sum of the squared distances between each data point and the centroid of its assigned cluster.

Next, we plot the intra-cluster sum of squares against the number of clusters. The resulting graph provides an indication of the appropriate number of clusters for our model. We look for a point on the graph where there is a bend or significant change in the rate of decrease in the sum of squares. This bend is often referred to as the "elbow" and represents the optimal number of clusters for our dataset.

Now we will fit that clusters into K-Means model and predict labels and also find centroids.

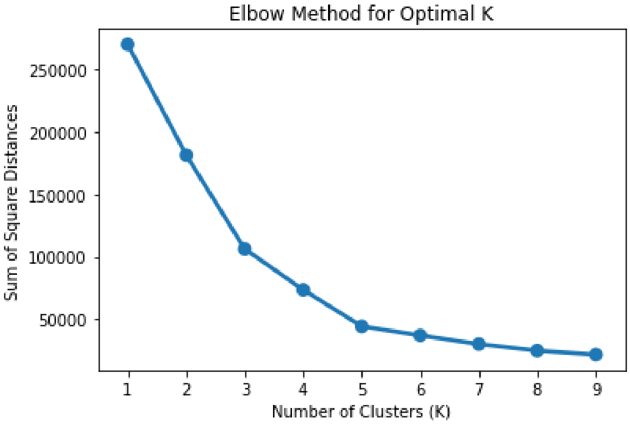
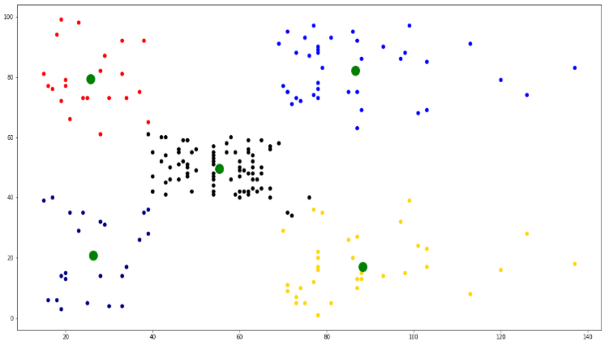


Fig. 5: Elbow Method Graph

Based on the graph above, it looks like K = 5, or 5 clusters is the correct number of clusters in this analysis. Now translates the customer segments provided by these components.

V. VISUALIZING THE CLUSTERS



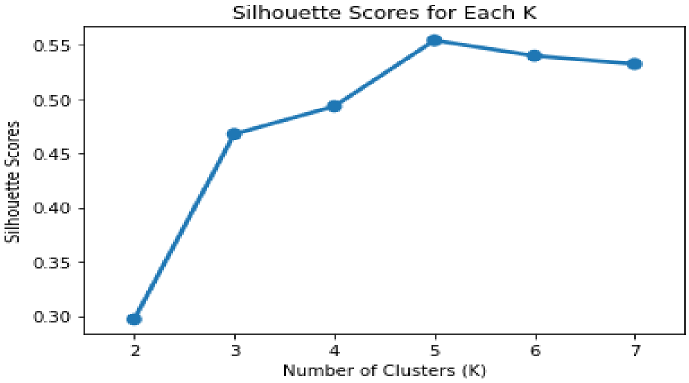


Fig. 6: Cluster of spending score and annual income Here each colorrepresents a different cluster

VI. SILHOUETTE SCORE

We have also used silhouette method which is used to measure how dense and well-separated the clusters are. Silhouette score takes into consideration the intra-cluster distance between the sample and other data points within same cluster (a) and inter-cluster distance between sample and next nearest cluster (b).

VII. THE COUPON GENRETOR UI

Next, We create a User Interface using Tkinter in python and using the different clusters we generate different coupons for the users. As different clusters represent different groups with similar characteristics.

This Python script that involves the use of various libraries and frameworks such as Pandas, Tkinter, PIL (Pillow), scikit-learn (sklearn), and Matplotlib. Here is a summary of the code: The necessary libraries and modules are imported. The data is read from a CSV file called "Mall_Customers.csv" using Pandas. Some columns are dropped from the Data Frame. The data is then clustered using the K-means algorithm from the scikit-learn library. A graphical user interface (GUI) is created using Tkinter.

The GUI consists of a login window and a main page with various features. The login window contains an image background and a home button. The main page displays an image background, an input field to enter a customer ID, and two buttons. When the "Generate" button is clicked, a coupon is generated based on the cluster label of the customer ID entered. The generated coupon is displayed on the GUI.

When the "Show Clusters" button is clicked, a new window opens displaying a scatter plot of the data points colored according to their assigned clusters.

Another button on the main page opens a shop window with an image background of a shopping pa. The script can be executed by calling the main() function. This button is a work in progress and has no functions as of now.

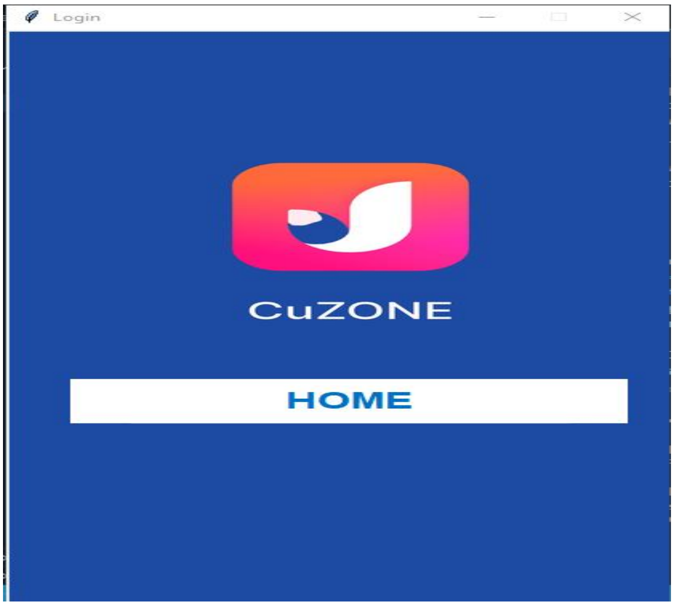


Fig. 7: Home Page

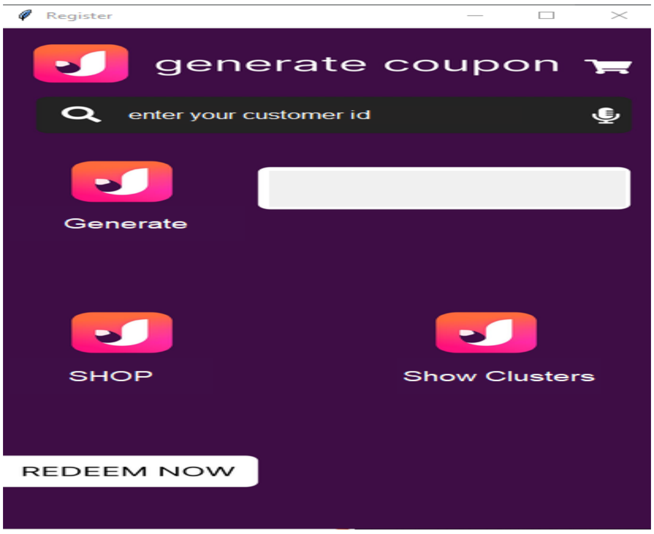


Fig. 8: Main page

VIII. CONCLUSION

In conclusion, this project demonstrates the application of clustering techniques to customer data in order to generate personalized coupons for a mall's customers. The code utilizes the K-means algorithm to cluster the customer data based on their spending behavior. By analyzing the customer clusters, the program assigns specific coupons to each cluster, allowing the mall to offer targeted discounts and promotions to its customers.

The graphical user interface (GUI) created using Tkinter provides a user-friendly platform for customers to interact with the system. Customers can enter their unique customer ID, and the program generates a coupon based on their cluster label. The coupon is then displayed on the GUI for the customer to view and utilize during their shopping experience.

Additionally, the project offers a visualization of the customer clusters by plotting the data points on a scatter plot. This feature helps the mall better understand the distribution of customer segments based on annual income and spending scores.

Overall, this project highlights the potential of data analysis and machine learning techniques in customer segmentation and targeted marketing. By leveraging clustering algorithms and providing personalized coupons, businesses can enhance customer satisfaction, drive sales, and improve overall marketing effectiveness.

In conclusion, customer segmentation using the K-means clustering algorithm is a powerful tool that can provide valuable insights for businesses. By understanding the steps involved in implementing K-means clustering for customer segmentation and addressing associated challenges, businesses can effectively leverage this approach to enhance their marketing strategies and drive business success.

REFERENCES

[1.] Customer Segmentation Using Machine Learning Prof. Nikhil Patankar a ,1,Soham Dixit a , Akshay Bhamare a , Ashutosh Darpel a and Ritik Raina a Dept.Of Information Technology Sanjivani College of Engineering,Kopargaon423601(MH),India

[2.] K. Windler, U. Juttner, S. Michel, S. Maklan, and E. K.¨Macdonald, "Identifying the right solution customers: A managerial methodology," Industrial Marketing Management, vol. 60, pp.173 –186, 2017.

[3.] "A Case Study on Customer Segmentation by using Machine Learning Methods", IEEE, Year: 2018

[4.] https://scikitlearn.org/stable/documentation.html

[5.] CUSTOMER SEGMENTATION USING MACHINE LEARNING a journal by AMAN BANDUNI*

[6.] ,Prof ILAVENDHAN A from School of Computing Science & Engineering,Galgotias University, Greater Noida, U.P

[7.] https://towardsdatascience.com/customer-segmentation-using-k-means-clustering-d33964f238c3

[8.] https://medium.com/mlearning-ai/customer-segmentation-using-k-means-clustering-ae73e3d82934

[9.] https://www.tutorialsteacher.com/python/create-gui-using-tkinter-python