Customer Segmentation Using K-Means Algorithm

Big Data and Analytics Mini Project

Submitted by:

Ananya Kapoor (20103104)

Dhairya Sachdeva (20103098)

Utkarsh Kathpalia (20103071)

Under the supervision of:

Dr. Bharat Gupta

Department of CSE

Jaypee Institute of Information Technology University, Noida

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and beliefs, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma from a university or other institute of higher learning, except where due acknowledgment has been made in the text.

Place: Jaypee Institute of Information Technology, Noida
Date: December, 2022

Name: Ananya Kapoor

Enrolment No.: 20103104

Name: Dhairya Sachdeva

Enrolment No.: 20103098

Name: Utkarsh Kathpalia

Enrolment No.: 20103071

Table of Contents

Abstract

Proper planning is essential for a long-term business. This can be done by implementing proper marketing strategies from time to time. Machine learning can play a key role in decision-making. Machine learning can play a key role in decision-making. This paper proposes a systematic approach which can help offline stores target their customers and obtain maximum profit by using the clustering application of machine learning. It helps offline stores get the benefits of the latest technologies in their business. The initial step for this system is to analyse the acquired sales data based on the purchase history, which will be used to group the customers. K-Means clustering is used to segment customers. Later, the most preferred product of each cluster is determined, and the result of this can be used by the shopkeepers to analyse their business and make good decisions for the long life of the business. It can assist offline stores in finding different groups of customers rather than viewing the entire customer as a single unit.

Chapter 1: Introduction

Customer relationship management (CRM) is a marketing approach that allows a store to learn about its customers' behaviour and wants to build a strong relationship and customer loyalty. It can help in increasing the sales and profit of the store. Advancements in technology can facilitate the above-mentioned objectives successfully and more efficiently. Stores may recognise their important customers and anticipate their future actions and their favourite items by using data mining and extraction of hidden patterns of client purchases from massive databases. This paper aims to use such technologies to improve the business of offline stores. This can help the stores make good decisions. The two intelligent components of Customer Relationship Management are customer clustering and buyer targeting. In this paper, it proposes an approach that can help offline stores cluster customers according to their purchase behaviour and find out the best-selling product in each group. It can help the stores analyse their customers and their needs. The stores can get an idea about the products their customers prefer and provide those products with high quality to satisfy them.

The most successful companies today are the ones who know their customers so well that they are able to anticipate their needs beforehand. This can better be achieved if we can segment the customers into different groups that reflect the similarities among the customers in each group. The goal of the segmentation is to foresee the needs of customers, get to know their interests, lifestyles, priorities and learn their spending habits so that to maximize the value of customers to the business. Customer segmentation has many advantages for the businesses which include:

- *Price Optimization:* Understanding the customers and their financial status will help to pace up with price optimization

- *Enhanced Competitiveness:* More the customer attention and more will be the revenue generated and this would in return enhances company competitiveness in the market. If the company can segment the market, it is well known to the customers and the company can come up with new products and optimize the existing products according to the changing preferences of the customers

- *Acquisition and Retention:* A personalized connection with the customers helps a company to win satisfied customers. Better segmentation of the customers will lead to better relationships with the prospective customers. About 75% of satisfied customers are more likely to stay with a company

- *Increased Revenues:* By fine tuning the marketing strategies will help generate more revenues because users will more likely to purchase when they are delivered exactly what they need. Personalized and segmented Emails increase the likelihood that more Emails will be opened. Infact the more emails are opened; the more sales will be made. Successful marketing not only require knowledge about who your customers are but also where exactly they are in the buying process and customer segmentation based on such information ensure that the marketing campaigns are truly effective

This system proposes an approach that helps the stores group the customers according to their behaviour and other patterns to enhance the existing marketing model.

Chapter 2: Related Work

The proposed method in [1] is about segmenting customers who have similar behaviours into similar segments and customers who have different patterns into different segments. This paper describes different clustering algorithms (k-Means, agglomerative, and mean shift) which can be implemented to segment the customers and finally compare the results of clusters obtained from the algorithms.

Murugeswari R. and Ramasakthi G. relate the process of classifying a small text piece into positive, negative, or neutral [2]. The process of sentiment analysis is carried out by performing a step-by-step process. First, the dataset is collected. Then, the dataset is loaded, and pre-processing is done. After that, the data is split. Then, the data is trained on the model. Finally, it categorises the comments as positive, negative, or neutral. Sentimental analysis consists of four types of classifications (Support Vector Machine, Gaussian Naive Bayes, Random Forest, and Multilayer Perceptron).

[3] clusters customers into different groups. Creating appropriate actions for each customer cluster is an excellent way of determining effective marketing tactics. Customers would be evaluated based on a number of factors, including their most recent visit (recency/R), purchase frequency (frequency/F), and total money spent on the product (monetary/M). The RFM method is a typical name for this method. The conclusion of this study is that machine learning technology may be utilised to cluster customers, and the clusters of customers can be used by marketing managers to develop appropriate marketing strategies.

Pranata Ilung and Skinner Geoff describe the process of selecting, analysing, and interpreting clusters of customers to evaluate annual spending on the products [4]. Many clusters are created with the k-means clustering algorithm, and analysis of these clusters is performed using a number of techniques to select the best cluster. An insight into various purchase behaviours is provided, and the best customer group to target is determined by assessing and interpreting clusters.

Maryani Ina and Riana Dwiza propose a method to execute customer clustering and profiling using the RFM model in order to provide customer relationship management recommendations to enterprises [5]. The findings of this study provide characteristics of each customer in order to carry out a customer relationship strategy.

Chapter 3: Proposed Method

The proposed system aims to help offline stores analyse their customers and their needs. The traditional method of analysing the customers used by the stores is not enough in the new world. The world is changing with the help of growing technology. So, it is the time of offline stores to change their methods for business analysis. In this system, it uses the K-means clustering algorithm to group the customers into different groups and helps to find the preferred product within each cluster. It requires the collection of data related to customers and their purchase history. In this way, the stores can divide the customers into different groups and prepare customised strategies for each group instead of considering the whole customer base as a single unit.

The existing base papers use Elbow method to find out the minimum optimal clusters for the K-means clustering. But elbow method does not work effectively in a few cases, for instance take the below scatter plot. Scatter plot for X1 vs X2Humans may be able to tell that the data originates from five different clusters, but we struggle to perceive high- dimensional data. The Elbow Method, as seen on the left, would most likely lead us to k = 4. The Elbow Method causes us to conclude that two of the clusters are one since they are so close together. This is due to the fact that establishing a centroid in the centre of both clusters reduces the relative distance between data points. As a result, calculating the appropriate number of clusters for our clustering job requires an approach that is more accurate, rigorous, and dependable.

Our dataset contains details for clustering and analysing the customers. It stores the customer id, gender, age, annual income, and spending score of each customer during the time of purchase. The spending score of a customer will be updated by the system based on the purchase pattern of the corresponding customer. These details are used for clustering the customers into different groups. The marketing_campaign dataset from Kaggle [6] is used for the implementation.

The system mainly consists of steps such as collection of data related to customers, their purchase details, clustering of the customers based on this information, and finally obtaining the popular product from each cluster.

The model of the system is represented by Fig 1. In the system, the initial step is to collect data from the customers. Later, this data has to be processed to obtain the necessary data for analysis. Then clustering is done, in which the clustering algorithm K-means will be used. Once the clustering is performed, then the final step is to find the preferred product for each cluster based on the purchase history of customers in the corresponding clusters. This provides enough information for the stores to understand their customers' needs. Based upon this, they can make decisions to improve their business. Its detailed description is given below.

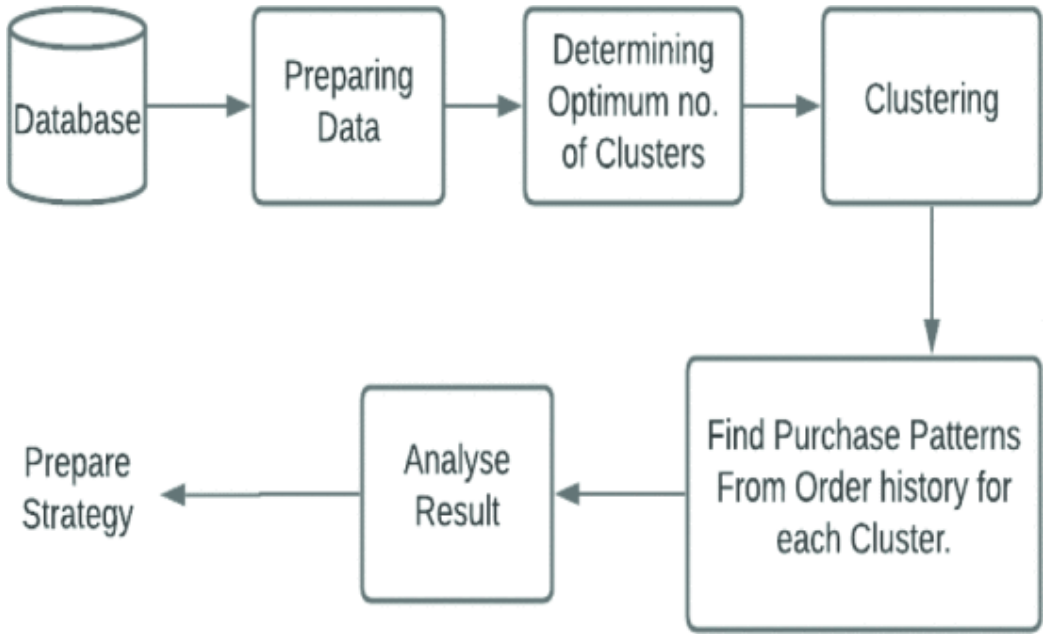


Fig. 1

The results obtained from the above steps will be used to analyse the business and the shopping behaviour of the customers. It helps to consider the customers as different groups and find their needs.

Chapter 4: Experimental Results

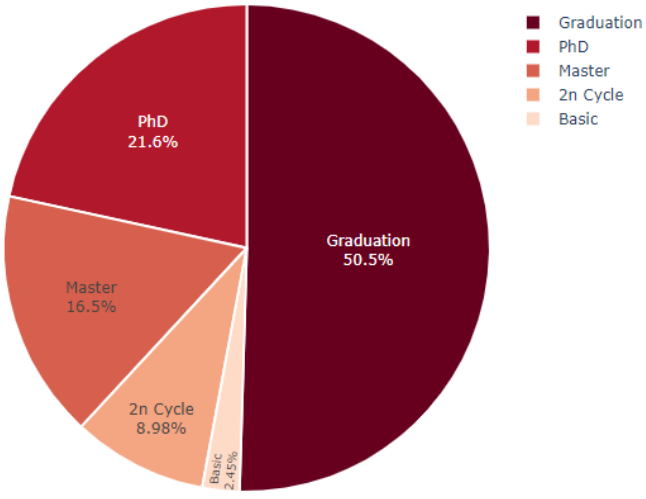
This system can be a helpful partner for retailers and stores to analyse and study their business. The results obtained from the analysis can help the store make decisions for a successful future.

For the given Dataset, we have performed the exploratory data analysis with the help of customer segmentation. Customer segmentation will be carried out with the help of K-Means algorithm. At the end of analysis, I would like to answer some questions by gaining some insights from the data, which are as follows:

1. What are the statistical characteristics of the customers?
2. What are the spending habits of the customers?
3. Are there some products which need more marketing?
4. How the marketing can be made effective?

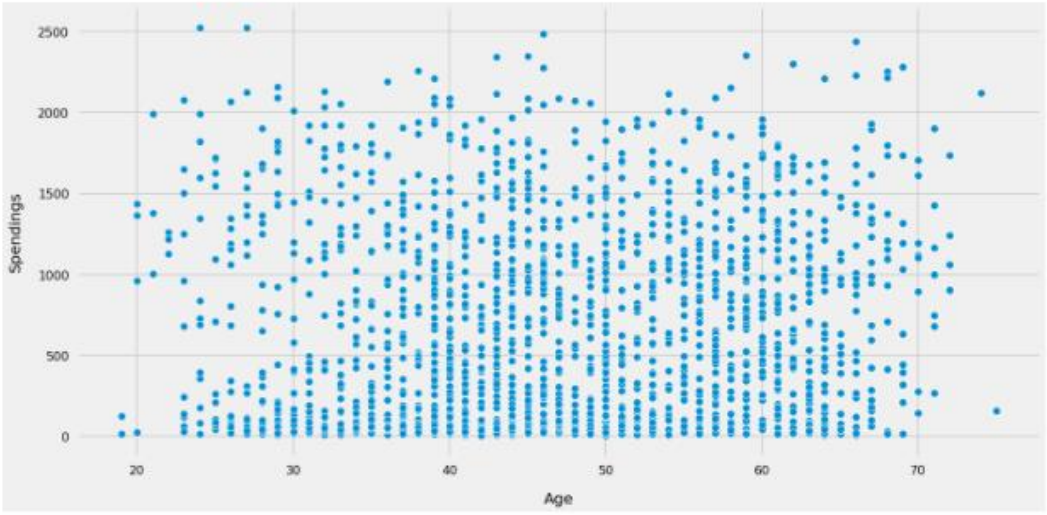
Exploratory Data Analysis:

Education Level



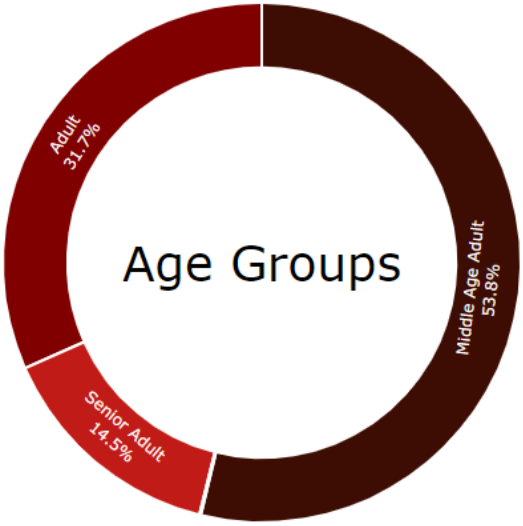
1. Half of the customers are University graduates
2. There are more customers who hold PhD degrees than the customers who did Masters

Relationship: Age vs Spendings



There doesn't seem to be any clear relationship between age of customers and their spending habits

Customers Segmentation: Age Group Wise

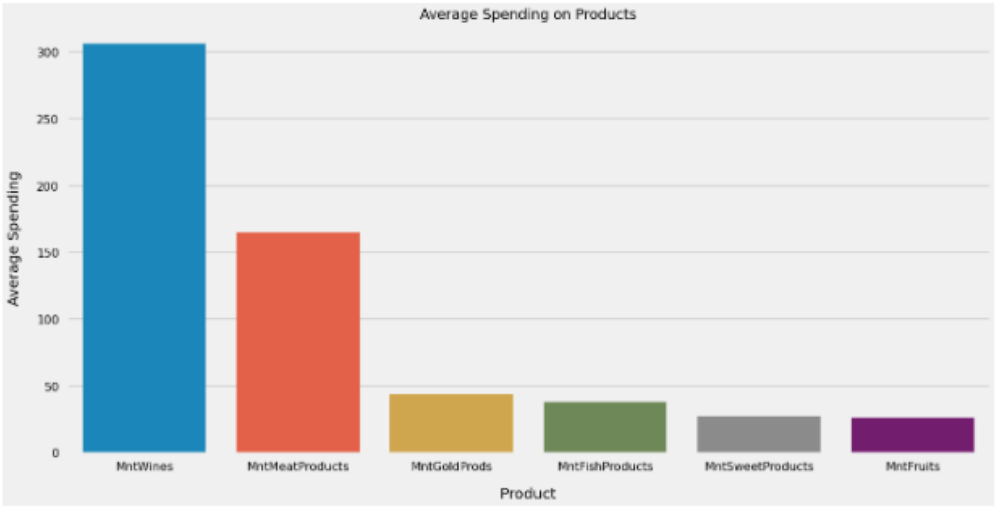


1. More than 50% of the customers are Middle Age Adults aged between 40 and 60
2. The 2nd famous age category is Adult, aged between 20 and 40

Income Distribution of Customers

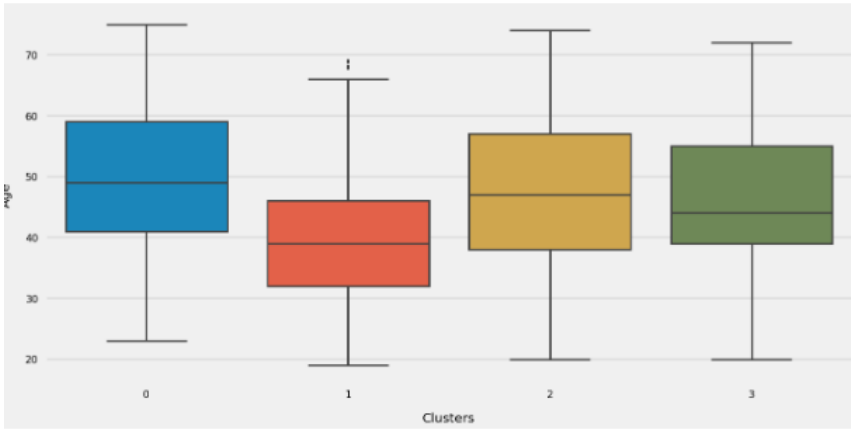
The salaries of the customers have normal distribution with most of the customers earning between 25000 and 85000

Most Bought Products



1. Wine and Meats products are the most famous products among the customers
2. Sweets and Fruits are not being purchased often

Clusters Identification



From the above analysis we can segment the customers into 4 groups based on their income and total spendings:

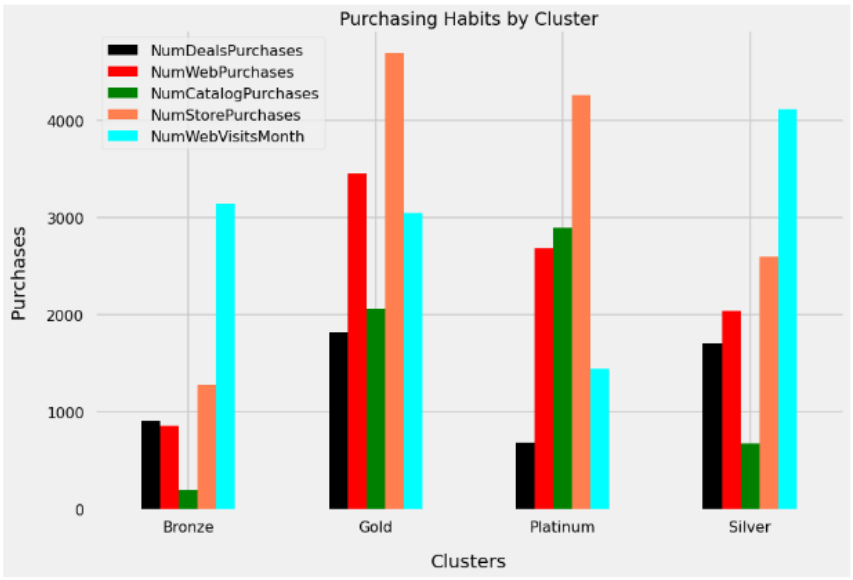
Platinum: The one's with highest earnings and highest spendings

Gold: The one's with high earnings and high spendings

Silver: The one's having low salary and less spendings

Bronze: The one's having lowest salary and least spendings

Purchase Habits by Clusters



1. Platinum and Gold Customers mostly likely to do store purchasing
2. Most of the web and catalogue purchases are also done by the customers from Platinum and Gold segments
3. Silver and Gold categories also like to buy from the stores
4. Deal purchases are common among the gold and silver customers
5. Silver category customers made the greatest number of web visits while customers from Platinum segment have least web visits

Chapter 5: Conclusion

The process of maintaining customer loyalty and attention span are the fundamental issues that most offline stores encounter. As a result, appropriate marketing strategies must be developed from time to time. This system helps offline stores to make strategies for identifying customers' needs and maximising profits for businesses. It is a methodical approach to attracting customers and increasing profits for offline stores. Client happiness may be achieved by creating a model that predicts customer preferences and assists businesses in maximising profits.

This study demonstrates that client segmentation in shopping malls is achievable despite the fact that this form of machine learning application is highly useful in the market, a manager can concentrate all of his or her attention on each cluster that has been discovered and meet all of their requirements. Mall

managers must be able to understand what customers require and, more importantly, how to meet those needs. analyse their purchasing habits, and establish frequent encounters with customers that make them feel comfortable in order to satisfy their demands.

It can be improved in the future by including more analytic techniques and methods that would provide more information. It can also be improved by linking multiple stores together, which would give information about the market as a whole. In short, with further improvement and by including more features, it can prove to be a great way for offline stores to evaluate their business and manage it to obtain more profit. The study of customer relationship management is a vast topic, and this system can play a great role in it.

References

[1] Tushar Kansal, Suraj Bahuguna, Vishal Singh and Tanupriya Choudhury, "Customer segmentation using K-means clustering", *2018 international conference on computational techniques electronics and mechanical systems (CTEMS)*, pp. 135-139.

[2] R Murugeswari and G Ramasakthi, "Conceptual Analysis of Product Evaluations Using Deep Learning", *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 820-825.

[3] Muhammad Ridwan Andi Purnomo, Abdullah Azzam and Annisa Uswatun Khasanah, "Effective Marketing Strategy Determination Based on Customers Clustering Using Machine Learning Technique", *2020 Journal of Physics: Conference Series*, vol. 1471, pp. 012023.

[4] Ilung Pranata and Geoff Skinner, "Segmenting and targeting customers through clusters selection & analysis", *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 303-308.

[5] Ina Maryani and Dwiza Riana, "Clustering and profiling of customers using RFM for customer relationship management recommendations", *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, pp. 1-6.

[6] kaggle kernels output yasirhussain1987/customer-segmentation-using-k-means -p /path/to/dest