Customer Segmentation and Management Strategy Optimization for Gyms Using K-Means Clustering

Ziyu Niu[a]
Faculty of Natural, Mathematical &Engineering Sciences, King's College London, London, U.K.

Keywords: K-Means Clustering, Customer Segmentation, Gym Management.

Abstract: This study aims to provide a robust segmentation strategy for the gym managers by using K-Means clustering algorithms. The objective is to help gym recognise valuable customer groups that align with their marketing strategies, thereby maximizing profitability. The dataset is from Kaggle which includes 4, 000 records related to gym membership. The dataset has features such as age. Data preprocessing includes standardization which is used to ensure that each variable contributed equally to the clustering process. Using Elbow Method to determine the best number of clusters and two clusters were identified as ideal. K-Means clustering, combined with Principal Component Analysis (PCA) for dimensionality reduction, revealed clear distinctions between the customer groups. The first cluster is consisted with old customers with low fitness frequency. While the second cluster includes younger, more active and higher income customers. These findings provide valuable concept for gym management to tailor service, such as providing low-intensity fitness programs for older members and high-intensity workouts or premium membership plans for younger, higher-income customers. The study proves the effectiveness of K-Means and PCA in customer segmentation while also suggesting that future research should explore more advanced algorithms and incorporate additional data to further refine the segmentation process.

1 INTRODUCTION

A gym is a space used as a workout space, fitted with fitness equipment, and some gyms will also have pools and boxing areas. Working out at the gym has become increasingly popular for several reasons, one of which is the less efficient atmosphere of home workouts. Unlike gyms, home settings typically lack professional equipment, contributing to a less effective exercise environment. Another reason is that going to the gym to work out is likewise a social activity, where people may meet like-minded others, including getting some guidance. In such situations, customer segmentation is an important task from the perspective of gym managers. Segmentation of customers can help gyms filter out more valuable customers who may fit into the gym's marketing strategy, thus gaining more profits.

Artificial Intelligence has made great progress in recent years, and there are a variety of representative algorithms, such as random forest, Logistic Regression, KMeans, etc. These methods have been widely used in various fields, such as chemistry, biomedicine, and especially in the business analysis. Random forests are commonly used for new material prediction in materials chemistry due to its ability to handle high-dimensional data and insensitivity to noise. Random forests are used to predict the properties of new materials, such as electrical conductivity and thermal stability. This method can more accurately predict chemical reaction results and material properties by combining the results of multiple decision trees (Rigatti, 2017).

Logistic regression is widely used in business analysis, especially in marketing, customer segmentation and risk management. A typical application is customer churn prediction. Companies use Logistic Regression to analyse customer data and predict the likelihood of churn based on historical behaviour and demographic information (Moro, 2015). The model identifies key factors such as purchase frequency, customer service interactions and payment history, helping organisations to proactively manage customer retention strategies.

[a] https://orcid.org/0009-0006-3988-9611

Niu, Z.
Customer Segmentation and Management Strategy Optimization for Gyms Using K-Means Clustering.
DOI: 10.5220/0013214000004568
In *Proceedings of the 1st International Conference on E-commerce and Artificial Intelligence (ECAI 2024)*, pages 239-242
ISBN: 978-989-758-726-9

Random forests, despite being powerful ensemble methods, often suffer from issues related to interpretability and can require significant computational resources for large datasets. Similarly, logistic regression, though effective for binary classification, is limited by its assumption of linear relationships between features and the target variable, making it less suitable for complex data structures. K-Means clustering offers advantages in terms of its simplicity, ease of implementation, and ability to partition data into distinct clusters, which can be beneficial for unsupervised learning tasks where the structure of the data is not predefined. K-Means clustering is often used for classification analysis of gene expression data. For example, it can be used to cluster patients based on the expression levels of specific genes to identify subtypes of disease, such as different forms of cancer (Golub, 1999). Such analysis helps to understand the molecular basis of disease and can drive the development of personalised therapies.

In order to achieve the above objectives, this paper uses data from Kaggle and uses K-means to cluster the variables. The number of clusters is determined using the elbow method, which ultimately proves the effectiveness of the approach. Customers are accurately classified into distinct categories, providing the gym management with valuable insights for effective decision-making.

2 METHOD

2.1 Dataset Preparation

The dataset comprises 4, 000 entries, each representing an individual record about gym membership prediction research based on professional status. The structure of these data includes different kinds of features related to analysing, such as age, gender, career and membership status. The number of features is 16. The mission is to cluster data using all input features in a dataset that lacks a target variable (Y). The preprocessing of these data is conversion from strings to numerical values and standardization. Data standardization is an important step of data preprocess, which can be used to adjust data variables from different scales or units to a uniform standard. It can make sure that the contribution of each variable in data analysis is equivalent (Han, 2011). This step is particularly critical for many machine learning and statistical models. The benefits of data standardization are improvement of algorithm performance, elimination of scale deviations and improved data consistency.

2.2 K-means

2.2.1 Elbow Method

Elbow Method is a method to determine the optimal number of clusters k in cluster analysis, which is particularly used in K-means clustering algorithm. This method is used to determine the optimal k by evaluating how the Sum of Squared Errors (SSE) within a cluster varies with the number of clusters k (Thorndike, 1953). The specific operation is that as the increasing value of k, SSE will usually taper, this is because samples will be divided into smaller clusters so that the distance from each point to the centre of its cluster.

After drawing series of k and corresponding SSE, the pattern will usually show a curve point which is like an elbow. SSE decreases rapidly before this point, but it decreases significantly slowly after reaching the point (Kodinariya, 2013). Therefore, this "elbow point" is considered as the optimal number of clusters k because adding more clusters at this point will not improve the performance of the model significantly but it will increase the risk of the complexity of computation and overfitting.

2.2.2 Introduction of K-means

K-means is a widely used clustering algorithm. It is mainly used in dividing dataset into a predetermined number of clusters, which maximises the similarity between data points in the same cluster and minimise the similarity between different clusters (Ahmed, 2020; Hamerly, 2003). It a form of unsupervised learning and is always used in analysis of statistical data, market research, image processing etc. The algorithmic principle has four steps. The first one is determining k data points as the initial clustering centre by using elbow method. The second one is assigning each data point to the nearest cluster centre to form k clusters. The third one is computing mean value of each cluster and set the mean value as the new cluster centre. Finally, repeating and updating the steps until the cluster centre does not change anymore or satisfy other stopping conditions (the number of repeating) (MacQueen, 1967). K-means usually shows high efficiency on large datasets. The structure of K-means is simple and easy to implement. K-means can be used for cluster analysis of different kinds of data types. The "sklearn" library's "K-Means" algorithm is employed for cluster analysis. Initially, k=3 is chosen by applying elbow method. Before applying the "K-Means"

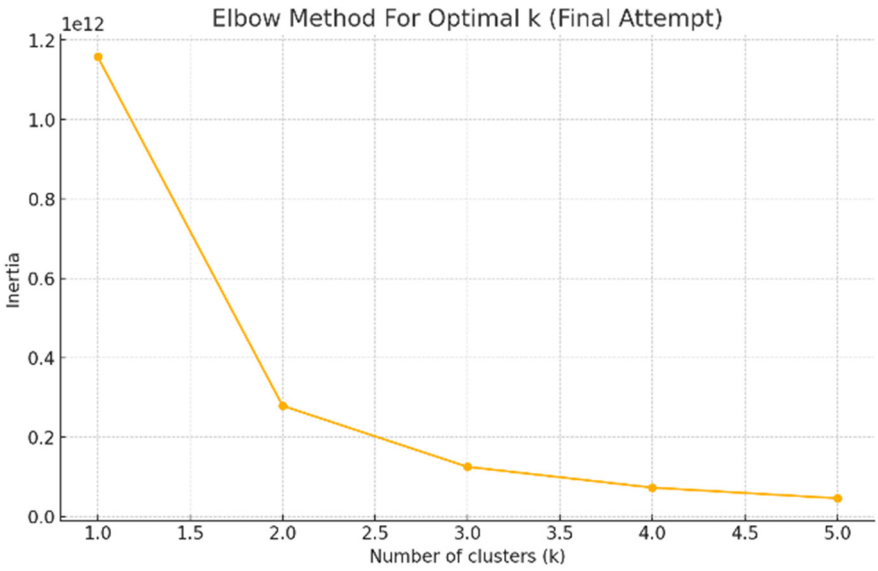

Figure 1: The elbow method for optimal k (Photo/Picture credit: Original).

Figure 2: Customer segmentation based on K-Means clustering results (Photo/Picture credit: Original).

algorithm, "random_state=42" is set to ensure the reproducibility of the results. Model is applied on the pre-processed data via the "fit" method. The internal of this method performs the core iterative process of the algorithm. Principal Component Analysis (PCA) is a statistical technique (Maćkiewicz, 1993), which is used to reduce dimension by exchanging to a new coordinate system, reducing number of variables in the dataset and retain as much information as possible in the original data variables the same time. PCA is particularly useful in data visualization because it can reduce multi-dimensional data to 2D or 3D, which makes it easy to view and analyse the structure and pattern of data in 2D or 3D.

3 RESULTS AND DISCUSSION

To ensure the best number of clusters, k. This study uses Elbow Method. Elbow Method is a commonly used method of cluster analysis. It finds the best number of clusters by plotting k against the cluster inertia (the sum of the square of distance from data

points to the centres of the clusters to which they belong). The smaller the inertia, the better the clustering is indicated. However, as the increase of k, the inertia induction becomes progressively smaller (Kodinariya, 2013). By analysing the Figure 1, k=2 is chosen.

This Figure 2 shows the results of customer clustering based on K-Means (k=2). Downscaling high-dimension data to two dimensions by PCA For easy visualisation and label the clustering result. Each point in the graph represents a data sample and each color represents a clustering result. There are two main clusters represent in yellow and purpleX axis and Y axis represent the first two principal components of PCA respectively. These two components capture the information of the largest variance in the original high dimension data, which makes different customer groups can be distinguished in the reduced dimensionality space. PCA Component 1 primarily reflects differences in characteristics associated with age or annual income, while PCA Component 2 is associated with membership duration or fitness frequency. The purple cluster represents a group of customers whose projections on PCA Component 1 and Component 2 are centred on the lift side. Based on the properties of PCA, these customers show similarity in the characteristics of being order or having a low fitness frequency. The yellow cluster is located on the right side, these customers are younger, having a higher fitness frequency or having a higher annual income compared with the purple cluster customers. Two advisers or gym managers. 1. As purple cluster customers are old and have a low fitness frequency, gym can provide more suitable low-intensity fitness programmes such as yoga, tai chi or course of health management. It can satisfy their need for fitness and improve their participation and loyalty as well. For the younger yellow cluster customers, they prefer high fitness frequency which means that gym can add high-intensity course or introduce more challenge programmes such as Crossfit, HIIT training etc., to keep their interest and activeness. 2. For higher annual income yellow cluster customers, gym can provide higher-end membership services such as personalised fitness coaching, nutritional advice or premium facility access. Also, the gym can consider launching high-end membership plan to increase the value-added experience for customers. For customers with relatively low annual incomes, gym can provide affordable membership program or promote special offer to attract and retain them to make sure their continued use of gym's service.

4 CONCLUSIONS

This study successfully applied machine learning algorithms to segment gym customers. By using K-Means clustering and PCA for visualization, this paper divided the customer data into distinct groups, providing valuable insights for gym management. The Elbow Method determined that the optimal number of clusters was 2. The analysis highlighted that while K-Means and PCA are effective tools for customer segmentation, the limitations of these techniques should be considered, especially when dealing with complex high-dimensional data. Future research should explore more advanced clustering algorithms and dimensionality reduction techniques to enhance the accuracy of clustering and the clarity of visualizations. Additionally, incorporating other customer data, such as behavioral patterns and preferences, could further refine the segmentation process and lead to more tailored marketing strategies.

REFERENCES

Ahmed, M., Seraj, R., & Islam, S. M. S. 2020. The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, 9(8), 1295.

Golub, T. R., et al. 1999. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science

Hamerly, G., & Elkan, C. 2003. Learning the k in k-means. Advances in neural information processing systems, 16.

Han, J., Pei, J., & Kamber, M. 2011. Data mining: concepts and techniques.

Kodinariya, T. M., & Makwana, P. R. 2013. Review on determining the number of cluster in K-means clustering. International Journal of Advance Research in Computer Science and Management Studies, 1(6), 90-95.

Maćkiewicz, A., & Ratajczak, W. 1993. Principal components analysis (PCA). Computers & Geosciences, 19(3), 303-342.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability.

Moro, S., Cortez, P., & Rita, P. 2015. A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.

Rigatti, S. J. 2017. Random forest. Journal of Insurance Medicine, 47(1), 31-39.

Thorndike, R. L. 1953. Who belongs in the family? Psychometrika, 18(4), 267–276.