Market Basket Analysis of Mall Customer Data using K-Means and GMM

Menaka Kollu
Arizona State University
mkollu@asu.edu

Nishanth Solomon
Arizona State University
nsolomo2@asu.edu

Sushilkumar Muralikumar
Arizona State University
smural32@asu.edu

***Abstract*-The Assignment focuses on Clustering Algorithms based on unsupervised learning. The goal is to perform Market Basket Analysis based on the Kaggle Dataset provided. This report serves as an all in resource, where we introduce the concepts of the clustering algorithms followed by the results of the problem that we worked on as part of the assignment. The report also includes suggestions to improve sales based on the optimum value of the number of clusters chosen.**

I. INTRODUCTION

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses[1]. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or groupings in data[1]. The clusters are modeled using a measure of similarity which is defined upon metrics such as Euclidean or probabilistic distance[1]. Common clustering algorithms include:

- **Hierarchical clustering**
- **K-Means clustering**
- **Gaussian mixture models**
- **Self-organizing map**
- **Hidden Markov models**

One of the major applications of unsupervised learning is in the retail sector, especially in e-commerce. Various techniques are used to analyze customer behavior. One of the key techniques used is Market Basket Analysis[11]. It is used by large retailers to uncover associations between items[11]. It works by looking for combinations of items that occur together frequently in transactions[11]. These relationships can be used to increase profitability through cross-selling, recommendations, promotions, or even the placement of items on a menu or in a store[12].

In order to perform the market basket analysis of the given Kaggle dataset, we use the following algorithms to perform the clustering: 1. k-Means 2. Gaussian Mixture Model(GMM) using the Expectation-Maximization(EM) Technique.

II. METHODS

2.1. Dataset Preparation:

1. The given dataset consists of information about Customers who visited a mall. the dataset gives information about the following parameters:
   - Customer ID(Unique ID number assigned to each customer)
   - Gender of the customer
   - Total annual income
   - Spending Score
   - Age of the customer.
2. The dataset has the following dimensions:200*5
3. The dataset was initially checked for null parameters using the data.isnull().sum() command provided in the pandas library.
4. The dataset had no null values and did not require any preparation/usage of techniques such as interpolation.
5. We have considered four columns of data for clustering the data and they are as follows: Gender, Annual income, spending score and age. The Unique Customer ID has not been considered as it does not add any value to the clustering problem. However, the unique customer ID has been used to analyze the clustered data.
6. The gender column consisted of two entries “Male” or “Female”. This data field was ‘One Hot Encoded’ and the resulting dataset(dataset.csv) was used for the clustering. The original dataset is contained in the kaggle_dataset.csv file.

2.2. K-Means:

The K-Means clustering algorithm was applied to the dataset(dataset.csv) for the following cluster sizes: k=4,6,8 and 10. The K-Means package provided in the Scikit Learn library was used for this application.

2.3. GMM:

The GMM clustering algorithm using the EM technique was applied to the dataset(dataset.csv) for the following cluster sizes: k=4,6,8 and 10. The GMM package used in the Scikit Learn library was used for this application.

2.4 Clustering Metrics:

In order to determine the optimal value of the cluster sizes, the following metrics have been used:

2.4.1.Elbow Method for KMeans:

The “elbow” method to help data scientists select the optimal number of clusters by fitting the model with a range of values

for *K[14]*. The line chart resembles an arm and the "elbow" (the threshold point) indicates that the underlying model fits best at that point[14]. We have plotted the graph of Distortion vs Cluster size (K). Distortion is the sum of squared errors (SSE), the 'error' in this case is the difference between each point coordinates and corresponding centroid coordinates. The distortion can be calculated by using *km.inertia_*, km is the KMeans model built using the SkicitLearn library. The threshold point is the point after which there is a rapid increase in distortion. Thus the optimal value of K can be determined.

2.4.2.Silhouette Score for GMM:

The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample[15]. The Silhouette Coefficient for a sample is (b - a) / max(a, b)[15]. To clarify, b is the distance between a sample and the nearest cluster that the sample is not a part of[15]. We can compute the mean Silhouette Coefficient over all samples and use this as a metric to judge the number of clusters[15]. Thus the optimal value of the number of clusters can be determined.

III. RESULTS

3.1.K-Means Results:

K-Means results are available under the *results/kmeans* folder. The results contain clusters formed by using 4 features *'Age'*, '*Annual Income*', '*Spending Score*' and *'Gender'* obtained by K-Means for K=4,6,8,10. Results also show the plot between Number of clusters Vs Distortion, Elbow method was used to determine the optimal number of clusters to be used.

3.1.1. K=4

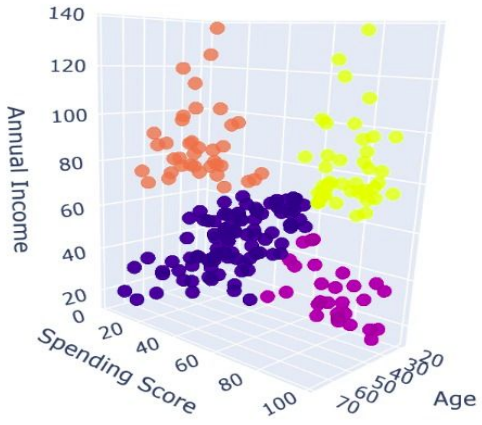

Figure 1: Plot of spending score vs age vs annual income for K=4

3.1.2. K=6

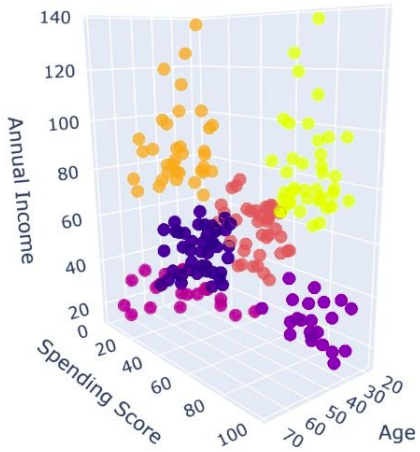

Figure 2: Plot of spending score vs age vs annual income for K=6

3.1.3. K=8

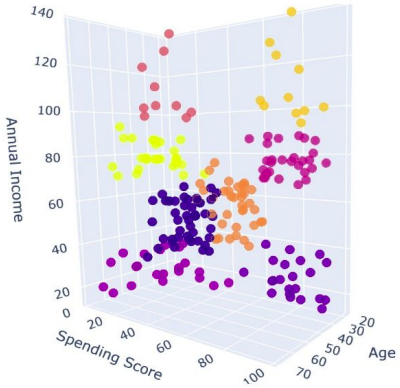

Figure 3: Plot of spending score vs age vs annual income for K=8

3.1.4. K=10

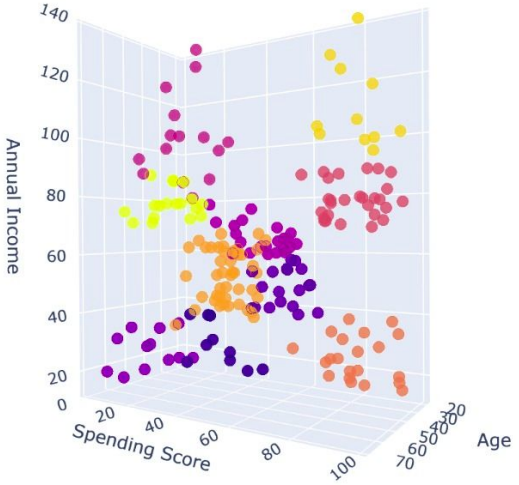

Figure 4: Plot of spending score vs age vs annual income for K=10

3.1.5. ELBOW DIAGRAM FOR K-MEANS:

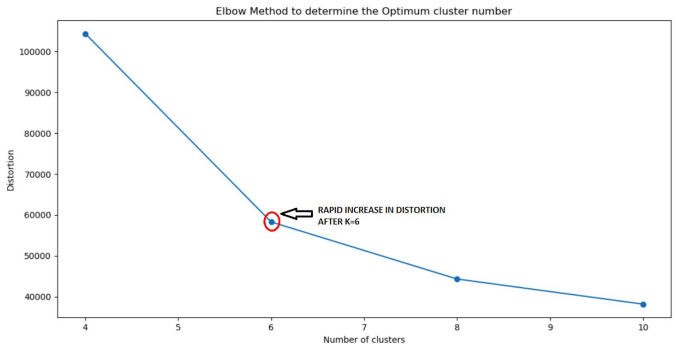

Figure 5: Plot of Number of clusters vs Distortion(Elbow Diagram)

3.2.GMM Results:

GMM results are available under the *results/gmm* folder. The results contain clusters formed by using 4 features *'Age'*, '*Annual Income*', '*Spending Score*' and *'Gender'* obtained by K-Means for K=4,6,8,10. Results also show the plot between Number of clusters Vs Silhouette diagram and the highest value of the Silhouette score was determined to be the optimum cluster size.

3.2.1. K=4

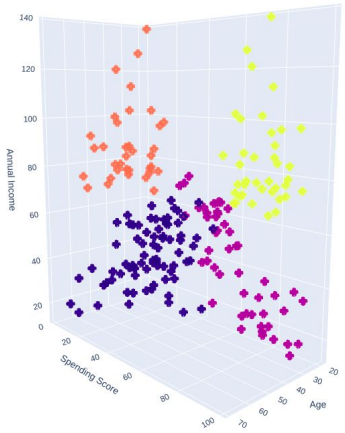

Figure 6: Plot of spending score vs age vs annual income for K=4

3.2.2. K=6

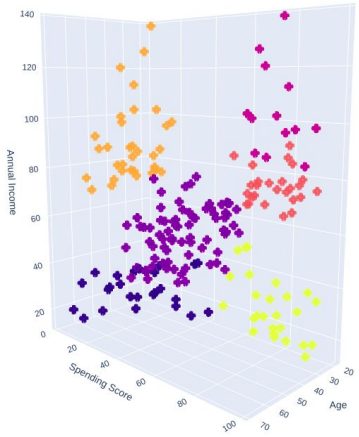

Figure 7: Plot of spending score vs age vs annual income for K=6

3.2.3. K=8

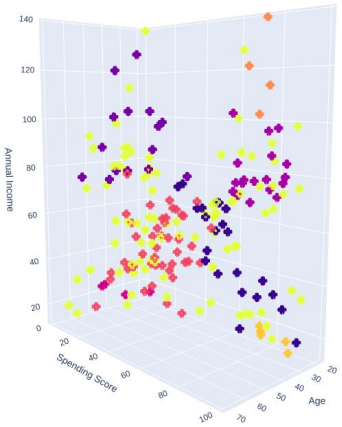

Figure 8: Plot of spending score vs age vs annual income for K=8

3.2.4. K=10

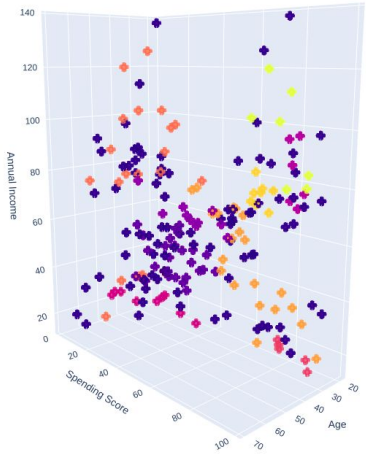

Figure 9: Plot of spending score vs age vs annual income for K=10

3.2.5. SILHOUETTE DIAGRAM FOR GMM:

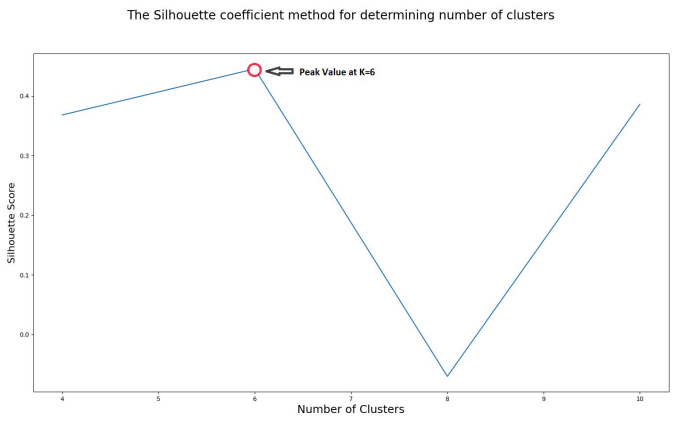

Figure 10: Plot of silhouette score vs Number of clusters

IV. Conclusions

- This assignment has focussed on using the following two Clustering Algorithms: 1. K-Means 2. EM Technique applied for GMM, for classifying the given dataset into clusters. The algorithms were used to classify the given dataset into 4,6,8 and 10 clusters respectively.
- The clustering was based on the four major features: age, income per annum, spending score and gender.
- In order to determine the optimal value of *K,* in the K-Means clustering algorithm, an Elbow Graph was plotted. After K=6, a sharp increase in the distortion was observed. Thus for K-Means, the optimal value was found to be K=6(Figure 5).
- Similarly for GMM, a Silhouette graph was plotted in order to determine the optimal value of the number clusters. It was observed that the Silhouette score for the cluster size 6 was the highest. Thus the cluster size 6 was found to be optimum(Figure 10).
- Thus based on the determined cluster sizes and the corresponding K-Means and GMM plots(Figure 2 and Figure 7), customer analysis was performed.
- Referring to Figure2, It was observed that the plot could be split into two sections, people with high income(Yellow and Orange clusters) and people with low income(pink, blue, violet, and red). This could be further subdivided into two classes: people with high spending scores(Yellow) and people with low spending scores(Orange).
- It was observed that the people who belonged to the Yellow scores, were people who already have high spending scores and proportionally high annual incomes. It was also observed that people who belonged to this group were predominantly below the age of 40.
- However, on the other hand, people who belonged to the Yellow class consisted of two groups. Approximately 30% of the people in this group were over the age of 40 and the rest were below the age of 40. Let the latter group be called “Target Group A”.
- Also, people in the Red group(Figure 2), are in the middle. These people have decent annual incomes and relatively low spending scores. This group can be called “Target Group B”.
- In Order to increase sales, our major target audience is the people who belong to Target Groups A and B. These people can easily be convinced to buy more products by employing targeted advertising techniques.

References

[1]https://www.mathworks.com/discovery/unsupervised-learning.html
[2]https://en.wikipedia.org/wiki/K-means_clustering
[3https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a
[4]https://stanford.edu/~cpiech/cs221/handouts/kmeans.html
[5]https://blogs.oracle.com/datascience/introduction-to-k-means-clustering
[6]https://docs.rapidminer.com/latest/studio/operators/modeling/segmentation/expectation_maximization_clustering.html
[7]https://towardsdatascience.com/gaussian-mixture-modelling-gmm-833c88587c7f
[8]https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm
[9]https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c
[10]https://select-statistics.co.uk/blog/market-basket-analysis-understanding-customer-behaviour/
[11]https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce
[12]https://smartbridge.com/market-basket-analysis-101/
[13]https://scikit-learn.org/stable/
[14]https://www.scikit-yb.org/en/latest/api/cluster/elbow.html
[15]https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6