Detecting Spending Score of Mall Customers

Prepared By: Erum Akmal

1. **Problem Statement for Project:**
   In this project, I have predicted spending scores of mall customers using Clustering K Means. Mall customers will be grouped into different clusters based on similar spending behavior using K-Means Clustering.

2. **Data Sources and Description:**
   For data sources I used the link below:
   https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python.

   For this project, I worked on a spending score of mall customers using Clustering K Means.

   My data set included variables, Customer ID, Gender, Age, Annual Income (k$) and Spending Score of the mall customers.

3. **Data Mining Operations:**
   -Data wrangling:
   After loading and describing the data set, I proceeded with preprocessing of data.

   I split the data into train and test sets. Customer ID was removed from the dataset and gender column was converted into numerical. Data was normalized using minmax scaler. 80% are training features and 20% are testing features. K Means was initialized, and graphs were plotted.

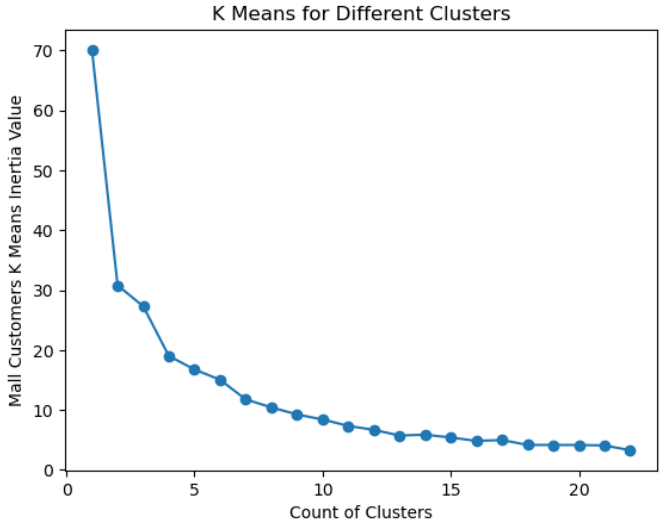
   In this project K means model was applied on 80% training features and assigns the clusters for each. While testing, the number of clustered were decided using elbow method. The Elbow method helped in identifying at which point clusters were meaningful. It shows that the value after which adding more clusters will not improve the performance of the model.

   -Libraries Used: I have used multiple libraries in this project including pandas, matplotlib and scikit-learn.

   -Data Modeling: I have performed the analysis below in my project:

   - Preprocessing of data
   - Splitting dataset into 80-20
   - Min Max Scaler
   - Clustering K Means

K Means for different clusters were plotted.



4. **Model Outputs:**
   Below is my evaluation of the 3 models for both testing and training data:

   Silhouette Score (Test): 0.3944
   Silhouette Score (Train): 0.3970

   Davies Bouldin Score (Test): 0.8680
   Davies Bouldin Score (Train): 0.9126

   Calinski Harabasz Score (Test): 29.4968
   Calinski Harabasz Score (Train): 111.13829

   Silhouette Score:
   Silhouette Score for both testing and training data is around 0.39. This implies that there are moderate cluster separations and model has reasonable cluster cohesion. Though the score is not very high, still it can be predicted that there

is some structure in mall customer data. This score usually ranges from -1 to 1 and a higher score is generally considered better. Silhouette Score measures how similar an object is to its own cluster as compared with other clusters.

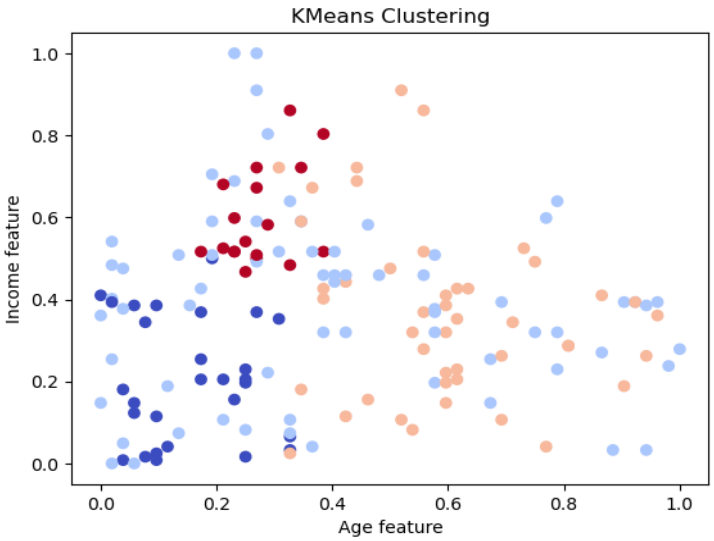
Davies Bouldin Score:
In this model, 0.86 on the test set implies well-separated clusters and clearly indicates the intra-cluster similarity is low. A low Davies-Bouldin score means that there is good compactness and separation. A lower Davies-Bouldin Index is considered better. It measures the similarity of the clusters and inter-cluster differences. If the score is close to 0, it means that clusters are well separated. In this study, the 0.86 index indicated good separation and is considered a good score since it is less than 1. This suggests that clusters are decent enough for practical use.

Calinski Harabasz Score:

For this model, a higher score in training (111.14) as compared with test (29.50) implies some overfitting or variance.

It can be concluded here that the clusters in the training set are more packed and better separated than in the test set. Training data has performed better than the testing data. Calinski-Harabasz Index estimates the ratio of cluster dispersion, and a higher score is considered better.

Graph was generated for visualizing clusters:



5. Limitations and Challenges:

Some of the limitations of this study are that there were limited features in the dataset. Only limited customer attributes were used here, other factors may affect the richness of clusters including location, online shopping trends etc.

The data used in this study is static, spending of mall customers can fluctuate over time. A time series approach can better analyze those trends.

In future, alternative clustering algorithms could be used and dimensionality reduction can be done.

6. Conclusion:

I was able to solve the problem partially. By adding more customer attributes, better clustering could have been done.

It can be concluded here that KMeans clustering model has worked well. Segmentation of mall customers based on other variables including spending behavior and demographics could be easily done. The scores for silhouette and Davies-Bouldin model imply that clusters are formed well with reasonable separation. The Calinski-Harabasz was low in the test set, but overall clustering performance is suitable for practical use.