|  |
| --- |
| Western Governors University |
| Data Mining II |
| Task 1 D212 |

|  |
| --- |
| Allison Casey  11-30-2024 |

**Part 1: Research Question**

**A1: PROPOSAL OF QUESTION**

Can customers be grouped using tenure and monthly charge to help make informed business decisions regarding targeting customers with certain products and marketing?

**A2: DEFINED GOAL**

One goal of the analysis would be to define at least two groups of customers from the data set.

**Part 2: Technique Justification**

**B1: EXPLANATION OF THE CLUSTERING TECHNIQUE**

The clustering technique used to analyze the churn dataset was K-Means. This technique works by clustering data points based on their distance from a cluster center. The number of cluster centers is manually set but can be analyzed to help find an optimal number of clusters. Once the number of clusters is defined the algorithm randomly assigns the centroids then each data point is assigned to its closest centroid then the mean is computed for all the points for each cluster and then the centroid is re-assigned. This process continues iteratively until the best clusters are found. Ideally the sum of distances between the assigned data points and their center is minimized. For this analysis this would be building clusters for monthly charge and tenure data to create categories of customers. An expected outcome would be that all data points are assigned to a cluster and these clusters help identify specific characteristics of customers that have a higher monthly charge and longer tenure to help identify customers to target that would fit the profile of this type of customer and increase the number of customers like this.

**B2: SUMMARY OF THE TECHNIQUE ASSUMPTION**

One assumption of K-Means clustering is that clusters are spherical and isotropic. (GeeksforGeeks, 2023)

**B3: PACKAGES OR LIBRARIES LIST**

|  |  |
| --- | --- |
| Library / Package | Justification |
| os and sys | Setting the working directory |
| pandas | Used for data frame capabilities |
| numpy | For array functionality |
| seaborn | For visualizations |
| pyplot from matplotlib | For visualizations |
| StandardScaler from sklearn.preprocessing | For scaling and standardizing the data for analysis |
| KMeans from sklearn.cluster | For building the kmeans model and fitting it to the data |
| silhouette\_score from | For analyzing the model and optimizing it |

**Part 3: Data Preparation**

**C1: DATA PREPROCESSING**

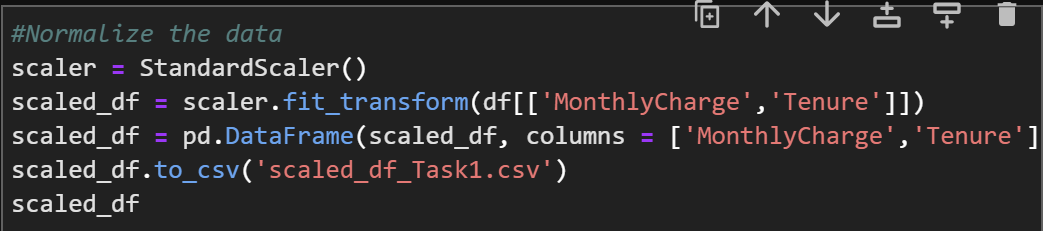
One data preprocessing goal relevant to the clustering technique is to standardize the data. This is done because K-Means creates clusters based on distance so if the two features have vastly different scales or were measured in different units then one feature might dominate how the clusters are set up and skew the results of the analysis.

**C2: DATA SET VARIABLES**

The data set variables that will be used for the initial clustering are tenure and monthly charge which are both continuous.

**C3: STEPS FOR ANALYSIS**

1. The data was read in and explored using df.info() to check for null values
   1. A black background with red and white text

      Description automatically generated
2. There were no null values, but one of the columns got read in incorrectly so this was fixed
   1. 
3. Columns were dropped that would not be needed for analysis
   1. *df = df.drop(columns=['Bandwidth\_GB\_Year','PaymentMethod','PaperlessBilling','StreamingMovies','StreamingTV','TechSupport','DeviceProtection','OnlineBackup','OnlineSecurity','Multiple','Phone','InternetService','Tablet','Population','Outage\_sec\_perweek','Email','Contacts','Yearly\_equip\_failure','Contract','Port\_modem','Customer\_id', 'Interaction', 'UID', 'City', 'State', 'County', 'Zip', 'Lat', 'Lng', 'TimeZone', 'Area', 'Job', 'Techie', 'Item1', 'Item2', 'Item3', 'Item4', 'Item5', 'Item6', 'Item7', 'Item8'])*
4. Tenure and MonthlyCharge were separated out and scaled for analysis so that the distances could be accurately compared since the units of measurement and the scale of the data for the two columns are different. This was done using StandardScaler() and then the data was converted back to a dataframe and finished cleaned and preprocessed data was exported to a csv.
   1. 

**C4: CLEANED DATA SET**

scaled\_df\_Task1.csv

**Part 4: Analysis**

**D1: OUTPUT AND INTERMEDIATE CALCULATIONS**

The optimal number of clusters was determined using two methods. A K-Means model was created using a for loop to create generate the model with several clusters between two and eleven then an array was filled with the inertias for each of the models and another array was filled with the silhouette score. These were both then potted to review the results. The elbow plot using the within cluster sum of squares or the inertia showed that there was an elbow to the graph at approximately 4 clusters and the silhouette score also showed a positive peak at four clusters so that was the chosen amount to use for the final model.

**A graph with a line

Description automatically generatedA graph with blue dots

Description automatically generated**

**D2: CODE EXECUTION**

*#Exaluate for the best number of clusters (wcss = Within-Cluster Sum of Square)*

*wcss = []*

*silhouette = []*

*for k in range (2, 11):*

*model = KMeans(n\_clusters = k, n\_init = 50, random\_state = 300)*

*model.fit(scaled\_df)*

*wcss.append(model.inertia\_)*

*silhouette.append(silhouette\_score(scaled\_df, model.labels\_))*

*wcss\_s = pd.Series(wcss, index = range(2,11))*

*silhouette\_s = pd.Series(silhouette, index = range(2,11))*

*#plot elbow with wcss*

*plt.figure(figsize = (12,10))*

*ax = sns.lineplot(y = wcss\_s, x = wcss\_s.index)*

*ax = sns.scatterplot(y = wcss\_s, x = wcss\_s.index, s = 200)*

*ax = ax.set(xlabel = 'Optimal Cluster Number (k)', ylabel = 'Within Cluster Sum of Squares')*

*#plot silhouette*

*plt.figure(figsize = (12,10))*

*ax = sns.lineplot(y = silhouette\_s, x = silhouette\_s.index)*

*ax = sns.scatterplot(y = silhouette\_s, x = silhouette\_s.index, s = 200)*

*ax = ax.set(xlabel = 'Optimal Cluster Number (k)', ylabel = 'Silhouette Score Average')*

*#Create Final Model*

*k\_model = KMeans(n\_clusters = 4, n\_init = 25, random\_state = 300)*

*k\_model.fit(scaled\_df)*

See D212\_Task1.ipynb for full executable code

**Part 5: Data Summary and Implications**

**E1: QUALITY OF THE CLUSTERING TECHNIQUE**

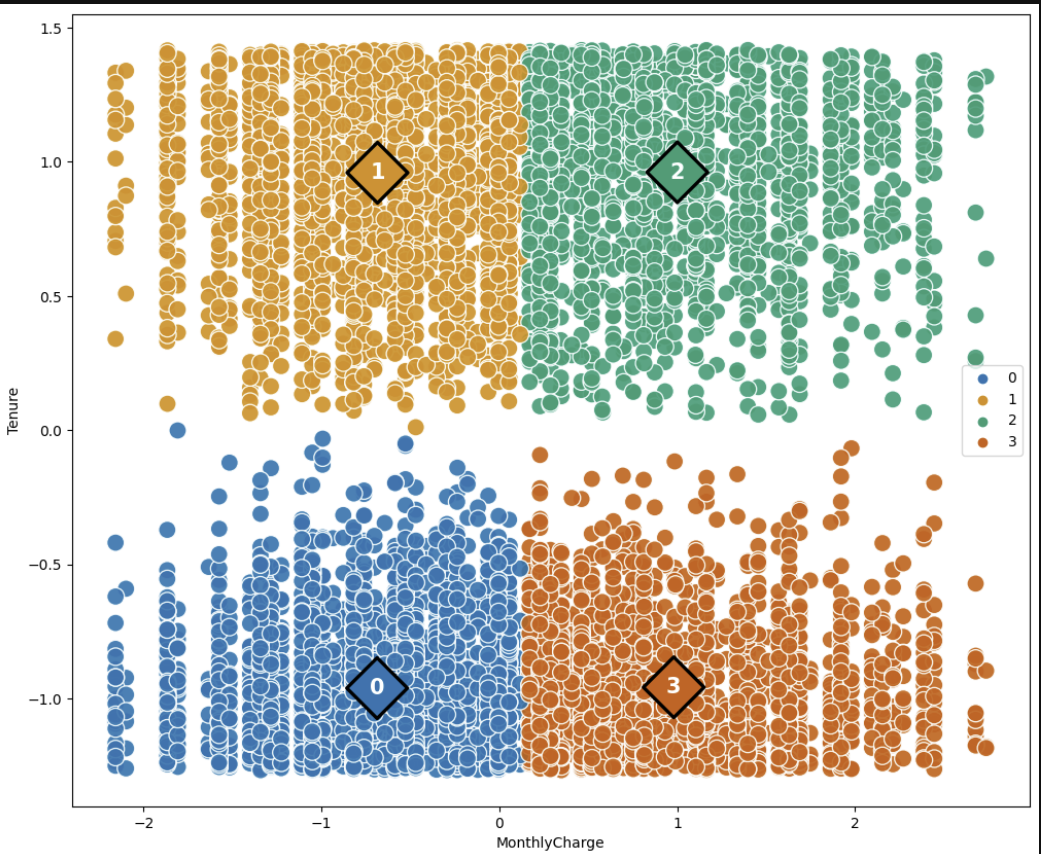
The quality of the clustering technique can be evaluated using the silhouette score of the final model. This measures how well the data points fit within their respective clusters compared to the other clusters, so a higher score indicates a better fitting model. The cluster score for the final model was .38 on a scale from -1 to 1 so the model may be useful as it is not negative, but the structure of the clusters may not be very strong.

**E2: RESULTS AND IMPLICATIONS**

Using the clusters, a cluster label was created and applied to the data frame containing the other remaining features for analysis. The categorical variables were then one hot encoded and the data aggregated based on cluster to gather information about the customers in each cluster. The final clusters appear to be grouped as customers with low monthly charge and low tenure, customers with high monthly charge and low tenure, customers with low monthly charge and long tenure, and customers with high monthly charge and high tenure. Many of the features don’t appear to have a lot of variance between the different clusters, but the ones that do stand out are the churn rate ones which appears to be significantly higher for the customers in the cluster with the lower monthly charge and low tenure and is especially high for customers with the higher monthly charge and low tenure so this is definitely a group of customers that could be targeted to prevent churn.

A group of images on a black background

Description automatically generated



**E3: LIMITATION**

One limitation of the analysis is that the silhouette score for the clusters is not particularly strong so there may be better features to base the clusters off to create better clusters and better understand the customers.

**E4: COURSE OF ACTION**

A course of action based on the analysis would be to target the customers in cluster three to prevent churn. This may mean specifically advertising certain services to them that some of the other clusters may have or finding other ways to try and lessen the churn rate for this cluster. It would also be advisable to do further cluster analysis using different features to try and get a better silhouette score as well as just be able to further explore relationships within the data.

**Part 6: Demonstration**

**F1: DASHBOARD ALIGNMENT**

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=2083af6c-6365-440c-a567-b239004929c9

**Sources**

GeeksforGeeks. “Demonstration of K-Means Assumptions.” *GeeksforGeeks*, 9 Dec. 2023, www.geeksforgeeks.org/demonstration-of-k-means-assumptions/. Accessed 30 Nov. 2024.

Kavlakoglu, Eda, and Vanna Winland. “What Is K-Means Clustering?” *IBM*, 27 Aug. 2024, www.ibm.com/topics/k-means-clustering. Accessed 30 Nov. 2024.

“Silhouette\_score.” *Scikit*, scikit-learn.org/dev/modules/generated/sklearn.metrics.silhouette\_score.html. Accessed 30 Nov. 2024.

“Standardscaler.” *Scikit*, scikit-learn.org/dev/modules/generated/sklearn.preprocessing.StandardScaler.html. Accessed 30 Nov. 2024.