­­­­­­­

D212 Clustering Techniques

Telecommunication Churn Data

Darien Nguyen

Western Governors University

D212 Task One: Clustering Techniques

June 6, 2023

**Part I: Research Question**

A.  Describe the purpose of this data mining report by doing the following:

1.  Propose **one** question relevant to a real-world organizational situation that you will answer using **one** of the following clustering techniques:

•  *k*-means

•  hierarchical

Can k-means clustering be used to identify distinct customer segments based on their purchasing behavior, and how can these segments inform targeted marketing strategies to increase customer engagement and retention?

2.  Define **one** goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

Our goal is to identify distinct customer segments based on their characteristics and behavior using K-means clustering.

**Part II: Technique Justification**

B.  Explain the reasons for your chosen clustering technique from part A1 by doing the following:

1.  Explain how the clustering technique you chose analyzes the selected dataset. Include expected outcomes.

The k-means clustering technique is a valuable tool for businesses to analyze customer data and gain insights that drive effective strategies. By examining the selected dataset, it helps answer important questions related to customer segmentation, enabling businesses to tailor their approaches accordingly.

For instance, imagine a retail company aiming to enhance its marketing strategies. By applying k-means clustering to their customer data, they can identify distinct groups of customers with similar buying patterns, demographics, or preferences. This analysis helps answer questions such as:

1. Which groups of customers behave similarly when making purchases?
   * The k-means algorithm identifies clusters of customers who share common traits, like frequent buyers, high spenders, or those with specific product preferences.
2. What are the defining characteristics of each customer segment?
   * By analyzing the data, businesses gain insights into the demographics, interests, and preferences associated with each cluster. This knowledge helps tailor marketing campaigns and product offerings to each segment's specific needs.
3. How can marketing strategies be customized for each customer segment?
   * With the identified clusters, businesses can develop personalized marketing strategies for each segment. This involves crafting relevant messages, promotions, and product recommendations that resonate with the unique needs of each group.
4. How can customer retention be improved?
   * Clustering allows businesses to identify segments with a higher churn rate. By understanding the common characteristics of customers who churn, strategies can be devised to proactively engage and retain them. This might involve offering loyalty programs, personalized incentives, or addressing specific pain points.

2.  Summarize **one** assumption of the clustering technique.

One assumption of the clustering technique in k-means, is that the clusters formed are round and have similar sizes. In simpler terms, the technique assumes that the data points within each cluster are closely grouped together and have similar characteristics.

This assumption is important because k-means works by finding the center points of clusters based on the distance between data points. It assumes that the clusters can be represented by a single point at the center and that the distances from the data points to the center reflect how similar or different they are from each other.

3.  List the packages or libraries you have chosen for Python or R, and justify how *each* item on the list supports the analysis.

1. Pandas (imported as pd): Pandas is a powerful library used for data manipulation and analysis. It provides convenient data structures and functions for handling structured data, making it ideal for working with datasets. In this analysis, Pandas was used to manipulate the data, perform operations like grouping and merging, and create the final dataset for clustering.
2. Matplotlib.pyplot (imported as plt): Matplotlib is a widely-used library for data visualization in Python. The pyplot module provides a comprehensive set of plotting functions and tools, allowing for the creation of various types of charts and graphs. In this analysis, Matplotlib was used to visualize the clustering results and display scatter plots of the age and income variables.
3. Seaborn (imported as sns): Seaborn is a higher-level interface built on top of Matplotlib. It offers enhanced functionality and aesthetic improvements for creating informative statistical graphics. In this analysis, Seaborn was used to enhance the visual representation of the clusters by providing a colormap for differentiating the clusters in the scatter plot.
4. Scipy.stats (specifically the zscore function): Scipy is a scientific computing library that includes various statistical functions. The stats module within Scipy provides the zscore function, which calculates the z-scores of the numeric columns. In this analysis, z-scores were computed using Scipy to standardize the variables before clustering. Standardization ensures that variables with different scales contribute equally to the clustering process.
5. Sklearn.cluster (imported as KMeans): Scikit-learn is a widely-used machine learning library in Python. The cluster module within Scikit-learn includes the KMeans class, which provides an efficient implementation of the K-means clustering algorithm. In this analysis, the KMeans class was used to perform the actual clustering of the normalized data, assigning cluster labels to each data point.
6. Tabulate: Tabulate is a helpful package for creating formatted tables from structured data. It simplifies the process of presenting tabular data in an organized and visually appealing manner. In this analysis, Tabulate was used to format and display the results of the cluster counts, churn percentages, and other relevant information.

Each of these packages brings specific functionalities and capabilities that contribute to different aspects of your analysis, such as data manipulation, statistical modeling, visualization, machine learning, and table formatting. By utilizing these packages, you can perform a comprehensive analysis of your k-means clusters, including data preprocessing, statistical modeling, visualizing the results, and presenting the findings in a structured and visually appealing manner.

**Part III: Data Preparation**

C.  Perform data preparation for the chosen dataset by doing the following

1.  Describe **one** data preprocessing goal relevant to the clustering technique from part A1.

One data processing technique used in this analysis was removing null or incomplete data values. Missing values can impact the clustering process by introducing biases or affecting the distance calculations between data points. Therefore, it is important to address missing values before performing clustering analysis.

In the code, the **dropna()** function is used to remove rows with null values from the dataset. Specifically, the line **churn\_data = churn\_data.dropna()** is responsible for this step. By calling **dropna()** on the **churn\_data** DataFrame, any rows that contain missing values are removed, resulting in a cleaned dataset with complete information.

This step ensures that the clustering algorithm operates on a dataset without missing values, enabling accurate distance calculations and cluster assignments. Removing null values helps to create a more reliable and representative dataset for clustering analysis.

By utilizing the **dropna()** function, the code effectively handles missing values and fulfills the data preprocessing goal of removing null values to clean the dataset for clustering with k clusters.

2.  Identify the initial dataset variables that you will use to perform the analysis for the clustering question from part A1, and label *each* as continuous or categorical.

1. Population (Continuous)
2. Age (Continuous)
3. Income (Continuous)
4. Outage\_sec\_perweek (Continuous)

These variables are used to analyze the churn behavior and perform clustering analysis. The variables are categorized as either continuous or categorical based on their nature.

The continuous variables (Population, Age, Income, and Outage\_sec\_perweek) represent numerical measurements that can take any value within a range. These variables are used to capture quantitative information about the customers.

3.  Explain *each* of the steps used to prepare the data for the analysis. Identify the code segment for *each* step.

Step 1: Reading the data into a DataFrame

The code reads the data from the 'churn\_clean.csv' file and stores it in a DataFrame called 'data'.

Step 2: Selecting relevant columns for cluster analysis

The code selects specific columns ('Population', 'Age', 'Income', and 'Outage\_sec\_perweek') from the 'data' DataFrame and creates a new DataFrame called 'churn\_data'. These columns are deemed relevant for the churn analysis.

Step 3: Removing rows with missing values

To ensure the integrity of the data used for analysis, this code removes any rows in the 'churn\_data' DataFrame that contain missing values. This step helps ensure the data is complete and avoids potential issues in the analysis.

Step 4: Extracting numeric columns

This code creates a new DataFrame called 'normalized\_data' by selecting only the numeric columns ('Population', 'Age', 'Income', and 'Outage\_sec\_perweek') from the 'churn\_data' DataFrame. These numeric columns will be used for further analysis.

Please note that in the provided code, there is no explicit normalization step performed on the data.

**Part IV: Analysis**

D.  Perform the data analysis and report on the results by doing the following:

1.  Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.

In this analysis, the K-means algorithm is applied to the normalized data, which consists of the numeric variables 'Population', 'Age', 'Income', and 'Outage\_sec\_perweek'. The algorithm iteratively assigns data points to the nearest centroid and recalculates the centroids until convergence. The number of clusters, k, is determined using the Elbow method, where the inertia (sum of squared distances of samples to their closest cluster center) is plotted against different values of k, and the "elbow" point is selected as the optimal number of clusters.

A graph with a line

Description automatically generated

A cluster of numbers and a few digits

Description automatically generated

An optimal number of clusters was determined to be 4 based on the curvature of the elbow curve, the K-means algorithm is applied again with the chosen k value. The resulting cluster labels are then assigned to the original dataset.

To analyze the results, the count of data points in each cluster is calculated and visualized. A scatter plot matrix was created, where our variables are plotted, and the points are colored according to their assigned cluster labels. This plot helps visualize how the data points are grouped into different clusters based on their age and income.

A collage of different colored dots

Description automatically generated

2.  Provide the code used to perform the clustering analysis technique from part 2.

import pandas as pd

import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

from tabulate import tabulate

import seaborn as sns

# Read the data into a DataFrame

data = pd.read\_csv('churn\_clean.csv')

# Select the columns relevant for analysis

columns = ['Population', 'Age', 'Income', 'Outage\_sec\_perweek']

churn\_data = data[columns]

# Remove any rows with missing values

churn\_data = churn\_data.dropna()

# Standardize the data

scaler = StandardScaler()

normalized\_data = scaler.fit\_transform(churn\_data)

# Determine the optimal number of clusters using the Elbow method

inertia = []

for k in range(1, 11):

kmeans = KMeans(n\_clusters=k, random\_state=42)

kmeans.fit(normalized\_data)

inertia.append(kmeans.inertia\_)

# Plot the Elbow curve

plt.plot(range(1, 11), inertia)

plt.xlabel('Number of Clusters')

plt.ylabel('Inertia')

plt.title('Elbow Curve')

plt.show()

# Choose the optimal number of clusters based on the Elbow curve

k = 4

# Perform K-means clustering

kmeans = KMeans(n\_clusters=k, random\_state=42)

kmeans.fit(normalized\_data)

# Assign cluster labels to the original data

churn\_data['Cluster'] = kmeans.labels\_

# Analyze the results

cluster\_counts = churn\_data['Cluster'].value\_counts().reset\_index()

cluster\_counts.columns = ['Cluster', 'Count']

# Generate the table using tabulate

table = tabulate(cluster\_counts, headers='keys', tablefmt='pretty')

# Print the table

print("Cluster Counts:")

print(table)

# Save the cleaned data to a CSV file

churn\_data.to\_csv('D212\_p1\_cleaned\_churn\_data.csv', index=False)

# Assign cluster labels to the original data

churn\_data['Cluster'] = kmeans.labels\_

# Create a scatter plot matrix

sns.set(style="ticks")

sns.pairplot(churn\_data, vars=columns, hue="Cluster", palette="viridis")

# Display the plot

plt.show()

**Part V: Data Summary and Implications**

E.  Summarize your data analysis by doing the following:

1.  Explain the accuracy of your clustering technique.

In the case of K-means clustering, the accuracy was assessed through the use of a Silhouette Coefficient. The values produced from this analysis revealed a 0.25 Silhouette Coefficient and a 21671.86 inertia value. Silhouette Coefficient describes the cluster separation and cohesion of our clusters based on 1 being well defined and 0 being poorly separated. With a coefficient of 0.25, we can see a moderate degree of separation between our clusters.

The Inertia value of 21671.86, represents the sum of squared distances of samples to their closest cluster center. Higher Inertia values indicate more scattered or less compact clusters. In this case, the relatively high inertia suggests that the clusters may have a considerable amount of spread or dispersion within them, indicating that the clusters might not be as compact as desired (Bhardwaj, 2020).

A screenshot of a computer code

Description automatically generatedOverall, the analysis reveals that the K-means clustering technique achieved a moderate level of separation and cohesion among the clusters.

2.  Discuss the results and implications of your clustering analysis.

The clustering analysis using the k-means algorithm resulted in the identification of four distinct customer clusters based on the selected continuous variables. Each cluster represents a group of customers with similar characteristics and preferences.

Cluster 0, characterized by higher population, older age, higher income, and longer outage duration, suggests a segment of customers who may have higher purchasing power and are more tolerant of service outages. Cluster 1, with lower population, younger age, lower income, and shorter outage duration, indicates a different demographic profile and potentially different needs and preferences. Cluster 2 and Cluster 3 also exhibit their own unique patterns and customer profiles.

These results have important implications for businesses. By understanding the distinct characteristics of each customer cluster, companies can tailor their marketing strategies, product offerings, and customer service approaches to better meet the needs and preferences of each segment. This targeted approach enhances the effectiveness of marketing campaigns, improves customer engagement, and increases customer satisfaction and loyalty.

3.  Discuss **one** limitation of your data analysis.

One limitation of the k-means clustering technique in our data analysis is the assumption of equal variances and isotropic clusters. K-means assumes that the clusters formed are spherical and have the same scatter in all dimensions. However, in real-world datasets, the clusters may have different shapes and varying levels of scatter. This limitation can impact the accuracy and validity of the clustering results.

4.  Recommend a course of action for the real-world organizational situation from part A1 based on your results and implications discussed in part E2.

With our four client clusters, we can formulate specific market strategies that target our clients to promote and enhance retention and spending rates.

Cluster 0 showcases customers with higher population, older age, higher income, and longer outage duration. This segment represents customers who may have higher purchasing power and are more tolerant of service outages. To retain and engage these customers, we can develop personalized loyalty programs and incentives that align with their preferences and needs. Offering exclusive benefits, tailored promotions, and exceptional customer service can foster a sense of loyalty and satisfaction among this cluster.

Cluster 2 represents a different demographic profile with lower population, younger age, lower income, and shorter outage duration. For this segment, our marketing strategies should focus on capturing their attention and addressing their unique needs. Targeted communication through digital channels, social media campaigns, and appealing promotional offers can attract and engage these customers. Additionally, by providing seamless and reliable services, we can build trust and loyalty among this cluster.

Clusters 2 and 3 exhibit their own distinctive patterns and customer profiles. Understanding the characteristics and preferences of these clusters allows us to tailor our marketing messages and offerings to meet their specific needs. For example, Cluster 2 may respond well to personalized recommendations, cross-selling, and up-selling strategies based on their preferences and past purchase behavior. Cluster 3, on the other hand, may appreciate a focus on affordability, value for money, and discounts.

Analyzing the reasons behind customer churn within each cluster can provide valuable insights for improving our services and customer experience. By addressing pain points specific to each cluster, such as outage durations, billing issues, or service quality concerns, we can proactively resolve customer issues and prevent churn. Implementing customer feedback mechanisms, conducting satisfaction surveys, and actively listening to customer concerns will allow us to continuously improve our services and retain our customers.

**Part VI: Demonstration**

F.  Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the programming environment.

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=0b8543ce-0af5-4772-8e1a-b03c00a43895#>

G.  Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

No Sources used

H.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

Bhardwaj, A. (2020, May 27). *Silhouette coefficient : Validating clustering techniques*. Medium.

<https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c>