***Clustering Techniques***

*WGU*

*Course Number: 603*

*Trayvonious Pendleton*

*011205284*

**B1: Propose one question relevant to a real-world organizational situation**

What distinct groups of customers can be identified based on their service usage, contract details, and monthly charges, and how can these clusters inform targeted retention strategies?

This question aims to help the company understand who its customers are and how they behave. This is crucial in crafting targeted marketing or retention efforts, especially in a high-churn industry.

**B2: Define one goal of the data analysis**

The goal of the analysis is to segment the customer base into distinct clusters using K-means clustering based on continuous variables such as tenure, monthly charges, and bandwidth usage so that the telecommunications company can better understand and target different types of customers for churn reduction strategies.

**C1: Explanation of clustering technique and expected outcomes**

K-means clustering analyzes the dataset by partitioning customers into a predefined number of clusters (k), where each customer is grouped with others with similar characteristics. It works by:

* Selecting k initial cluster centroids randomly
* Assigning each customer to the nearest centroid based on Euclidean distance.
* Recomputing the centroids by averaging the feature values of all customers in a cluster.
* Repeating the assignment and centroid update steps until the clusters stabilize

This analysis uses continuous variables such as tenure, monthly charges, and bandwidth usage to represent customer behavior. The expected outcome is a set of distinct customer segments that reflect usage patterns, pricing sensitivity, and duration with the company. These insights will help the company develop tailored retention strategies, upselling, and service improvements.

**C2: Summarize one assumption of the clustering technique**

One key assumption of K-means clustering is that the data is isotropic and clusters are spherical and equally sized. In other words, K-means assumes that clusters are roughly similar in size and density and are well separated from each other in the feature space. This can affect performance if the real-world customer segments are not naturally spherical or have overlapping characteristics.

**C3: Python Libraries and Justification**

|  |  |
| --- | --- |
| Library | Justification |
| Pandas | Used to load, explore, clean, and manipulate the dataset (e.g., filter columns, handle missing values). |
| Numpy | Provides mathematical operations, especially for distance computations and numerical transformations. |
| Scikit-learn | Includes the KMeans clustering algorithm, preprocessing tools (e.g., StandardScaler), and evaluation tools like silhouette score. |
| Matplotlib | Used for visualizing cluster results in 2D or 3D plots. |
| Seaborn | Enhances data visualization, especially for pair plots, distributions, and heatmaps. |

**D1: Describe one data preprocessing goal**

The goal is to scale the continuous features (e.g., tenure, monthly charges, bandwidth usage) so that each variable contributes equally to the distance calculations in the K-means algorithm. Without scaling, variables with more extensive ranges could dominate the clustering results.

**D2: Identify the initial dataset variables**

|  |  |  |
| --- | --- | --- |
| Variable | Description | Type |
| Tenure | Number of months the customers has been with the company | Continuous |
| MonthlyCharge | Average amount charged to the customer per month | Continuous |
| Banwidth\_GB\_Year | Average GB used in a year | Continuous |

**D3: Explain each of the steps used to prepare the data**

**Step 1: Import libraries**

import pandas as pd

from sklearn.preprocessing import StandardScaler

from sklearn.cluster import KMeans

**Step 2: Load the dataset**

df = pd.read\_csv(“churn\_clean.csv”)

**Step 3: Select relevant continuous features**

Features = df[[‘Tenure’, ‘MonthyCharge’, ‘Bandwith\_GB\_Year’]]

**Step 4: Handle missing values**

features = features.dropna()

**Step 5: Scale the data**

scaler = Standard scaler()

scaled\_features = scaler.fit\_transform(features)

**E1: Determine the Optimal Number of Clusters**

**A graph with a line

AI-generated content may be incorrect.**

The Elbow Method was used to determine the optimal number of clusters for the K-means clustering algorithm. This method involves plotting the Within-Cluster Sum of Squares (WCSS) or inertia against different values of k (number of clusters). The goal is to find the point at which the inertia decreases slower, known as the “elbow point.”

As shown in the elbow curve, the inertia drops sharply between k = 1 and k = 4, after which the decrease slows significantly. This suggests that k = 4 is the optimal number of clusters, as it provides a good balance between minimizing intra-cluster distance and avoiding overfitting.

By selecting k = 4, we aim to segment the customer base into four groups based on their tenure, monthly charges, and bandwidth usage. These segments will allow the business to understand customer behavior more deeply and potentially develop targeted retention strategies.

**F1: Visualize the clusters and explain the quality of the clusters created**

A diagram of a number of colored dots

AI-generated content may be incorrect.

The customer segments were visualized using a scatter plot of standardized Tenure vs. Monthly Charge, with each point colored by its assigned cluster (Figure 2). The visualization shows a clear separation between the four clusters, confirming the effectiveness of the K-means algorithm in distinguishing groups within the data. Each cluster occupies a distinct region of the feature space, indicating that K-means with **k=4** produced quality clusters that reflect meaningful customer behavior patterns.

**F2: Discuss the results and implications**

The results revealed four customer clusters:

* **Cluster 0**: Short tenure, high charges, likely new premium users.
* **Cluster 1**: Long tenure, high charges, likely loyal, high-value customers.
* **Cluster 2**: Short tenure, low charges, potentially price-sensitive or disengaged users.
* **Cluster 3**: Long tenure, low charges, long-term budget-conscious users.

These segments provide insights for targeting. For instance, Cluster 2 may be at higher risk of churn and could be targeted with onboarding or engagement campaigns. Cluster 1, being high-value customers, could be nurtured with loyalty rewards.

**F3: Discuss one limitation of your data analysis**

A key limitation is that K-means only handles continuous variables, which means omitting categorical features like contract type or internet service. These categorical variables could have provided valuable insights if encoded and included in another clustering approach like Gower distance with hierarchical clustering.

**F4: Recommend a course of action**

Based on the analysis, the telecom company should:

* Create retention programs for Cluster 2 customers, such as promotions or personalized outreach.
* To encourage continued engagement, offer loyalty rewards or bundled upgrades to Cluster 1 customers.
* Explore cluster behavior over time to detect early churn patterns or emerging customer trends.