D212 Clustering Technique

Nhi Le

Western Governors University

October 19, 2023

**Part I: Research Question**

**A1. Proposal of question**

How can we cluster customers to get insight into their data (GB) usage habits?

**A2. Defined goal**

The goal of this analysis is to cluster customers based on their bandwidth (GB) usage and tenure to find the characteristics in each group. From this analysis, the stakeholders can understand customers better and come up with any data providing capacity plan to increase customer’s tenure.

**Part II: Technique Justification**

**B1. Explanation of the clustering technique**

I will be using k-means as my clustering technique. This is an unsupervised machine learning algorithm that sorts unlabeled dataset into clusters (GeeksforGeeks, 2023). K is called mean, cluster centers or centroids. After k point is randomly picked, each item is categorized into the closest cluster. Once all items have been grouped into clusters, the centroids are updated. These newly calculated centroids serve as the new centers for their respective clusters. This iterative process continues until the centroids no longer exhibit any alterations, that is when we have our clusters (GeeksforGeeks, 2023). The expected outcomes are that optimal number of clusters will be chosen, and the unsorted data will be grouped based on patterns, parallels, and variations.

**B2. Summary of the technique assumption**

The assumption of k-means clustering technique is that simple Euclidean distance is used to compute distance. It is important to normalize the values to reduce misleading results. Lastly, the technique assumes that an appropriate number of clusters is chosen, and all relevant variables need to be included in the analysis (IBM, n.d.).

**B3. Packages or Libraries list**

Here are the packages and libraries I will use in Python:

* NumPy is used for working with arrays.
* Pandas is used for working with a data set. Example, we can use .read.csv() to load the data set to Python. Or I can use .info() to get the information of the data set.
* Matplotlib is a comprehensive library for creating visualizations. I will use matplotlib.pyplot submodule for creating histograms. They will help me detect outliers. This package also helps me to create bivariate visualization such as scatter plots. I can also use it to create plots to find the optimal K values with the elbow method.
* Seaborn is also used to create boxplots to detect outliers and create univariate/ bivariate visualizations.
* Scikit-learn is used for machine learning, especially for k-means clustering in this analysis. Example, StandardScaler is for normalizing the data, KMeans is for k-means clustering, metrics is for silhouette scores.

**Part III: Data Preparation**

**C1. Data preprocessing**

My data pre-processing goal is to detect missing data, duplicate data, and outliers, then decide to treat them with appropriate methods.

**C2. Data set variables**

The 2 continuous variables including:

* ‘Bandwidth\_GB\_Year’ (the average yearly amount of data used, in GB, per customer) is a continuous variable.
* ‘Tenure’ is continuous variable to record how many months the customer has been with the provider.

**C3. Steps for analysis**

Data preparation steps:

* Import dataset churn\_clean.csv into Jupyter Notebook.
* Get information (column names, data types).
* Pick only necessary variables for this analysis: Tenure and Bandwidth\_GB\_Year.
* Get statistical details (count, min, max, mean, std, percentile) of the variables.
* Detect duplicates and delete the duplicated records if there are any.
* Find missing data and impute missing data with meaningful measures of central tendency (mean, median, or mode).
* Find outliers and treat them by removing them, retaining them, excluding them, or imputing them with the median.
* Run univariate and bivariate visualizations to see the spread of data.
* Extract the cleaned dataset as CSV file named ‘churn\_prepared.csv’.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A close up of a number

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated



**C4. Cleaned data set**

The cleaned data set will be submitted as ‘churn\_prepared.csv’ along with this doc file.

**Part IV: Analysis**

**D1. Output and intermediate calculations**

I will be using the elbow method to find the optimal number of clusters. In this method, I will create a plot demonstrating the number of clusters along with the sum of squared errors (SSE) of the model. The plot created has the point where the SSE begins to bend or level off, looking like an elbow. The elbow point is the optimal K or the number of clusters (Zach, 2022).

**D2. Code execution**

A screenshot of a computer program

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a computer

Description automatically generated

**Part V: Data Summary and Implications**

**E1. Quality of the clustering technique**

To evaluate the quality of the clusters, I will be calculating the silhouette scores. The silhouette score is between -1 and 1. A good score should be positive. ‘A negative silhouette score symbolizes that a point is closer to the centroid of a different cluster than the cluster it’s currently assigned to’ (Kaplan, 2023).

In this analysis, the average silhouette score is positive, at 0.813, with no data point with negative silhouette score. The cluster is excellent with a high positive average silhouette score and has positive scores for all data points.

A screenshot of a computer program

Description automatically generated

**E2. Results and implications**

For the cluster of the set Tenure and Bandwidth (GB) Usage, I could sort the data into 2 groups by using k-means clustering. The first group is for customers with lower tenure and lower data usage. The second group is for customers with higher tenure and higher data usage. With the high average silhouette score, at 0.813, the clustering is significantly practical. Through this analysis, the stakeholder can see that when the customers stay with the company longer, they tend to use more data.

**E3. Limitation**

The limitation affecting the results of my data analysis is that the data set did not come from a data warehouse. Since I did not know the data set’s owners, I could not confirm with them the quality of the data set. Especially, when dealing with outliers, I could not be certain whether these outliers are legitimate entries or errors. According to Google Developers (n.d.), outliers can drag centroids. Therefore, by not removing or treating them right, the analysis results might not be absolutely accurate.

**E4. Course of action**

From the analysis’s results, loyal customers used more data than the customers with low tenure did. The recommendation is to make sure that we can provide sufficient data for customers to keep them with the company longer. Moreover, the stakeholders should consider a data limit strategy in different ranges of tenure to make sure customers are satisfied with the smooth services. Along with that, they can consider a marketing campaign to suggest that if the customers stay longer with the company, they will get higher data limit.

**Part VI: Demonstration**

**F. Panopto video**

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=244984ae-1ef3-4e4a-ba02-b09f00f96b6e>

**G. Sources for third-party code**

GeeksforGeeks. (2023, August 25). *K means clustering - introduction*. GeeksforGeeks. https://www.geeksforgeeks.org/k-means-clustering-introduction/

Zach. (2022, August 31). *K-means clustering in Python: Step-by-step example*. Statology. https://www.statology.org/k-means-clustering-in-python/

Kaplan, D. (2023, March 1). *K-means accuracy python with silhouette method " EML*. EML. https://enjoymachinelearning.com/blog/k-means-accuracy-python-silhouette/?expand\_article=1

**H. Sources**

IBM (n.d.). *K-Means Cluster Analysis*. IBM. https://www.ibm.com/docs/en/spss-statistics/saas?topic=features-k-means-cluster-analysis

Google Developers. (n.d.). *K-means advantages and disadvantages*. Google Developers. https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages

GeeksforGeeks. (2023, August 25). *K means clustering - introduction*. GeeksforGeeks. https://www.geeksforgeeks.org/k-means-clustering-introduction/

Zach. (2022, August 31). *K-means clustering in Python: Step-by-step example*. Statology. https://www.statology.org/k-means-clustering-in-python/

Kaplan, D. (2023, March 1). *K-means accuracy python with silhouette method " EML*. EML. https://enjoymachinelearning.com/blog/k-means-accuracy-python-silhouette/?expand\_article=1