**CLUSTERING TECHNIQUES (OFM1 TASK 1)– D212**

**Performance Assessment**

**Western Governors University**

**By: Christian LeBlanc**

***Part I: Research Question***

**A-1**

**What insight into churn customers would using a k-means technique on the age and monthly charge variables in the dataset?**

**A-2**

**My goal in doing the k-means analysis on the churn customers is to find clusters of characteristics in the variables age and monthly charge. This information could be used on the current customers to identify potential high risk for becoming churn and allow for management to come up with a course of action to keep more customers from leaving.**

***Part II: Technique Justification***

**B-1**

**The clustering technique I chose is the k-means technique. The k-means technique groups cluster of similar data together to find a useful pattern. It does this by clustering the data into a preset number, “k”, of clusters. The clusters are determined by minimalizing the sum of the squared distances between the data point and their assigned cluster mean.**

**I am expecting to see high monthly charge to be included in the largest clusters, as high price seems like a reason someone would switch to a new company. I would also expect that the churn customers with younger ages to be a strong cluster. My thinking on this is younger people will want to try different companies before settling on staying at one for long periods.**

**B-2**

**The k-means requires the assumption that the shapes of the clusters are spherical. The radius of the spheres is from the centroid to the furthest data point.**

**B-3**

**I used the Python language in a Jupyter notebook. I will be importing the following packages/libraries:**

**Pandas- for data manipulation such as uploading the csv**

**Matplotlib and seaborn- for creating graphs**

**Sklearn- the abilities to do the k-means, get the silhouette score, and the scaling of my variables.**

***Part III: Data Preparation***

**C-1**

**A data preprocessing goal that is relevant to the k-means clustering technique is normalizing the data. By normalizing the data using the z-score it will set the means to zero and the standard deviation at one. This is done to make sure the variables with larger numbers do not influence the clustering disproportionately.**

**C-2**

**Age and Monthly charge are the variables that will be used to perform the analysis to answer the question in part A-1. The k-means clustering technique requires for the variables to be continuous variables. Both age and monthly charge satisfy this requirement as they are continuous variables.**

**C-3**

**I used the .info command to check for nulls and none were present. I used .value\_counts to see the number of “Yes” responses that I will focus on. Then I created a new data set by using .loc on the** **Churn column with response of “Yes” to focus only on the churn customers. I ran .info to make sure I got the expected 2,650 entries as expected from the value count on churn. I used .describe to look at the statistics on Age and MonthlyCharge. I checked for any outliers by using** the rule of thumb of z-value is either greater than 3, or less than -3 to identify outliers(Larose & Larose, 2019). No outliers where detected in Age or MonthlyCharge. From there I created a scatter plot to visually inspect Age vs Monthly charge. I used StandardScaler to scale the two variables and then pd.DataFrame to get my scaled dataset. I used .describe on the scaled data to make sure I got the expected results such as a 0 mean and standard deviation of 1.

**C-4**

**Use scaled data as told by Dr. Kesselly Kamara in Constructing and Fitting the K-means on Dataset. It is uploaded as a csv file named d212scaled.**

***Part IV: Analysis***

**D-1**

**In looking to determine the optimal number of clusters in my data set I used both the elbow method and the Silhouette score. They both gave a result of 4 clusters is the best fit. The elbow method graph shows an elbow at 4. The silhouette score of 0.38552720 for 4 is the highest of the tested “k” values.**

**D-2**

**Below are screenshots of the code used to do the k-means technique:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a computer program

Description automatically generated

***Part V: Data Summary and Implications***

**E-1**

**The quality of the clusters created are not ideal. The silhouette scores range from approximately 0.352 to 0.386. The high score of just approximately 0.386 is still a very low silhouette score and this indicates a low-quality cluster.**

**E-2**

**The results of this k-means clustering analysis does not offer much usefulness. The low silhouette score indicates that it is a low-quality clustering. Looking at the scatter plots from both 2 and 4 clusters you can see it splits the data almost evenly. This is telling me that each quadrant is almost just as likely as the others. The four quadrants are low age and low monthly charge, low age and high monthly charge, high age and low monthly charge, and high age and high monthly charge. The only thing I am seeing that this implies is that age and monthly charge being high or low does not increase or decrease the likelihood of a churn customer.**

**E-3**

**One limitation on my data analysis was my assumption that the data would have followed closer to my personal experience in telecommunication sales. I picked to look at these two variables because I expected three clusters to be present. I expect to see low age low charge, middle age high charge, and high age low charge. Younger people usually have lower wages so they must spend less and will jump to the cheapest option. Middle age would have more income to spend but tend to add so many add-ons that they have highest monthly charges leading them to look for a switch. Older age use telecommunications less and start to become on a fixed income so they look for the cheapest options. This makes me wonder if my experiences were just unique or if this dataset was not based off real results from a telecommunications company since it is for educational purposes.**

**E-4**

**I was hoping that the analysis would have given clusters that could have been applied to the current customer base that would have allowed to monitor potential churn customers and try to come up with a way to keep them. This can not be applied because of the results of the analysis showing us that it is almost an equal chance that a customer will become a churn customer no matter age or monthly charge. If this was in a real-world organization, I would suggest that this analysis did not give useful insight to the question posed and that we need to investigate different variables to find useful clusters.**

***Part VI: Demonstration***

**F**

Please see my Panopto video Christian LeBlanc D212 Task 1. Link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=40a8ac50-25eb-48b2-b6c6-b1590184dc50>

**G**

No web sources were used to acquire data or segments of third-party code.

**H**

Chantal D. Larose, & Daniel T. Larose. (2019). *Data Science Using Python and R*.

Wiley.

**I**

The content in this Performance Assessment is set up and presented with the highest professional standards.