­

**Data Mining II, Task I: Churn Data**

Stephen E. Porter

College of Information Technology, Western Governors University

Dr. Keiona Middleton

September 27, 2022

**Data Mining II, Task I: Churn Data**

In this paper, I will use the provided data set containing cleaned customer data from a fictional telecommunications company. I will use k-means clustering to analyze the data set to better understand the characteristics of our customers. The goal is to determine if any subgroups of our customers exist and what characteristics they share.

# Part I: Research Question

## A1. Proposal of Question

The key question I would like to answer is how many subgroups exist within our customer base and what characteristics do they share?

## A2. Defined Goal

I will prepare the data by selecting only continuous data variables, removing outliers, and splitting the data set into a training data set and testing data set. Then, I will perform a k-means analysis on the training set to split the customer data set subgroups and analyze the characteristics they share. I will use the testing data set to confirm my findings.

# Part II: Technique Justification

## B1. Explanation of Clustering Technique

There are several clustering techniques overall; however, the one chosen for this analysis is k-means. This clustering technique is simple yet handles large data sets well, and the churn data set is relatively large. (Bruce et al., 2020)

The algorithm “divides the data into *K* clusters by minimizing the sum of the squared distances of each record to the *mean* of its assigned cluster.” (Bruce et al., 2020) More specifically, centers for each cluster are randomly assigned and the distance from the center to the nearest points is determined until *K* clusters of comparable size are created. New centers are then determined based on the centroid mean of all points within the cluster. This process repeats until there is no change the elements for each cluster. (Analytics, 2017)

The ideal number of clusters, or *K* value, is unknown, but certain functions can provide a quality starting point. By running a k-means clustering on the data set with increasing values of *K* (say, from 1 to 10) and finding the sum of the *withinss* value (a measure of similarity within each group) for each *K*. When plotted, there will typically be a value of *K* where the steepness between each successive point is notably less, and that *K* value should be the best choice. This is known as the “elbow rule.” (Fonseca, 2019) Fonseca provides a custom function; however, in the *factoextra* package, the built-in *fviz\_nbclust()* function provides a comparable visualizaition. (Kassambara, 2018)

The expected outcome of the k-means clustering is to find underlying trends in customer data related to the continuous variables to determine if any sub-groups of customers exist within the data set. We would expect to have two or more clusters identified (ideally with clear boundaries) and their means will describe the group as a whole.

## B2. Summary of Technique Assumption

The main assumptions of k-means clustering are the clusters are spherical and of similar size. Because the algorithm is based on Euclidean distance, the algorithm is most successful in determining the boundaries when they are circular/spherical in shape. The assumption of similar sizes, clearly, allows the algorithm to determine the number of observations in each cluster. (Analytics, 2017)

## B3. Packages or Libraries List

The libraries included, and their justification are as follows:

|  |  |
| --- | --- |
| **Library** | **Justification** |
| tidyverse | Contains core functionalities for data analytics |
| dplyr | Contains the pipe operator %>% and other useful functions |
| ggplot2 | For graphs & visualizations |
| factoextra | Contains functions for clustering analysis |
| cluster | Contains the k-means cluster function |

# Part III: Data Preparation

## C1. Data Preprocessing

To perform k­-means clustering on this data set, I will need to ensure that there is no missing data, only continuous variables are retained, outliers are removed, and the data is normalized to reduce the effect of varying units of measurement.

## C2. Dataset Variables

The variables remaining after preparation, and their type, are as follows:

|  |  |
| --- | --- |
| **Variable** | **Type** |
| Income | Continuous |
| Outage\_sec\_perweek | Continuous |
| Tenure | Continuous |
| MonthlyCharge | Continuous |
| Bandwidth\_GB\_Year | Continuous |

## C3. Steps for Analysis

In preparing the data, first I confirmed there were no null values. I then selected only continuous variables as k-means clustering can only process continuous data. These variables were analyzed for outliers, and Outage\_sec\_perweek and Income were found to contain several. I repeatedly collected outliers and removed the corresponding rows until none remained. Because the units of measurement for the columns have vastly different ranges (Income especially) I normalized the data to reduce the impact of relative scale on the analysis. The data set is now ready for cluster analysis.

To determine the appropriate number of clusters, I used the function *fviz\_nbclust()* from the *factoextra* package using the method *wss* (within sum of squares). The visualization should have a clear bend (or elbow) which indicates the ideal number of clusters. In this situation, the graph indicates 2 clusters is ideal. (See Figure 1) (Banerji, 2022)

Chart, line chart

Description automatically generated

***Figure 1: The Elbow Rule***

Another option is to run the *fviz\_nbclust()* function using the *silhouette* method. The silhouette value is a measure of similarity of points within their own cluster compared to points from other clusters. A formula for the silhouette of a point in a cluster can be defined as

where represents the average distance between and all points from its own cluster, represents the average distance between and all other points. The maximum silhouette value is 1, and the value of *K* that has a silhouette value closest to 1 will be optimal. Figure 2 shows the silhouette values for the churn data and also implies that 2 clusters is optimal. (Banerji, 2022)

Chart, line chart

Description automatically generated

***Figure 2: Silhouette***

## C4. Cleaned Dataset

See the attached *churn\_kmeans\_train.csv* and *churn\_kmeans\_test.csv* files.

# Part IV: Analysis

## D1. Output and Intermediate Calculations

Now with the data set prepared and the *K* value determined, k-means cluster analysis can begin. Partial outputs will be provided here, but the full code output is provided in the attached PDF document. For the training data set, the centers and accuracy metric are given in Figure 3.

Text

Description automatically generated***Figure 3: Training k-means Analysis***

These values indicate the means of Income, MonthlyCharge, and Outage\_sec\_perweek are very close between the two groups, but while one group has a high Tenure and Bandwidth\_GB\_Year, the other is quite low. The ratio of *betweenss* to *totss* serves as a measure of accuracy for the model. The calculation is described in more detail in a later section. For the training data set, the value is 59.97% which, while low, is likely due to the closeness of the centers for the three variables listed at the beginning of this paragraph.

The cluster analysis is then visualized using the *fviz\_cluster()* function from the *factoextra* package with the primary principal component along the horizonal axis and the secondary along the vertical. The output shows two distinct clusters with possibly some overlap in the middle. (See Figure 4)

Chart, scatter chart

Description automatically generated

***Figure 4: Training k-means Visualization***

The testing set will now be analyzed and compared to the training set. Two clusters was confirmed to be the optimal value using both the Elbow Rule and the Silhouette values.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

***Figure 5: Elbow & Silhouette for Testing Data***

The full output of the k-means clustering can be found in the attached PDF while the centers and accuracy measurements are given here.

Text

Description automatically generated

***Figure 6: Testing k-means Analysis***

The means of each variable closely resemble those of the training data set as well as the accuracy metric. The k-means algorithm arbitrarily assigns the cluster value so the fact that they are reversed here is of no consequence.

Chart

Description automatically generated

***Figure 7: Testing k-means Visualization***

## D2. Code Execution

Please see the attached *d212\_task1\_revision1* file (provided in R, PDF, and TXT formats).

# Part V: Data Summary and Implications

## E1. Accuracy of Clustering Technique

The ratio of *between\_ss* to *total\_ss* serves as a measure of accuracy for the analysis. The “ss” in each stands for “sum of squares” therefore *tot\_ss* is the total sum of squares for the data set while *between\_ss* is the difference between the total sum of squares and the total (sum of) *within\_ss* for all clusters. (Kassambara, 2018)

The accuracy metric for both models was approximately 59-60%. The lower value is likely due to the closeness of the means for Outage\_sec\_perweek, MonthlyCharge, and Income for both clusters meaning the algorithm had difficulty determining clear boundaries between the groups. The accuracy of the model increases to nearly 92% when restricted to only Tenure and Bandwidth\_GB\_Year. This output can be seen in its entirety in the attached pdf, but the centers and accuracy are given below, along with the cluster plots.

Text

Description automatically generated

***Figure 8: Final Cluster Compairson***

Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated

***Figure 9: Final Cluster Plots for Training (left) and Testing (right)***

## E2. Results and Implications

The analysis has determined that there are two sub-groups of customers based on the continuous variables in the data set – one group that has high tenure and high bandwidth usage while the other has low tenure and low bandwidth usage. The other continuous variables are nearly identical in both groups and, while they provide additional context, they do not help distinguish between the clusters.

## E3. Limitation

While k-means clustering is a useful tool for its simplicity and general applicability, it does have limitations. Arguably the most notable is that the algorithm cannot identify the optimal k value on its own. There are packages and functions that can aid the analyst in identifying a k value, but it is ultimately up to them to choose. This limitation can be mitigated by running the cluster analysis for several k values to determine which is best for the given data set and goal. (Kassambara, 2018)

## E4. Course of Action

The analysis has identified two sub-groups of customers: those with high tenure and high bandwidth usage and those with low tenure and low bandwidth usage. While not directly related to customer satisfaction, a customer with high tenure has decided to remain with the company for an extended period. As a result, they appear, in general, to use more bandwidth. This likely implies that they utilize our streaming service(s) and/or perform other activities that would require a higher bandwidth usage (like online gaming/streaming, working from home, or running a business out of their home). From the other perspective, customers who use more bandwidth tend to stay with our services longer – so encouraging customers to use our streaming services or upgrade internet (for those who need it) may improve customer satisfaction/retention.

We could work with the sales and marketing teams to offer free trials of our streaming services with upgraded internet to incentivize customers to use more of our services and, in turn, may stay with us longer.

# Part VI: Demonstration

## F. Panapto Recording

Please see attached video link.

## G. Sources of Third-Party Code

No third party code was used in the execution of this script.

## H. Sources

**References**

Analytics, P. (2017, August 7). Exploring assumptions of k-means clustering using r: R-bloggers. R. Retrieved September 9, 2022, from https://www.r-bloggers.com/2017/08/exploring-assumptions-of-k-means-clustering-using-r/

Banerji, A. (2022, July 19). *K-mean: K means clustering: Methods to find the best value of k*. Analytics Vidhya. Retrieved September 28, 2022, from https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/#:~:text=The%20silhouette%20coefficient%20is%20a,for%20each%20value%20of%20K.

Boehmke, B. (2017, March 3). *K-means cluster analysis*. K-means Cluster Analysis · UC Business Analytics R Programming Guide. Retrieved September 28, 2022, from https://uc-r.github.io/kmeans\_clustering

Bruce, P., Bruce, A. G., & Gedeck, P. (2020). Practical statistics for data scientists: 50+ essential concepts using r and python. O'Reilly.

Kassambara, A. (2018, October 21). *K-means clustering in R: Algorithm and practical examples*. Datanovia. Retrieved September 24, 2022, from https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/

Jiang, J. (2022, May 26). *K-means clustering in R*. Medium. Retrieved September 24, 2022, from https://towardsdatascience.com/k-means-clustering-in-r-feb4a4740aa