­

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

Stephen E. Porter

College of Information Technology, Western Governors University

Dr. Keiona Middleton

September 8, 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 what characteristics (if any) might indicate if a customer is at-risk of ending their service with our company.

# Part I: Research Question

## A1. Proposal of Question

The key question I would like to answer is can we determine which customers are at risk of churning? Meaning, they share characteristics (are clustered) with customers with a churn value of “Yes” but themselves have a churn value of “No.” These are the customers we need to attempt to keep.

## A2. Defined Goal

I will perform a k-means analysis to split the customer data set into two groups. I will then view the distribution of Cluster and Churn as a bar chart to determine the groups of users. There will be essentially four sub-groups:

* Churn: Yes & Cluster: A – Those who have churned and would expect to churn
* Churn: Yes & Cluster: B – Those who have churned but would *not* expect to churn
* Churn: No & Cluster: A – Those who have *not* churned but would expect to churn
* Churn: No & Cluster: B – Those who have *not*churned but would *not* expect to churn

The customers in the third group are of greatest concern based on this analysis.

# 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 20) 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)

## 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, non-numeric variables are removed, and the data is normalized to reduce the effect of relative outliers based on the varying units of measurement.

## C2. Dataset Variables

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

|  |  |
| --- | --- |
| **Variable** | **Type** |
| Children | Categorical |
| Age | Categorical |
| Income | Continuous |
| Outage\_sec\_perweek | Continuous |
| Email | Categorical |
| Contacts | Categorical |
| Yearly\_equip\_failure | Categorical |
| Tenure | Continuous |
| MonthlyCharge | Continuous |
| Bandwidth\_GB\_Year | Continuous |
| Response | Categorical |
| Fix | Categorical |
| Replacement | Categorical |
| Reliability | Categorical |
| Options | Categorical |
| Respectful | Categorical |
| Courteous | Categorical |
| Listening | Categorical |

## C3. Steps for Analysis

In preparing the data, first I checked for null values (and found none). I then replaced the ambiguously named columns of “Item1” through “Item8” to be more informative. Once the basic preparation was completed, I dropped all non-numeric columns as k-means clustering can only process a data frame of numeric values, and I dropped several columns I determined would not be useful in the analysis. Because the units of measurement for the columns led to vastly different ranges (Income, Number of Children, Age, Tenure, etc.) I normalized the data to reduce the impact of relative outliers 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. The result indicated that two clusters is ideal which makes logical sense for the question I am trying to answer – what do customers who *have* churned look like so that I can better determine who is at risk of churning but has not? (See Figure 1)

Chart, line chart

Description automatically generated

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

The image also seems to imply that four clusters may provide a valuable analysis (which also makes sense based on the sub-groups identified in Section A2). To be thorough, I will perform a k-means analysis using *K = 2* and *K = 4*.

See Section D1 for the code output and intermediate calculations for the analysis.

## C4. Cleaned Dataset

See the attached *churn\_clean.csv* file.

# Part IV: Analysis

## D1. Output and Intermediate Calculations

*Two-Cluster Analysis:*

Now with the data set prepared and the *K* value determined, k-means cluster analysis can begin. The output for two clusters is shown below (see Figure 2).

Text

Description automatically generated

***Figure 2: Two-Cluster Analysis***

The output of the means for each cluster and the corresponding variable (before normalizing) can be seen using the aggregate function below (Figure 2).A screenshot of a computer

Description automatically generated with medium confidence

***Figure 3: Two-Cluster Mean Analysis***

The cluster analysis is then visualized using the *fviz\_cluster()* function from the *factoextra* package. The output shows two distinct clusters with some overlap in the middle. (See Figure 4)

A picture containing chart

Description automatically generated

***Figure 4: Two-Cluster Visualization***

The cluster value was added back to the original data frame so Churn value colored by cluster could be created as a bar chart. This will be useful to identify users who are similar to those who have churned (the same cluster) but have not yet churned (Churn = No). (See Figure 5)

Chart, bar chart

Description automatically generated

***Figure 5: Churn Bar Chart Colored by Cluster (Two)***

*Four-Cluster Analysis:*

Because there appears to be another “elbow” at *K=4*, I also performed a cluster analysis for four clusters. I followed the same process as for two clusters. The output for four clusters can be seen below (Figure 6).

Text

Description automatically generated

***Figure 5: Four-Cluster Analysis***

The output of the means for each cluster and the corresponding variable (before normalizing) can be seen using the aggregate function below (Figure 2

Graphical user interface, text

Description automatically generated

***Figure 6: Four-Cluster Mean Analysis***

The cluster analysis is then visualized using the *fviz\_cluster()* function from the *factoextra* package. The output shows four clusters, but pairs of two that are overlapped (See Figure 7).

Map

Description automatically generated

***Figure 7: Four-Cluster Visualization***

The cluster value was added back to the original data frame so Churn value colored by cluster could be created as a bar chart. Clusters 2 and 4 comprise a greater percentage of the values of Churn = Yes while Clusters 1 and 3 do the same for Churn = No. (See Figure 8)

Chart, bar chart

Description automatically generated

***Figure 8: Churn Bar Chart Colored by Cluster (Four)***

## D2. Code Execution

Please see the attached *d212\_task1.R* file.

# 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) For the Two-Cluster analysis, this ratio is 30.8% and for the Four-Cluster analysis, it is 39.9%. This level of accuracy is notably low; however, it could be improved by reducing the number of variables. I purposely kept as many as what seemed reasonable to provide the widest view of our customers. Understandably, the more features used in the analysis, the more challenging it is to fit them neatly into groups. By reducing the number of variables, I would expect a greater level of accuracy in the clustering.

## E2. Results and Implications

The two-cluster analysis, while lower in its accuracy metric, still provides useful context for the question posed in section A1 – are there customers at risk of churning? By reviewing the corresponding bar chart (Figure 5), there clearly appears to be an association between Cluster value and Churn value. Cluster 2 comprises nearly all of the “Yes” bar which then allows us to identify those customers from the “No” group who are also in Cluster 2. These users I would deem are at risk of churning because they share more features of users who have already churned (as determined by the cluster analysis).

The four-cluster analysis showed significant overlap between pairs of clusters (see Figure 7), and similar conclusions can be drawn from it. Clusters 2 and 4 appear in a larger percentage of the total bar for “Yes” while Clusters 1 and 3 do the same for “No” (see Figure 8). Therefore, our attention should be on users with a “No” for churn but are in Cluster 2 or 4.

## 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 a group of users who appear to be at risk of churning; however, it has been noted that the accuracy metric is relatively low. If accuracy is a concern and the level of effort involved is worth revisiting the model, the first step would be to reduce dimensionality, keeping only variables deemed most important and repeat the cluster analysis to achieve a higher accuracy metric. At which point, the next step would be the same: contact the customers who have been identified as at risk of churning and gauge their level of satisfaction with their service. In preparation for this step, further review of each customer may be beneficial – do they have services they are not using? Do they have multiple services we could instead offer to them in a “bundle” at a discount? Have they experienced a higher-than-expected number of issues with their service? These types of questions are best answered by the sales/marketing team(s) to determine what is reasonable to offer.

# 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/

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