­

**Performance Assessment for D212: Data Mining II  
Task 1: Clustering**

Christopher Fischer - 011933891

College of Information Technology, Western Governors University

September 25, 2024

Performance Assessment for D209: Data Mining I – Task 1

This document contains the tasks and outputs required for the “OFM4 TASK 1: CLUSTERING TECHNIQUES” assessment. All work is original to the author unless otherwise indicated by a citation.

# A - Research Question and Goal

For years, BigTel has treated all it’s customers (and potential customers) the same. However, best practices demand that BigTel understand that different customer segments have different needs, motivations, and desires. To this end, management has commissioned this study to ask: can tenure and annual bandwidth consumption be used to cluster BigTel’s customers?

If successful, this analysis will form a basis for modernizing BigTel’s marketing and services to appeal to discrete subsets of the population. Improving marketing to attract different groups of customers and tailoring service offerings to what different groups desire will all contribute to increased satisfaction and, therefore, lower churn.

# B – Method Justification

This study will discover patterns among BigTel’s customer data, of which none are currently known. This is the kind of problem solved by unsupervised learning, a form of exploratory data analysis where the response variable, or label, is not known. In particular, we will use a clustering algorithm to group similar customers together. While clustering is not itself a means of prediction, a successful clustering activity may help jumpstart various forms of supervised learning by providing a set of labeled data from which to begin building a prediction model. (Bruce, Bruce, & Gedeck, 2020)

Of the many clustering algorithms available to us, this study will use the *k-means* method. This is a simple, distance-based clustering algorithm. The algorithm starts by randomly assigning all records to *k* number of clusters, where the analyst selects *k*. Then, recalculate the cluster centers and repeat the process until no further movement of the cluster centers is detected. At completion, all records will be labeled with one of the *k* clusters. (Larose, 2015)

One of the key assumptions for *k-means* is that all variables studied must have the same scale and variance. (Urbonas, Lacroix, & Nazary, n.d.) Consider a study on two variables: number of children and monthly income. The scale of likely variables for the number of children might be 0-10, while the monthly income scale might be $500-$10000. As the *k-means* algorithm measures Euclidean distance between values, even a minor difference between two observations of monthly income (e.g., $500 and $750 would utterly conceal a comparably large difference in the number of children (e.g., two children and nine children).

The following table lists the Python libraries which will support this study.

|  |  |
| --- | --- |
| Module | Purpose |
| Pandas | Provides DataFrame used for data management |
| Missingno | Displays a graphic of missing data used in the cleaning process |
| Matplotlib.pyplot | Tools for visualizing data and results |
| Seaborn | Additional visualization capabilities |
| Sklearn.preprocessing | Implements scaling |
| Sklearn.cluster | Provides the *k-means* clustering algorithm |
| Sklearn.metrics | Provides several metrics and tools for evaluating model performance |

Table - Python Libraries to be Used

# C – Data Preparation

As mentioned above, *k-means* requires all variables to have the same scale and variance. Therefore, all study variables will be scaled with the RobustScaler before fitting the model.

This table details the variables included in our clustering study.

| Variable Name | Data Type |
| --- | --- |
| Tenure | Numeric/continuous |
| Bandwidth\_GB\_Year | Numeric/continuous |

Table - Study Variables

These sections detail the data preparation steps performed for the study.

A screenshot of a computer program

Description automatically generated

No duplicate observations were noted.

A screenshot of a computer

Description automatically generated

No variables exhibit missing data.

A black screen with colorful lines

Description automatically generated

No outliers were detected.

A screenshot of a computer

Description automatically generated

Substantial deviation in scale and variance are noted.

A screenshot of a computer program

Description automatically generated

Data set scaled with RobustScaler.

The prepared data set will be uploaded with this report. See file *churn\_prepared.csv*.

# D – Analysis

As the *k-means* clustering algorithm requires the analyst to specify the desired number of clusters, we employed the elbow method to identify the optimal number of clusters for our data set. With this method, a KMeans model is fitted iteratively with a different n\_clusters value. Then, the inertia values for each model are plotted. Lastly, the analyst examines the plot, looking for the point at which additional clusters yield only minor improvements in inertia. (Bobbitt, 2023)

A screen shot of a computer screen

Description automatically generated

Examination of the elbow plot suggests that two is the optimal number of clusters for our data set.

A computer screen shot of a program

Description automatically generated

Reviewing the standard performance measures for a clustering model indicates that our resulting clusters are tightly grouped and adequately separated.

# E – Data Summary and Implications

Additional visualizations help assess the quality and usefulness of the clusters resulting from this study.

A screen shot of a computer program

Description automatically generated

A graph of a cluster of dots

Description automatically generated with medium confidence

This graph plots our scaled study variables along with the centroids of the resulting clusters. The observations are tightly grouped around the centroids.

A graph of a bar chart

Description automatically generated with medium confidence

This graph plots the same data only with the original, unscaled values. This view will help business leaders grasp the implications of the data more easily.

A black background with white text

Description automatically generated

A diagram of a cluster

Description automatically generated A diagram of a cluster

Description automatically generated

The boxplots of the study variables grouped by cluster also demonstrate a good separation of values. For *Tenure*, cluster 0 has a mean of about 9, while cluster 1 has a mean of just over 60. For *Bandwidth\_GB\_Year*, cluster 0 has a mean of about 1200, while cluster 1 has a mean of close to 6000.

These results are encouraging. They demonstrate that service usage patterns for our customers are not homogenous but may vary widely among different groups of customers. Specifically, these results show that our customers tend to use more data bandwidth each year as their length of tenure increases.

One limitation of this study is that only two continuous variables were included in the analysis. While these two provided one valuable insight, adding more variables to the clustering model could reveal patterns and insights that are more difficult to recognize without clustering.

To this end, the team recommends that management conduct further clustering studies with additional variables. Beyond this, the marketing team should consider conducting customer surveys to understand why customers use more data as they stay with BigTel longer.

# F – Recorded Code Review

A recording of the code review presentation was uploaded with this submission. For quick reference, that video may be found here: [Panopto Recording](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6c45536d-8d75-405c-9eaa-b1f800f57797)

# G – Third-Party Code References

Amelia, A. (2018, 09 27). *K-Means Clustering: From A to Z*. Retrieved 09 24, 2024, from Towards Data Science: https://towardsdatascience.com/k-means-clustering-from-a-to-z-f6242a314e9a

Kamara, K. (n.d.). *Analyze and Interpret K-means*. Retrieved from Panopto: https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3fe13831-fe4b-4c6b-a3eb-b0ee018754bc

Nalcin, S. (2022, 10 11). *StandardScaler vs. MinMaxScaler vs. RobustScaler: Which one to use for your next ML project?* Retrieved 09 24, 2024, from Medium: https://medium.com/@onersarpnalcin/standardscaler-vs-minmaxscaler-vs-robustscaler-which-one-to-use-for-your-next-ml-project-ae5b44f571b9

Savare, K. E. (2024, 03 21). *Clustering Metrics in Machine Learning*. Retrieved 09 23, 2024, from Geeks for Geeks: https://www.geeksforgeeks.org/clustering-metrics/

# H – Referenced Works

Bobbitt, Z. (2023, 01 03). *How to Use the Elbow Method in Python to Find Optimal Clusters*. Retrieved 09 24, 2024, from Statology: https://www.statology.org/elbow-method-in-python/

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data scientists : 50+ essential concepts using r and python.* O'Reilly Media, Incorporated. Retrieved 09 25, 20241, from https://ebookcentral.proquest.com/lib/westerngovernors-ebooks/detail.action?docID=6173908

Larose, D. T. (2015). *Data Mining and Predictive Analytics.* John Wiley & Sons. Retrieved 09 25, 2024, from https://ebookcentral.proquest.com/lib/westerngovernors-ebooks/detail.action?docID=7104155

Urbonas, K., Lacroix, H., & Nazary, M. (n.d.). *Data Preprocessing for Clustering*. Retrieved 09 25, 2024, from DataCamp: https://campus.datacamp.com/courses/customer-segmentation-in-python/data-preprocessing-for-clustering?ex=1