

|  |
| --- |
| PERFORMANCE ASSESSMENT OF  DATA MINING II  TASK1  D212  BY KOFFI M. GANU |
| MSDA MAY 2022 AT WGU  INSTRUCTOR: KESSELLY KAMARA  Phone  Email |

**INTRODUCTION**

People often say in business: “It takes a month to find a customer but a second to lose one.” The primary goal of any company is to maintain its customers as long as possible. It is with this in mind that, especially with the advent of the use of data, these companies hire data specialists (Data analysts, Data scientists) for strategies and recommendations for the improvement of their company’s services. In the rest of our project, we will use a K-means clustering method to identify the number of distinct segments of customers and the characteristics of each one.

**PART I/**

A1-

How many distinct customer groups can be identified, and what are the characteristics of each in terms of tenure, monthly payment, income, and other relevant attributes?

A2-

Our goal is to group similar data points together into clusters while keeping the clusters as distinct from each other as possible using the k-means method.

**PART II/**

B1

K-means clustering is a popular unsupervised machine learning algorithm to partition data into distinct groups or clusters based on their similarities. First, the k-means analysis starts by choosing the number of clusters, denoted as k. Indeed, choosing the appropriate value of k is critical and can significantly impact the quality and interpretability of the result. But the Elbow method will help us for this purpose. The Elbow method involves plotting the within-cluster sum of squares against different values of k, and the optimal value is often identified as the elbow point. After choosing k, each data point in the dataset is assigned to the nearest centroid. The distance metric commonly used is the Euclidian distance. This assignment creates k clusters, where each data point belongs to the cluster associated with the nearest centroids. After the initial assignment, the centroids of the clusters are recalculated based on the data point within each cluster. The new centroid is computed as the means of all data points in the cluster along each dimension. This ensures that the centroids represent the center of the data points in each cluster. We have this until the centroids no longer significantly change between iterations or a maximum number of iterations is reached. Furthermore, the cluster assignments and centroids gradually converge to a stable state. At that point, the algorithm found a data partitioning that minimizes the sum of squared distances between data points and their respective centroids.

B2-

One assumption of the K-means algorithm is that the clusters in the cluster in the dataset are spherical and have roughly equal sizes. This assumption implies that their centroids can represent the cluster, the mean values of the data points within each cluster.

B3-

For analyzing the data set provided using the K-means algorithm method in Python, we can use the following pack

* Pandas: Pandas provide a function to create a DataFrame by reading data from various files type. In addition, pandas help to manipulate and preprocess the data set provided by handling the missing value and converting categorical variables to numerical form.
* Numpy: This library can perform various mathematical operations for implementing the K-Means algorithm, such as calculating distances between instances.
* Matplotlib: This is a data visualization library that can be used to visualize the data and the analysis results.
* Sklearn.cluster: This is a tool for data analysis and modeling. It includes a KMeans that can be used to build a K-Means model for the provided data set.
* Sklearn.preprocessing: From this library, we will import StandardScaler, which will help us to scale the data using z-score.
* Scipy.spatial.distance: used to compute distances.
* Yellowbrick.cluster : used to visualize the elbow curve and the silhouette scores.

PART III/

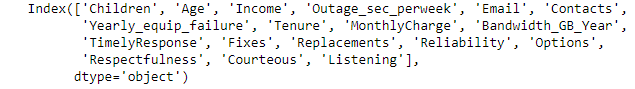
C1

One data preprocessing goal of K-means clustering is variable scaling. Indeed, K-means clustering is sensitive to the scales of variables in the dataset. If the variables have different scales, it can lead to biased results as K-means rely on calculating distances between data points. So, by performing variable scaling, the variables are brought to the same scale, and the influence of one variable is minimized.

C2

As k-means relies on calculating distances between data points, the primary type of variable used by k-means is continuous numerical data.

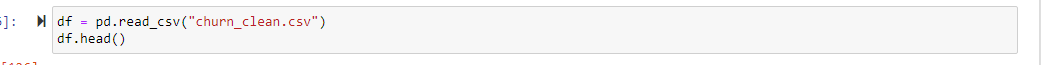
List of the continuous variable:



C3

The following are the steps to prepare the data set provided for analysis.

* Import the necessary libraries: The necessary libraries for our analysis are Pandas, Numpy, Matplotlib, Sklearn.cluster, Yellowbrick.cluste, Scipy.spatial.distance, Sklearn.preprocessing.
* Load the data set: We load the data set into our working directory using the read\_csv( ) fuction.



* Rename columns of a survey to easily recognizable features.

Text

Description automatically generated

* Get information on the data set, structure, and data types.



* Drop irrelevant columns: We can drop some columns we have seen unnecessary for our analysis.



* Check and remove outliers that are several standard deviations above the mean.

Graphical user interface, text

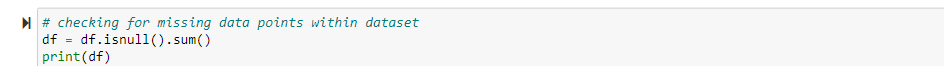
Description automatically generated

Graphical user interface, text, application, email

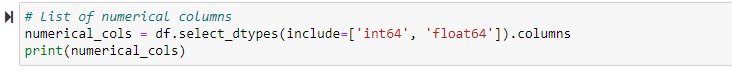
Description automatically generated



* Check for records with missing data and impute missing data with meaningful measures of central tendency.



* Select appropriate variables suitable for k-means analysis.



C4

A copy of the prepared data set:

We will extract the prepared data set with this python code.



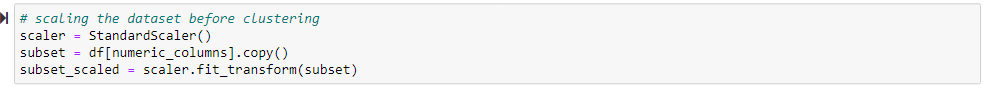
The data set will be attached and named “clean\_D212\_task1”.

PART IV/

D1

The K-means algorithm is a popular unsupervised machine-learning technique for clustering data points into groups. It aims to partition a provided dataset into k clusters, where each data point belongs to the cluster with the nearest centroid. First, we will choose k using the elbow method. The elbow method involves plotting the within-cluster sum of squares against different values of k, and the optimal value is often identified as the elbow point. Also, for each data point, calculate the distance to each centroid and assign the point to the cluster with the nearest centroid. Furthermore, recalculate the centroids of each cluster based on the data points assigned to it until convergence, which occurs when the centroids no longer change significantly. The final centroids represent the clusters, and each data point is assigned to a specific cluster.

Before performing the K-means analysis, we scaled the dataset. Here is the result.





A screenshot of a computer

Description automatically generated

Here is the elbow method code and the result.

A screenshot of a computer

Description automatically generated

A graph with a red line

Description automatically generated

D2

Here we will provide the code used to perform the K-means analysis.

A screenshot of a computer program

Description automatically generated

PART V/

E1

The k-means cluster model doesn’t have a straightforward accuracy measure like in supervised learning; we evaluate its effectiveness using various metrics and methods that assess the compactness of clusters and the separation between them. For our case, we have chosen the elbow method to evaluate our model. According to the elbow method, we plotted the Within-Cluster Sum of Square (WCSS) for different values of k and looked for the elbow point on the plot. The elbow point is where the rate of decrease in WCSS starts to slow down. Adding more clusters doesn’t significantly improve the model’s fit.

E2

After analysis, we have four groups of customers.

* Group 0: Customers with a high email sent and high outage second per week.
* Group 1: Customers with high monthly charges and low tenure.
* Group 2: Customers with high income, high tenure, and their customer experience is excellent.
* Group 3: Customer with high contact, high yearly equipment failure, and high number of customers in the dataset

E3

One limitation of the K-means algorithm is its sensitivity to the initial placement of cluster centroids. The algorithm starts by randomly initializing the centroids of the cluster, and the final clustering outcome can vary based on the initial positions. Due to this sensitivity, K-means may converge to different solutions if run multiple times with different initialization. This can lead to unstable results because K-means may converge to a local minimum rather than the global minimum of the clustering objective function.

E4

After analysis of our model created by the K-means methodology, it can be concluded that customers with a high risk of churning have the same profile, such as short-term contracts and fair customer support. Therefore, we recommend that telecommunication companies sign at least a one-year contract with their new customers and provide good customer support to keep their customers for as long as.

**PART VI/ DEMONSTRATION.**

F- PANOPTO VIDEO

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3e2dfa23-56da-4e74-9bad-b06500e257ec

G-

Data camp course

D212 Predictive modeling WGU textbook

D212 Predictive modeling WGU course webinar

H-

No in-text references were used.

I-

We used professional communication in the presentation of the submission.