D212 Kmean Project

Jacob Polomsky

July 27, 2024

A1. Are there any hidden groups in our churn data that we might not see using Kmean Clustering?

A2. Our goal is to use Kmean clustering to find the Inertia of the data and the elbow method to find the ideal number of clusters for this dataset.

B1. The cluster method works by marking a number of clusters on a graph. Once the mark has been put on, it assigns the nearest points to that cluster, and then you move the mark to the center point of all the points. After that, some of the points might change and get assigned to new clusters, and you keep moving the mark to the center point until no points move and the mark doesn’t move anymore. Kmeans allows you to group points together by looking at the distance on points from another point or center, allowing you to find hidden patterns you might not have seen before and assigning each data point into that given group.

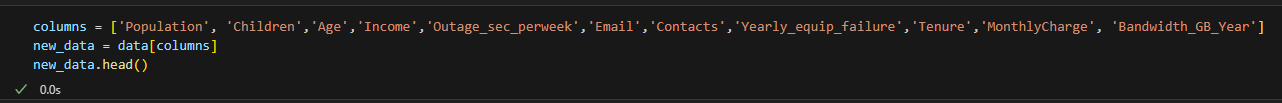
B2. One assumption of the Kmean model is that the clusters are spherical in shape, saying that the variances should be spread out in all directions equally

B3. The packages I used were pandas to import and manipulate the data, sklearn standscaler to scale the data to its standard deviation so that one column doesn’t have a bigger effect than another and sklearn Kmeans to run the model, matplotlib to visualize the Inertia to see the ideal number of clusters.

C1. One goal of the data cleaning process is to make the data usable for the model. First, we check for duplicates, null values, and outliers. Then, we limit the columns used to be only continuous variables, and then we scale these scales.

C2. All of the variables used in this dataset can be seen as continuous. The variables that are continuous are Age, Income, Outages\_sec\_perweek, tenure, monthly\_charge, and bandwidth\_gb\_year. Age can be discrete or continuous, and in this dataset, it is more of a discrete variable than a continuous one.

C3. The first was to limit the columns to the columns we are going to use in our dataset



next step was to look for null values and which there were not any

A screenshot of a computer

Description automatically generated

Next was to check for duplicates which there were not any

A screen shot of a computer

Description automatically generated

Then to look for any outliers

A screen shot of a computer

Description automatically generated

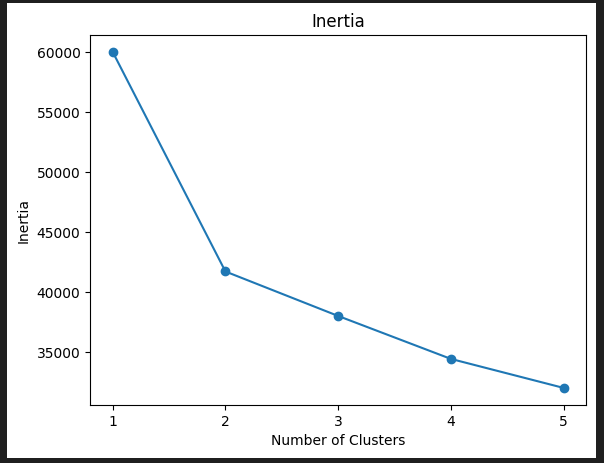
Finally it was to scale the data and

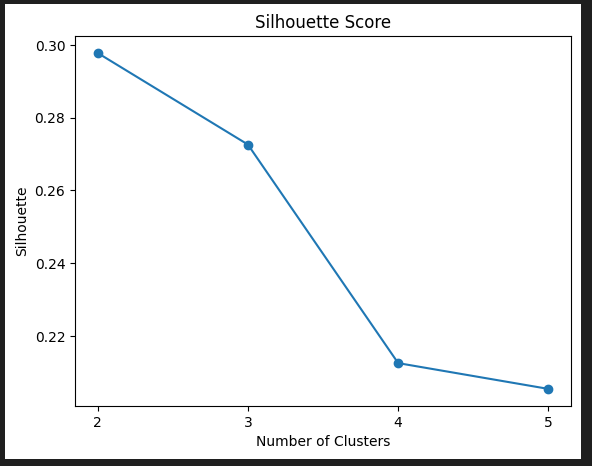
A screen shot of a computer

Description automatically generated

C4. scaled\_data.csv

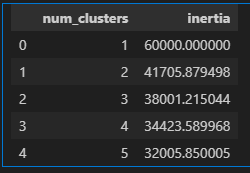
D1. I found that the optimal number of clusters for the dataset is 2. The Methods I used to determine these values were plotting the Inertia of the model by the number of clusters and by looking at the silhouette score. Plotting the Inertia, you can use the elbow method to determine the optimal number of clusters, which you can see with the visualization is 2. With the Silhouette Score, we are looking to see what the highest silhouette score is, and with the visualization, we can see it is 2



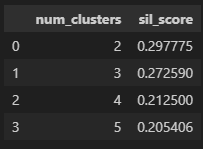


D2. Kmean\_code.ipbny

E1. I used Inertia to find the quality of the cluster, which measures how far each data point is from the center point when you add them all together. The picture below shows that there is a big decrease from 1 to 2 clusters, about a 20000 inertia drop, while all of the points after are close to a 4000 inertia drop or less after 2 clusters.



With the Silhouette scores, see how close the point is from the center, ranging from 1 to -1, where 1 is right on top of the center and -1 is on top of another center. So, a higher silhouette score means closer clustering or better clustering. See that a cluster of 2 has the highest silhouette score of 0.3.



E2. The result is that there should be ideally two groups when we cluster the data that align with some of the columns we have in our dataset that are yes-no questions, but I think a column that stands out with our dataset that might explain the two groups is whether a customer churns or not.

E3. One limitation is that the points picked are based on their distance from each other, so it would not be a good model for predicting which group belongs to which group with good accuracy, only that there are two groups.

E4. The organization can look more into possible groups for this cluster. Still, I would recommend they use this analysis to look more into customer churn, as I suspect that for two groups. Do more research on what causes customer churn.