Taylor Cina

D212 – Data Mining II

***Part I: Research Question***

*A.  Describe the purpose of your data mining report by doing the following:*

*1.  Propose****one****question relevant to a real-world organizational situation that you will answer by using PCA.*

How many principal components can be extracted from the customer dataset?

*2.  Define****one****goal of the data analysis. Ensure your goal is reasonable within the scope of the selected scenario and is represented in the available data.*

The goal of this analysis is to use PCA to determine how many new features can be extracted from the customer dataset. The components should provide a similar amount of information to the original variables with a smaller set of new variables.

***Part II: Method Justification***

*B.  Explain the reasons for using PCA by doing the following:*

*1.  Explain how PCA analyzes the selected data set. Include expected outcomes.*

PCA analyzes the dataset by using variables with high variance to extract new features, or principal components. The new features reduce dimensionality and make the dataset easier to analyze but are less directly intuitive than the original variables. The result of this method is a smaller set of principal components that can be used to analyze the dataset in place of the original variables that loses the least amount of information.

*2.  Summarize****one****assumption of PCA.*

PCA assumes that the relationship between variables is linear as it is based on variance. Nonlinear relationships between variables might not be captured well by PCA.

***Part III: Data Preparation***

*C.  Perform data preparation for the chosen data set by doing the following:*

*1.  Identify the continuous data set variables that you will need to answer the PCA question proposed in part A1.*

The dataset mostly consists of categorical variables, with only a few relevant continuous variables. The list is as follows:

-Income

-Outage\_sec\_perweek

-Tenure

-MonthlyCharge

-Bandwidth\_GB\_Year

Lat and Lng will not be considered for PCA.

*2.  Standardize the continuous data set variables identified in part C1. Include a copy of the cleaned data set.*

Scaled dataset is attached as the file “TCina D212 T2 Scaled.csv”

***Part IV: Analysis***

*D.  Perform PCA by doing the following:*

*1.  Determine the matrix of all the principal components.*

The loading matrix is pictured below:

A black background with white numbers

Description automatically generated

*2.  Identify the total number of principal components, using the elbow rule or the Kaiser criterion. Include a screenshot of the scree plot.*

The dataset has 2 principal components by the kaiser rule. The scree plot is pictured below:

A graph with a line and a red line

Description automatically generated

*3.  Identify the variance of each of the principal components identified in part D2.*

The first component has a variance of 1.993 and explains 39.8% of the variance of the set of variables.

The second component has a variance of 1.024 and explains 20.5% of the variance of the set of variables.

*4.  Identify the total variance captured by the principal components identified in part D2.*

The two components together explain 60.3% of the total variance.

*5.  Summarize the results of your data analysis.*

The analysis found that there are 5 principal components. The first two principal components explain 60.3% of the variance of the original 5 continuous numerical variables analyzed. The last three principal components fall below the 1.00 eigenvalue line and are excluded, with the fifth component being particularly useless. Using the kaiser rule, the first two components would be considered useful. However, given the scree plot, the first three or four components could be included and still succeed in dimensionality reduction given the extremely low score of the fifth component.

***Part V: Attachments***

*E.  Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.*

No web sources used.

*F.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.*

No third party sources used.