Jordan Tompkins

wgu d212  DATA MINING II TASK 2

D212, Data Mining II Task 2

A1, Proposal of Question: The question for this analysis is are we able to determine an appropriate number of principal components of the dataset using Principal Component Analysis (PCA) for dimensionality reduction?

A2, Defined goal: The goal of this analysis is to perform dimensionality reduction on the data set and be able to appropriately reduce the total components used for analysis of the data, while maintaining enough information that it is still useful for the company.

B1, Explanation of PCA: Principal Component Analysis is very helpful when it comes to the dimensionality reduction of a large dataset. It is basically the transformation of the large set of variables into a smaller one, while maintaining most of the information from the larger set (“*A step-by-step explanation of principal component analysis, nd*). Using PCA, we can essentially reduce the dataset to a smaller set without losing a large amount of information. There are a few steps that must be taken to perform principal component analysis. The first step is the standardization of the data. You want to ensure that each variable in the data contributes equally to the analysis, and in order to do that, you must first standardize the data set. After standardizing the data, you want to see if there are any relationships between variables by looking at a covariance matrix. This will allow you to see correlations between different and see if you might be able to remove any redundant information. After getting the covariance matrix, the eigenvalues and eigenvectors will need to be identified. This will help with determining the principal components of the data. After the eigenvectors and eigenvalues are created, a feature vector must be determined. The feature vector is another matrix, this time with all the principal components that you wish to keep after determining if you want to keep all or disregard ones based on a certain criteria. Finally, you will recast all the data along the principal component axes, and reorient the data from the original axes to the ones that are represented by the PC. This was just a short summary of what happens in Principal Component Analysis. Zakaria Jaadi, in his article titled “A Step-by-Step Explanation of Principal Component Analysis (PCA)” explains these steps much more in-depth as to how to perform PCA. We expect the outcome of this analysis to be a set of principal components that allows for the reduction of dimensionality of the dataset to make analysis of the larger data set a lot easier.

B2, PCA Assumption: There are many assumptions of PCA. One of those assumptions is: “PCA assumes a linear relationship between features” (*A Guide to Principal Component Analysis (PCA) for Machine Learning*. (n.d.).). Using PCA, you want to use variables that have linear relationships. It doesn’t work well with non-linear relationships. If non-linear features or relationships are used, you can always change it to linear using methods such as log transforms.

C1, Continuous Dataset Variables: In order to perform the PCA, we needed to look at specifically continuous variables within the dataset. The numeric variables such as zip code, latitude and longitude were excluded from the analysis, and other identifiable information such as customer ID. These were simply deemed unnecessary in performing the PCA. This left the variables of ‘Children’, ‘Age’, ‘Income’, ‘Tenure’, ‘MonthlyCharge’, ‘Bandwidth\_GB\_Year’, ‘Outage\_sec\_perweek’, ‘Email’, ‘Contacts’, and ‘Yearly\_equip\_failure’. These were the final continuous variables that were used to perform the PCA for dimensionality reduction.

C2, Standardization of dataset variables: See csv file uploaded separately for a copy of the standardized dataset.

D1, Principal Components: After the data was standardized, it was ready for use in the principal component analysis. See the loading matrix below for the matrix of all the principal components identified in part C1:

A table with numbers and letters

Description automatically generated

D2, Identification of Total Number of Components: There was initially 10 principal components that were determined by looking at the continuous variables in the dataset. However, the whole purpose of PCA for dimensionality reduction is to essentially reduce the amount of components used. One of the best ways to do that is to utilize the Kaiser Rule, which states “… the number of factors to retain should correspond to the number of eigenvalues greater than one…” (Braeken, and van Assen, 2015). This rule helps determine the most meaningful PCs to keep, so we aren’t keeping them all. To help with this, the eigenvalues were determined, and graphed as shown below. As stated, any PC with an eigenvalue greater than 1 should be kept, which leads to the first 5 PCs being kept, when accounting for rounding the values to 2 decimal places. The visualization below shows a graph of the eigenvalues with a line at y equal to 1. It is noted that the graph drops below 1 at around the 5th PC (as shown in the jupyter notebook attached, the 5th PC has an eigenvalue of 0.9969. This was treated as 1.00 due to the decision that all eigenvalues would be rounded to 2 decimal places.)

A graph with a red line

Description automatically generated

D3, Total Variance of components: After it was determined that the first five PCs would be kept, PCA was performed again, this time keeping only the 5 components. After this was performed, the variance attributed to each PC was calculated using ‘.explained\_variance\_ratio\_’ function allowing for the calculation of the variance that is specifically attributed to that one PC. The variance attributed to each PC is as follows:

PC1: 19.94

PC2: 10.53

PC3: 10.27

PC4: 10.12

PC5: 10.0

See jupyter notebook for code and calculations there

D4, Total Variance Captured by Components: The total variance of the 5 PCs used is 60.87. The five components used attributed to 61% of the variance in the data.

D5, Summary of Data Analysis: Through Principal Component Analysis, PCA, we were able to determine which principal components in the dataset were the most useful, and thus reduce the dimensionality of the dataset. We were able to determine that the first 5 principal components were the ones that were most impactful, and were the ones to be looking at.

E, Sources for Third-Party Code: N/A as no third-party code was utilized.

F, Sources:

*A Guide to Principal Component Analysis (PCA) for Machine Learning*. (n.d.). Www.keboola.com. Retrieved July 15, 2023, from https://www.keboola.com/blog/pca-machine-learning#:~:text=PCA%20assumes%20a%20linear%20relationship

‌Braeken, Johan, and Marcel A. L. M. van Assen. *An Empirical Kaiser Criterion - American Psychological Association*. American Psychological Association, 18 Dec. 2015, https://www.apa.org/pubs/journals/features/met-met0000074.pdf.

PCA in Machine Learning: Assumptions, Steps to Apply & Applications | upGrad blog. (n.d.). *PCA in Machine Learning: Assumptions, steps to apply & applications*. upGrad blog. https://www.upgrad.com/blog/pca-in-machine-learning/

*A step-by-step explanation of principal component analysis (PCA)*. Built In. (n.d.). https://builtin.com/data-science/step-step-explanation-principal-component-analysis