|  |
| --- |
| Western Governors University |
| Data Mining II |
| Task 2 D212 |

|  |
| --- |
| Allison Casey  11-6-2024 |

**Part 1: Research Question**

**A1: PROPOSAL OF QUESTION**

Which combinations of features contribute most to the variation in the data?

**A2: DEFINED GOAL**

One goal would be to reduce the features of the churn data set.

**Part 2: Method Justification**

**B1: EXPLANATION OF PCA**

PCA is used to convert data with high dimensionality to lower dimensionality. PCA reduces the dimensionality of a data set by creating linear combinations of the initial features that maximize variance with the first component and then decreases with following components. The PCA algorithm does this by creating a covariance matrix of standardized data then eigenvalue decomposition performed on the matrix which then allows the eigenvalues to be ranked in descending order to determine the most important principal components (Ibm, 2024). The expected outcome of this would be to decrease the dimensionality of the churn dataset continuous numerical columns.

**B2: PCA ASSUMPTION**

One assumption of PCA is that there is a linear relationship between features.

**Part 3: Data Preparation**

**C1: CONTINUOUS DATA SET VARIABLES**

Continuous data set variables: (Note location variables were removed Lat, Lng, and Zip along with survey questions even though they were numeric as well as non-continuous numeric columns)

1. Population
2. Age
3. Income
4. Outage\_sec\_perweek
5. Tenure
6. MonthlyCharge
7. Bandwidth\_GB\_Year

**C2: STANDARDIZATION OF DATA SET VARIABLES**

See scaled\_df\_Task2.csv for scaled data and D212\_Task2.ipynb for code

A computer screen shot of a code

Description automatically generated

**Part 4: Analysis**

**D1: OUTPUT AND INTERMEDIATE CALCULATIONS**

Principal components matrix:

**A screenshot of a computer screen

Description automatically generated**

**D2: IDENTIFICATION OF THE TOTAL NUMBER OF COMPONENTS**

Both methods were used to see both options, but the number of principal components was identified using the Kaiser criterion which identified 4 total components.

A graph with a line

Description automatically generated

**D3: VARIANCE OF EACH COMPONENT**

|  |  |
| --- | --- |
| Component | Variance |
| PC1 | 1.99341957 |
| PC2 | 1.02577429 |
| PC3 | 1.01309161 |
| PC4 | 1.0029618 |

**D4: TOTAL VARIANCE CAPTURED BY COMPONENTS**

Total variance: 0.3571428571428571

**D5: SUMMARY OF DATA ANALYSIS**

The results of the PCA indicate that the churn data set can be reduced to an optimum number of five principal components that could then be used in a model. The data was first cleaned to remove location data and non-numerical features and non-continuous features. Then the data was standardized and transformed into a loading matrix using PCA. The number of principal components to be retained was then determined using the Kaiser criterion which is that components with eigenvalues greater than one are significant.

**Sources**

GeeksForGeeks “Implementing PCA in Python with Scikit-Learn.” *GeeksforGeeks*, 23 Sept. 2021, www.geeksforgeeks.org/implementing-pca-in-python-with-scikit-learn/. Accessed 30 Nov. 2024.

Ibm. “What Is Principal Component Analysis (PCA)?” *IBM*, 23 Aug. 2024, www.ibm.com/topics/principal-component-analysis. Accessed 30 Nov. 2024.