D212 Dimensionality Reduction Methods

Nhi Le

Western Governors University

October 25, 2023

**Part I: Research Question**

**A1. Proposal of question**

Can we use Principal Component Analysis to identify the optimal number of components of the customer base?

**A2. Defined goal**

The goal of this analysis is to find the principal components of the customers and to identify the total explained variance by those components. By getting insights into the amount of variance the principal components account for, the company will get a more profound understanding of the customer's characteristics. Through that, stakeholders can improve effectiveness of their marketing plans.

**Part II: Method Justification**

**B1. Explanation of PCA**

Principal component analysis (PCA) is a technique used to reduce the dimensionality of large datasets while retaining most of the essential information. PCA includes 5 steps: normalizing continuous variables, calculating covariance matrix to check correlation, calculating the eigenvectors and eigenvalues of the covariance matrix to identify the principal components, constructing a feature vector for selecting the principal components, and transforming the data according to the axes of the principal components (Whitfield, 2023). The expected outcomes are to find the optimal number of components and to identify the total explained variance by those components.

**B2. PCA Assumption**

The assumption of PCA is that there is correlation among the variables. PCA is most effective for datasets where there is strong correlation among variables, allowing it to effectively reduce the dimensionality of the dataset. In the absence of such correlation, PCA will be able to achieve dimensionality reduction (Keboola, 2022).

**Part III: Data Preparation**

**C1. Continuous data set variables**

The 9 continuous variables including:

* ‘Age’
* ‘Income’
* ‘Email’
* ‘Contacts’
* ‘Outage\_sec\_perweek’
* ‘Bandwidth\_GB\_Year’
* ‘Tenure’
* ‘MonthlyCharge’
* ‘Yearly\_equip\_failure’

**C2. Standardization of data set variables**

The cleaned data set will be submitted as ‘churn\_prepared1.csv’ along with this doc file.

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

**Part IV: Analysis**

**D1. Principal components**

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

**D2. Identification of the total number of components**

I will identify the total number of principal components using Kaiser criterion:

A screen shot of a computer

Description automatically generated

**D3. Variance of each component**

The explained variance of the first principal component is 22.15% and of the second principal component is 11.52%.

A screenshot of a computer code

Description automatically generated

**D4. Total variance captured by components**

The total variance is 33.67%,

A close-up of a white background

Description automatically generated

**D5. Summary of data analysis**

After performing PCA, I could reduce the data set’s dimensions from 9 original variables to 2 principal components. The optimal number of principal components is 2 and the total explained variance captured by those components is 33.67%. The first principal component (PC1) alone could explain 22.15% variance, which is a lot more than a single original variable could explain on average. Tenure and Bandwidth\_GB\_Year have a high correlation with PC1, over 0.70. The stakeholders should investigate those features more to see how much of their impact within customer’s characteristics.

**Part V: Attachments**

**E. Sources for third-party code**

Soriano, P., & Kebabci, C. (2023, July 31). *Principal Component Analysis in Python (Example Code)*. Statistics Globe. https://statisticsglobe.com/principal-component-analysis-python

**F. Sources**

Whitfield, B. (2023, March 29). *A step-by-step explanation of principal component analysis (PCA)*. Built In. https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Keboola. (2022, April 2). *A guide to principal component analysis (PCA) for Machine Learning*. Keboola. https://www.keboola.com/blog/pca-machine-learning