D212 Dimensionality Reduction Methods

Telecommunication Churn Data

Darien Nguyen

Western Governors University

D212 Task Two: Principal Component Analysis

June 14, 2023

**Part I: Research Question**

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

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

How can PCA help us understand the key customer characteristics that drive the profitability of our telecommunications company?

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

Our goal is to utilize PCA to reduce the dimensionality of the telecom churn dataset, enabling the identification of essential patterns and relationships among variables while simplifying the data for analysis and gaining insights into key factors influencing churn.

**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 characteristics of our telecom customers by applying the technique to the relevant customer data set. This data may include demographics, usage patterns, preferences, and other variables that describe customer behavior and characteristics. By utilizing PCA on this data, we can gain insights into the key customer variables that contribute to the overall variability in the dataset.

The expected outcomes of applying PCA to our telecom customer data are as follows:

1. Identification of Principal Components: PCA will identify the principal components that capture the most significant variations in the customer characteristics. Each principal component represents a combination of the original variables and contributes to the overall variability in the data.
2. Variance Explained: PCA will provide information about the proportion of variance explained by each principal component. This allows us to understand the relative importance of each component in capturing the overall variability in customer characteristics. By examining the cumulative proportion of variance explained, we can determine the number of components to retain.
3. Reduction in Dimensionality: PCA reduces the dimensionality of the customer data by transforming it into a smaller set of uncorrelated components. This reduction simplifies the representation of the customer characteristics, making it easier to interpret and analyze the underlying patterns and relationships.
4. Loadings and Variable Relationships: PCA provides loadings that indicate the correlations between the original customer variables and the principal components. These loadings help us understand the relationships and contributions of each variable to the principal components. Variables with higher absolute loadings have a stronger relationship with a particular component.

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

One assumption of PCA is linearity. PCA assumes that there exists a linear relationship among the variables in the dataset. It identifies the directions of maximum variance in the data by considering linear combinations of the original variables. However, if the relationship between variables is nonlinear, our PCA may not accurately capture the underlying structure of the data.

**Part III: Data Preparation**

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

1.  Identify the continuous dataset variables that you will need in order to answer the PCA question proposed in part A1.

numerical\_columns = ['Population', 'Children', 'Age', 'Income', 'Outage\_sec\_perweek',

'Yearly\_equip\_failure', 'Tenure', 'MonthlyCharge', 'Bandwidth\_GB\_Year']

This code snippet holds the continuous dataset used for the PCA.

2.  Standardize the continuous dataset variables identified in part C1. Include a copy of the cleaned dataset.

In order to standardize the continuous dataset variables, the z-scores were calculated for

each variable. This was achieved by subtracting the mean of each variable from its values and then dividing by the standard deviation. The z-scores represent the number of standard deviations each data point is away from the mean. Standardization through z-scores ensures that the variables have a mean of zero and a standard deviation of one, making them comparable and removing the influence of scale. The **dropna()** function was used to remove any rows containing missing values, ensuring that only complete cases were considered in the analysis. Lastly, scaler.fit\_transform was used to further standardize our data.

**Part IV: Analysis**

D.  Perform PCA by doing the following:

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

Principal Components:

[[-2.04063881e-02 1.15399397e-02 -6.43986035e-04 3.23927720e-03

8.63049058e-03 1.06097565e-02 7.05399118e-01 4.28425184e-02

7.06985893e-01]

[-1.65993405e-01 6.10057311e-01 -5.50261612e-01 1.46114301e-01

4.02291979e-01 3.36418394e-01 -2.90901450e-02 -1.61132417e-02

4.12198348e-03]

[-4.18861093e-01 -3.48766355e-02 -1.75934846e-01 4.41778342e-01

-5.35601972e-01 -6.71207032e-03 2.88790777e-02 -5.56295425e-01

-2.16944583e-03]

[-2.03343203e-01 -1.97191217e-03 5.05971010e-01 2.88422561e-01

-8.76729349e-02 7.23152948e-01 -1.92493524e-02 2.97006595e-01

-1.52719277e-02]

[ 7.27302890e-01 3.08427954e-01 2.15184846e-01 3.45312090e-02

-1.03192770e-01 2.38776057e-01 3.82937947e-02 -5.09152910e-01

6.31884928e-03]

[ 2.09116607e-01 1.30450034e-01 1.46978666e-01 7.92856597e-01

1.79273879e-01 -4.40946380e-01 -1.33797024e-02 2.49534522e-01

3.06477700e-03]

[-7.81753452e-02 6.64356349e-01 1.39150148e-01 -2.24424019e-01

-5.60439965e-01 -2.34539500e-01 -2.87614135e-02 3.35880519e-01

6.75863684e-03]

[ 4.20030960e-01 -2.69426467e-01 -5.67879075e-01 1.41785589e-01

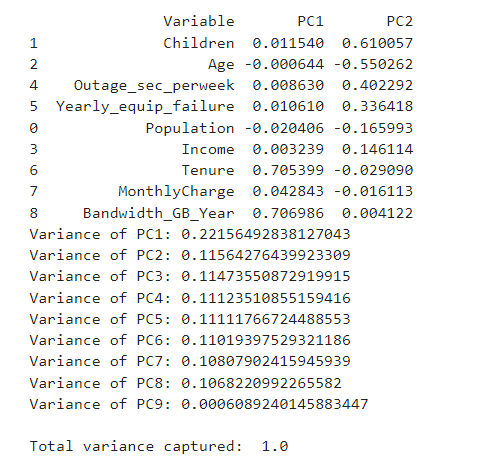
-4.32029899e-01 2.39287202e-01 -2.23509402e-02 4.04680517e-01

1.48152079e-02]

[ 2.34868794e-04 -1.92781007e-02 2.22566187e-02 -1.07682958e-03

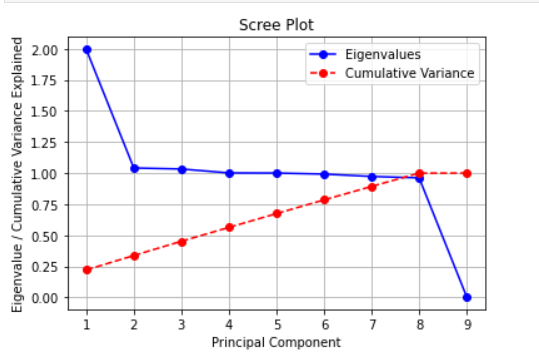
4.28365016e-05 3.45332462e-05 -7.05258176e-01 -4.62726368e-02

7.06824963e-01]]



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

From the scree plot, we can identify up to two principal components.



Eigenvalue of PC1: 1.9942977634318015

Eigenvalue of PC2: 1.0408962649612044

Eigenvalue of PC3: 1.0327300900759093

Eigenvalue of PC4: 1.001223116944759

Eigenvalue of PC5: 1.0001660320666828

Eigenvalue of PC6: 0.9918519148154514

Eigenvalue of PC7: 0.9728153175226171

Eigenvalue of PC8: 0.9615017824753505

Eigenvalue of PC9: 0.00548090263773039

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

Variance of PC1: 0.22156492838127034

Variance of PC2: 0.11564276439923317

Variance of PC3: 0.1147355087291991

Variance of PC4: 0.11123510855159402

Variance of PC5: 0.11111766724488543

Variance of PC6: 0.1101939752932119

Variance of PC7: 0.1080790241594593

Variance of PC8: 0.10682209922655825

Variance of PC9: 0.0006089240145883452

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

Total variance captured: 1.0

5.  Summarize the results of your data analysis.

The data analysis using PCA on the telecom churn dataset has provided several key findings. The analysis began by preparing the dataset, including removing outliers and null values, and standardizing the continuous variables. PCA was then performed on the standardized data.

The scree plot was used to determine the number of principal components to retain. The plot showed that the cumulative explained variance plateaus after the first few components, suggesting that a smaller number of components can capture a significant portion of the data's variability.

A total of eight principal components were extracted from the data, and the loadings of these components were analyzed to understand the impact of each variable on the components. The variance of each principal component was also calculated, revealing the proportion of total variance explained by each component. The first two components explained a substantial amount of variance, followed by a gradual decrease in variance explained by the subsequent components.

The analysis demonstrates that PCA is a powerful tool for dimensionality reduction and can effectively capture the underlying patterns in the data. The findings highlight the importance of a smaller set of principal components in explaining the majority of the dataset's variance, providing valuable insights for future analysis and decision-making in the context of telecom churn prediction.

**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 sources used.

F.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.  
No sources used.

G.  Demonstrate professional communication in the content and presentation of your submission.

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=38a21b4a-a8f7-4ade-9afa-b0440010ed38#>