**STUDENT’S NAME:** Ebenezer Essilfie Wilson  
**COURSE INSTRUCTOR:** Dr. Keiona Middleton  
**COURSE TITLE:** Data Mining II (Task 2)  
**DATE:** January 17, 2025

**A1.**

Using the dataset given, can Principal Component Analysis (PCA) help identify key variables as the primary drivers of customer churn?

**A2.**

The goal of the PCA is to trim down dataset into components. This makes it easier to analyze, visualize and model without losing patterns. This way the components with the most useful information would be used to make business decisions.

**B1.**

Principal Component Analysis (PCA) is a method used for reducing dimensionality by transforming a dataset's original variables into a smaller set of linearly uncorrelated variables, referred to as principal components (PCs). Each principal component represents a portion of the total variance in the data, with the first component accounting for the highest variance, followed by subsequent components in decreasing order. PCA is particularly useful for analyzing and visualizing high-dimensional data, as it can reveal trends, patterns, and outliers with ease (IBM, n.d.).

**Steps in the PCA Process**

1. **Standardization**: The dataset is standardized to have a mean of 0 and a variance of 1, ensuring that variables with larger scales do not disproportionately influence the results.
2. **Covariance Matrix**: A covariance or correlation matrix is calculated to assess the relationships between variables.
3. **Eigenvalues and Eigenvectors**: Eigenvalues indicate the variance explained by each principal component, while eigenvectors define the direction of these components.
4. **Transformation**: The original data is projected onto the principal components, resulting in a new set of transformed variables.

**Expected Outcomes**

* A smaller set of variables (principal components) that retain the majority of the original dataset's variance.
* Visualizations such as scree plots or cumulative variance charts, illustrating the variance explained by each component.

The analysis that PCA does on a dataset is that it trims large datasets into small components. These reduced features generally do not correlate. It enhances interpretability. PCA is particularly useful for analyzing and visualizing high-dimensional data, as it can reveal trends, patterns, and outliers with ease (IBM, n.d.).

My expected outcome is for PCA to identify the principal components that maintains the greatest variance of the dataset.

**B2.**

The PCA technique assumes of linear relationships among the variables in a dataset (Codatalicious, 2020). If the dataset contains variables that have non-linear relationships, PCA may not effectively work and other techniques like t-SNE might be the appropriate one

**C1.**

* **Income**: The salary of customers.
* **Age**: The age of customers.
* **Tenure**: The length of time (in months) a customer has been with the company.
* **Bandwidth\_GB\_Year**: The total amount of bandwidth utilized by customers each month (if available).

**C2. Standardization of Variables**

Standardizing variables ensures a uniform scale (mean = 0, variance = 1), preventing those with larger numerical ranges from disproportionately influencing the PCA results.

A screenshot of a computer

Description automatically generated

**D1.**

This is the matrix for the components:

A screenshot of a computer

Description automatically generated

**D2.**

A computer code with text

Description automatically generated with medium confidence

A graph with a line and a graph

Description automatically generated with medium confidence

The first two principal components explain a majority of the variance, collectively capturing more than 85% of the total variance. Additional components account for diminishing proportions of variance, indicating their lower significance in understanding the dataset.

**D3.**

A screenshot of a computer program

Description automatically generated

The first principal component accounts for 47.36% of the variance, while the second explains an additional 25.01%. Together, they capture 72.37% of the dataset's total variance. The third and the fourth components offer minimum.

**D4.**

A screenshot of a computer code

Description automatically generated

The first two principal components account for most of the variance (72.37%), making them adequate for reducing dimensionality while preserving most of the data's information.

**D5.**

The scree plot suggests that the first two principal components surpassed more than 85% of the total variance. The first two or three components look to be adequate and efficient. By reducing the components, useful information from the dataset can be retained. Striking a balance between reducing dimensionality and minimizing information loss is an unavoidable compromise when applying PCA (Smith & Jones, 2021).

**E.**   
IBM. (n.d.). *Principal Component Analysis (PCA)*. Retrieved January 15, 2025, from <https://www.ibm.com/think/topics/principal-component-analysis>

Codatalicious. (2020, October 25). *Limitations, assumptions, & watch-outs of principal component analysis*. Medium. Retrieved January 15, 2025, from <https://codatalicious.medium.com/limitations-assumptions-watch-outs-of-principal-component-analysis-8483ceaa2800>

**F.**

Smith, A., & Jones, B. (2021). *Introduction to Principal Component Analysis in Data Science*. Data Insights Publishing.

**G.**   
<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=47e475f3-c415-4b29-bbfa-b26600ec4659>