D212: Data Mining II Task 2

Justin Huynh

Student ID: 012229514

M.S. Data Analytics

A1. Proposal of Question

My research question for this project is "Can we reduce the dimensionality of customer usage data to identify key factors that contribute to customer churn?" We will use the Principle Component Analysis (PCA) to answer our research question.

A2. Defined Goal

The goal of the data analysis is to identify the most significant factors influencing customer churn by reducing the dimensionality of the dataset using Principal Component Analysis (PCA). This will help to simplify the complexity of the data, making it easier to focus on the primary components that explain the most variance in customer behavior and churn risk. By understanding these key components, an organization can improve its efforts to reduce churn and customer retention rate.

B1. Explanation of PCA

Principal Component Analysis (PCA) is a technique that reduces the number of dimensions in large data sets to principal components while retaining most of the original information. It does this by identifying the directions (principal components) in which the data varies the most. These principal components are linear combinations of the original variables and are uncorrelated to each other.

Using the churn data set, PCA will take the variables such as MonthlyCharge, Bandwidth_GB_Year, and potentially others, and reduce them to a smaller number of components that capture the most significant variations in the data. The expected outcome is that PCA will reveal a few key components that explain most of the variability in customer behavior and churn risk. These components can then be

analyzed to identify the underlying factors that contribute most significantly to customer churn.

B2: PCA Assumption

One assumption of PCA is that the data has a linear structure, meaning that the relationships between the variables can be captured through linear combinations. This assumption is important because PCA works by finding the directions (principal components) that maximize the variance in the data through linear transformations. If the relationships among the variables are non-linear, PCA may not be able to capture the true underlying structure of the data effectively, which could lead to less meaningful components.

C1: Continuous Data Set Variables

To perform PCA for the research question, we need to focus on continuous variables that represent customer usage patterns and possibly other relevant metrics. The continuous variables from the data set that are most relevant to our research question are as follows:

- **MonthlyCharge**: The amount charged to the customer monthly.
- **Bandwidth_GB_Year**: The average amount of data used by the customer in a year.
- **Outage_sec_perweek**: Average number of seconds per week of system outages in the customer's neighborhood.
- **Tenure**: Number of months the customer has stayed with the provider.

These continuous variables highlight different aspects of customer behavior, which makes them suitable variables for PCA to better understand the factors contributing to churn.

C2: Standardization of Data Set Variables

A copy of the fully prepared data set will be submitted as 'prepared_data_d212_task2.csv'.

See code attached in WGU_D212_Task_2.ipynb.

D1: Principal Components

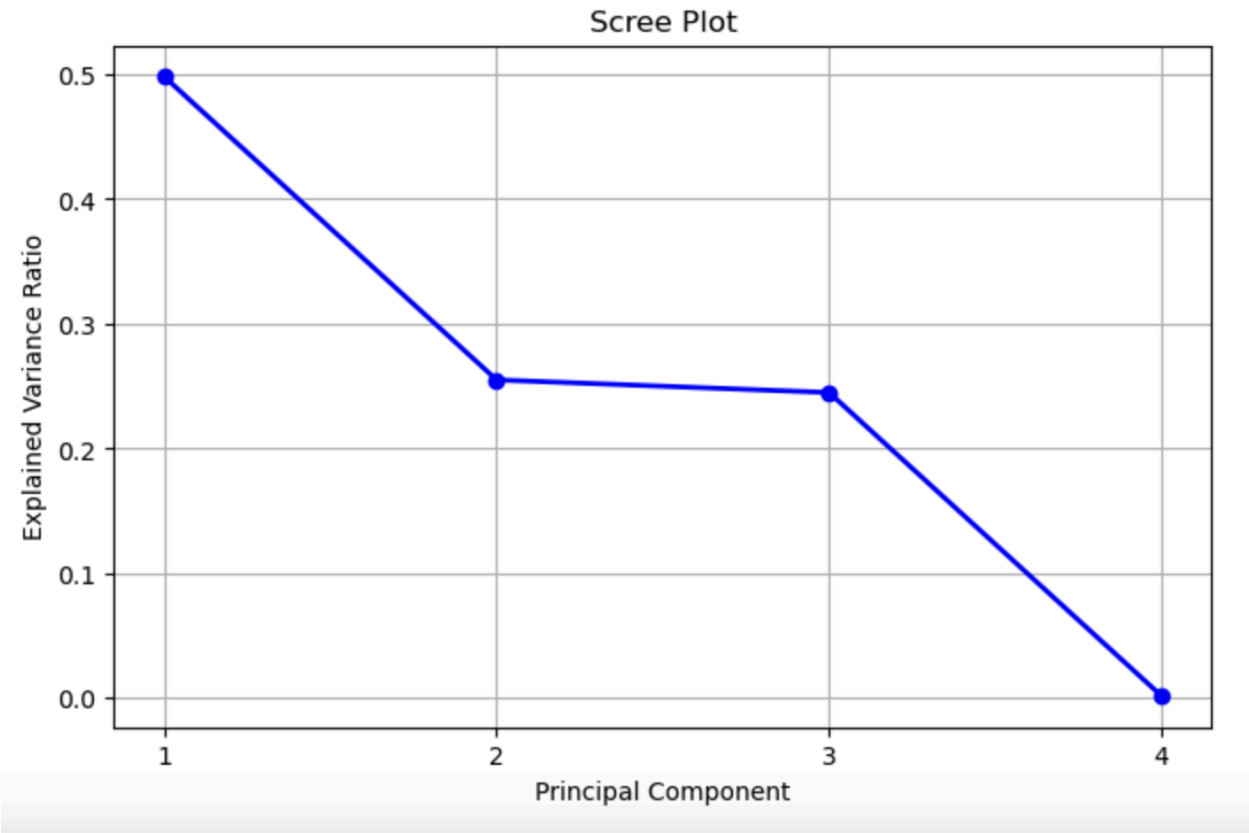
	MonthlyCharge	Bandwidth_GB_Year	Outage_sec_perweek	Tenure
0	0.040761	0.707163	0.005898	0.705850
1	0.709435	-0.000646	0.703255	-0.046197
2	0.702127	0.005274	-0.710914	-0.039890
3	-0.045358	0.707031	0.000047	-0.705727

See code attached in WGU_D212_Task_2.ipynb.

D2: Identification of the Total Number of Components

There are 4 total numbers of principal components using the elbow method in this analysis. The scree plot shows that the explained variance drops sharply after the first component and then gradually decreases before another small drop after the third component. The 'elbow' appears at the second component, where the rate of decrease in variance begins to slow down significantly. This suggests that the first two components capture the majority of the significant variance in the data. Retaining these two components would effectively summarize the original data set while minimizing the

loss of information. The third and fourth components contribute less to the total variance, indicating that they add less new information. However, including the third component might still be beneficial, as it captures additional variance before the curve levels off.



See code attached in WGU_D212_Task_2.ipynb.

D3: Variance of Each Component

The variance of each principal component are as follows:

- PC1: 49.83%
- PC2: 25.51%
- PC3: 24.50%
- PC4: 0.16%

See code attached in WGU_D212_Task_2.ipynb.

D4: Total Variance Captured by Components

The total cumulative variance captured by the principal components are as follows:

- First Component (PC1): 49.83%
- First Two Components (PC1 + PC2): 75.34%
- First Three Components (PC1 + PC2 + PC3): 99.84%
- All Four Components (PC1 + PC2 + PC3 + PC4): 100%

See code attached in WGU_D212_Task_2.ipynb.

D5: Summary of Data Analysis

In this analysis, the Principal Component Analysis (PCA) was applied to reduce the dimensionality of a data set containing customer usage patterns and service-related variables. The goal was to identify the most important factors contributing to customer churn. The PCA produced a matrix where each principal component represents a linear combination of the original variables. This matrix reveals how each original variable contributes to the principal components. For example, PC1 variables like Bandwidth_GB_Year and MonthlyCharge drove almost 50% of the variance, indicating a strong contribution to the variance captured by this component.

Using the scree plot and the elbow method, it was determined that the first three components capture the majority of the variance. The first three components capture almost all the variance (99.84%), suggesting that retaining these three components would effectively summarize the original dataset without much loss of information. The fourth component contributes minimally to the total variance, indicating that it does not add substantial new information. Since PC1-PC3 reflect the primary underlying factors

in customer behavior, these principal components can be used in predictive models or further analyses to develop targeted strategies for reducing customer churn.

E: Sources for Third-Party Code

1. “How to Create a Scree Plot in Python” Retrieved from https://www.statology.org/scree-plot-python/
2. “PCA Using Python: A Tutorial” Retrieved from https://builtin.com/machine-learning/pca-in-python

F: Sources

1. “PCA” Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
2. “In Depth: Principal Component Analysis” Retrieved from https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html