Telecom Churn Analysis

Dimensionality Reduction Methods

Javier Lopez

D212 - Data Mining II

Table of Contents

Part I: Research Question

A1. Research Question

How can we reduce the dimensionality of customer data while retaining the most important information, in order to better understand the key factors that contribute to customer behavior and preferences?

A2. Objective

Our goal is to use PCA (Principal Component Analysis) to reduce customer data by identifying the principal components that capture the largest amount of variance in the data. The reduced dataset can then be used as input for other machine learning algorithms, such as clustering or classification, to improve the performance of targeted marketing campaigns and strategies.

Part 2: Method Justification

B1. PCA Analysis

PCA transforms the original variables in a dataset into a new set of uncorrelated variables called principal components. The new variables capture the maximum variance in the data while reducing the number of dimensions. The first step in PCA is calculating the covariance matrix of the data fed to it. This expresses the relationship between each variable in the dataset to the others. The algorithm then computes the eigenvectors and eigenvalues of the covariance matrix, the directions of the principal components, and the magnitude of the variance captured by each principal component respectively. PCA then sorts the eigenvectors in descending order based on their eigenvalues. The eigenvector with the highest eigenvalue captures the most variance in the data and thus is assigned to the first principal component, the process continues until all eigenvectors have been assigned. The number of principal components specified to the PCA algorithm can be determined using the elbow method on a scree plot or the Kaiser criterion. When transforming the original data to reduce dimensionality, PCA projects it onto the selected principal components by multiplying the standardized dataset by the selected eigenvectors to obtain the principal component scores.

B2. Assumption

PCA (principal component analysis) relies on an important assumption that the most significant information in a dataset is captured by the principal components with the highest variance. PCA assumes that the directions where the data varies the most are the most important. While this assumption generally works well, it may not always be accurate, especially if the components with low variance contain important information for the specific context being analyzed. Additionally, PCA assumes that there is a linear relationship between the variables in the dataset, as it seeks to identify linear combinations of variables in the dataset that capture the most amount of variance in the data.

Part III. Data Preparation

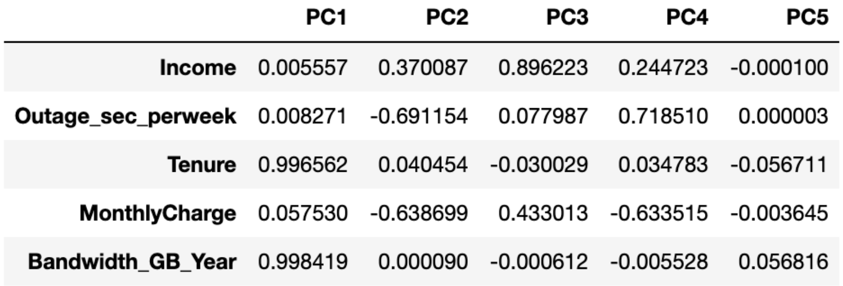
C1. Variables

Variable Name	Variable Name	Variable Name
Income	Tenure	Bandwidth_GB_Year
MonthlyCharge	MonthlyCharge	

Part IV: Analysis

D1. Principal Component Matrix

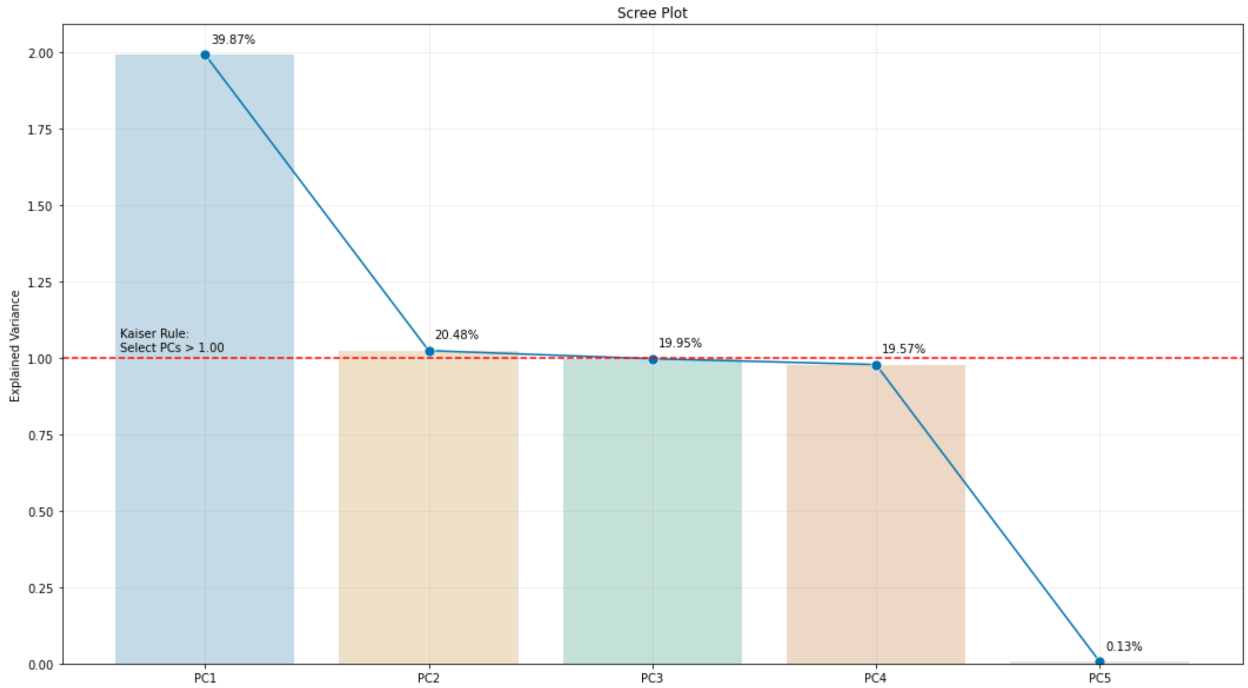
Below is the Principal Component Matrix representing the correlation between the continuous variables (rows) and principal components (columns). Each value in the matrix represents the loading of the variable on the respective principal components. The loading can be interpreted as a measure of the contribution of the variable to the principal component. The higher the absolute value of the loading, the stronger the correlation between the variable and the principal component. From our principal component matrix, we can see that Bandwidth_GB_Year has the highest correlation to PC1 followed by Tenure, while Income has the highest correlation to PC3.



	PC1	PC2	PC3	PC4	PC5
Income	0.005557	0.370087	0.896223	0.244723	-0.000100
Outage_sec_perweek	0.008271	-0.691154	0.077987	0.718510	0.000003
Tenure	0.996562	0.040454	-0.030029	0.034783	-0.056711
MonthlyCharge	0.057530	-0.638699	0.433013	-0.633515	-0.003645
Bandwidth_GB_Year	0.998419	0.000090	-0.000612	-0.005528	0.056816

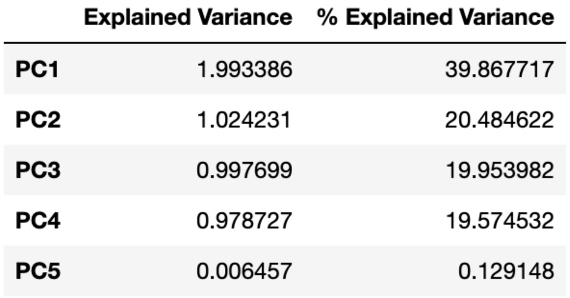
D2. Scree Plot

The scree plot below shows the percentage of explained variance for each principal component. We based our principal component selection on the Kaiser Rule, which states that any principal component with an explained variance, or eigenvalue, over 1.00 should be kept. Based on the Kaiser Rule, we kept the first two principal components, which had an explained variance of 1.99 and 1.02 respectively. The Elbow Method shown on the scree plot confirms that two principal components, chosen using the Kaiser Rule, is the best number of principal components.



D3. Principal Component Variance

The table below shows the explained variance and the percentage of explained variance for each principal component. From the table we can gather that PC1 explains the most variance in our data, capturing 1.99 units, or 39.87% of the variance in the dataset, when the data is standardized.



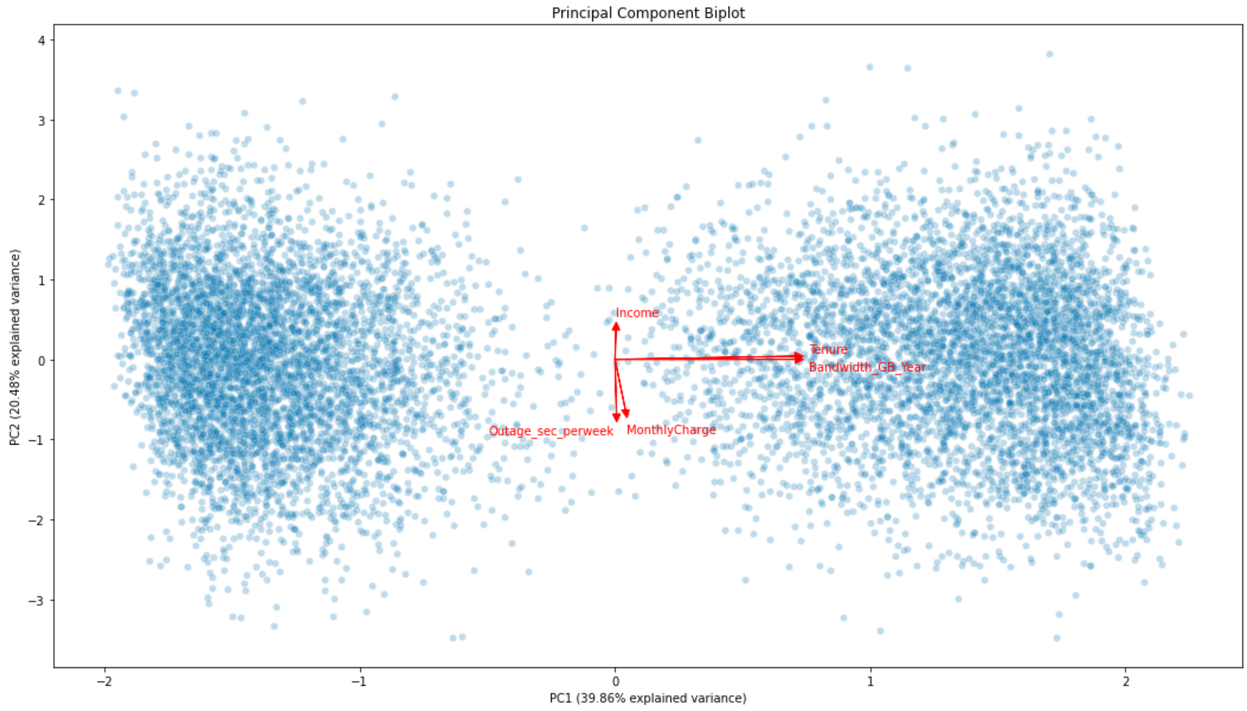
	Explained Variance	% Explained Variance
PC1	1.993386	39.867717
PC2	1.024231	20.484622
PC3	0.997699	19.953982
PC4	0.978727	19.574532
PC5	0.006457	0.129148

D4. Total Principal Component Variance

The three principal components we selected captured a total variance of 60.35%. The "% Explained Variance" column in the table above shows a breakdown of the percentage of explained variance for each principal component. The total percentage of explained variance is calculated by multiplying the sum of the explained variance of the selected principal components by one hundred.

D5. Results

Our research question sought to reduce the dimensionality of customer data while retaining the most important information, in order to better understand the key factors that contribute to customer behavior and preferences. We used PCA to reduce the dimensionality of our data and determined that two principal components were the most important in our data, as they captured the most variance. All in all, by transforming our data using PC1 and PC2 we were able to reduce the dimensionality of our dataset while still capturing 60.35% of the total variance. When plotting the PCA results we were able to begin visualizing two distinct clusters within our dataset. The biplot below shows how the original features contributed to the shape of the data by representing their relationships with the selected principal components (red arrows). This visualization helps us gain insights into the underlying structure of the data and the role of the original features in forming the observed clusters.



Part V. Attachments

E. Sources

No third-party code or in-text citations were used to complete our research assessment.