D212 Principal Component Analysis

The purpose of this Data Mining Report is to be able to identify the principal components of customers. The goal of the report is to find valuable and actionable insights to satisfy business needs. The principal component is ideal since the data provided has many variables that require a process to determine any possible relationship. Once we have performed the PCA then one question we would ask is what marketing techniques can be applied or implemented.

Principal Component Analysis is the process in which we reduce the feature space by eliminating and creating new features (Brems, 2017). However, we create these new independent variables in a specific way and order these new variables by how well they predict our dependent variable. The expected outcomes are to begin with a large number of predictor variables and have them reduced to a substantial amount in order to assist your model in processing the data properly. One assumption of PCA is that your data needs to be scaled (Principal Components Analysis (PCA) using SPSS Statistics, n.d.). Scaling the variables consisted of determining the highest value in those columns and then dividing each value by the highest amount. This allows the ranges of the particular column to have a minimum value of 0 and the highest value to be 1. All other values are a decimal number below 1. This is extremely important in ensuring all features could be properly computed by the model in balanced way.

From the original 50 variables that were in the dataset only 10 were used to perform Principal Component Analysis. The continuous variables that were used in the model were Population, Age, Income, Outage\_sec\_perweek,Email, Contacts, Yearly\_equip\_failure, Tenure, MonthlyCharge, Bandwidth\_GB\_Year. Once the categorical variables were removed, we were ready to perform the feature reduction.

Once the data was prepared, a function was created to calculate how many components were going to be needed in order to capture 95% of the variance. The screenshot below shows us that we would require 7 components. The elbow method was also used to confirm the number of components required. Attached is a screenshot. Text

Description automatically generated

Chart, line chart

Description automatically generated

Once PCA was performed on the dataset, we were able to create a matrix that showed us both the amount of each individual component and the sum of all components. The first columns in the screenshot below identifies each component. The second column shows us just how much variance each individual component contains from the dataset. Here we are ablet to see how PC1 contains 51% of the total variance while PC2 has 18%, PC3 9%, PC4 4%, PC5 has 4%, PC6 has 4%, and PC7 has 4%. The third column shows us the total variance captured by principal component analysis. This shows us that the 7 components make up 95% of the total variance.

Table

Description automatically generated

After performing the analysis, one result that we were able to take from the data is how much PC1 contributes to identifying whether someone will churn, leave the cable company, or not. Attached is a screenshot that highlights this fact. When a scatterplot is created to show PC1 against PC2, we see how a member churns can be explained by PC1 and PC2. Since these two components were the ones that contained the majority of the variance, then we would be correct in ensuring we processed the information to see how it can contribute for better market results.

Chart

Description automatically generated

# Works Cited

Amelia, A. (2018, September 27). *K-Means Clustering: From A to Z*. Retrieved from Towards Data Science: https://towardsdatascience.com/k-means-clustering-from-a-to-z-f6242a314e9a

Brems, M. (2017, April 17). *A One-Stop Shop for Principal Component Analysis*. Retrieved from Towards Data Science: https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

*Principal Components Analysis (PCA) using SPSS Statistics*. (n.d.). Retrieved from Laerd Statistics: https://statistics.laerd.com/spss-tutorials/principal-components-analysis-pca-using-spss-statistics.php