**Principal Component Analysis of Churn Dataset**

Kelseyann Wright

Student ID: 00565814

College of Information Technology, Western Governors University

Data Mining II - D212

Kesselly Kamara

January 13, 2024

**Principal Component Analysis of Churn Dataset**

**A1. PROPOSAL OF QUESTION**

The proposed question is, “How many principal components can be used to adequately describe the churn customer dataset?”

**A2. DEFINED GOAL**

The goal of this analysis is to use principal component analysis (PCA) to identify important features of customers to support strategic business decision-making. PCA is useful because it helps reduce the number of features in the dataset while retaining important information, which can improve efficiency of models and make it easier to identify trends and relationships in datasets with high dimensionality.

**B1. EXPLANATION OF PCA**

PCA uses the covariance of dataset features to identify relationships and create new variables, called principal components, in order to reduce dataset dimensionality and simplify data exploration and analysis. The principal components created during the analysis are mixtures of the initial variables. The principal components are largely uncorrelated but retain most of the information from the initial variables in the dataset. The expected outcome of PCA is the identification of principal components that can be used to reduce the dimensionality of our dataset while retaining as much information as possible to explain the variance in the dataset.

**B2. SUMMARY OF THE TECHNIQUE ASSUPTION**

One assumption of PCA is that the features are correlated. If features are not correlated, PCA will not be able to identify principal components. The identification of principal components relies on correlation and covariation between dataset features.

**C1. CONTINUOUS DATASET VARIABLES**

The principal component analysis (PCA) included seven continuous variables: 'Lat', 'Lng', 'Income', 'Outage\_sec\_perweek', ‘Tenure’, ‘MonthlyCharge’, and “Bandwidth\_GB\_Year’. These variables were isolated for use in the analysis, while discrete and categorical variables were not included.

**C2. STANDARDIZATION OF DATASET VARIABLES**

The data were standardized using the StandardScaler() function from the sci-kit learn library in python (Pedregosa et. al 2011). This function scales features so that they will have a Gaussian normal distribution with a mean of zero and a standard deviation of 1. Scaling the data is important for PCA so that variables with larger numerical values do not dominate the analysis or appear as more important than variables with smaller values. The prepared dataset used for PCA is attached (churn\_processed\_T2.csv). A summary of the scaled dataset is provided below.

Lat Lng Income Outage\_sec\_perweek

count 1.000000e+04 1.000000e+04 1.000000e+04 1.000000e+04

mean 4.526157e-16 -5.998757e-16 5.222489e-17 9.521273e-17

std 1.000050e+00 1.000050e+00 1.000050e+00 1.000050e+00

min -3.823984e+00 -5.338407e+00 -1.399303e+00 -3.327464e+00

25% -6.282263e-01 -4.157121e-01 -7.299042e-01 -6.665728e-01

50% 1.173845e-01 1.889583e-01 -2.353430e-01 5.615783e-03

75% 6.160141e-01 7.056100e-01 4.765941e-01 6.611971e-01

max 5.863971e+00 1.657146e+00 7.769694e+00 3.765413e+00

Tenure MonthlyCharge Bandwidth\_GB\_Year

count 1.000000e+04 1.000000e+04 1.000000e+04

mean 2.273737e-17 -2.529532e-16 9.094947e-17

std 1.000050e+00 1.000050e+00 1.000050e+00

min -1.267917e+00 -2.157520e+00 -1.481263e+00

25% -1.006306e+00 -7.602435e-01 -9.865847e-01

50% 3.420043e-02 -1.197020e-01 -5.162246e-02

75% 1.019358e+00 6.546178e-01 1.003942e+00

max 1.417195e+00 2.737145e+00 1.723716e+00

**D1. PRINCIPLE COMPONENTS**

The loadings matrix for the principal components is provided below.

PC1 PC2 PC3 PC4 PC5

Lat -0.023929 0.699473 -0.122083 0.024641 -0.034349

Lng 0.007948 -0.706476 0.003629 0.072063 0.030741

Income 0.003751 0.072029 0.330376 0.904620 0.257294

Outage\_sec\_perweek 0.005784 -0.026424 -0.687545 0.045709 0.721765

Tenure 0.705612 0.020180 0.038144 -0.028709 0.034436

MonthlyCharge 0.040770 -0.071391 -0.633843 0.415871 -0.639926

Bandwidth\_GB\_Year 0.706941 0.015428 -0.001818 -0.000478 -0.006245

PC6 PC7

Lat -0.702476 0.001028

Lng -0.703332 0.000755

Income 0.033327 -0.001254

Outage\_sec\_perweek 0.059290 0.000025

Tenure -0.014295 -0.705716

MonthlyCharge 0.083492 -0.045372

Bandwidth\_GB\_Year -0.007080 0.707038

**D2. IDENTIFICATION OF THE TOTAL NUMBER OF COMPONENTS**

Based on the Kaiser Criterion (see plot below), the total number of principal components is three.

A graph with a red line

Description automatically generatedA graph with a line

Description automatically generated

**D3. VARIANCE OF EACH COMPONENT and D4. TOTAL VARIANCE CAPTURED BY COMPONENTS**

Table 1 shows the variance of each principal component identified in part D2, as well as the calculated total variance captured by the components.

|  |  |  |
| --- | --- | --- |
| **Table 1: Variance Explained by principal components** | | |
|  | *Explained Variance* | *Proportion of Total Variance Explained (%)* |
| **PC1** | 1.99398 | 28.48258 |
| **PC2** | 1.10128 | 15.73099 |
| **PC3** | 1.02448 | 14.63402 |
| **TOTAL** | 4.11974 | 58.84758 |

**D5. SUMMARY OF DATA ANALYSIS**

From the PCA analysis on the churn dataset, it was concluded that no more than 3 principal components are required. Table 2 lists the selected principal components and the highest correlated features from the “loadings” table for each. Based on this information, we can select which features would be best for future modeling of the dataset.

Table 2: Selected PCA components and correlated features

|  |  |
| --- | --- |
| **Component** | **Top Correlated Features (loading > 0.5).** |
| PC1 | Bandwidth\_GB\_Year, Tenure |
| PC2 | Outage\_sec\_perweek, MonthlyCharge |
| PC3 | Lng, Lat |

PCA is useful because it helps reduce the number of features in the dataset, which can improve efficiency of models and reduce feature correlation when modeling. Large amounts of feature correlation can negatively affect model predictions, so reducing feature correlation is incredibly important. PCA can also make it easier to visualize and explore the data, as it identifies correlation between variables and reduces the number of features for analysis.

**E. SOURCES OF THIRD-PARTY CODE**

No sources of third-party code were used.

**F. SOURCES**

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.