­

**Performance Assessment for D212: Data Mining II  
Task 2: Dimensionality Reduction**

Christopher Fischer - 011933891

College of Information Technology, Western Governors University

September 27, 2024

Performance Assessment for D209: Data Mining I – Task 1

This document contains the tasks and outputs required for the “OFM4 TASK 2: DIMENSIONALITY REDUCTION METHODS” assessment. All work is original to the author unless otherwise indicated by a citation.

# A - Research Question and Goal

This study aims to answer the question: how can the number of features in the data set be reduced while still retaining most of the variance of the given data set? A key goal is to improve the outcomes of downstream studies that use this data by reducing the chances of overfitting and reducing the computational load imposed by excess dimensionality. (Dey, 2023)

# B – Method Justification

This study aims to reduce the number of dimensions in the dataset. This will be accomplished by using Principal Component Analysis (PCA). While some dimensionality reduction techniques seek to eliminate dimensions from the data set, PCA extracts new composite features that combine varying subsets of the original features. In this way, much of the variation within the data set may be captured by fewer dimensions. (Larose, 2015)

PCA is a “scale variant” model. This means that variables with comparatively large scales will overpower variables with smaller ones. Therefore, PCA assumes all input variables have been scaled and/or normalized. (Jain, 2021)

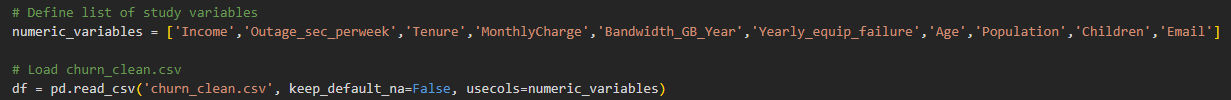
# C – Data Preparation

This table details the variables included in our dimensionality reduction study.

| Variable Name | Data Type |
| --- | --- |
| Population | Numeric |
| Children | Numeric |
| Age | Numeric |
| Income | Numeric |
| Outage\_sec\_perweek | Numeric |
| Email | Numeric |
| Yearly\_equip\_failure | Numeric |
| Tenure | Numeric |
| MonthlyCharge | Numeric |
| Bandwidth\_GB\_Year | Numeric |

Table 1 - Study Variables

These sections detail the data preparation steps performed for the study.



The desired columns of the data set were loaded into a data frame.

A screenshot of a computer program

Description automatically generated

No duplicate observations were noted.

A screenshot of a computer

Description automatically generated

No variables exhibit missing data.

A screenshot of a computer

Description automatically generated

A small number of potential outliers were identified. However, none of the values is clearly out of the range of plausibility. Therefore, all values will be retained as-is.

A screen shot of a computer program

Description automatically generated

Population :

count 10000.000000

mean 9756.562400

std 14432.698671

min 0.000000

25% 738.000000

50% 2910.500000

75% 13168.000000

max 111850.000000

Name: Population, dtype: float64

Children :

count 10000.0000

mean 2.0877

std 2.1472

min 0.0000

25% 0.0000

50% 1.0000

75% 3.0000

max 10.0000

Name: Children, dtype: float64

Age :

count 10000.000000

mean 53.078400

std 20.698882

min 18.000000

25% 35.000000

50% 53.000000

75% 71.000000

max 89.000000

Name: Age, dtype: float64

Income :

count 10000.000000

mean 39806.926771

std 28199.916702

min 348.670000

25% 19224.717500

50% 33170.605000

75% 53246.170000

max 258900.700000

Name: Income, dtype: float64

Outage\_sec\_perweek :

count 10000.000000

mean 10.001848

std 2.976019

min 0.099747

25% 8.018214

50% 10.018560

75% 11.969485

max 21.207230

Name: Outage\_sec\_perweek, dtype: float64

Email :

count 10000.000000

mean 12.016000

std 3.025898

min 1.000000

25% 10.000000

50% 12.000000

75% 14.000000

max 23.000000

Name: Email, dtype: float64

Yearly\_equip\_failure :

count 10000.000000

mean 0.398000

std 0.635953

min 0.000000

25% 0.000000

50% 0.000000

75% 1.000000

max 6.000000

Name: Yearly\_equip\_failure, dtype: float64

Tenure :

count 10000.000000

mean 34.526188

std 26.443063

min 1.000259

25% 7.917694

50% 35.430507

75% 61.479795

max 71.999280

Name: Tenure, dtype: float64

MonthlyCharge :

count 10000.000000

mean 172.624816

std 42.943094

min 79.978860

25% 139.979239

50% 167.484700

75% 200.734725

max 290.160419

Name: MonthlyCharge, dtype: float64

Bandwidth\_GB\_Year :

count 10000.000000

mean 3392.341550

std 2185.294852

min 155.506715

25% 1236.470827

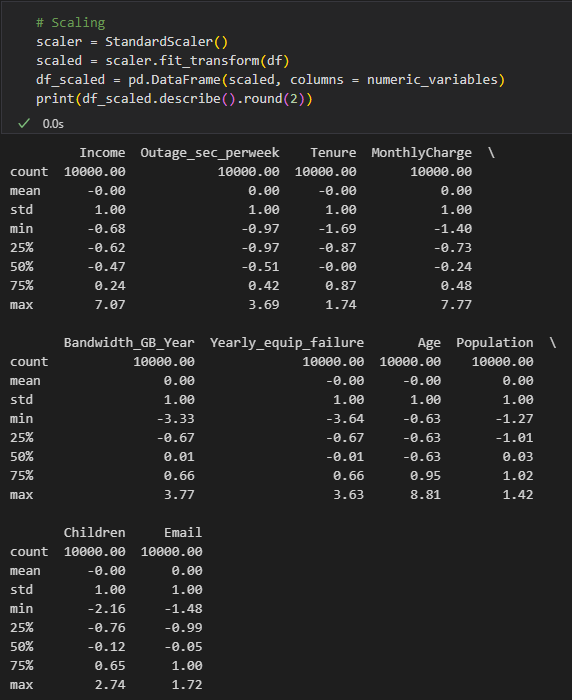
50% 3279.536903

75% 5586.141370

max 7158.981530

Name: Bandwidth\_GB\_Year, dtype: float64

Substantial deviation in scale and variance are noted.



Data set scaled with StandardScaler.



The prepared data set will be uploaded with this report. See file *churn\_prepared.csv*.

# D – Analysis

This section will outline the steps of our analysis as well as a summary of the results.

A screenshot of a computer

Description automatically generated

The model was initially fit using ten components. The loadings of each variable in each component are displayed.

A computer screen shot of a program

Description automatically generated

A graph with a line in the middle

Description automatically generated

A scree plot of the model’s eigenvalues is shown above. The horizontal line represents the threshold specified by the Kaiser rule. This rule recommends that principal components with an eigenvalue less than 1.0 be discarded. (Soriano & Kebabci, n.d.)

A screenshot of a computer program

Description automatically generated

We selected the number of components with eigenvalues greater than or equal to one and refitted the model. Here are the final loadings.

A screen shot of a computer program

Description automatically generated

In this final step, each component’s percentage of variance is shown along with the total. This means that instead of using ten features to explain 100% of the variance in the data set, just five principal components explain nearly 61% of the variance. This result is somewhat disappointing. Statistically, five variables out of ten could explain 50% of a data set’s variance. In this case, our results are about 20% better than this. While some improvement is better than no improvement, it might be necessary to identify additional numeric features that could be added to the PCA study to achieve better results.

# E – Third-Party Code References

Bathelt, J. (2017, 02 01). *Pandas - Compute z-score for all columns*. Retrieved from StackOverflow: https://stackoverflow.com/questions/24761998/pandas-compute-z-score-for-all-columns

Soriano, P. V., & Kebabci, C. (n.d.). *Scree Plot of PCA in Python*. Retrieved from Statistics Globe: https://statisticsglobe.com/scree-plot-pca-python

# F – Referenced Works

Dey, R. (2023, 10 06). *Understanding Principal Component Analysis (PCA)*. Retrieved from Medium: https://medium.com/@roshmitadey/understanding-principal-component-analysis-pca-d4bb40e12d33

Jain, S. (2021, 05 15). *Limitations, Assumptions Watch-Outs of Principal Component Analysis*. Retrieved from Medium: https://codatalicious.medium.com/limitations-assumptions-watch-outs-of-principal-component-analysis-8483ceaa2800

Larose, D. T. (2015). *Data Mining and Predictive Analytics.* John Wiley & Sons. Retrieved 09 27, 2024, from https://e`bookcentral.proquest.com/lib/westerngovernors-ebooks/detail.action?docID=7104155

Soriano, P. V., & Kebabci, C. (n.d.). *Scree Plot of PCA in Python*. Retrieved from Statistics Globe: https://statisticsglobe.com/scree-plot-pca-python