­

**Data Mining II, Task II: Churn Data**

Stephen E. Porter

College of Information Technology, Western Governors University

Dr. Keiona Middleton

September 24, 2022

**Data Mining II, Task II: Churn Data**

In this paper, I will use the provided data set containing cleaned customer data from a fictional telecommunications company. I will use principal component to reduce the dimensionality of the data set to better understand the characteristics of our customers. The goal is to determine what characteristics (if any) might indicate if a customer is at-risk of ending their service with our company.

# Part I: Research Question

## A1. Proposal of Question

The key question I would like to answer is can we determine which characteristics are essential to describing our customers (and, therefore, which can we safely ignore)?

## A2. Defined Goal

I will perform a Principal Component Analysis (PCA) to determine which of the numeric variables capture the greatest variance in the data. The other variables can then be dropped to reduce dimensionality.

# Part II: Method Justification

## B1. Explanation of PCA

As dimensionality increases, so does the challenge of visualizing trends in these data. With two or three dimensions, we can at least create a graphical representation of the data points. What Principal Component Analysis provides is a method for understanding the overall trends of the separate variables. These trends are groupings of related variables (or “components”) that will then serve as a general value for the group.

Using the the *mtcars* data set (a public, built-in data set in R) as an example, the *cyl* (number of cylinders), *disp* (displacement; combined volume of the engine’s cylinders), and *wt* (weight) are all highly correlated. The heavier a vehicle is, the more cylinders an engine needs to move it safely and the greater displacement value (more cylinders = more volume they can displace). (Hayden, 2018) These three could be reduced to a single Principal Component (PC) thus reducing the dimensionality of the dataset overall. This also reduced the undesirable effect of overfitting allowing for the trends observed to be more widely applicable.

To relate back to the *churn* data set used in this analysis, *Bandwidth\_GB\_year* and *Tenure* are highly correlated (although the explanation for why is not exactly apparent). However, that means a single Principal Component could serve as a substitute for the values.

The actual process is more technical than what is described above. We use eigenvectors and eigen values to measure the direction and variance, respectively, in the data. PCA is a linear combination of variables where the greatest variance in the data is captured by the first PC and fit to a new coordinate system where the first PC serves as the “x-axis.” Each PC is orthogonal (perpendicular) to the previous until all variance is captured. The greater the eigenvalue, the greater the variance of the PC. (Hayden, 2018)

## B2. PCA Assumption

There are several assumptions made when performing a PCA. The main one for this scenario is that the relationships between variables are linear as the process involves linear combinations of variables in the data set. Other assumptions include multiple continuous variables, the data set contains independent data that is sufficiently large, the data set has correlations that allow for reduction, and no significant outliers. (Laerd Statistics)

# Part III: Data Preparation

## C1. Continuous Data Set Variables

The continuous variables from the data set are listed below. All numeric variables are used to search for the most complete view of our customers as possible.

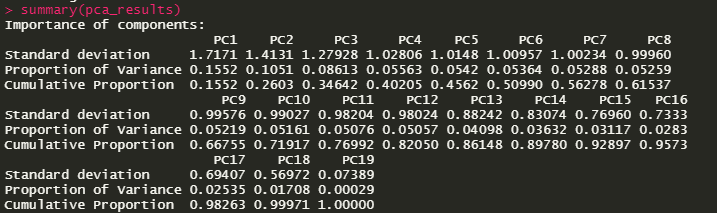
|  |
| --- |
| **Variable** |
| Children |
| Age |
| Income |
| Outage\_sec\_perweek |
| Email |
| Contacts |
| Yearly\_equip\_failure |
| Tenure |
| MonthlyCharge |
| Bandwidth\_GB\_Year |
| Response |
| Fix |
| Replacement |
| Reliability |
| Options |
| Respectful |
| Courteous |
| Listening |

## C2. Standardization of Dataset Variables

The data set was standardized using the *scale()* function in R. Please see the attached csv file for the scaled data output.

# Part IV: Analysis

## D1. Principal Components

Below is a summary of the environment variable *pca\_results* which holds the results of the *prcomp()* function. The summary output provides a readable list of PC1 through PC19 with the Standard deviation, Proportion of Variance (for that individual PC), and Cumulative Proportion.

***Figure 1: Summary of PCA Results***

For individual values for each variable in the PCs, please see the code output below:

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Text

Description automatically generated

***Figure 2: Individual PCA Results***

## D2. Identification of Total Number of Components

There are 19 total Principal Components that capture 100% of the variance. There are several methods to determine how many PCs are necessary to optimize the capture of data without significantly increasing dimensionality. The Scree Plot below can help identify the number of PCs worth retaining. Based on the Scree Plot below, at PC4 there is a clear “elbow” where the next series of PCs remain nearly even in percentage of variance; however, from PC12 to PC13 there is a notable drop where less variance would be captured. Arguments can be made for a few PC cutoff points:

* PC4 – the first and clearest “elbow” in the plot
* PC7 – the final PC with an eigenvalue > 1
* PC12 – the point at which cumulative variance passes 80%

(Mangale, 2020)

Because the goal is to capture variance while reducing dimensionality, PC12 appears to best accomplish both goals without significantly sacrificing one. The eigenvalues of PC8 through PC12 is still relatively close to 1, but PC13 drops to 0.77 (see Figure 4 for the full list). The cumulative variance captured by PC4 is only 40% and at PC7 is 56%.

Chart, histogram

Description automatically generated

***Figure 3: Scree Plot (% of Variance)***

Text

Description automatically generated

***Figure 4: Eigenvalues***

## D3. Total Variance of Components

The variance of each component is shown in the image below in the column labeled “variance.percent.”

Text

Description automatically generated

***Figure 5: get\_eignvalue of pca\_results for PC1-PC12***

## D4. Total Variance Captured by Components

The total variance captured by the four components can be seen in the bottom right corner of the table above – cumulative.variance.percent (at Dim. 12). This value is 82%.

## D5. Summary of Data Analysis

The PCA process has identified 19 total PCs to mirror the original 19 continuous variables. By using the Scree Plot to determine the point at which increasing the number of PCs captures an inconsequential amount of variance, we have effectively reduced dimensionality while capturing 82% of the variance in the model.

While there are multiple interpretations of the number of PCs to keep and arguments can be made for others (as described in Section D2), I believe that PC12 is the appropriate cutoff point. Regardless of the argument, it is safe to say, that after PC12 there is little to be gained by increasing the number of principal components used.

The Variables – PCA visualization in Figure 6 provides some clarity on which variables are highly correlated and contribute to the first two Principal Components (as it’s difficult to visualize 12 at once!). Notice that Dim.1 (or PC1) is along the x-axis and Dim.2 (or PC2) is along the y-axis.

In a review of Figure 2, six of the eight questions from the survey (Response, Fix Replacement, Respectful, Listening, and Courteous) all have a relatively high value. These six values can be summarized by PC1 and can be seen in the circle visualization in Figure 6 as a grouping of arrows pointing to the right. The higher the color on the scale, the greater the contribution to the PC. In reviewing PC2 in Figure 2, both Bandwith\_GB\_year and Tenure have values of around -0.7. Therefore, these values can be summarized by PC2 and are colored orange/red in the visualization in Figure 6 showing they are significant contributors to their PC.

Another useful conclusion that can be drawn from the circle graph in Figure 6 is that Reliability (one of the survey questions) and Options (the eighth and final survey question) are negatively correlated because their arrows point in opposite directions.

Chart

Description automatically generated

***Figure 6: Variables – PCA***

# Part V: Attachments

## E. Sources for Third-Party Code

No third-party code was used in this script. Color selection for some visualizations was based on Kassambara’s work, though the functions themselves are documented in the *factoextra* package and not strictly their own work.

## F. Sources

**References**

Hayden, L. (2018, August 9). *R PCA tutorial (principal component analysis)*. DataCamp. Retrieved September 24, 2022, from https://www.datacamp.com/tutorial/pca-analysis-r

Kassambara. (2017, September 23). *PCA - Principal Component Analysis Essentials*. STHDA. Retrieved September 24, 2022, from http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/#eigenvalues-variances

Laerd Statistics. (n.d.). *Principal components analysis (PCA) using SPSS statistics*. How to perform a principal components analysis (PCA) in SPSS Statistics | Laerd Statistics. Retrieved September 24, 2022, from https://statistics.laerd.com/spss-tutorials/principal-components-analysis-pca-using-spss-statistics.php

Larose, C. D., & Larose, D. T. (2019). *Data Science using python and R*. Wiley.

Mangale, S. (2020, August 28). *Scree plot*. Medium. Retrieved September 24, 2022, from https://sanchitamangale12.medium.com/scree-plot-733ed72c8608