PCA Project

Jacob Polomsky

July 27, 2024

A1. Can We Reduce the number of columns while retaining above 80% of the variance in those models?

A2. The goal of the analysis is to see how many columns we need to keep 80% of the variance using PCA.

B1. PCA analysis uses linear combinations or multi-dimensional datasets to explain the variance in the dataset as much as possible. PCA can help reduce the number of columns needed by looking at the variance that each PCA column provides and looking at where the PCA variance has a minimal explanation for the variance

B2. One assumption of PCA is that the data is standardized, or the data points are compared to the mean and the standard deviation. This allows the data with high variance and low variance to have equal weight for the model.

C1. All of the variables used in this dataset are continuous. The variables used are Population, Children, Age, Income, Outages\_sec\_perweek, email, contacts, yearly\_equip\_failure, tenure, monthly\_charge, and bandwidth\_gb\_year.

C2. scaled\_data.csv, the Code I used to standardize the data

A screen shot of a computer

Description automatically generated

D1.

D2. Using the elbow Kaiser method, you can see two main drops at PC2 and PCA 10. Either of those would work, but you have to keep in mind that PC2 explains 27% of the variance, and PC10 explains 99% of the variance with a total of 11 PCA rows in total

A line graph with numbers and a line

Description automatically generated

D3. PCA 1 can explain 18% of the variance in the data set from 2 to 10. The variance stays between 9% and 8% in the variation in the data, and the final PCA variable explains almost no variance in the data.

A screenshot of a black screen

Description automatically generated

D4. PCA 1 explains 18% of the variance in the data; looking at the last question, we see from 2 to 10, there is an increase of 9% to 8%, so 2 PCA 27%, 3 PCA 37%, 4 PCA 46%,5 PCA 55%, 6 PCA 64%, 7 PCA 73%, 8 PCA 82%, 9 PCA 91%,10 PCA 99%. PCA 11 explains 100 percent of the data

A screenshot of a black screen

Description automatically generated

D5. So, with this model, it would be best to keep 10 PCA variables to maintain the most variance while getting rid of a row. This doesn’t answer the question we set out to ask. If we wanted to keep 80% of the variance without the dataset, we should keep 8 PCA variables to reach our goal of 80% of the variance in the data.