Anthony Marcolongo

D212

2/3/2021

PCA Analysis

The goal of this project was to determine if PCA could be used to reduce the number of factors in a dataset containing continuous and categorical variables. The dataset used for this project was made up of 50 factors and 10,000 observations. The dataset was prepared for PCA. This involved removing unnecessary factors and then scaling the remaining data. This left the dataset with nine factors from the original fifty. Once the data was scaled a covariance matrix could be created. The values of the covariance matrix are the eigen values of the original nine factors. After acquiring the eigen values a Scree plot was created. The Scree plot showed that the eigen values dropped below one once the number of components went above seven.

The next thing to do was to create the PCA model using the prcomp function in R. The output of the prcomp function is the standard deviation for the PCA’s, as well as the individual feature values for the PCA’s. Applying the summary function to the output of the prcomp function shows the proportion of the variance for each PCA and the cumulative variance. This shows that 97% of the variance is contained in the first seven PCAs. From this it is clear that the original 9 features are more than are needed for this analysis. Seven components will suffice for the purpose of this project.

Looking at the values contained in the first few PCAs it is clear that the three most important factors in determining hospital readmission are Age, Total Charge, and Initial Days. The value of Age is negative which shows that older patients are less likely to be readmitted. On the surface this seems odd, but when you think about it, it does make sense. Older patients tend to get their health insurance from the government rather than their employer. This insurance is much less likely to approve readmittance then a better policy would be. The next factor is Total Charge, this is the amount the patient was charged during their initial visit. Higher values of this imply patients that are suffering from more serious medical conditions. These patients are much more at risk for readmittance due to their medical conditions lasting past their first visit. The last factor is Income. Patients with higher income are significantly more likely to have better health insurance. This better insurance is more willing to approve additional visits than a worse insurance plan would be.

My recommendation would be for the hospital to target at risk patients during their initial visit. If whatever condition was causing people to come to the hospital initially could be taken care of, then patients would not need to be readmitted. The hospital would need to work with insurance companies to increase patients initial stay so that issues could be fully treated, and patients would not need to come back a second time for the same thing.