**Part I: Research Question**

*Research Question:* Which variables or characteristics are the most important to consider when predicting patient success, status, or potential to be readmitted?

The goal of this data analysis is to learn the most significant patient features in our dataset. We can isolate these features and further analyze their correlation with readmission status or other aspects of the patient’s medical situation. If we know which features are correlated with readmissions, for example, we can identify patients at higher risk, adjust our care and plan accordingly. This will allow us to be more prepared, avoid penalties for readmissions, and hopefully, provide better overall healthcare to patients.

**Part II: Method Justification**

Principal Component Analysis, or PCA, is a dimension reduction technique “that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set” (Jaadi, 2021). It first “decorrelates” the data by shifting and rotating the datapoints so they are aligned with the coordinate axes and have means of zero, and then reduces the dimension of the data (*Visualizing the PCA transformation: Python*). This can make large datasets, including the medical dataset that is part of this course, easier to explore, visualize, and analyze.

PCA operates by making some fundamental assumptions. One key assumption is that there are multiple continuous (and not categorical) variables, all of which have linear relationships with one another (*Principal components analysis (PCA) using SPSS statistics*).

**Part III: Data Preparation**

Because PCA is only applicable with continuous variables, I will use the ones listed below, which constitute all continuous variables in the dataset:

*Lat:* GPS coordinate (Latitude) of patient residence as listed on the billing statement.

*Lng:* GPS coordinate (Longitude) of patient residence as listed on the billing statement.

*Population*: Population within a mile radius of patient, based on census data

*Age*: Age of the patient as reported in admissions information.

*Children*: Number of children in the patient’s household as reported in the admissions information (might not be children of the patient).

*Income*: Annual income of the patient (or primary insurance holder) as reported at time of admission.

*VitD\_levels*: The patient’s vitamin D levels as measured in ng/mL.

*Doc\_visits*: Number of times the primary physician visited the patient during the initial hospitalization.

*Full\_meals\_eaten*: Number of full meals the patient ate while hospitalized (partial meals count as 0, and some patients had more than three meals in a day if requested).

*vitD\_supp*: The number of times that vitamin D supplements were administered to the patient.

*Initial\_days*: The number of days the patient stayed in the hospital during the initial visit.

*TotalCharge*: The amount charged to the patient daily. This value reflects an average per patient based on the total charge divided by the number of days hospitalized. This amount reflects the typical charges billed to patients, not including specialized treatments.

*Additional\_charges*: The average amount charged to the patient for miscellaneous procedures, treatments, medicines, anesthesiology, etc.

Because the ranges of our variables will largely vary, we want to standardize this data so that all features are on similar scales. To do this, we will use the StandardScaler and Pipeline packages in sklearn. A copy of this scaled data is attached separately.

**Part IV: Analysis**

Using the below scree plot and the elbow rule, we can see that there appear to be three (3) principal components in this dataset. This is evident in the fact that after the third component, the variance does not decrease at the same rate.

Chart, bar chart, histogram

Description automatically generated

Below are principal component values of the first five and last five rows (patients) of our dataset.

|  | **PC1** | **PC2** | **PC3** | **ReAdmis** |
| --- | --- | --- | --- | --- |
| **0** | -1.136196 | 0.619518 | 0.642796 | No |
| **1** | -0.778151 | 0.565780 | 1.150359 | No |
| **2** | -1.633870 | 0.663045 | -0.487327 | No |
| **3** | -1.866283 | 1.025758 | -0.969875 | No |
| **4** | -2.081208 | -1.792136 | 0.334074 | No |
| **...** | ... | ... | ... | ... |
| **9995** | 0.775728 | -1.477410 | 0.204750 | No |
| **9996** | 2.052583 | 2.620933 | -0.181685 | Yes |
| **9997** | 1.853719 | -0.257625 | -0.146843 | Yes |
| **9998** | 1.403092 | -0.999949 | -1.513676 | Yes |
| **9999** | 2.017897 | 0.167547 | 1.342103 | Yes |

The principal components and their eigenvalues are in the following table. The highlighted numbers tell us which variables the corresponding principal component places the most weight on. PC1 is more strongly influenced by TotalCharge and Initial\_days, PC2 is more influenced by Age and Additional\_charges, and PC3 is most influenced by Lat, Lng, and Population. It is these variables that are most influential to the dataset, and therefore should be considered when predicting patient behavior.

| **PC1** | | **PC2** | **PC3** |
| --- | --- | --- | --- |
| **Lat** | -0.018834 | 0.000913 | -0.715570 |
| **Lng** | -0.011011 | 0.009716 | 0.274895 |
| **Population** | 0.028719 | -0.029027 | 0.626046 |
| **Children** | 0.034537 | 0.017244 | -0.034510 |
| **Age** | 0.084650 | 0.700793 | 0.011244 |
| **Income** | -0.019701 | -0.019176 | 0.075776 |
| **VitD\_levels** | -0.001995 | 0.020340 | -0.020176 |
| **Doc\_visits** | -0.006991 | 0.015446 | 0.017291 |
| **Full\_meals\_eaten** | -0.020712 | 0.031960 | -0.103248 |
| **vitD\_supp** | 0.025381 | 0.014511 | 0.029741 |
| **Initial\_days** | 0.699994 | -0.089859 | -0.022902 |
| **TotalCharge** | 0.701146 | -0.079267 | -0.020888 |
| **Additional\_charges** | 0.085029 | 0.700745 | 0.013730 |

The variances of each of the three principal components are below:

[**PCA1:** 0.15349968 **PCA2:** 0.13194741 **PCA3:** 0.09438151]

The total variance explained by these three components is 0.3798286, or about 37.98%.

These results imply that about 38% of the information in our data can be explained by the three principal components. Thus, 62% of the information in our data was lost in this principal component analysis. At the same time, the 3-dimensional scatterplot below showing readmitted (orange dots) and non-readmitted (blue dots) patients shows us that these three components do a rather good job at predicting readmission status. Our organization can benefit from this information by using the features highlighted in the matrix above to track and correlate with readmissions.

Chart, scatter chart

Description automatically generated

**References**

Jaadi, Z. (2021, April 1). *A step-by-step explanation of principal component analysis (PCA)*. Built In. Retrieved December 10, 2021, from https://builtin.com/data-science/step-step-explanation-principal-component-analysis.

*Principal components analysis (PCA) using SPSS statistics*. How to perform a principal components analysis (PCA) in SPSS Statistics | Laerd Statistics. (n.d.). Retrieved December 10, 2021, from https://statistics.laerd.com/spss-tutorials/principal-components-analysis-pca-using-spss-statistics.php.

*Visualizing the PCA transformation: Python*. Datacamp. (n.d.). Retrieved December 10, 2021, from https://campus.datacamp.com/courses/unsupervised-learning-in-python/decorrelating-your-data-and-dimension-reduction?ex=1.