Performance Assessment 2 – Revision 4

D212 – Data Mining II

Jessica Hosey

MSDA, College of Information Technology

Western Governors University

November 16th, 2024

Part I: Research Question

A. 1. Propose one question relevant to a real-world organizational situation that you will answer using PCA.

There are several reasons why patients are readmitted to the hospital. The research inquiry is about finding patient characteristics that affect company costs associated with patient readmission. So, the research question for this project is, what patient characteristics correlate to readmission using a decision tree after using PCA to reduce the dataset?

A. 2. Define one goal of the data analysis. Ensure your goal is reasonable within the scope of the selected scenario and is represented in the available data.

The analysis aims to predict the characteristics of patients that affect company costs. Hospitals are directly involved financially when patients are readmitted. This will allow the hospital to understand better their patient's needs and how to provide specific care to individuals with these critical characteristics to lower the cost of care in the long term.

Part II: Method Justification

B. 1. Explain how PCA analyzes the selected data set. Include expected outcomes.

The tool we are using is Principal component analysis (PCA). PCA identifies Principal Components (variables) that have the most variance between them without knowing anything about these variables (GeeksforGeeks 2024). The first step in using PCA is that the variables need to be standardized, standardization ensures that the features (variables) "have a mean of zero and a standard deviation of one" (James et al. 2013). Standardization is necessary to ensure that one value in the dataset is not weighted differently from another value. After standardization, we use a covariance matrix to identify features with similar covariance values (identical in linear terms). These variables need to be removed to ensure we do not have variables identical to one another. PCA then uses eigenvalues and eigenvectors to select the best features with the highest variance (known as the principal components). We then look at these eigenvalues to decide how many principal components we want to select. These chosen PCs need to explain around 95% of the variance in the dataset. Then, the original data is projected onto the selected PCs to create a new list of statistically different features. PCA reduces the dataset by selecting the variables (PCs) with higher variance and removes those that are insignificant in terms of variance. We will use these characteristics in our analysis to find which patient attributes will help the hospital lower readmission fines. The outcome of PCA will be a reduced dimensionality dataset containing all the information the original dataset had but with the variables with most characteristic variance and connections within the dataset (GeeksforGeeks 2024).

B. 2. Summarize one assumption of PCA.

One assumption of Principal component analysis (PCA) is that the sample of data is large enough. Typically, the bigger the dataset, the more reliable the results are; therefore, a larger dataset will benefit PCA in effectively selecting principal components most correlated with patient readmission. Datasets that are not large enough will make PCA results unreliable and invalid.

Part III: Data Preparation

C. 1. Identify the continuous data set variables needed to answer the PCA question.

Column Name	Description of Continuous Variable
Lat	Latitude of Patient
Lng	Longitude of Patient
Age	Age of Patient
Income	Patient's Income
VitD_levels	Patient's Vitamin D Levels
TotalCharge	The total charge for a hospital visit
Additional_charges	Additional Charges for the hospital visit

C. 2. Standardize the continuous data set variables identified in part C1. Include a copy of the cleaned data set.

See the attached files in my submission.

Part IV: Analysis

D. 1. Determine the matrix of all the principal components.

Loading Matrix

	PC1	PC2	PC3	PC4	PC5	PC6
Lat	0.707758	0.043688	-0.011824	-0.084460	0.099423	-0.692829
Lng	-0.698786	-0.023200	-0.134689	-0.128656	0.093751	-0.683869
Income	-0.085710	0.436801	0.473664	0.729930	-0.119732	-0.174262
VitD_levels	0.058327	-0.246033	-0.732909	0.576021	-0.254160	-0.050115
TotalCharge	-0.003639	-0.609034	0.409425	-0.043671	-0.663482	-0.138997
Additional_charges	-0.003983	-0.612620	0.229296	0.331374	0.679842	0.010550

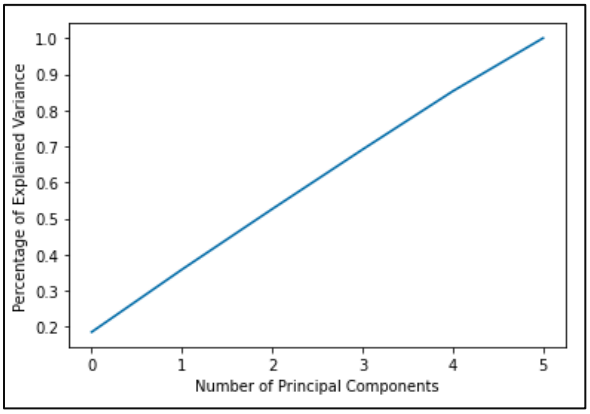
Correlation Matrix

	Lat	Lng	Age	Income	VitD_levels	TotalCharge	Additional_charges
Lat	1.000100	-0.112359	-0.007270	-0.019371	0.001494	-0.010760	-0.002283
Lng	-0.112359	1.000100	0.007494	-0.006666	-0.006390	-0.008831	0.000079
Age	-0.007270	0.007494	1.000100	-0.012229	0.010316	0.016877	0.716925
Income	-0.019371	-0.006666	-0.012229	1.000100	-0.013116	-0.014347	-0.009826
VitD_levels	0.001494	-0.006390	0.010316	-0.013116	1.000100	-0.001403	0.008291
TotalCharge	-0.010760	-0.008831	0.016877	-0.014347	-0.001403	1.000100	0.029259
Additional_charges	-0.002283	0.000079	0.716925	-0.009826	0.008291	0.029259	1.000100

The covariance matrix outputs two nearly identical variables; therefore, I chose to reduce the dataset and remove the Age variable, keeping the Additional_charges data.

D. 2. Identify the total number of principal components using the elbow rule or the Kaiser criterion. Include a screenshot of the scree plot.

Using the Elbow Rule, this PCA prediction has five principal components (see screenshot below). In addition to the PC loadings, a scree plot is provided below.



The scree plot identifies 5 (the 6th one is not shown on the graph; see explanation below) PCs that show 100% of the variability in this dataset. As we add each additional PC, we gain more information about the variability of this dataset. Using fewer components might be sufficient if less precision is acceptable. PC6 is not even on the OBscree plot; this indicates that PC6 does not add meaningful information to the model in terms of explained variance. This alone justifies the removal of PC6 and reduces the dataset. To see if reducing to 3 or 4 PCs, we tested whether it would affect the model's variance, and it was not significant compared to a model with 5 PCs. Therefore, the final PCA will have 5 Principal components. This still aligns with the elbow rule and will explain 85% of the dataset's variance (as seen below in the following parts). PC6 is not even on the scree plot, this indicates that PC6 does not add meaningful information to the model in terms of explained variance. This alone justifies the removal of PC6 and reduces the dataset. To see if reducing to 3 or 4 PCs, we tested whether it would affect the model's variance, and it was not significant compared to a model with 5 PCs. Therefore, the final PCA will have 5 Principal components. This still aligns with the elbow rule and will explain 85% of the dataset's variance (as seen below in the following parts).

D. 3. Identify the variance of each principal component in part D2.

Before reducing from 6 principal components:

```
The contribution of each principal component to the total can be seen here:
For PC1, the contribution is 18.558%
For PC2, the contribution is 17.326%
For PC3, the contribution is 16.759%
For PC4, the contribution is 16.491%
For PC5, the contribution is 16.163%
For PC6, the contribution is 14.704%
```

After reducing to 5 principal components:

```
The amount of variance for by each principal component can be seen here:
For PC1, the contribution is 18.558%
For PC2, the contribution is 17.326%
For PC3, the contribution is 16.759%
For PC4, the contribution is 16.491%
For PC5, the contribution is 16.163%
```

D. 4. Identify the total variance captured by the principal components in part D2.

After reducing the dataset to 5 PCs:

```
These 5 principal components explain 85.296% of variance in the data.
```

D. 5. Summarize the results of your data analysis.

Reminder that PCA is used to reduce the dimensionality of a dataset and to find variables (PCs) with the most variance in the dataset. Our analysis reduced the dataset to 5 PCs, which explained 85.296% of the variance in the dataset. As seen in D3, the second screenshot shows the variance for each of our PCs in the reduced dataset. PC1 explains the 18.56% variance in the dataset. PC2 explains the 17.33% variance in the dataset. PC3 explains the 16.76% variance in the dataset. PC4 explains the 16.49% variance in the dataset. PC5 explains the 16.16% variance in the dataset. PC 3, 4, and 5 have similar variance percentages.

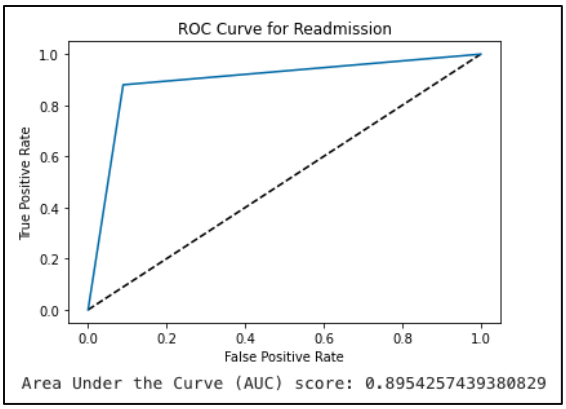
Looking at the correlation coefficients from D1, age positively correlates with vitamin D levels and total charge. Another positive correlation was found between total charges and additional charges. Several negatively correlated variables were in the dataset, indicating that these variables have an indirect relationship. Latitude and Longitude had the strongest negative correlation in the dataset. The total charge had many negatively correlated variables, such as Latitude and Income. A few examples of variables with little to no correlation to one another are as follows: Longitude and additional charges (0.000079), Latitude and Vitamin D levels (0.001494), Vitamin D Levels and Total Charge (-0.001403), and Income and Additional Charges (-0.009826).

Now, we will focus on the loading matrix values from D1, which is the first screenshot. For PC1, Latitude has a high favorable loading of 0.707758, which indicates that it contributes to the variance in PC1. Also, Longitude has a high negative loading of –0.698786, which suggests that it is inversely related to PC1. For PC2, Income and Longitude have positive loadings that positively contribute to the PC2. Additional and Total Charges have strong negative loadings, indicating they are negatively associated with PC2. For PC3, Vitamin D levels had a strong negative loading of –0.732909, which significantly negatively influenced PC3. Income had a favorable loading of 0.473664, which indicates that it positively influences PC3. For PC4, income influences PC4 and majorly contributes to PC4 with a loading value of 0.729930. Vitamin D Levels also had a favorable loading of 0.576021 and contributed positively to PC4. For PC 5 and PC6, there is less variance with these components, indicating that they are more complex in interpreting their influence over PC5 and PC6. Total charge and additional charges had high positive loadings in PC5, and Latitude and Longitude had strong negative loadings in PC6.

```
Decision tree accuracy: 0.8995

The confusion matrix for this Decision Tree model:
Predicted No Readmission  | Predicted Readmission
                    [1153  113] Actual No Readmission
                     [ 88 646] Actual Readmission
```

After completing the PCA analysis and reducing the number of PCs used in the prediction model, we have around 89.95% accuracy for the decision tree predictions (see screenshot to the left).



The ROC Curve for Readmission shows that the models' predictions are above 50% random prediction value (dotted line). This is a good sign; however, a successful model would show a line curved to the upper left side of the graph, close to 100% rate. The curve above is close to the upper left corner but is much lower than the 100% rate. So, the model's accuracy is not great, and its predictions are untrustworthy. The Area Under the Curve (AUC) score is 89.54%, which is close to the perfect score of 100%; however, again, its success rate would be more beneficial with a 95-100 % range. , which is close to the perfect score of 100%; however, again, its success rate would be more beneficial with a 95-100 % range.

Overall, I believe this model is the correct direction for the company to continue gathering data and using it to test whether the prediction model is reliable. More data (more than one day of hospital operation) would create a larger sample size and make the prediction model more effective.

Part V: Attachments

E. Sources

Data to Fish. (2021, October 15). How to create a covariance matrix using Python. https://datatofish.com/covariance-matrix-python/

GeeksforGeeks. (2024, September 10). Principal component analysis (PCA). GeeksforGeeks. Retrieved October 26, 2024, from https://www.geeksforgeeks.org/principal-component-analysis-pca/

Hosey, Jessica. (2024). Performance Assessment 2, D208 – Predictive Modeling. [Unpublished manuscript]. WGU.

Hosey, Jessica. (2024). Performance Assessment 1, D209 – Data Mining I. [Unpublished manuscript]. WGU.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer.

Shlens, J. (2014). A tutorial on principal component analysis. arXiv Preprint. https://arxiv.org/abs/1404.1100