**Clustering of Medicare Physician and Other Supplier Data**

Charles Marshall

Prof. Klabjan

IEMS 308

27 January 2020

**Exploratory Data Analysis:**

After initially cleansing the data of outliers, Figure 1 was completed to show the distributions of the seven numerical features in the data set. None of the features are normally distributed. However, when a log transformation is applied in Figure 2, four of the features display an approximately normal distribution. These features are:

average\_Medicare\_allowed\_amt, average\_submitted\_chrg\_amt,

average\_Medicare\_payment\_amt, and average\_Medicare\_std\_amt.

A close up of a map

Description automatically generated A close up of a map

Description automatically generated

Figure 2: Distribution of Numerical Features in Medicare Provider Data Set after the application of a log transformation.

Figure 1: Distributions of Numerical Features in Medicare Provider Data Set.

In the analysis of categorical features, it was found that many would not be suitable for clustering. For instance, the provider credentials were interesting but required lots of effort to clean to gain any insight. Features such as participation and drug indicators were highly skewed in favor of participating in Medicare and the drugs not being used in the procedure.

The conclusion of exploratory data analysis lead to the forming of the following business question to answer during modeling:

*How do providers in certain regions of the United States differ in terms of unique patients and proportion of their allowed amount compared to the submitted charge for common outpatient services?*

The dataset needed to be made smaller in order to perform clustering. In order to achieve this, HCPCS codes between 99201 – 99215 which represent common outpatient care procedures, some of the most common treatments performed by Medicare providers, were This grouping guaranteed that there would be a significant amount of each procedure performed in each region. Zip codes were preferred to other methods such as states to split the US into regions. According to Figure 3, the American zip code system is already used to split the US into 10 regions based on the first digit of the zip code.

A close up of a map

Description automatically generated

Figure 3: Zip Code zones sorted by first digit (1)

**Modeling:**

Clustering was performed on two numerical variables: “Pay Ratio” and bene\_day\_srvc\_cnt, which represents the number of distinct Medicare beneficiaries per day. The Pay Ratio variable was created by dividing the log transform of the average\_Medicare\_allowed\_amt, representing how much on average was paid to a provider by Medicare, coinsurance, and other avenues, by the log transformation of the average\_submitted\_chrg\_amt, which was how much on average a provider charger for a procedure. These three numerical variables were chosen because of their relatively low correlation between each other. as displayed in Table 1.

Table 1: Correlation Between Pay Ratio and Unique Service Count variables

A screenshot of a cell phone

Description automatically generated

In order to select the number of clusters, a SCREE (Figure 4) and Silhouette Plot (Figure 5) were constructed on the training data.

A screenshot of a cell phone

Description automatically generated A close up of a mans face

Description automatically generated

*Figure 5: Silhouette Plot of Training Data for clustering on Pay Ratio and Unique Service Count variables.*

*Figure 4: SCREE Plot of Training Data for clustering on Pay Ratio and Unique Service Count variables.*

The SCREE Plot in Figure 4 displays a joint at k = 5. The Silhouette Plot in Figure 5 shows spikes at both k = 3 and k = 5. K-means clusters were completed and compared with k=3 in Figure 6 and k=5 clusters in Figure 7.

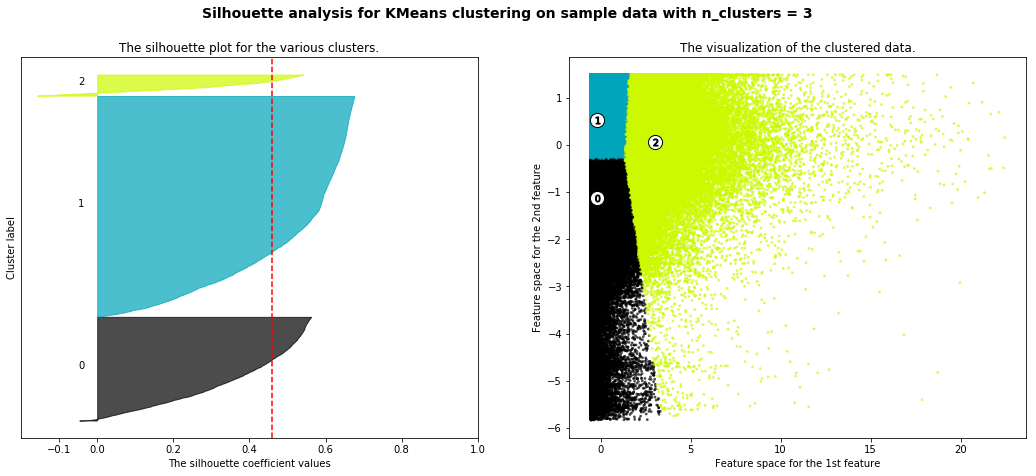


Figure 6: Visualization of Clustering Silhouette Plot and Physical Clusters for k = 3 clusters

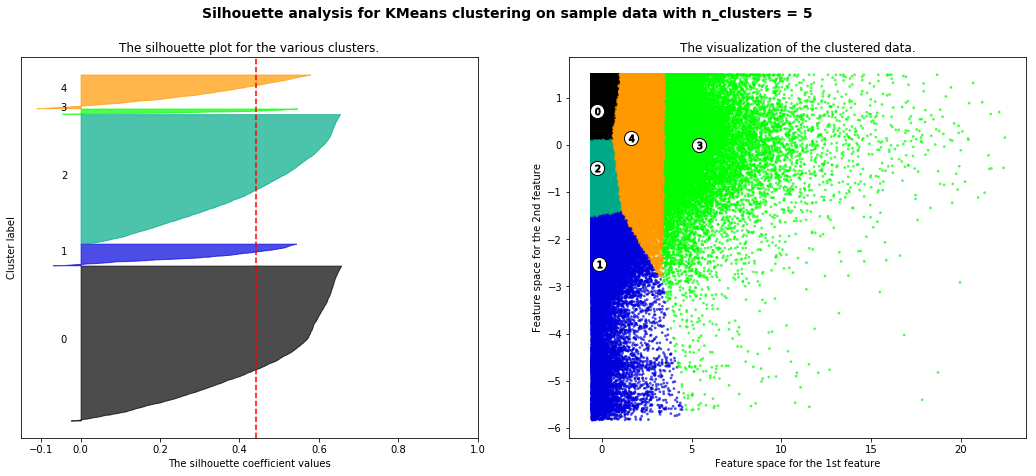


Figure 7: Visualization of Clustering Silhouette Plot and Physical Clusters for k = 5 clusters

From looking at the graphs, it is clear that k=3 clusters is best. Not only is the Silhouette score higher for 3 clusters, but the visualization of the clustered data shows a much more natural break in where clusters start.

**Insights:**

After clustering on the two variables, three distinct clusters were revealed in Table 2:

Table 2: Mean Values of Pay Ratio and Unique Service Count in each cluster

A screenshot of a cell phone

Description automatically generated

The breakdown of the clusters is as follows:

Cluster 0: Average service count, relatively low pay ratio.

Cluster 1: Average service count, relatively high pay ratio.

Cluster 2: Relatively high service count, average pay ratio.

The zip code region (0-9) was concatenated onto the data with the hope of being able to draw conclusions about where these clusters might occur in the United States. Table 3 shows the relative frequency of each postal region in each individual cluster.

Table 3: Relative Proportion of data points in individual clusters in each Postal Region

A screenshot of a cell phone

Description automatically generated

Analyzing Table 3, the mean amount tells us how often a certain region of the country falls within a specific cluster. The assumption is that the relative proportion in each cluster will be consistent, and therefore based on the population in each region. The sum of the relative proportions in each region should approximately add to 1.

In all regions, the relative proportions are very consistent between clusters 0 and 1 despite the difference in pay ratio. This leads to the conclusion that the difference in pay ratio is not due to the people in a certain region being more susceptible to need that treatment.

However, one valuable insight from Table 3 is the difference between the relative frequencies in clusters 0 and 1 compared to cluster 2. It seems that based on the relative frequencies, providers in zip regions 0, 1, 4, and 5 are less likely to fall into cluster 2. However, providers in zip regions 2, 3, and 7 are more likely to fall into cluster 2.

Table 4: Relative Proportion of data points in individual clusters for each HCPCS codeA screenshot of a cell phone

Description automatically generated

Looking at Table 4, we also conclude that the difference in pay ratio is not due to a difference in treatment type because the relative proportion of each treatment type is very similar across clusters. This leads to the conclusion that the difference in the clusters is probably due to a difference in provider which is not included in the data. For instance, I believe the relatively large difference for the unique service count between clusters 0 and 1 and cluster 2 is due to a difference in provider. For instance, I would suppose that certain providers such as pediatricians or walk-in clinics are more likely to provide these common outpatient services such as general checkups, and therefore would have a higher unique service count than more specialized providers. Combining our intuition from Tables 3 and 4, we might conclude that there is better access to these high-volume providers in zip regions 2, 3, and 7 than in other regions across the country.

Works Cited:

1. “U.S. ZIP Codes: Free ZIP Code Map and Zip Code Lookup.” *UnitedStatesZipCodes*, 2020, www.unitedstateszipcodes.org/.