Kathleen Young
IEMS 308, Professor Klabjan
25 January 2018

Medicare Through the Lens of K-Means Clustering

Executive Summary

The Medicare public file provided a plethora of information, including everything from the average numbers of patients needing treatments at a facility per day to the average amount Medicare covers for their patients' medical visits. The following report digs deeper into a vital concern of patients requiring medical services—the amount that he or she will have to pay out of pocket. As a Medicare recipient it is important to understand which types of services are thoroughly covered by Medicare, and which are not.

K-means clustering is an algorithm that illustrates similarity among points in a data set. The algorithm determines "similarness" by distance—in this case, Euclidean distance. K-means is a very strong tool for unsupervised machine learning. In this case, K-means clustering was used to detect patterns in what insurance pays for, and what it doesn't.

More specifically, the clustering model was trained on two features of the Medicare data—the average submitted charge amount for a service, as well as the average amount Medicare covered for that service. That is, the following report discusses the presence (or absence) of distinct groups of provided services that are cheap but well covered by insurance, or very expensive and poorly covered by insurance.

After analyzing the results, it is apparent that within the four most common medical services, none of them appear to be severely over or under covered. Rather, Medicare meets the patient in the middle, paying more for more expensive treatments.

Problem Statement

The Medicare database was explored in order to discover if any factors are common amongst different levels of treatment cost and Medicare coverage.

Assumptions

1. While digging into the data, it became clear that although there are 91 provider services, the top four most common alone accounted for 40% of the data. That is, in a dataset with nearly 10,000,000 data points, 4,000,000 of those were either Diagnostic Radiology, Internal Medicine, Family Practice, or Radiology. It is assumed that reducing the dataset by 60% is appropriate for the following analysis.
2. It is assumed that although one of the features in the clustering (average Medicare payment amount) included a note that in April of 2013 there was a 2% reduction in Medicare payment due to changes in Federal spending, the data Medicare payment amount was not jeopardized.
3. While performing analysis, it was intentionally chosen *not* to normalize or standardize the features (average Medicare payment amount and average submitted charge amount). Both of these features are in the same unit, dollars, and although the

Medicare payment amount is lower than the submitted charge in most cases, the difference is not significant enough to normalize for.

4. For the silhouette values and plots, the kernel simply could not handle the data set. Because of this, random samples were drawn from the data, and the silhouette values were based on the random sample instead. It is assumed that this is an appropriate fix.
5. It is assumed that the provider type provided in the data is representative of that particular service as a whole.
6. It is assumed that the features used are correct and reliable.

Methodology

1. The process began with exploratory data analysis. Histograms were constructed for every feature included in the data. The two features that were chosen to cluster the data set were average Medicare payment amount and average submitted charge amount. The average Medicare payment amount represents the cost that Medicare pays (the charge minus what the patient pays in deductibles and coinsurance), while the average submitted amount is the total amount the provider charged for the service (the amount the patient would have to pay if they had no form of insurance). The features appeared to be fairly normal after applying logs to them. The histograms of the features are included below. Additionally, the two features began to tell an interesting story. They begged the question of which types of providers are undercovered are overcovered. Would the clusters split the data into different provider types—a cluster for internal medicine, a cluster for family medicine, and so on?

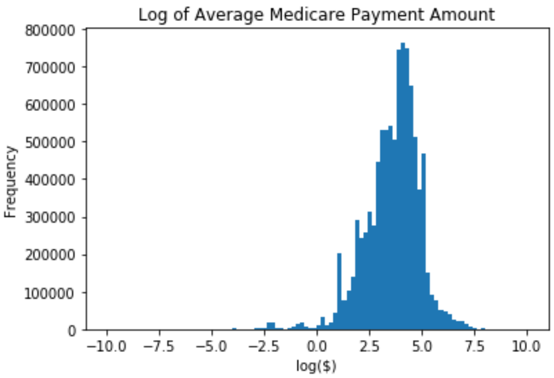

Figure 1

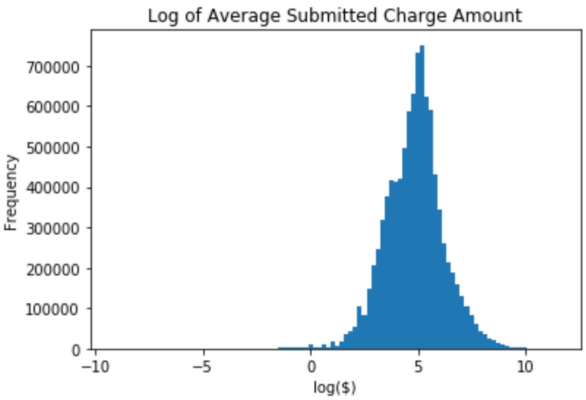

Figure 2

2. With these questions in mind, the data was dug into deeper. The plots of the two features were color coded based on provider type and examined closer. A benefit of selecting only two features allows for the clusters to be directly mapped onto the graph

of the raw data.

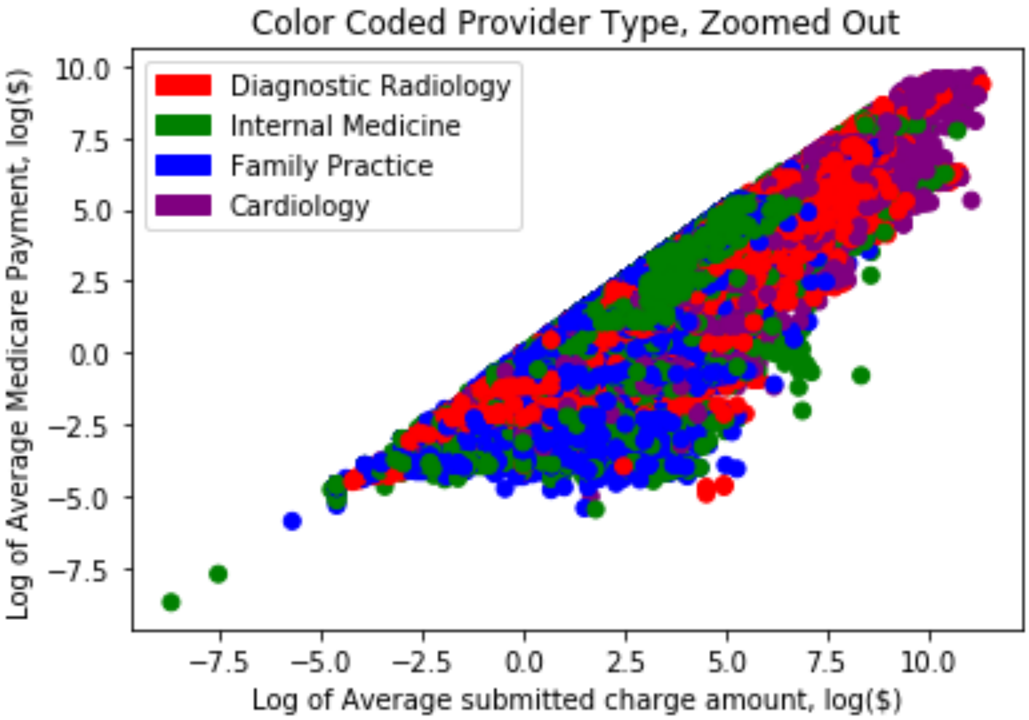

Figure 3

3. Having chosen a compelling set of features, the k-means clustering algorithm was run nine times—once for each grouping of cluster, two through ten. A random seed was utilized in order to ensure repeatability. The absolute value of the “score” (objective) of each cluster iteration was plotted against its number of clusters. This is the Scree plot—based on this plot, it appeared that the graph began to level out around 7 clusters. Thus,

7 clusters was chosen as the best number to move forward with.

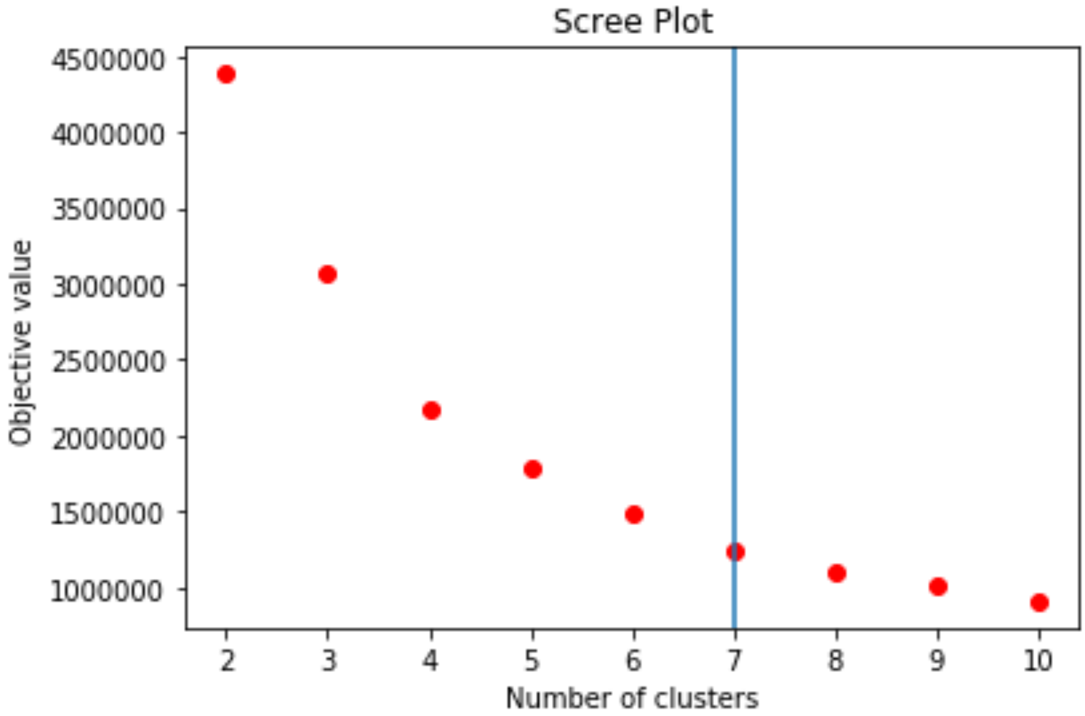

Figure 4

4. Silhouette plots gave insights on the effectiveness of the clustering. The final average silhouette value was 0.39. The silhouette plots are included below. They demonstrate how appropriate each cluster was for each point. As noted in the assumptions, in order to run code on the silhouettes, the data set had to be parsed down considerably. A random sample of size 0.1% of the data set was used to create the silhouette values and plots.

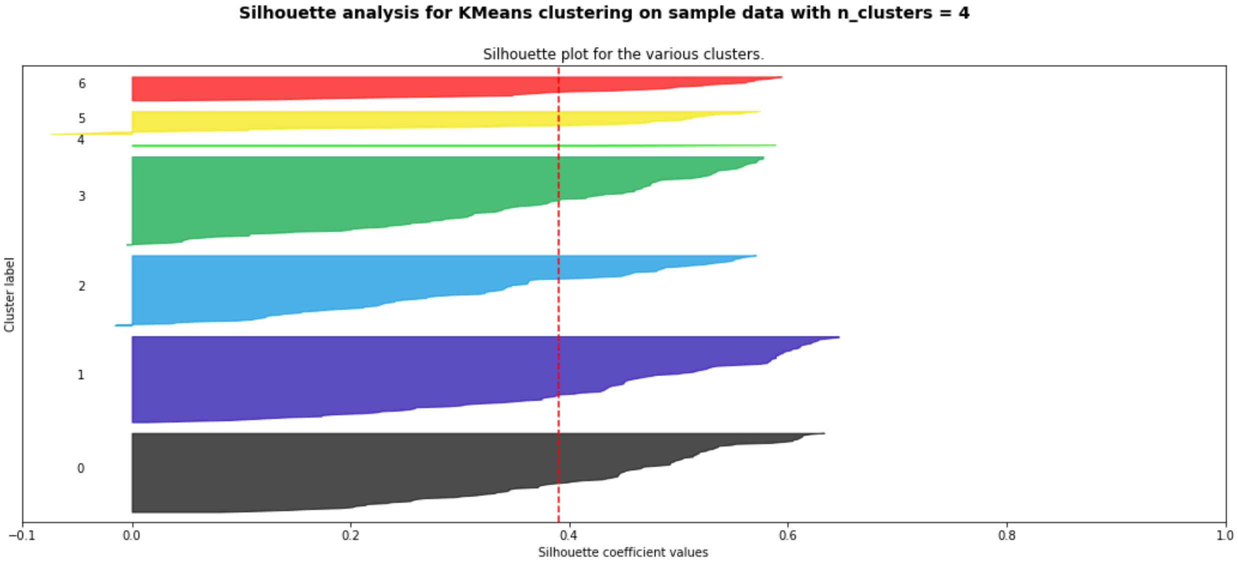

Figure 5

5. Finally, the clusters were plotted in color, as featured below.

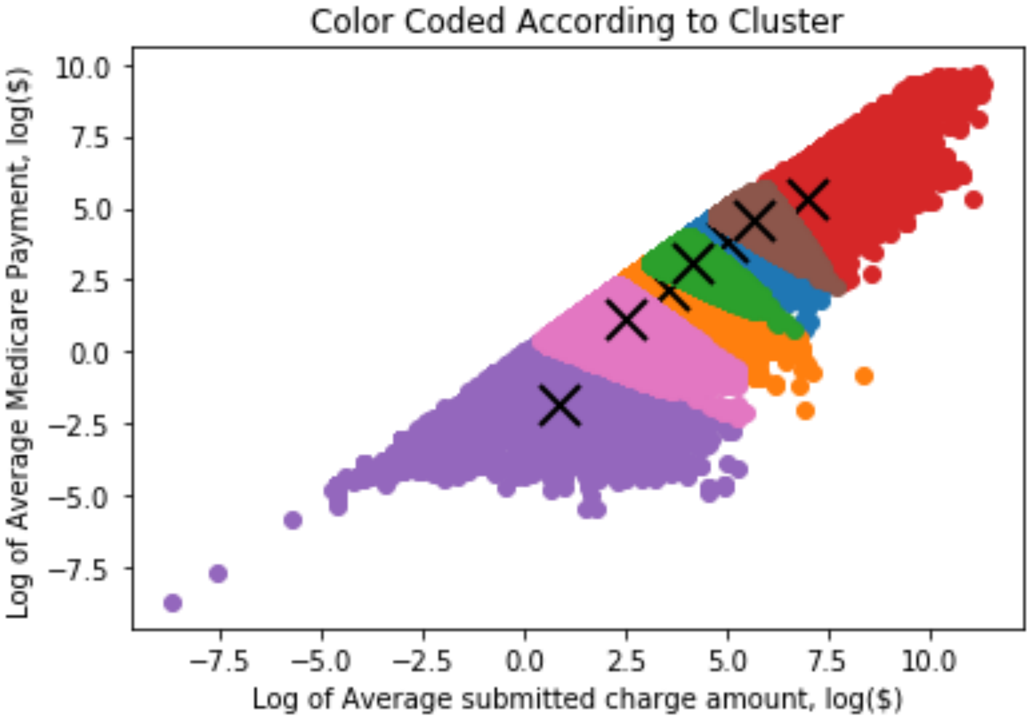

Figure 6

Analysis

According to the average silhouette value, this is not a bad way to cluster the data, though it is mediocre at best. A value closer to one would be preferable. The silhouette plots are also telling of this fact—though some of the clusters are large and reach well over the red vertical line indicating the average, clusters 6, 5, and especially four are thinner and don't reach quite as far over the average line.

When examining the cluster plot itself, it appears that both of the features were at least relevant when dividing up the clusters—that is, the different clusters aren't split by perfectly horizontal or vertical lines. Everything is covered pretty well.

Conclusions

Although the implications from clustering could have been stronger, the analysis still appears to disconfirm the concept that the points might naturally and clearly cluster based on the service they belong to—diagnostic radiology, internal medicine, family practice, or cardiology.

Next steps

The most interesting potential for this project is to dig deeper into the other types of services provided. It would be fascinating to see if more diverse types of medical care fall into different categories. Perhaps dermatology, OBGYN, and medical oncology all carry similarities and would be clustered together. This would address one of the shortcomings of this project—one might assume that the most common types of care would be well covered by insurance. Thus the four types of providers chosen weren't too dissimilar from one another, because although some were more expensive than others, Medicare tends to pay a good amount. It would be more interesting to see more uncommon medical needs represented as well. Additionally, adding more features could introduce more depth to the data and analysis as well.