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 number of patients needing treatments at each facility per day to the average amount Medicare covers for its 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. It is important for patients to understand which services will be thoroughly covered by Medicare, and which will 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 (average_submitted_chrg_amt), as well as the average amount Medicare covered for that service (average_Medicare_payment_amt). *The following report is intended to discuss the presence (or absence) of distinct groups of provided services that are cheap but well covered by insurance, or that are very expensive but 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 database included a vast amount of information on treatment costs, as well as the amount of that cost that is covered by insurance. The Medicare database was explored in order to discover if certain types of medical care are over or under covered by Medicare.

Assumptions

- 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.
- One of the features in the clustering process (average Medicare payment amount, average_Medicare_payment_amt) included a note that in April of 2013 there was a 2%

reduction in Medicare payment due to changes in Federal spending. It is assumed that the Medicare payment data was not jeopardized.

- While performing analysis, it was intentionally chosen *not* to normalize or standardize the features (average_Medicare_payment_amt and average_submitted_chrg_amt). Both of these features are in the same unit, dollars, and although the values for average_Medicare_payment_amt is generally lower than the submitted charge, the difference is not significant enough to normalize the data.
- For the silhouette values and plots, the kernel simply could not handle the sheer size of 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.
- It is assumed that the provider_type noted in the data is representative of that particular service as a whole.
- It is assumed that the features used are correct and reliable.

Methodology

The process began with exploratory data analysis. Histograms were constructed for every feature. The two features that were chosen to cluster the data set were average Medicare payment amount (average_Medicare_payment_amt) and average submitted charge amount (average_submitted_chrg_amt). 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 logging the data. The histograms of the features are included below (figure 1 and figure 2).

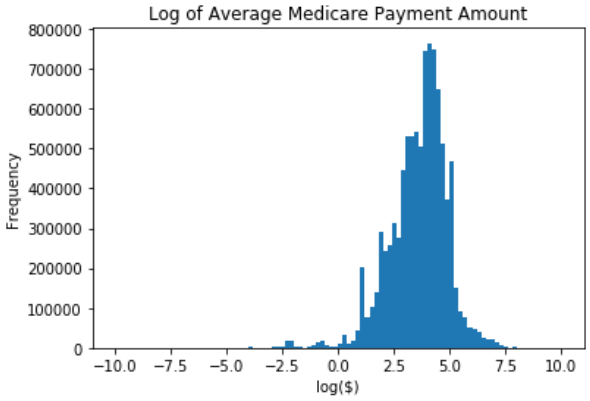

Figure 1

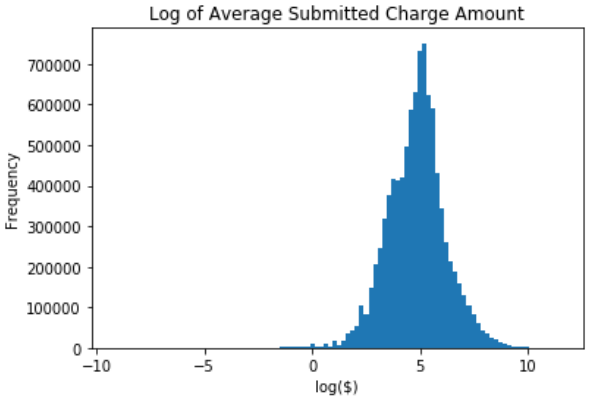

Figure 2

A plot of the two features were color coded based on provider type and examined more closely. A benefit of selecting only two features allows for the clusters to be directly mapped onto the graph of the raw data.

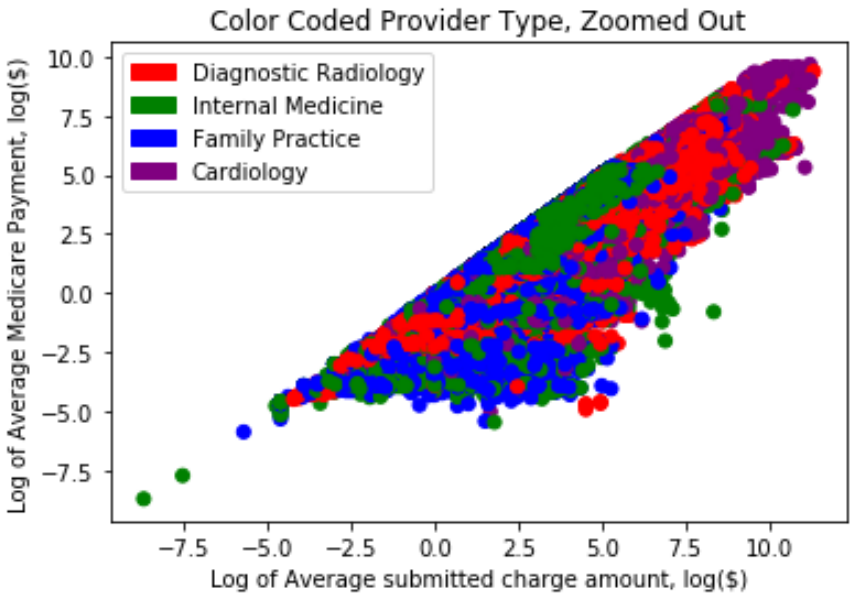

Figure 3

Having chosen a compelling set of features, the k-means clustering algorithm was run nine times—once for each number of clusters, 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. A Scree plot was then created—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.

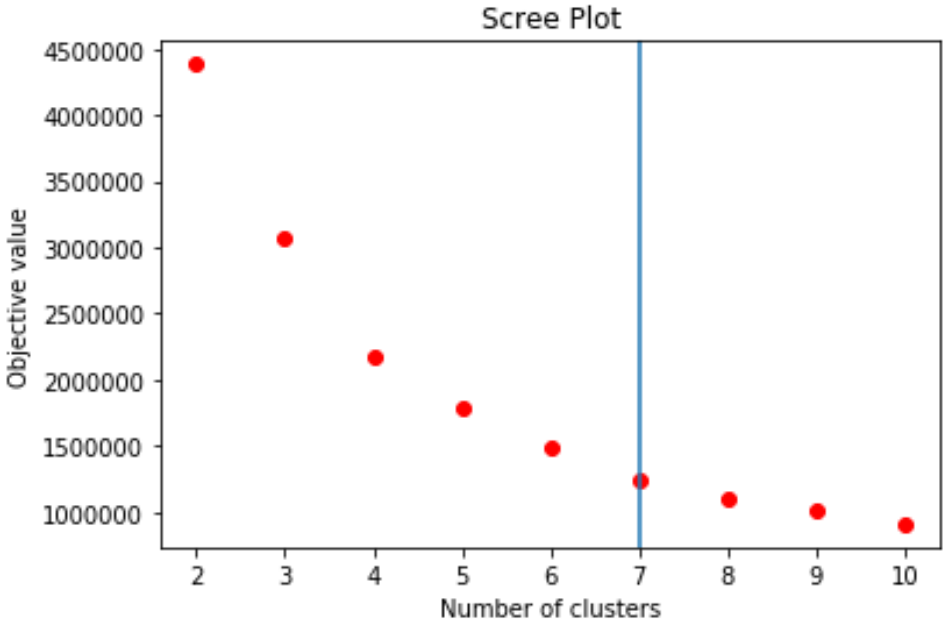

Figure 4

The clusters were plotted in color based on cluster, as featured below.

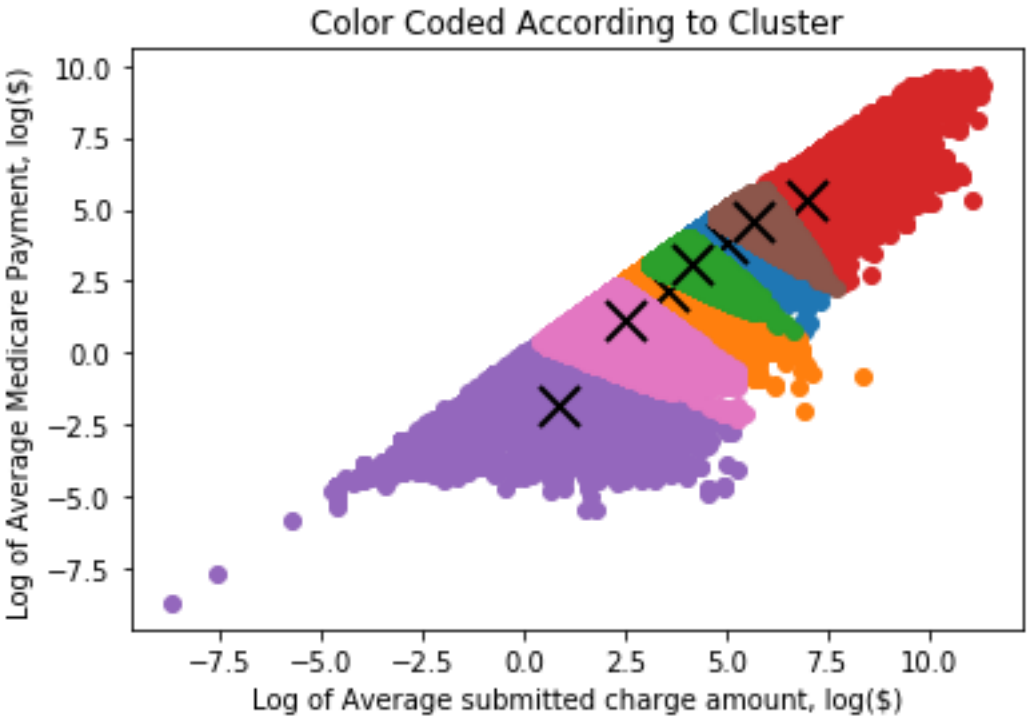

Figure 5

Analysis

A silhouette plot 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.

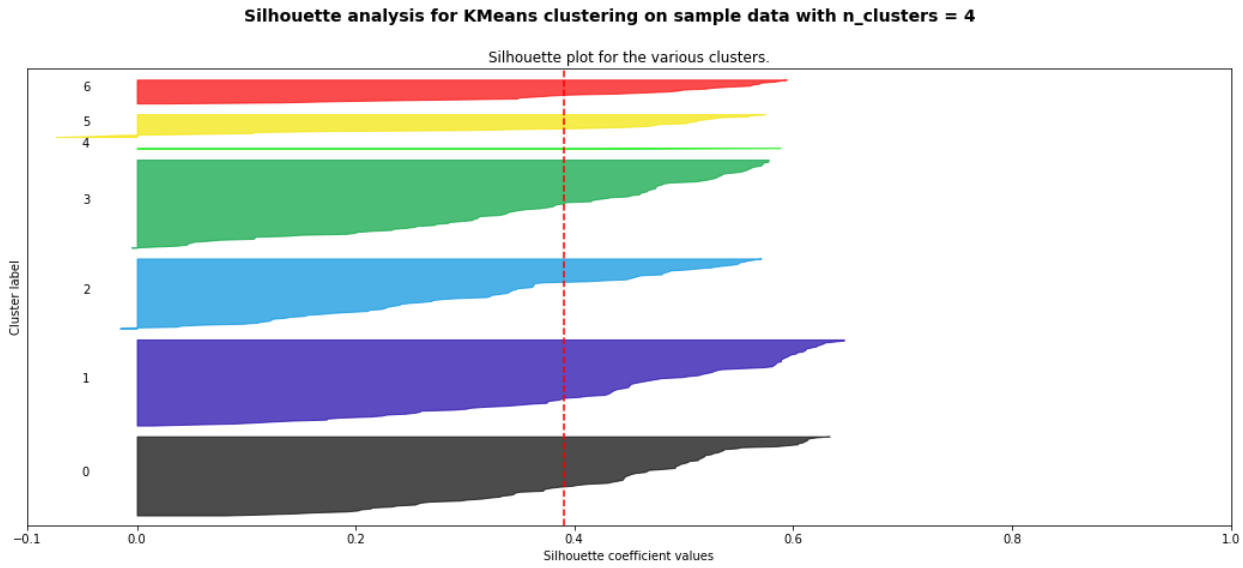

Figure 6

According to the average silhouette value, this clustering 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 five, six, 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. In fact, the clusters seem to move linearly with a positive slope. This seems to indicate that Medicare pays more for more expensive treatments, and less for less expensive treatments.

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. That is, it was posited that perhaps certain services are over- or under- covered. However, based on the clustering, it does not appear that this is true. Rather, the data is divided into clusters in which expensive treatments are covered more thoroughly than inexpensive treatments.

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 are well covered by insurance. Thus, the four types of providers chosen weren't dissimilar enough from one another, because although some were more expensive than others, Medicare tends to cover a proportionally similar amount for all services. Additionally, adding more features could introduce more depth to the data and analysis as well. This analysis is somewhat limited by its use of only two of the available features. Exploration into other attributes of the data could be enlightening as well.