Brent Claypool

IEMS 308 HW 1

Executive Summary:

The dataset is US Government released Medicare data. It contains information on services provided to Medicare beneficiaries in 2018. This data’s features include qualitative information about the provider such as: provider’s name, provider’s location, and the type of provider. The quantitative features of this dataset describe the type of service that was administered by a provider. These features are the number of services completed, the number of Medicare beneficiaries, number of distinct Medicare beneficiary/per day services, average Medicare allowed amount, average submitted charge amount, and the average standardized payment amount. The dataset has multiple entries for each healthcare provider. Each entry represents a different type of service, marked by a unique HCPCS code, accompanied by a description for the code.

The range of providers represented is immense. In total, over 9.9 million services by providers in 27 countries are represented. For this project, I scaled back the scope a bit in order to both make the data manipulation more manageable, and to make insights stronger, as it is difficult to generalize findings for healthcare providers all over the world that operate in a diverse set of conditions. For this clustering project, I am analyzing services from providers located in Illinois, USA. This set describes 404,554 services in 1242 zip codes and 782 cities.

This project segments services based on the potential revenue each one could yield. The algorithm choses 5 clusters that describe the trade-off that exists between providing a service that many people will pursue and providing a service that costs top-dollar to receive.

Problem Statement:

This project aims to segment the types of services provided by potential revenue. The potential revenue within the scope of this project is defined by the product of the amount charged by the provider for the service and the amount of Medicare beneficiaries who have used the service.

Assumptions:

The biggest assumption used in this project is that the potential revenue of each service as defined by this project only takes into account how much beneficiaries are being charged. There is no consideration into how much revenue is being collected by the service in any other way.

Another assumption is that each provider has similar costs for their services. The amount charged by each provider will be significantly swayed by the amount it costs to operate their systems. Factors that influence the difference in costs are the tax rate in each county and city, city regulations, size of the practice, and medical suppliers.

Methodology:

After careful data exploration and consideration of the physical properties of the features of this project, only a few quantitative features were chosen for use in clustering. In the end, two were deemed most important. These were:

1. The total number of distinct beneficiaries that used the service
2. The average amount charged by the provider to perform the service

The physical nature of the relationship between these two variables was the most interesting. Intuition says that the rarest of health conditions requires a high level of expertise which is expensive. Thus, the amount charged by the provider will be high and the total number of unique beneficiaries that require this service will be very low. On the other hand, very common health conditions, such as treatment for an allergy, may not require as much expertise, which leads to many beneficiaries using a low-cost service. These clusters may be easy to spot out, but the clusters in-between will be interesting to find. Such an example is a pathology examination of tissue using a microscope. This service provided by Dr. Mark Wang at Feinberg treated 1059 beneficiaries, charging an average of $267.40.

Below are histograms for each variable:

Chart, histogram

Description automatically generated

Figure 1. Distribution of Average Amount Charged

Figure 1 shows are clear skewness, which is expected. The distribution is also multi-modal, which suggests that there is segmentation within the variable.

Chart, histogram

Description automatically generated

Figure 2. Distribution of Distinct Beneficiary Count

Figure 2 shows a clear skewness. The importance to note here is that there are seemingly more *rare* services than there are common services. The frequency of the number of distinct beneficiaries seems to drop off after about 20. The overall shape stands out but notice the small spikes along the distribution. There is a strong spike at 50 distinct beneficiaries, and at 100. This suggests that there may be some segmentation within just this variable.

The skewness of these two variables highlights the importance for preprocessing. In the code, I used a built-in preprocessing function from sklearn to normalize the skewness. After analyzing the distribution more, I found the presence of very extreme outliers. For the unique number of beneficiaries, there was a mean of 84, however the max number in the dataset was over 250,000. This type of outlier, whether valid or not has no place in this model, because it so far outweighs the others. For the charged amount, the mean was $394, but the max value in the sample was $98,000. These types of values would point to high revenue potential, but really only help to throw the model off. I decided to only analyze values within 2 standard deviations of the mean. This would mean only considering services with an average charge of $2,606.48 and 310 distinct beneficiaries.

I used a K-means clustering algorithm to segment the services into categories. I used SSE to determine the number of clusters.

Results:

The following is a table of the result. The algorithm found 5 clusters to segment. The results reflect a bit of the intuition assumed in the project. Cluster 2.0 represents the case where there is a common health condition that has an inexpensive treatment. For this cluster, it has the highest average count of unique beneficiaries along with the lowest average price charged. Cluster 1.0 is worth looking into for a provider because it has the second highest average distinct beneficiary count and the second highest average price charged. This is a case where cost should be heavily considered, because this analysis suggests that a provider could benefit greatly from providing services in this cluster. A full breakdown of each service, along with its cluster is in a data frame named ‘data’ in the .ipnyb filed titled, “IEMS308HW1”

Table

Description automatically generated