IEMS 308 Assignment 1 Report - JunHwa Lee

Executive Summary + Introduction

One of the fundamental values of Medicare is in fairness. Although different medical conditions and age of the service receivers (SRs) may cause some differences in Medicare coverage, the coverage, in general, should be in a relatively similar range. However, Medicare fraud is still prevalent; some service providers (SPs) charge SRs more by adding services that were not provided. On the other hand, SPs may also find a way to charge less on SRs, because they personally know each other.

Both of those cases should be addressed by Medicare. If someone is getting charged more than he/she should be, Medicare must intervene and make sure that everyone is paying a fair amount. Contrarily, if someone is getting charged less than others, and if Medicare is paying more for that service, Medicare should also intervene and save its spending on that service. Then, those savings can be used for funding other services that need more supports from Medicare. No matter what the condition is, the intervention might lead to a better and fairer financial support to everyone.

Therefore, detecting abnormalities in services and charges by service providers would help Medicare to identify rooms for improvements in its service and its fairness. With K-mean clustering, I identified a set of 15 SPs who behave slightly differently from other SP. That group purely consisted of SPs doing Chiropractic treatments. Some specialties of that group were: (1) they have comparatively high Medicare coverage and (2) they are located in the zip codes where there are not many SPs from other clusters. With this finding, Medicare can start to reevaluate whether high coverage of Chiropractic treatment is reasonable. Also, with further investigations on those 15 SPs (such as proofs of Medicare fraud, the correlation between remote locations and high Medicare coverage), Medicare might be able to find a room to save its fundings, redistribute money for the people who are really in need, and ultimately improve the fairness in its service.

Data Preparation & Preprocessing

(1) Filtering data by country and entity code

I first limited the scope of this analysis to individual SPs in the United States. I made this decision to conduct more focused analysis and prevent external factors from affecting the analysis. For example, the innate difference between the individual service providers and the organization service providers could be excluded.

(2) Creating a new column "medicare_perc"

Then, I created a new column called "medicare_perc." This represents the proportion that (amount that Medicare paid after deduction of deductible and coinsurance) out of (total charges that the provider submitted for the service). It represents the proportion of Medicare actually covered out of total costs.

(3) Aggregating by NPI

At this point, there are multiple rows for each NPI. With multiple rows for each NPI, it is difficult to say some person is an outlier. Therefore, to find which service provider is the outlier, I needed only one row for each unique NPI. Thus, because I am interested in the case where there is the least coverage from Medicare (meaning smaller medicare_perc value), I only used the row with the lowest medicare_perc value for each NPI.

(4) Filtering data by state

For each state, I calculated the mean and standard deviation (std) of medicare_perc. Then, I found that Wisconsin has the highest std/mean ratio, meaning SPs in Wisconsin have the highest variability in medicare_perc. Based on the assumption that higher variability means the higher likelihood of the existence of the outliers, I chose to use data only from Wisconsin for further analysis.

Exploratory Data Analysis (EDA)

(1) Checking unique values of each column

With the preprocessed data, I tried to understand more about the data by counting unique values in each column. One apparent issue was with the zip code; while Wisconsin has approximately 709 zip codes, there were 3,931 unique zip codes in the dataset. To handle this issue, I used the first five letters of the data as zip codes. That ended up with 462 zip codes, which is more reasonable.

(2) Checking distribution of each numerical column with a box-and-whisker plot

I drew a box-and-whisker plot to check the distribution of each numerical column. All the graphs showed an extremely right-skewed distribution. This trend is reasonable in that smaller and cheaper services are more accessible, while services that cost a lot are rare. Because this trend represents the reality and there were no strange values, no specific point has been excluded.

(3) Checking correlation between numerical columns

Both Pearson and Spearman correlations were calculated to understand the relationship between variables. For both correlations, (average_Medicare_allowed_amt and average_Medicare_payment_amt) are highly correlated with a correlation higher than 0.98. It was the same case for (average_Medicare_payment_amt and average_Medicare_standard_amt) because the only difference between them is whether the value is standardized or not.

(4) Transforming each numerical column

As we saw in (2), all the numerical variables are right-skewed. Therefore, before going through normalization, I transposed every numerical column with Box-Cox transformation. With Box-Cox transformation, I could conduct more advanced transformation than simple logarithmic and reciprocal transformation and ultimately make most columns to be normally distributed.

Clustering (K-means)

(1) Selecting columns for clustering & one-hot encoding

For feature selection, I tried to include as many categorical variables as possible to characterize each SP. However, some columns, such as nppes_credentials, nppes_provider_zip are not included because they are so many unique values. For those chosen categorical variables, I also conducted one-hot encoding so that I can include categorical variables for the clustering without any bias. Similarly, for the numerical variables, I tried to use as many variables as I can while excluding highly correlated variables. Features I used are 'nppes_provider_gender', 'nppes_entity_code', 'place_of_service', 'medicare_participation_indicator', 'hcpcs_drug_indicator', "bene_day_srvc_cnt", "bene_unique_cnt", and "average_Medicare_allowed_amt".

(2) Normalizing every column

To prevent one column from dramatically affecting cluster formations, I normalized all the columns, including the ones created through one-hot encoding.

(3) Finding appropriate K

I tested different K values from 2 to 20 and drew a scree plot. With eyeballing, somewhere near 7 seemed appropriate value for K. To validate my thought, I plotted the silhouette score respective to each K. Having 6 as a K value led to the highest silhouette score of 0.4228, making me use 6 as the number of clusters for cluster analysis.

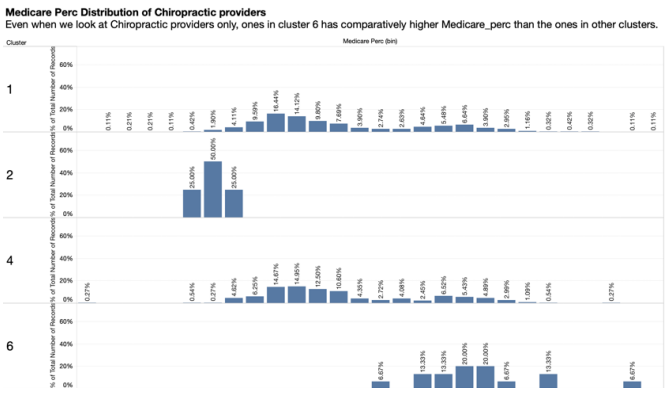
Discussion

By looking at the distribution of different variables in each cluster, I could characterize each cluster like below. The comparatively small size of cluster 6 makes it highly likely to become a cluster of outliers.

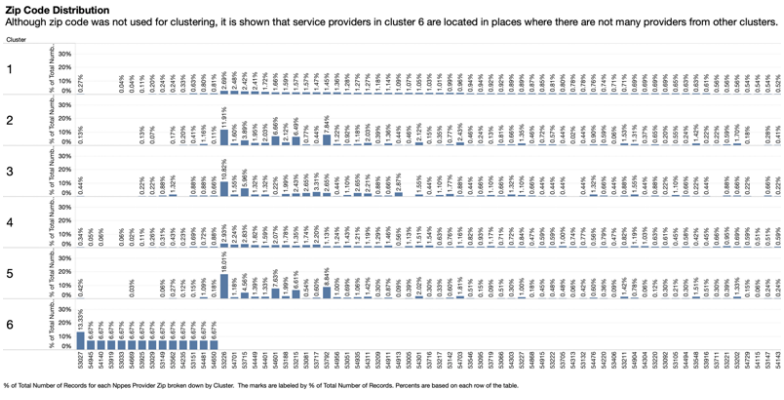
Cluster #	Samples #	Description
1	5,529	Male M.D.s that provide service, especially family practice, at facilities
2	5,424	Male M.D.s that provide service at non-facilities
3	453	M.D.s that provide service that is listed on Medicare Part B Drug ASP at the facility with low Medicare coverage rate
4	6,219	Female M.D.s, mostly nurse practitioners, that provide service at facilities
5	3,315	Female M.D.s, mostly nurse practitioners, that provide service at non-facilities
6	15	D.C.s that provide Chiropractic manipulative treatment at facilities that are located in places not covered by other clusters with a high Medicare coverage rate

Further analysis even validates that insight that cluster 6 are outliers. One interesting finding is that, although medical credentials were not used for the clustering, the clustering could successfully create cluster 6 which is a pure group of D.C.s (Doctor of Chiropractic) providing Chiropractic manipulative treatment. However, considering that D.C. SPs are in other clusters too, it made me explore different features that differentiated these Chiropractic SPs in Cluster 6 from ones in other clusters.

Currently, Medicare only covers spinal subluxation. Therefore, all the services provided by all Chiropractic SRs must be relatively the same. Consequently, it is expected to have a similar Medicare coverage rate from all SRs. However, as shown in the bottom row of the right graph, the coverage of cluster 6 is distributed in 0.595 - 0.735 range, which is higher than the corresponding spectra of other groups.



Another difference is that, as shown in the bottom row of the left graph, SPs in cluster 6 are distributed in 14 zip codes where there are not many SPs from other clusters. That difference is identified even though we did not use zip code for the clustering. That means that there must be some relationship between providing services in unpopular locations and high Medicare coverage. However, with the data we have, it is hard to identify what led to that correlation.



With these findings, Medicare can work on the following to improve its service. First, Medicare can reevaluate whether high coverage of Chiropractic treatment is reasonable. If it is not, it can cut down the coverage of that treatment and redistribute its funding to other services that need more federal funding. Besides, it would be essential to investigate SPs in cluster 6. By comparing Chiropractic SPs in cluster 6 with similar SPs in other groups, it would be possible to identify why SPs in cluster 6 have a high Medicare coverage rate and whether remote location affected the Medicare coverage. If some proofs of Medicare fraud are found, Medicare can penalize those SPs and offer fairness to SRs who were getting services from those SPs.