Surgery Health Care Provider Clustering

Yuxiang (Alvin) Chen

Executive Summary

As a Non-profit organization which focuses on promoting more accessible health care services in the US (excluding US territories), ABC is recently focusing on the health care providers which provide surgery services. This research is conducted to figure out how ABC could best group surgery health care providers based on their surgery specialty type, service place type, average charge on patients & Medicare health insurance coverage extent. This grouping is important because in the next step, ABC would like to figure out what subset of surgery healthcare providers they should focus on to promote health care accessibility through either promoting a reduction in medical charge or promoting an increase in insurance coverage.

The research team accessed the 2018 Medicare Provider Utilization and Payment Data: Physician and Other Supplier dataset from the Centers for Medicare & Medicaid Services website (cms.gov). The dataset consists of information on services and procedures provided to Medicare beneficiaries by physicians and other healthcare professionals in 2018.

The team focused on data of surgery service providers and grouped the data by each health care provider to get aggregated information related to the research goal. Exploratory data analysis is applied and K-clustering, a widely used clustering algorithm is used to figure out the appropriate grouping of health providers. The number of groups are decided based on Elbow Method, Silhouette score and business need.

The result shows that current Medicare insurance coverage fact has its merit but has improvement space. The result also gives 7 clusters of health care providers and the organization could further analyze the clusters separately to mark the priority and design customized future strategies.

Problem Statement

ABC has access to 2018 Medicare Provider Utilization and Payment Data: Physician and Other Supplier dataset. The team's job is to group surgery providers based on surgery specialty type, service place type, average charge on patients & Medicare health insurance coverage for further strategy design

Assumptions

1. The dataset documentation does not specifically clearly show the calculation equation of all average numeric columns. We assume the average amount per service times the service count of the same row is the corresponding total amount.
2. We assume that all data entry values are correct and there is no error there.
3. We assume that a healthcare provider primarily offers surgery treatment to patients, the provider type data from the dataset will have surgery as part of the health care provider type entry.

Methodology & Intermediate Analysis

The detailed step by step process taken is clearly recorded in the project jupyter notebook file with codes as well. Here I will present the steps we took.

1. **EDA1** - We carefully read through the documentation of the dataset and conduct exploratory data analysis on the whole dataset to figure out columns that have satisfactory quality (with 99% entries valid) and provide information related to our research question. All columns used are presented in Jupyter Notebook file (cell 28).

2. **Data Selection** - We selected these columns and filter out rows that we do not want to include in further analysis based on the data quality and business need. These rows include 1 row with non-valid data in columns we want to use, rows that are not providers in the United States (excluding US territories), rows that are not on surgery type health care providers. This process leaves xxx rows.

3. **Data Grouping** - Since in the original dataset, each row presents information about one hcpcs service type offered by one health care provider, to reach our goal of grouping providers instead of services, we would like to group the selected data rows by the health care providers ID column (npi). After grouping, there are in total xxx health care providers considered for further analysis.

4. **EDA2** & Feature Creation - We conducted EDA on the grouped dataset again. We excluded 'nppes_entity_code' information in further analysis because all remaining data is individual health care providers with this information as 'I'. We found positive correlation among two groups of three numerical columns, ('line_srvc_cnt', 'bene_unique_cnt', 'bene_day_srvc_cnt') and ('average_Medicare_allowed_amt', 'average_submitted_chrg_amt', 'average_Medicare_payment_amt'). This result is matching with our hypothesis based on their definitions from dataset documentation.
    a. For the three columns related to count, since unique beneficiary number and exclusion of multiple services from one beneficiary do not add much value for our research purpose, the information from line_srv_cnt column about total service number is sufficient for our grouping, we decide to use it alone.
    b. For the three columns related to average service payment amount, to better capture the insurance coverage information we would like to focus on, we generated two coverage percentage columns to use for clustering instead of using all 3 original columns related to average service payment amount.

5. **One-hot Encoding & Normalization** - We used one-hot encoding for the two categorical columns used ('provider_type' and 'place_of_service'). We also plotted the distribution of 4 numerical columns and applied log transformation to two of them because of their right skew. The graphs in Jupyter Notebook file (cell 86 & 190) show the distribution before and after comparison.

6. **Outliers Addressing** - We removed outlier rows with any one of the numerical column values more than 3 standard deviations away.

7. **K-Clustering** - We run SSE Elbow method & Silhouette score on group number from 2 to 30 and the two graphs do not match very well on the selection of the best cluster number. Thus, we made a decision to use k=7 because it is a jump increase in Silhouette score and it's close to the elbow of the SSE Elbow curve.

8. **K-Clustering Result Evaluation** - We generated a feature main table of each cluster & also a pairwise scatter plot to see whether the clustering result is reasonable and good. We found the quality of clustering to be satisfactory according to the table and the graph. There is no cluster

with an extremely small number of samples or cluster that should be an outlier. There are also no two or more clusters that obviously need to be merged.

Result Analysis

The final mean statistics of each feature for each cluster & the pairwise scatter plot with cluster marked are presented at the end of the same attached Jupyter Notebook file. From the result we find 3 important observations:

1. Focusing on how place of service (facility or non-facility), among 7 clusters, there are 3 that are more than 85% non-facility, they are the bottom 3 clusters if we look at the average submitted charge. There are 2 clusters that are more than 90% operated in facility and they are also the two clusters which have the highest average service charge. The remaining 2 clusters have similar share of two types and a common feature is that about 50% of the providers in these 2 clusters are General Surgery (provider type) providers and these two clusters have the charged cost in the middle of 7 custers.

2. Focusing on health care provider types distribution in each cluster from pairwise distribution graphs (in github repo 'sns_plot3.png'), we found that for each cluster, there is no extreme distribution of one type. And specifically, for the highest cost cluster,different types of surgeries are well blended without one type dominance.

3. Focusing on the two percentage columns, we find that the paid over allowed percentage stays similar across all clusters (between 69-77 percent). However, allowed over charged percentage ranges largely between 21 and 63 percent. And there is an observable negative relationship between charged amount and allowed over charged percentage.

Conclusion & Next Step

From the observations, we see that place of service plays an important role in grouping. Different surgery types put different financial burdens but the impact is suprisingly less than the place of service. Medicare insurance service is doing a good job in covering higher percentages for lower cost health care providers because people who have the larger financial pressure tend to go to health care providers that charge less (potentially because of facility quality). Thus the insurance coverage is helping to an extent helping the right people with limited resources.

With this grouping result, ABC is able to see that although Medicare is targeting well on resource allocation, it is still not covering a high percentage of the total charge (with the highest cluster with mean 63%). ABC would like to promote a higher insurance coverage here to promote surgery service accessibility for people who can not afford them.

ABC is also now able to further look into and analyze each cluster and decide which cluster(s) of healthcare providers should it prioritize in contacting to promote accessibility because more people who go there need the help and which clusters mostly contain providers which mostly serve people who have less financial barriers so they are less important for the immediate next step strategy planning. ABC can also customize its events and strategies for health care providers in different clusters after further analysis.