K-MEANS CLUSTERING

Physician and Other Supplier Data 2018

James Jia, sjs8610, Northwestern IE308

EXECUTIVE SUMMARY

The 10 million lines of data released by the Centers for Medicare & Medicaid Services offers information on services and procedures provided to Medicare beneficiaries by physicians and other healthcare professionals. The US Department of Health and Human Services (HHS) and various health organizations such as hospitals and health advocacy groups can take advantage of this information to better understand the trend and characteristics of services performed in Medicare, and providers who offer these services.

Clustering is one of the potential analytical strategies that can deliver important insights using this data. Clustering is a machine learning method that segments the population of data into a number of groups (clusters) such that members in the same group are more similar to each other, than members from the other groups. In this analysis assignment, clustering is used to segment Medicare physicians (or other providers) into different groups. Such decisions will be based on important features such as their credentials, genders, average Medicare payment and so on. By analyzing the characteristics of these groups and their features, relevant stakeholders can provide target services such as training and publicity.

During the analysis process, we first aggregate 10 million lines of data to around 1 million so that each line of data represents a unique provider. By computing z-score for continuous data and the frequency of occurrence for categorical data, outliers are removed from the datasets. After one-hot encoding and data normalization, K-means clustering is performed to determine the segmentation of providers.

Taking a brief look at the results, it is determined that the providers are best segmented into *26* different groups. The most crowded groups have *123268* physicians, other healthcare professionals and organizations. The least crowded groups have *9386* entities. Out of all the features included in the clustering, *provider_entity* is the most exclusive feature: Each cluster is either composed of all individual professionals or all organizations. Additionally, among all the credentials, *DPT, FNP, PT and O.D*. are the only 4 that form their own clusters (i.e. each cluster either has none or all those credentials). For continuous variables, the characteristics are rather distinguishing as well. Some clusters have their average standardized Medicare amount (*avg_std_amt*) as low as *$21.78*, while others have as high as *$92.89*. The average number of services undertaken also vary from *120+* to *1400+*.

Overall, the results have sufficient characteristics to be distinguished from each other, and by using the information, CMS should be able to design specific training programs and marketing materials that are suitable to the groups, and reduce time, money, and increase the efficiency of its potential outreach.

PROBLEM STATEMENT

The Centers for Medicare & Medicaid Services (CMS) provide the payment and utilization data of healthcare providers, services performed and how much services were charged. As the new Biden Administration promised reform on the Affordable Care Act and Medicare, CMS wants to initiate a new wide-ranging targeted publicity campaign and training program. The goal is to segment providers with similar characteristics into different groups, so that CMS can design specific training materials and targeted email campaigns, instead of universal but ineffective approaches.

METHODOLOGY

Data Aggregation:

As each row of the data provided by CMS represents a unique provider-HCPCS (service procedures) pair, in order to solve the proposed problem, the data will have to be aggregated to represent individual providers instead. Using the National Provider Identifier (NPI) as the unique identifier, 10 million original lines of data are collapsed to around 1 million, representing 1 million distinct providers in the datasets. For categorical variables (such as gender, credentials, operating facility), we assumed that the information should be the same across all entries from the same provider. Information such as HCPCS code will be lost because the data is collapsed. For continuous data (such as number of services and average Medicare amount), summation is applied. One additional column is created to calculate the average Medicare standardized payment, taking into account that different service procedures have different numbers of users. Hence, instead of a simple average of different rows:

$$Total\ Medicare\ Standardized\ Fee/provider\ (total\_std\_amt)$$

$$= \sum_{all\ HCPCS} (\ average\_medicare\_standard * bene\_day\_srvc\_cnt)$$

$$Average\ Fee\ per\ provider\ (avg\_std\_amt)\ =\ total\_std\_amt\ /\ \sum(bene\_day\_srvc\_cnt)$$

Outliers Analysis & Recategorization:

Using the newly aggregated data, an effort is made to identify both continuous variable outliers and categorical variable outliers, as a preprocessing step before performing clustering algorithm. Identifying outliers is always tricky -- in this project, outliers are identified not as “wrong” data, but rather entries that are significantly different from the rest of the dataset.

For continuous variables, z-score tests are run against each variable for the entire dataset. For example, if provider A’s total number of services provided has z-score over 3 (3 standard deviations away from mean), A’s data will be removed from the dataset. We assume that for a provider, as long as one of the entries is outlier, the entire line of data will be removed. By doing so, around 12% of the entire dataset is removed.

For categorical variables, general features are primarily understood through the frequency of occurrence. For the convenience of the one-hot encoding process during K-means clustering later, any entries appearing in less than 1% of the total population will be categorized as “others”. Additional measures are taken to standardized apparent typos such as “MD” and “M.D.”.

Histogram & Correlation Analysis:

To further understand the variables, histograms and correlation matrix are used to gain insights. The continuous variable features are selected primarily based on the results of these analyses.

One-hot Encoding, Normalization, K-selection & K-means:

After gaining sufficient understanding of the dataset, additional preprocessing steps are conducted before performing the K-means clustering. For categorical variables selected for K-means, one-hot encoding was used to create dummy variables that express the data entry in multiple columns with values of 1 and 0. For example, the original gender column is replaced by gender_M, gender_F, gender_Others columns. Normalization is also conducted to the selected continuous variables to achieve more accurate clustering results, as all variables are left-skewed based on histograms, as seen in Figure 1 below.

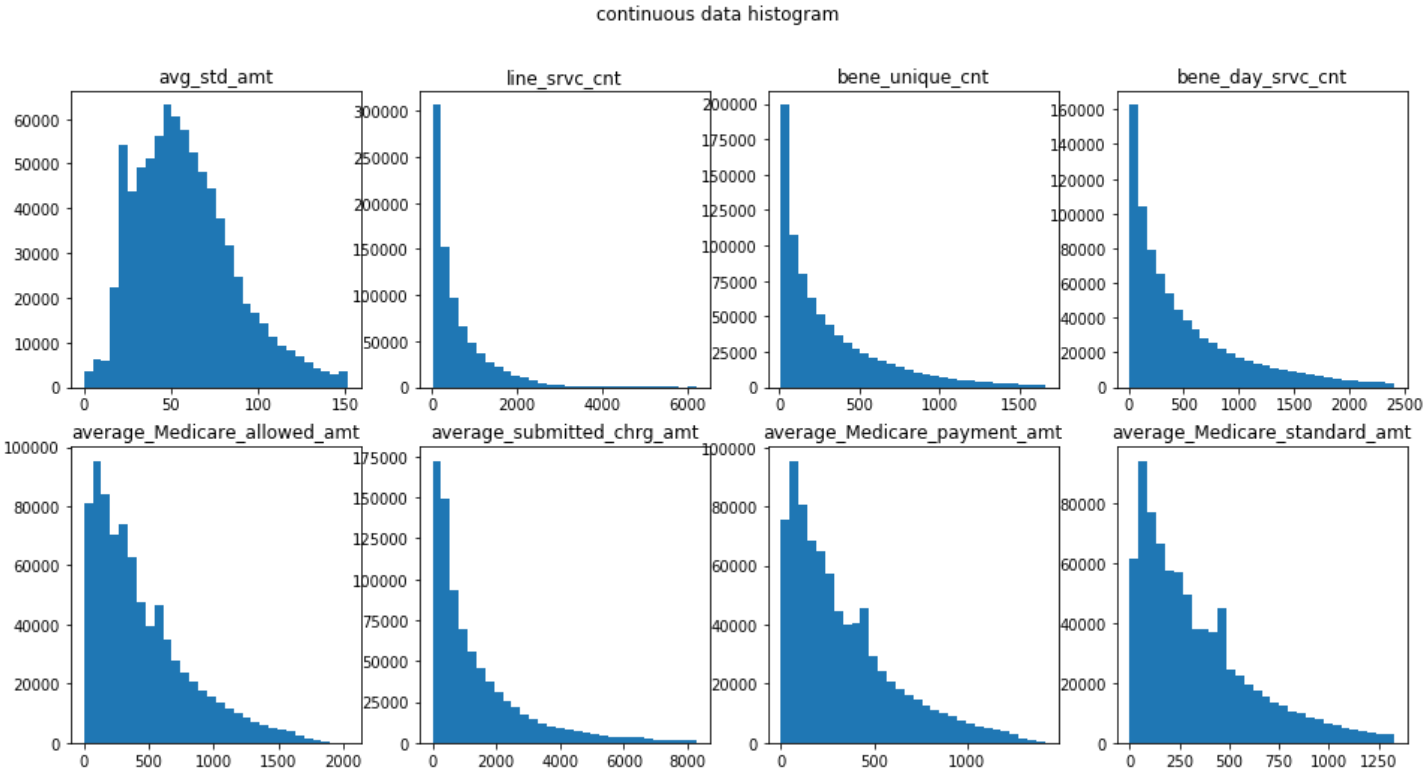

Figure 1

Finally, both scree plot and silhouette plot are used to cross-reference and determine the most suitable number of clusters (K). With the K determined, K-means clustering algorithm from Python's Scipy package. The grouping results are then analyzed.

ANALYSIS OF RESULTS

Feature Selection:

Selecting the features that will be used in the clustering algorithm is one of the challenges in decision-making for this project. Based on research, one of the algorithm-based methods is to run K-means clustering against each potential feature, evaluate the quality of the clustering (such as by calculating the silhouette value). However, due to constraints in time and computational power, this method was not adopted. Instead, the features are selected based on correlation analysis, feature observation and close examination of the nature of variables.

Firstly, for continuous variables, it is necessary to select variables that do not have strong correlations. For correlations with strong correlations (in extreme cases, such as r^2 = 1), including both in the clustering will simply add very similar data and the repeats will result in a waste of computational resources and less meaningful clustering process. Based on the correlation matrix in Figure 2, all the fee information are strongly

correlated (r^2 > 0.8), and beneficiary numbers are strongly correlated, though number of services are not strongly correlated with any other variables. Therefore, by calculating the average standardized Medicare fee (avg_std_amt , ***equations above)***, we are able to obtain a set of valuable continuous variables. $average\_medicare\_standard$ and $bene\_day\_srvc\_cnt$ was chosen because they are adjusted for geographical differences and repeated services.

Secondly, categorical variables are somewhat more tricky. However, by means of data aggregation, various variables are rendered meaningless (such as HCPCS-related data). Any geographical information (such as zip codes, cities, states, countries) are not included because of its complexity in one-hot encoding, and because the standardized fee variable is already adjusted for geographical differences.

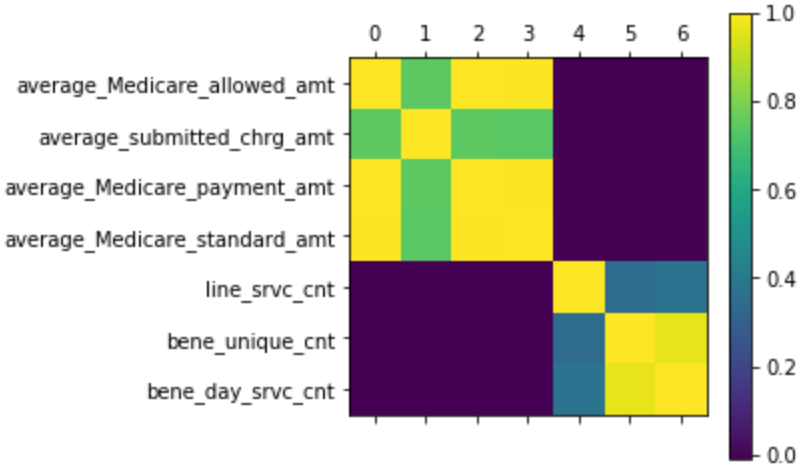

Figure 2

K-selection:

Because the number of clusters (K) has a significant impact on the quality of the clustering, both Scree plot and Silhouette plot are used to cross-reference and determine the most suitable K value. The scree plot (Figure 3) shows a significant kink in the plot of inertia value against K at around k = 26. Due to the time complexity of the Silhouette program *(silhouette_score)*, a plot was unable to be produced. However, the code has been included in the source code. K = 26 is thus chosen for the clustering.

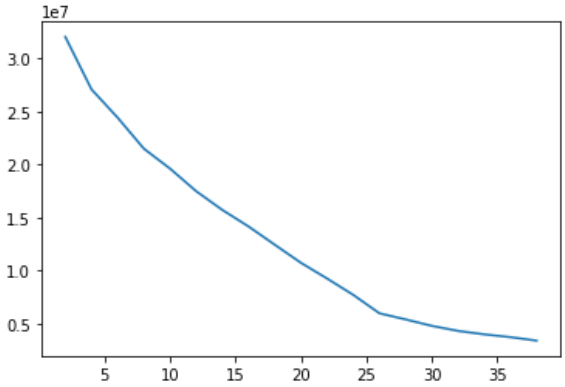

Figure 3

Clustering Results:

Based on the clustering results:

- In terms of the size of the groups:
  - The most crowded groups have *123268* physicians, other healthcare professionals and orgs.
  - The least crowded groups have *9386* entities.

    - Although K-means tend to produce groups of the same size, in this project, the size of the group has non-trivial variance (the max size is 13 times the min size). Therefore, it is important to pay attention to the distribution of resources due to the lopsided group size.
- Out of all the categorical features included in the clustering:
    - *provider_entity* is the most exclusive feature: Each cluster is either composed of all individual professionals or all organizations. This is probably most ideal, as the training program / publicity materials for individuals will likely differ significantly with organizations.
    - *Gender* does not seem to be a critical distinguishing feature, as different groups have varying male-female ratios. This is probably desired as there should be a mix of professionals with different gender to be in the same group.
- For continuous variables, the characteristics are rather distinguishing as well:
    - Some clusters have their average standardized Medicare amount (*avg_std_amt*) as low as *$21.78*, while others have as high as *$92.89*.
    - The average number of services (*line_srvc_cnt*) undertaken also vary from *120+* to *1400+*.

With reference to the table below (Table 2), Cluster #8 and Cluster #9 are selected to more explicitly discuss the success for this clustering model. For Cluster #8, the group as a whole performed way more services (*line_srvc_cnt*) than #9. However, its contribution to standardized Medicare charge is much lower. By closely observing the rest of the data, it can be seen that Cluster #8 is overwhelming in physical training private practice, in non-facilities. It likely results in a large amount of patients expenses unclaimable by Medicare.

The clustering result effectively segments providers into groups with distinct characteristics. *It will definitely help to aid the process of targeting training and effectively designing different messaging.*

Cluster	avg_std_amt	line_srvc_cnt	CRNA	Other	PT	F	M	entity_I	entity_O	CRNA	Other	PT(Private Practice)	facility_F	facility_O
8	21.98	1392.14	0.0%	0.0%	100.0%	60.5%	39.5%	100.0%	0.0%	0.0%	0.3%	99.7%	0.0%	100.0%
9	88.04	125.98	82.5%	17.4%	0.0%	59.9%	40.1%	100.0%	0.0%	100.0%	0.0%	0.0%	97.6%	2.4%

Table 1

LIMITATIONS

Over the course of the project, several limitations are also found. For example, in terms of outlier detection, there is no good way to detect outliers from categorical data. A value appearing only once does not necessarily mean the data itself is outlier. For continuous variables, z-score is not a very robust method to identify true, practical outliers (not just statistical outliers).

Additionally, there is also a big room of improvement for the feature selection process. Not only a more scientifically sound and universal approach is yet to be found, feature selection also has a large impact on the clustering time complexity and final outcome. In this project, the silhouette plot (for K selection) did not display a desired exponential decay shape due to its high time complexity demand. The reason for that might be because of the large number of features (~50) in the dataset, in addition to the large dataset size (1M).