Roger Mei

Illinois For-Profit Hospital Locations based on Medicare Payment Information Clustering

Executive Summary

The Medicare fee-for-service provider utilization & payment data contains information about services provided to Medicare beneficiaries by physicians and other healthcare professionals. This information can be used to determine the most profitable locations in Illinois to build a For-Profit Hospital.

Clustering is an unsupervised learning technique used to group instances of data with similarities together. By using a specific algorithm of clustering called K-means, certain instances of the data can be lumped together to possibly determine the various locations with the highest service costs and Medicare payment amounts.

After removing all features that have no relevance to the business problem, the only categorical feature that remained is the zip code of a provider. The zip code is the main metric in determining location for this problem and is reduced down to its first 3 digits. It is then one-hot-encoded into 45 different features, segmenting Illinois into 45 different areas by 3-digit zip codes.

Taking a look at the clustering results, the entities with the highest number of services provided also has the highest number of distinct Medicare beneficiary/per day services. Zip code 601 has the highest average Medicare allowed amount and zip code 605 has the highest average submitted charge amount. Although the clusters are not very significant, it can be seen that the entities with the highest number of services provided also has the highest number of distinct Medicare beneficiary/per day services. Thus to have more business in a day, a hospital should provide more services.

Problem Statement

A new for-profit hospital is to be built in Illinois. The job was to determine the most profitable location in Illinois by zip-code to build this new hospital. Because the solution looks for higher profits, the features to be looked at in the provided Medicare payment information data are number of services provided, number of distinct Medicare beneficiary/per day services, average Medicare allowed amount, and average submitted charge amount.

Assumptions

- All average payment amounts in the data are accurate and are paid for. Sometimes insurance providers can adjust payment amounts by paying for more or reducing charges to beneficiaries.
- All average submitted charges by the provider amounts are accurate and final. Sometimes there can be adjustments and law suits over medical charges.

Methodology

The first step is to filter only the instances where the state of the providers is Illinois. After doing this, filtering is also done to look at only instances of entities that are organizations that accept Medicare. Only organizations are to be looked at because a hospital is being built, not an individual physician’s office. Only examples with the

'MEDICARE_PARTICIPATION_INDICATOR' column equal to ‘Y’ are kept as Medicare is to be accepted at the new hospital and this problem relies on Medicare payment information.

Next, features with no relevance to the business problem are manually dropped. The city column is dropped as this study focuses mainly on using 3-digit zip code as the primary metric of location in Illinois. Other columns that are not relevant include information that are related to the provider’s name, gender, address, and HCPS code. The provider type, place of service, and entity type features are dropped too because this problem is focused on costs and payments.

The only categorical variable that remains is the 3-digit zip code, created by reducing the zip code column to only its first 3 digits. This column is one-hot-encoded into 45 separate columns. After examining the multi-collinearity of the numerical features and log transforming the ones that remain after further dropping features, all features can be normalized and clustering can begin.

Analysis

To further reduce the number of features used, multi-collinearity among numerical features is looked at. Correlation between every numerical feature is shown in Figure 1 and the features with the highest correlations are shown in Table 1 below.

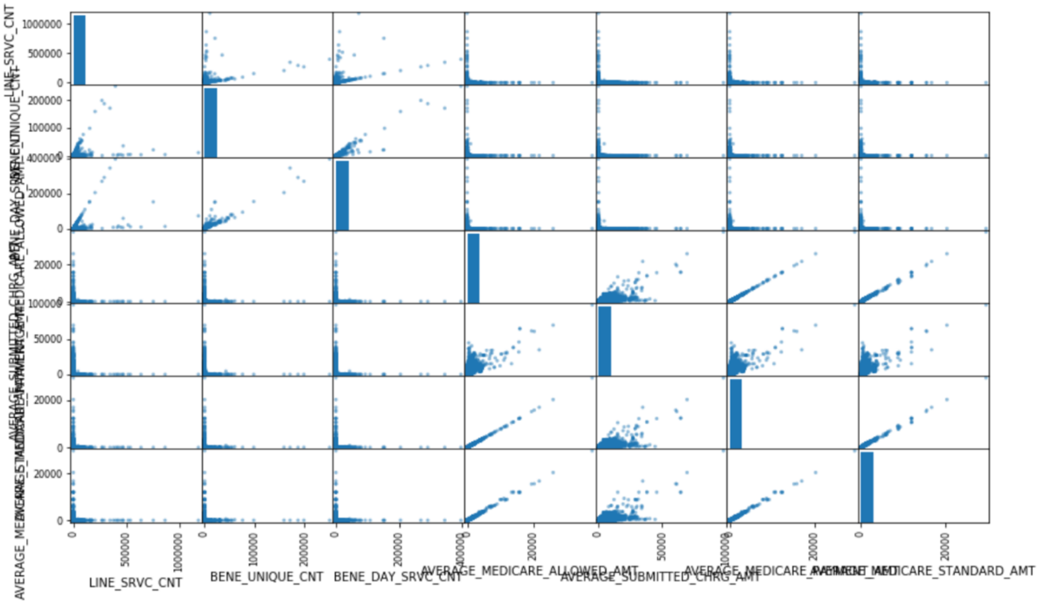


Figure 1

Feature 1	Feature 2	Correlation Coefficient
BENE_UNIQUE_CNT	BENE_DAY_SRVC_CNT	0.945993
AVERAGE_MEDICARE_ALLOWED_AMT	AVERAGE_MEDICARE_PAYMENT_AMT	0.998709
AVERAGE_MEDICARE_ALLOWED_AMT	AVERAGE_MEDICARE_STANDARD_AMT	0.996288
AVERAGE_MEDICARE_PAYMENT_AMT	AVERAGE_MEDICARE_STANDARD_AMT	0.996854

Table 1

'BENE_UNIQUE_CNT' has high correlation with 'BENE_DAY_SRVC_CNT'. 'BENE_DAY_SRVC_CNT' is kept since it removes double-counting of the similar services on the same day as well as describes the daily rates of service and not simply the total count.

Because the features "AVERAGE_MEDICARE_ALLOWED_AMT", "AVERAGE_MEDICARE_PAYMENT_AMT", and "AVERAGE_MEDICARE_STANDARD_AMT" have high correlations, two of these features can be removed. "AVERAGE_MEDICARE_STANDARD_AMT" is simply the standardized data of "AVERAGE_MEDICARE_PAYMENT_AMT" with geographical differences in payment rates removed. However, both features, "AVERAGE_MEDICARE_PAYMENT_AMT" and "AVERAGE_MEDICARE_STANDARD_AMT", don't include what the beneficiary is paying for and are thus removed. The feature "AVERAGE_MEDICARE_ALLOWED_AMT" is kept to explore what Medicare pays, including deductible and coinsurance amounts that the beneficiary is responsible for paying.

Looking at the distributions of the numerical features with histograms of 50 bins as shown in Figure 2, the distributions seem to be heavily skewed. A log transformation is made on all numerical data and the distribution becomes more normal and spread out as shown in Figure 3. Because there are 393,123 instances, the transformed data is very normally distributed. This dataset is then mean normalized.

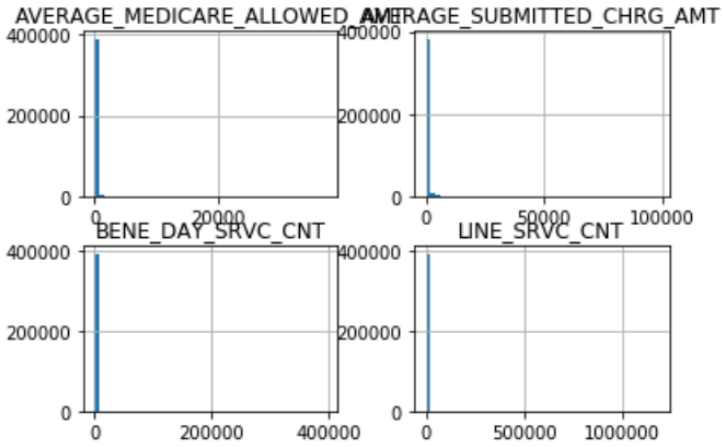


Figure 2

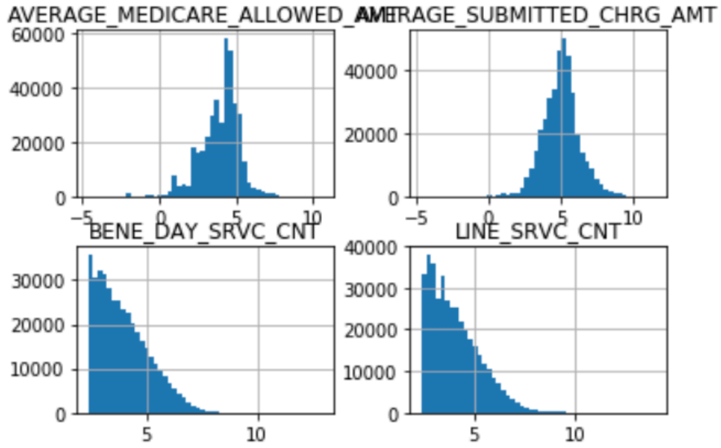


Figure 3

From the scree plot in Figure 4, there seems to be a “kink” at 8 clusters. Silhouette scores are calculated to further explore the number of clusters around 8. It seems that the higher number of clusters have a better silhouette score. Out of the number of clusters tried, 11 clusters have the best silhouette score of 0.35. Although the scree plot shows a steady decline as the number of clusters increase, the number of clusters used should not be so high for better interpretability so 11 clusters are chosen. Partitioning Illinois into any more than 15 clusters is difficult to interpret. Any number of clusters above 20 is also too computationally expensive and time consuming when calculating silhouette scores as well.

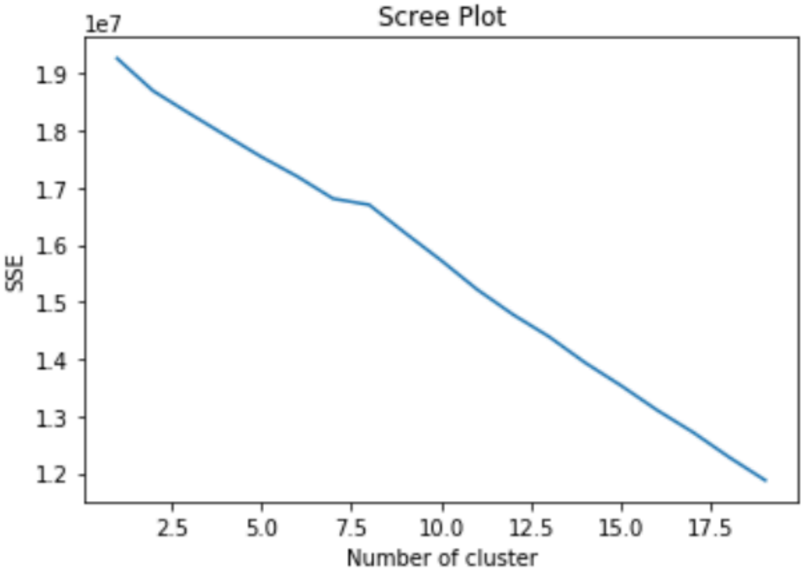
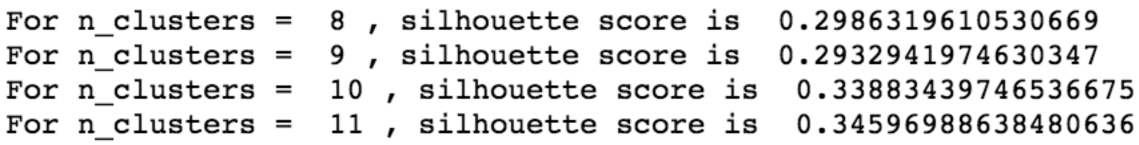


Figure 4



```
For n_clusters =  8 , silhouette score is  0.2986319610530669
For n_clusters =  9 , silhouette score is  0.2932941974630347
For n_clusters =  10 , silhouette score is  0.33883439746536675
For n_clusters =  11 , silhouette score is  0.34596988638480636
```

Figure 5

The main features used in clustering are zip code, number of services provided (‘LINE_SRVC_CNT’), number of distinct Medicare beneficiary/per day services (‘BENE_DAY_SRVC_CNT’), average Medicare allowed amount(‘AVERAGE_MEDICARE_ALLOWED_AMT’), and average submitted charge amount (‘AVERAGE_SUBMITTED_CHRG_AMT’).

The number of services provided allows us to see how the payments and charges correspond to the number of services provided. The average payments and charges could be high in an area, but this may be due to a provider having a large number of services as well. It is also important to determine the number of services the new hospital to be built will provide.

The number of distinct Medicare beneficiary/per day services shows how busy a certain location will be. A higher value for this shows that a certain area has more business and more profitability. The average Medicare allowed is how much Medicare pays including deductible and coinsurance amounts the beneficiary is responsible for paying. The average submitted charge amount is what the provider charged for the service. Both these values directly correlate to the profits of a provider.

The highest number of services and highest number of distinct Medicare beneficiary/per day services has the same cluster centroid. The zip code column with the highest value in the centroid is selected as the optimal zip code of a cluster. The cluster centroid with the highest average Medicare allowed amount has the optimal zip code of 601. The cluster centroid with the highest average submitted charge amount has the optimal zip code of 605.

Conclusions

In order to have a higher number of distinct Medicare beneficiary/per day services, a hospital must provide a higher number of services. By providing more services, more patients will use the hospital in a single day. This is shown by a single cluster centroid with both the highest number of services and highest number of distinct Medicare beneficiary/per day services.

Illinois can be split up into 11 clusters based on this study. However, these clusters are not very significant as demonstrated by the relatively low silhouette scores. The silhouette score for 11 clusters is 0.35. Thus, this study is inconclusive on which zip code can be the most profitable in terms of highest average Medicare allowed amount and highest average submitted charge amount. However, if you had to choose areas to build a hospital, it’d be in areas of either 3-digit zip codes 601 or 605.

Next Steps

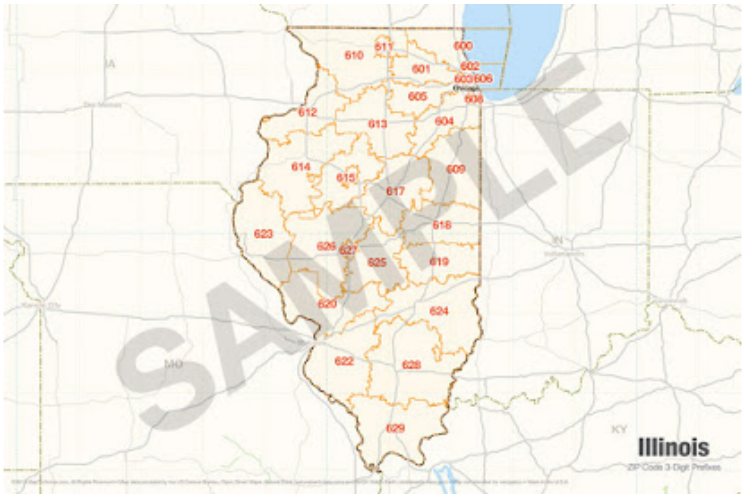


Figure 6

Looking at the 3-digit zip code map of Illinois, 601 and 605 are right next to each other in the Northern part of Illinois, west of Chicago. It may be good to perform a study with cities instead of zip codes and seeing if the Northern cities of Illinois are more profitable than the Southern cities. It would also be important to see if proximity to major cities such as Chicago matter.

Another important thing to look at is the relationship between numerical features. Linear regression could be done on the numerical features of this study to see if there are relationships between number of services provided, number of distinct Medicare/per day services, average Medicare allowed amount, and average submitted charge amount.