Clustering Analysis
Eric Xi
IEMS 308
January 23, 2019

Executive Summary

An article written on *Radiology Business* on October 04, 2018 titled "Diagnostic Imaging Facility operator indicted for $284M in fraud" revealed a California-based owner and operator of diagnostic imaging facilities operated a "fee-per-scan" system which gave incentives to recommend more scans that were medically necessary (Radiology Business).

Detecting Medicare fraud has been a focus on the Center for Medicare & Medicaid Services. Therefore, Medicare officials at the CMS have hoped that the release of Medicare Provider Utilization and Payment Data will expose fraud, inform consumer and lead to improvements in care (EMDs). Therefore, Given the recent exposure of fraud, our primary objective is to utilize the data provided by the CMS and implement a clustering algorithm that will help the state of California identify potential Medicare frauds in the Diagnostic Radiology provider type. Doing so will help the regulatory agencies prioritize their time by reviewing providers might have a higher potential of committing fraud.

Clustering is an unsupervised machine learning algorithm that can help group a set of objects in a way that objects in the same cluster are more similar to each other than to those in other groups. In the case of fraud detection, we hope that the algorithm could take in features specific to each provider's service (i.e. physical location, utilization and payment data etc.), and ultimately detect clusters that may be indicative of a provider committing fraud.

Taking a brief look at the result, we were able to identify two clusters that each potentially represents two kinds of fraud, namely "Billing for unnecessary medical services" and "Charging excessively for services or supplies.

Problem Statement

The Medicare Provider Utilization and Payment Data contains information on services and procedures provided to Medicare beneficiaries by physicians and other healthcare professional. Our job is to utilize the k-means algorithm to cluster the data based on categorical, utilization, and payment features. Using the algorithm, we hope to be able to identify clusters that can help government officials in the **State of California** detect physicians in the **Diagnostic Radiology** practice that might be committing Medicare Fraud.

Assumptions

1. The data provided by CMS is accurate and truly reflects each provider service's actual utilization and payment regarding Medicare reimbursements.

2. Medical procedures performed in each cluster are similar in type.

Methodology

Our first step to begin clustering was to filter the data so that only relevant information to the business problem is kept. We imported the data into R and only kept data where 'NPPES_PROVIDER_STATE' == 'CA', and 'PROVIDER_TYPE' == 'Diagnostic Radiology'. We also limited our scope of the problem by only looking at services that are non-drug related, hence we also set 'HCPCS_DRUG_INDICATOR' == 'N'. To speed up the query speed of the data, we also dropped some unnecessary features that are likely not to be important to clustering: 'NPI','NPPES_ENTITY_CODE', 'NPPES_PROVIDER_LAST_ORG_NAME', 'NPPES_PROVIDER_FIRST_NAME', 'NPPES_PROVIDER_MI', 'NPPES_PROVIDER_STREET1', 'NPPES_PROVIDER_STREET2', 'HCPCS_DESCRIPTION'

After reloading the data back into Python, we utilized the OneHotEncoder function within the sklearn.preprocessing package to encode categorical data such as GENDER and PLACE_OF_SERVICE. We then further filtered data to only keep features that will be feed into the clustering algorithm. This includes: 'LINE_SRVC_CNT', 'BENE_DAY_SRVC_CNT', 'AVERAGE_SUBMITTED_CHRG_AMT', 'AVERAGE_MEDICARE_STANDARD_AMT', 'GENDER' ,'PLACE_OF_SERVICE'

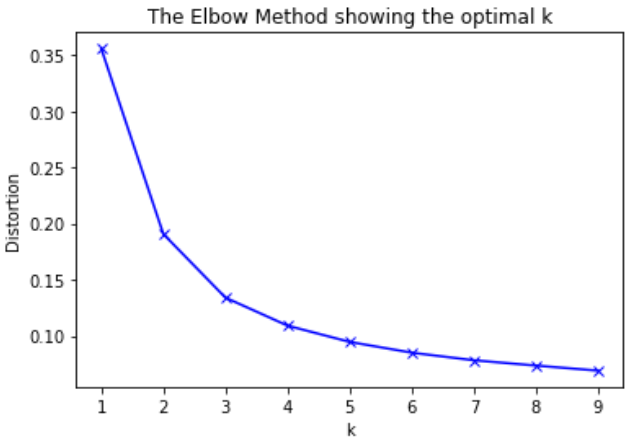
The next steps then involves normalizing the data and feeding it into a for loop that runs the KMeans algorithm through different number of clusters and creates a scree plot to find the kink that indicates a relative optimal number of clusters, which turns out to be 4. We then plotted a Silhouette graph using the Yellowbrick SilhoetteVisualizer package and also created a scatter plot after utilizing TSNE for dimensionality reduction.

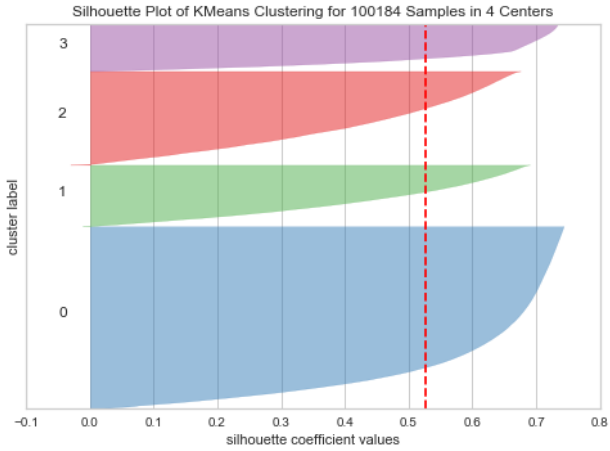
We then computed summary statistics for each of the clusters so that we can identify what differentiates each cluster.

Analysis

1. Search for a relative optimal number of clusters (k)

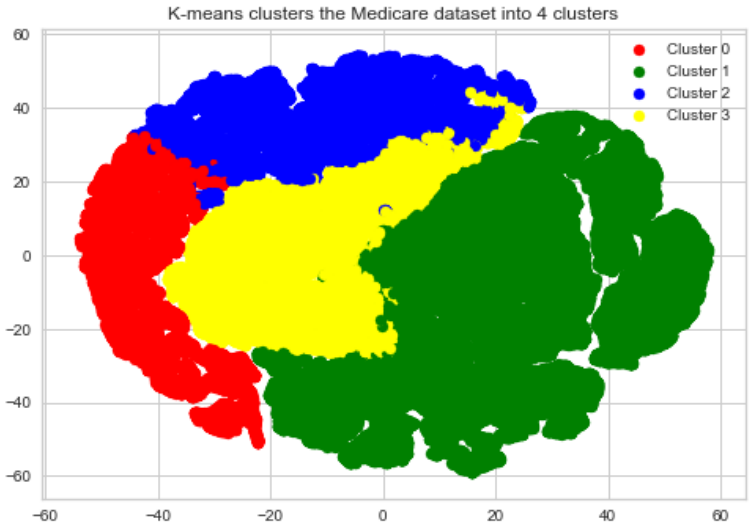
By constructing a scree plot and employing the elbow method, we found that 4 clusters seem to be the point beyond which reduction in distortion is diminished. We also created a Silhouette plot of KMeans clustering using 4 clusters and it returned an Average Silhouette score of approximately 0.52.





2. Visual Inspection of clusters after TSNE dimensionality reduction

After using TSNE for dimensionality reduction, we were able to plot the data and its respective labels on a 2-D graph. Upon initial visual inspection, there are no immediately clear outliers, but cluster 0 appears to have the lowest amount of data points. Therefore, the next step would be to compare summary statistics of these clusters to highlight their differences.



3. Summary Statistics of Clusters

Using red for minimum, and green for maximum, we were able to highlight the min and max clusters for each feature. While Cluster 2 and Cluster 3 are quite similar, Cluster 0 and Cluster 1 sits at two ends of the spectrum.

Physician services in Cluster 0 tends to occur at a higher frequency but charged lower amounts and consequently requires less payment from Medicare. While physician services in Cluster 1 occurs less frequently, have a higher rate of non-facility place of service, and are priced higher, which requires higher payments from Medicare.

Relating back to the central problem of identifying potential fraud. This result might indicate that physician services in Cluster 0 are overprescribing unnecessary procedures, while physician services in Cluster 1 are overcharging for its services.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Count	**12285**	**47410**	16039	24450
Average Service Count	**313**	**27**	69	45
Average Distinct Medicare Benefeciary/per day services	**302**	**26**	66	44
Average Charged Amount	**58**	**595**	83	138
Average Medicare Standardized Payment	**13**	**91**	17	27
% Male	80%	81%	82%	82%
% Place Of Service = Facility	82%	**58%**	82%	76%

Conclusions

Using the k-means algorithm on the Medicare data, we were able to identify 4 clusters within Diagnostic Radiology physician services in the State of California. Since the main objective was to provide regulatory agencies with the ability to prioritize their time in reviewing potential fraud, we were able to further analyze the features of each cluster and provide guidance as to which clusters should be reviewed first.

Cluster 0 and Cluster 1 represents two potential types of Medicare Fraud and thus should be priortized. Cluster 0 have high service counts and low amount charged, which might indicate the physician service committing fraud through overprescribing services. While Cluster 1 have low service counts but high amount charged, which might indicate the physician service committing fraud through overcharging services and supplies.

Next Steps

Since the k-means algorithm has a tendency to create similar size clusters, it is difficult to utilize this algorithm to identify smaller size clusters which might be more indicative of a physician committing fraud. Therefore, as a next step, it will be helpful to identify another algorithm that has a stronger ability to identify outliers clusters.

Another next step would be to incorporate data that indicates fraud that actually occurred. Through my research, some data scientist has in fact utilized a dataset called “List of Excluded Individuals and Entities” from the Office of the Inspector General. Using labels from this dataset, we would be able to employ other supervised machine learning techniques to actually help predict the likelihood of a person committing fraud.

Reference

"Diagnostic Imaging Facility Operator Indicted for $284M in Fraud." *Radiology Business*, 4 Oct. 2018, www.radiologybusiness.com/topics/healthcare-economics/california-diagnostic-imaging-operator-indicted-284m-fraud.

"Historic Release of Medicare Provider Utilization and Payment Data." *EMDs*, 10 Apr. 2014, www.e-mds.com/historic-release-medicare-provider-utilization-and-payment-data.