BA 820 - Unsupervised Machine Learning - Final Report - Team 15

Team Members: Jingjing Lu, Scott McCoy, Tiam Moradi, Tsung Yen Wu

For our project, we decided to take the perspective of a consulting company marketing services to nursing homes. Our goal was to use unsupervised learning and text analysis to identify distinct segments of customers, and develop unique marketing strategies for different groups.

**Dataset**

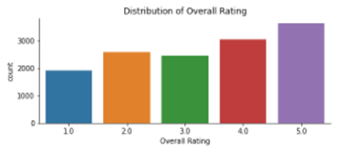
Our primary dataset is the Centers for Medicare and Medicaid Services (CMS) Nursing Home Inspection data. It contains 18 tables containing a variety of information on over 15,000 US nursing homes. Our analysis focuses on one particular table “Provider Info”, which contains aggregated annual figures for all nursing homes in the dataset.

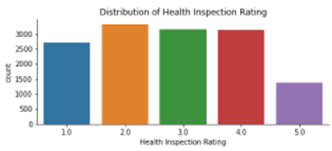
Even just this one table contains a massive amount of information with 87 columns containing a variety of numeric and categorical data. A brief description of the types of columns is listed below:

● Provider information columns- (name, address, federal id #, provider type, date established)

● Ratings calculated by CMS - (overall, staff, quality management, health inspection)

● General statistics columns - (fines, penalties, complaints, incidents, staff hours per resident)



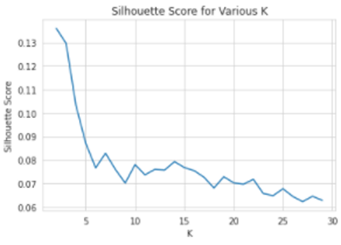
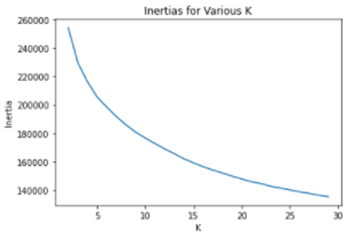


The above tables show the distribution of various ratings calculated by CMS. Detailed information on how these figures are calculated can be found in [this document](https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/CertificationandComplianc/Downloads/usersguide.pdf).

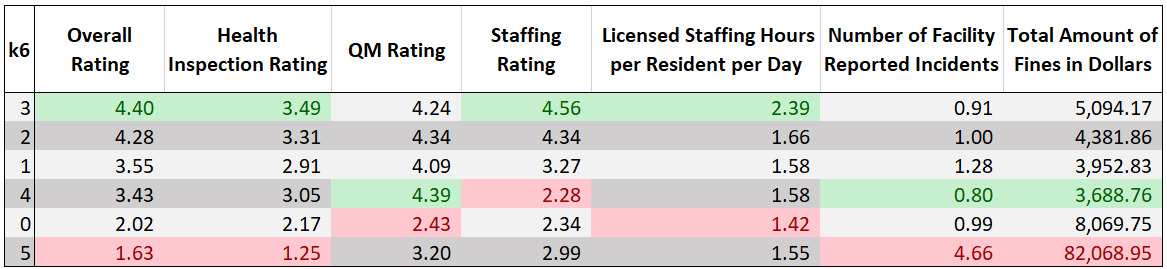
**Segmenting Potential Customers**

One issue with this dataset comes from the vast number of features, many of which are highly correlated or generally very similar (e.g., *reported* Nursing Staff hours vs *adjusted* nursing staff hours). Therefore, before running our clustering analysis, we decorrelated our data by running our columns through Principal Component Analysis. For the 31 numeric columns we analyzed, we were able to explain 95% of the variance of the rows with 15 columns, markedly reducing the feature space. However, the five most useful principal components, still explain less than 2/3rds of total variance, indicating that there is much more information in our data than can be accurately summarized in just a few columns.

Next, we ran our PCA columns through a range of K-Means clustering algorithms with different values of K to determine the ideal number of clusters to identify. The figures below show the inertia and silhouette score of the various values of K. While the silhouette scores are not very high at any value of K, we ultimately decided on segmenting into 6 clusters since that seems to be about where the elbow point comes in the inertia plot, and because it would be a reasonable number of segments to identify and profile.



We then fit a K-Means model with K=6 and generate labels to attach to our original dataframe. The table below shows the mean scores of each cluster for a few key columns.



With our market segmented, we can develop strategies tailored to a specific cluster, and apply these strategies to similar nursing homes. For example, providers in cluster 4 on average have the highest QM rating, but lowest overall staff rating of any cluster. We could target companies in this cluster with a strategy aimed at improving their staffing.

See below for a brief profile of each identified cluster.

Cluster 3 (1179 facilities):

● Has highest average overall rating of any group (4.40) out of 5

● Highest share of licensed staffing hours per resident (almost 50% higher than next closest group)

● Much more likely to be run by a church (6.5% of facilities in this group vs 2.2% in all other groups)

Cluster 2 (2345 facilities):

● Has average overall rating almost as high as cluster 5 (4.28)

● Achieves high rating with significantly less staff time per resident than cluster 5

● Highest average RN staff rating

Cluster 4 (2332 facilities):

● Middle of the pack average overall rating (3.43)

● Has highest average Quality Measure (QM) rating of any cluster

● Lowest average staffing rating of any cluster drags down high quality rating

● Highest average number of residents per day of any cluster

Cluster 1 (2447 facilities):

● Middle of the pack average overall rating (3.55)

● Highest average number of facility reported incidents of any cluster, other than cluster 0

Cluster 0 (2480 facilities):

● Very low average overall rating (2.02)

● Has the lowest average QM rating of any group by far

● Lowest average licensed staffing hours per resident

Cluster 5 (1176 facilities):

● Lowest average overall rating of any cluster (1.63)

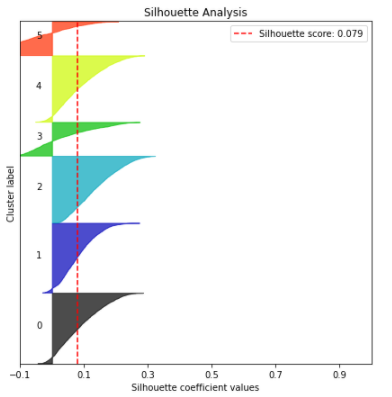
● Abysmal average health inspection rating (1.25)

● Average total fines for this group is $82,000, more than 10x the average for any other group

● Does have higher average staffing rating than cluster 4

**Cluster Evaluation**

With our segments identified, the next question is how well our clustering analysis has done identifying distinct groups. To answer this question, we use a Silhouette Score to and plot to measure how close each facility is to its next closest cluster.



As shown in the plot above, most of the points have a positive (though not very high) Silhouette score. However, two of the clusters (5,3) account for almost all of the negative scores, while the other four clusters are much more distinct.

**Review Mining and Text Analytics**

In addition to the datasets used, we tried to incorporate reviews of nursing homes. Our goal being to see if we can use text to get the overall sentiment of reviews of each nursing home, and see if there is a relationship with the score of the nursing homes. We used scrapy to gather nursing home reviews from caring.com, and selected random nursing homes from Los Angeles, California. Our sample had 11 nursing homes with a total of 57 reviews. To extract sentiment from the reviews, we used lexicon approaches with Afinn and Vader. For Afinn scores, we summed the scores of every word from the output to get an overall score of the review. When comparing the different nursing homes, Hollenback Palms and Broadway Manor Care had the highest sentiment scores; that being said, their overall scores were 4 and 3, respectively, out of 5. This could be explained for several reasons. One reason could be that lexicon based approaches may not be the best in this situation ; one review said that their mother was covered in her feses for over 4 hours, but because that word is not covered in the dictionary, afinn gave the review a score of 0. To improve upon this work, one can try to create document term matrices and try a supervised learning approach to classify sentiment. A second reason could be that different nursing homes have more reviews than others, which could help improve a nursing homes sentiment score, while also putting others at a disadvantage.

**Conclusions**

We set out to solve a very ambitious problem with analyzing the enormous amount of information captured in our dataset. While the solutions we generated were less obvious than a lot of the examples covered in class, we feel we still generated critical insight with a much more real-world set of data.

Our clustering solution identified a few segments of customers with distinctive characteristics that would allow us to develop a marketing strategy tailored to each specific group. While our evaluation of the clustering solution showed that the distinction between facilities on average was not very high, we also saw in our attempt at dimension reduction that our dataset contained a vast amount of information in many columns. With so many distinct features, it is not surprising that using all of them to develop clusters would lead to potential overlap.

We also saw how we could use text analytics to build our own dataset of reviews and sentiment scores for nursing homes. We were able to merge some reviews with our main dataset and see how their sentiment scores compared with their overall rating as calculated by CMS. As a next step, we could gather a larger corpus of reviews and see the average sentiment scores for each of our identified clusters. Finally, we could identify nursing homes with relatively negative sentiment compared to their actual performance and offer advice on how to improve online ratings.