Steffany Bahamon

IEMS 308, F16

Clustering Homework

Executive Summary

More than 53 million of Americans are Medicare beneficiaries; this is nearly 20% of the current US population. With such a massive amount of people, it’s easy to commit fraud or other unethical actions. One of these potential unethical actions is paying women systematically less than men, something that may be measured by comparing the pay of women and men who work the same amount of hours.

Clustering is an analytical strategy that groups observations with similar attributes together; it’s often used in market segmentation and outlier detection applications. The US Government provides all Medicare Provider information per year online in the form of a large csv file containing over 200,000 providers . This dataset provides a plethora of information on every provider, from location to percentage of patients with specific illnesses to number of patients treated. By clustering providers based on their standardized amount paid, number of patients treated, and gender, the US government can ensure it isn’t systematically underpaying its female Medicare providers, and in doing so avoid a potentially costly class action lawsuit. After comparing correlation and histogram data for each feature, the dataset was clustered to observe if women were being paid less than men. Unsurprisingly, they were not. Clusters appeared to split women and men each into 6 clusters mainly among number of patients served and payments received; this was consistent with the correlation data and heatmap produced. However, some of the clusters seemed to indicate a connection between few services and high pay; this is not gender related, and is more likely related to the physician’s type.

Problem Statement

In the working world, women are often presumed to work for less pay and work for less hours than their male counterparts. Our job was to cluster the dataset given among gender, number of services worked, and total standardized amounts charges, to see if there Ire clear groupings of low paid mostly women physicians. With this information, the government can ensure that at least the “pay gap” won’t affect a group of workers paid by the government.

Assumptions

* All physicians can be classified as male or female
* Standardized payment given is correct - the data in the dataset is correctly standardized with regards to geographic location.
* All services cost roughly the same
* All services take roughly the same amount of time to perform
* Physician specialization doesn’t affect time servicing patients

Methodology

Our first step to begin clustering was to reduce the dataset given by selecting only observations that belonged to individuals and not entities along with reducing the set of features to only gender, number of patients seen, and standardized amount paid. I converted the gender column of the dataset into a binary 1/0 column where 1 was equivalent to male, and I filled in every blank value with a 0. I used Python’s pandas and numpy packages in order to convert this information into a array saved into a csv file that was reused in every separate instance. After creating a purely numerical dataset that could easily be recalled as a csv file for other functions, my dataset was ready to begin analysis by comparison of statistical measures and k-means clustering

Analysis

The first step of my analysis involved comparing the attribute metrics to each other statistically. ￼￼￼

|  |  |  |  |
| --- | --- | --- | --- |
| Statistic | Gender of the Provider | Number of Services | Total Medicare Standardized Payment Amount |
| count | 924491.000000 | 924491.000000 | 924491.000000 |
| mean | 0.599420 | 2.225114e+03 | 8.187080e+04 |
| Standard deviation | 0.490016 | 1.218669e+04 | 1.885941e+05 |
| min | 0.000000 | 1.100000e+01 | 0.000000e+00 |
| 25% | 0.000000 | 1.970000e+02 | 1.173207e+04 |
| 50% | 1.000000 | 5.770000e+02 | 3.384939e+04 |
| 75% | 1.000000 | 1.657500e+03 | 8.759365e+04 |
| max | 1.000000 | 3.452292e+06 | 1.520900e+07 |

The table indicates there is a lot of variability in number of services provided as well as payment.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Gender of the Provider | Number of Services | Total Medicare Standardized Payment Amount |
| Gender of the Provider | 1 | 0.06175806 | 0.1630152 |
| Number of Services | 0.06175806 | 1 | 0.64138337 |
| Total Medicare Standardized Payment Amount | 0.1630152 | 0.64138337 | 1 |

The correlation matrix indicates a tie between number of services provided as well as pay.

The clustering data supports the correlation connection. Due to the size of the data, I chose to use K means clustering with 6 clusters; I determined 6 clusters by looking at the elbow plot.

The clustering centroids are below. The histogram below that indicates most points are in centroid 1.

Centroids = [[ 8.78355179e-01 1.06999350e+04 4.12173501e+05]

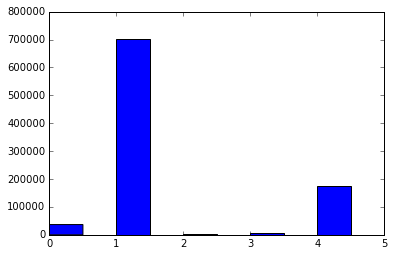
[ 5.31137209e-01 6.40607021e+02 2.85667002e+04]

[ 8.76777251e-01 1.13637825e+05 2.19553923e+06]

[ 8.49919153e-01 4.39196593e+04 1.04256551e+06]

[ 8.01133836e-01 3.43735897e+03 1.54416419e+05]

[ 9.29539295e-01 1.97111448e+05 4.54256064e+06]]



Conclusions

**Male and female medicare providers do not get paid statistically significantly different.**

Neither statistical correlation nor clustering showed that men and women medicare providers get paid differently.

**There is a strong connection between number of patients seen and total payment.**

Providers were clustered mainly on those two features; there was also a strong correlation factor relating those two attributes.

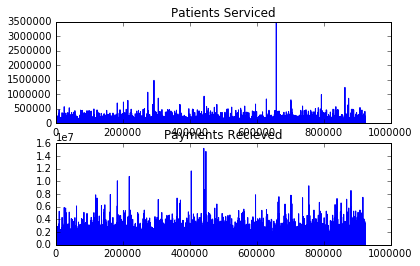
As such, there is no basis of gender discrimination by the federal government on how it pays its medicare providers.

Next Steps

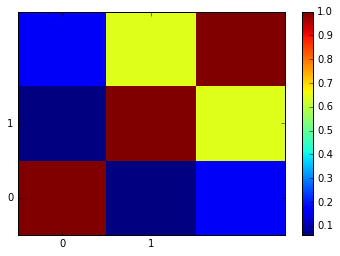
Despite the correlation between seeing more patients and payment, there is still uncertainty there that isn’t accounted for; I suspect this is due mainly to the physician type. Additional analysis would incorporate that as well as procedure provided into clustering analysis.

APPENDIX

Histogram of features per physician



Correlation Heatmap:



Elbow Plot:

