Matthew Shipley

IEMS 308 – Professor Klabjan

10/12/2016

**Clustering Medicare Data**

**Executive Summary:**

**Problem Statement:**

At the end of each year, the Recovery Audit Program (within Medicare) identifies and corrects improper payments (either underpayments or overpayments) to or from healthcare providers (“Recovery Audit Program”, 2016). Currently, there are four Recovery Auditors for four different regions in the United States:

Region A: Performant Recovery Region C: Cotiviti Healthcare

Region B: CGI Federal, Inc. Region D: HealthDataInsights, Inc.

Because of recent budget cuts, the Recovery Audit Program is looking for ways to optimize their method for detecting and distributing the healthcare providers going through audit to these four Recovery Auditors. The Recovery Audit Program has come to us, 308 Partners, to determine four new geographical regions in order to minimize the number of providers under audit and evenly distribute the audits to each of the four companies.

**Assumptions:**

* All information submitted to Medicare by the providers is accurate. In reality, if a healthcare provider is committing fraud, they may be playing with the numbers to tell a different story.
* While the Recovery Audit Program runs audits on both large health organizations and individual providers, only individual providers are considered in the outlier detection phase in order to simplify the problem.

**Methodology:**

In order to determine the geographical regions requested by the Recovery Audit Program, 308 Partners split the project into two phases. The first phase was determining which healthcare providers were to be singled out for audit for the upcoming year. This was accomplished by running an outlier detection clustering algorithm on the full *Medicare Physician and Other Supplier Aggregate Table* dataset (“Physician and Other Supplier Data CY 2014”, 2016). The second phase was plotting the outlier healthcare providers geographically, and using a second clustering algorithm on geographical location in order to determine the geographical regions covered by each auditing firm.

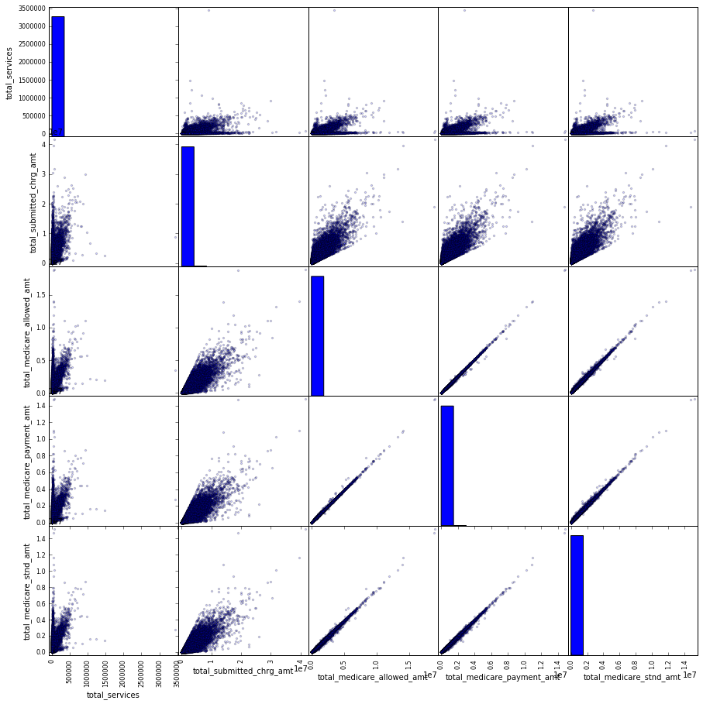
In the first phase of the project, several (mainly economic) variables were used in order to detect outliers in Medicare payment data. These variables were:

* Total provider services (*total\_provider\_services*)
* Total submitted charge amount (*total\_submitted\_chrg\_amt*)
* Total Medicare allowed amount (total\_medicare\_allowed\_amt)
* Total Medicare payment amount (total\_medicare\_payment\_amt)
* Total Medicare standard amount (*total\_medicare\_stnd\_amt*)

Since we are focusing the analysis only on individual healthcare providers, the data was further distilled by removing any tuple with “O” as the entity type. Each of the variables was first normalized in order to reduce the impact of differing scale of variables. To perform the clustering, the K-means algorithm from the scikit-learn Python package was utilized. The K-means algorithm was chosen because of the large size of the dataset: an algorithm using a distance matrix would consume too much memory to analyze almost a million records. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ different parameters for number of groupings were tested and plotted for \_\_\_\_\_\_\_\_\_\_\_\_ error measure. Based on where the kink in the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ plot was located, it was determined that the final algorithm should cluster into \_\_\_\_\_\_ segments. Based on examination of the scatterplot matrix, it was determined that cluster \_\_\_\_\_\_\_\_ represented the outliers.

The data points from the outlier cluster were then used in a second clustering algorithm in order to determine the geographical boundaries. To do this, the zip codes for each data point in the outlier cluster were converted to longitude and latitude coordinates. Another K-means algorithm was performed on this outlier group using four clusters (one for each of the four auditing firms). These clusters were plotted on a map to determine the geographical territories for each of the auditing firms under the Recovery Audit Program.

**Analysis:**

**

**Conclusions:**

**Next Steps:**

"Physician and Other Supplier Data CY 2014." *Physician and Other Supplier Data CY 2014*. CMS.gov, 5 May 2016. Web. 8 Oct. 2016.

"Recovery Audit Program." *CMS.gov*. Centers for Medicare and Medicaid Services, 15 Sept. 2016. Web. 8 Oct. 2016.