Matthew Shipley

IEMS 308 – Professor Klabjan

10/12/2016

Assignment 1: Clustering Medicare Data

Executive Summary:

Through this two phase clustering project for Medicare's Recovery Audit Program, we were able to determine four different regions to be covered by the four different auditing firms working under the program. Running a K-means algorithm first on the original dataset allowed us to determine financial anomalies in individual healthcare practices, i.e. the practices that need to be singled out for audits. These practices were then mapped geographically using the zip code translated into longitudinal and latitudinal coordinates. A second K-means algorithm was used to cluster these "high-risk" practices into four roughly equal geographical areas, which is the final deliverable to the client.

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 B: CGI Federal, Inc.

Region C: Cotiviti Healthcare

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.
- We're assuming that the first phase of the project is just an estimate, and will be used to determine the geographical boundaries in the second phase.

- The zip code coordinates correspond to the exact location of a practice. In order to convert to longitudinal and latitudinal coordinates, it was necessary to approximate the locations of each practice to the nearest 5-digit zip code.

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*)

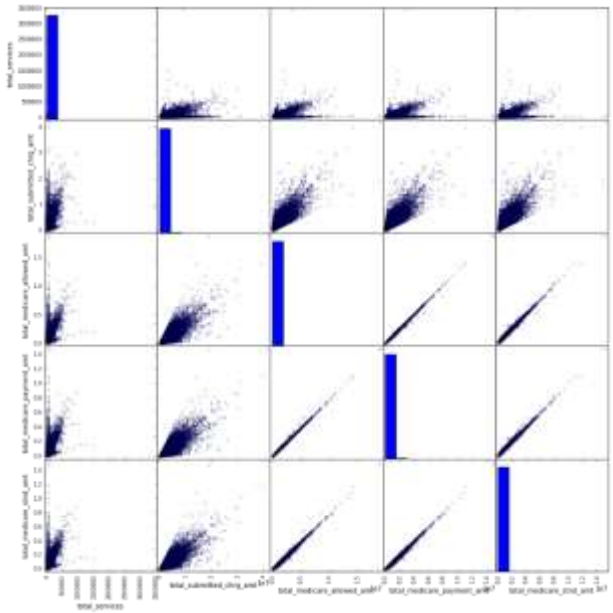
Due to the high correlation between the latter three variables, only the *total_medicare_stnd_amt* was used, as it corrects for geographical differences in payments. Since we are focusing the analysis only on individual healthcare providers, the data was further distilled by removing any tuple with "O" 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. After testing several values for the number of clusters in the K-means algorithm, it was decided to use 10 clusters to select our outlier cluster.

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.

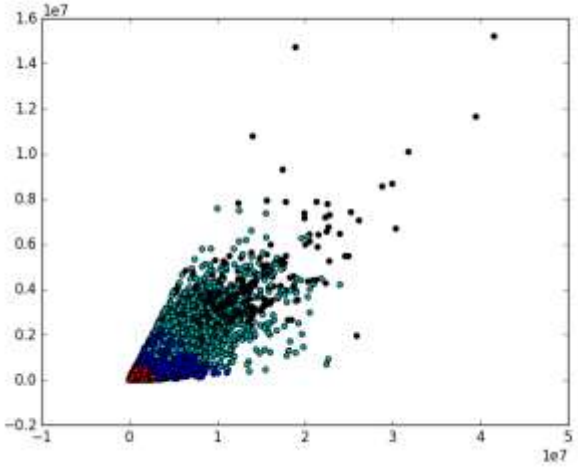
Analysis:

Below is a matrix scatterplot for the five variables that were initially chosen to examine. Due to the incredibly high correlation between *total_medicare_allowed_amt, total_medicare_payment_amt, and*

total_medicare_stnd_amt, only *total_medicare_stnd_amt* was utilized as it is corrected for geographical differences:



The following is a scatterplot of the original cluster analysis, with *total_submitted_chrg_amt* on the x-axis and *total_medicare_stnd_amt* on the y-axis.



And finally, the following is a scatterplot of the longitudinal and latitudinal addresses of the sample from the outlier cluster:

Next Steps:

To make a more detailed and accurate determination of the boundaries for each auditing firm, it would be beneficial to have information about previous audits and actual instances of fraud in Medicare reporting. If we had this information, we could use it to make more accurate predictions about which

healthcare practices need to be examined further. Additionally, having access to data of where previous audits occurred would prove even more useful as it would allow us to skip the first estimation-based part of this project.

Additionally, because of computational limitations, we had to take a further random sample of one of the outlier clusters, ideally we would use the entire sample to find all coordinates.

Sources:

"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.