Patrick Hall
Solution Abstract

2014 CCP: Data Scientist – Medicare Claim Anomaly Detection Challenge

Summary

I approached detecting anomalous patients, procedures, providers and regions in the Medicare Claim Anomaly Detection Challenge data using many different techniques – techniques that validated and augmented one another whenever possible. I also used several different types of data visualizations to explore the Challenge data and to assess my results.

Methods Summary

- Part 1: Descriptive statistics and straightforward data manipulation; box plots and pie charts
- Part 2: Deep neural networks, clustering, Euclidean distances, and linear regression; scatter plots
- Part 3: Association analysis, clustering, graph representations, matrix factorization; constellation plots, and donut charts

Technologies Summary

While size was not the most significant difficulty the Challenge data presented, the patient data was large enough to require special consideration. Moreover, producing the requested deliverables efficiently required the appropriate use of software tools and hardware platforms. I used disk-enabled, multi-threaded software tools coupled with a solid state drive (SSD) for data preprocessing and analysis in the first two parts of the Challenge. For part three of the challenge I used the same local environment and an MPP database appliance. Hadoop MapReduce and SAS are probably the most well-known disk-enabled data preprocessing and analysis software tools. I did not feel the moderate size of the Challenge data necessitated Hadoop.

- Platforms
  - Local: 24-core blade server with 300 GB SSD
  - Distributed: 24-node MPP Teradata appliance
- Source code management: Git
- Data preprocessing: bash scripting and Base SAS on local platform
- Part 1: Base SAS, SAS/GRAPH, and SAS/STAT on local platform
- Part 2: Base SAS, SAS/GRAPH, SAS/STAT, and SAS Enterprise Miner on local platform
- Part 3: Base SAS and SAS Enterprise Miner on local platform; SAS High Performance Data Mining on distributed platform

The following sections of this abstract describe in greater detail the approaches I used for data preprocessing and for completing the Challenge deliverables.

Data Preprocessing

I downloaded the summary CMS data in CSV format and imported it into SAS using standard DATA step programming. The PNTSDUMP.xml file contains numerous tables. I split the large file into separate tables using the bash applications grep, head, sed, and tail. I then imported each table into SAS using the automated XML LIBNAME engine. I used a brute force approach to import the ASCII delimited text files. My code read each character of the files, caching lines and tokenizing them using the respective ASCII record and unit delimiters. Importing single files took no longer than several minutes in all cases. All imported Challenge data was then validated with conventional techniques such as building frequency tables and analysis of missing and extreme values.

Part One Total Time Spent: 15 hours

Methods

To complete part one of the Challenge I used straightforward data manipulations and descriptive statistics to find anomalous procedures, providers and regions in the CMS summary data.

Testing and Validation

I deployed a simple checksum scheme to validate data manipulations. Subparts B, C, and D required either providers or regions to be grouped by procedure codes. One hundred-thirty unique procedure codes were available in the CMS summary data. My code counted distinct levels of procedure codes in tables built from several sorts and joins, always ensuring they summed to 130 distinct levels.

Discussion of Results

Part One – A

The three procedures with the widest variance in cost to the patient, whether the procedure was expensive or not to begin with, are Level I Excisions and Biopsies, Level I Hospital Clinic Visits, and Level II Eye Tests and Treatments. The results in Fig. 1 indicate that some providers are charging extremely large amounts for certain procedures, despite each procedure code being associated with a set level of severity and complexity and each procedure having relatively low mean and median costs.

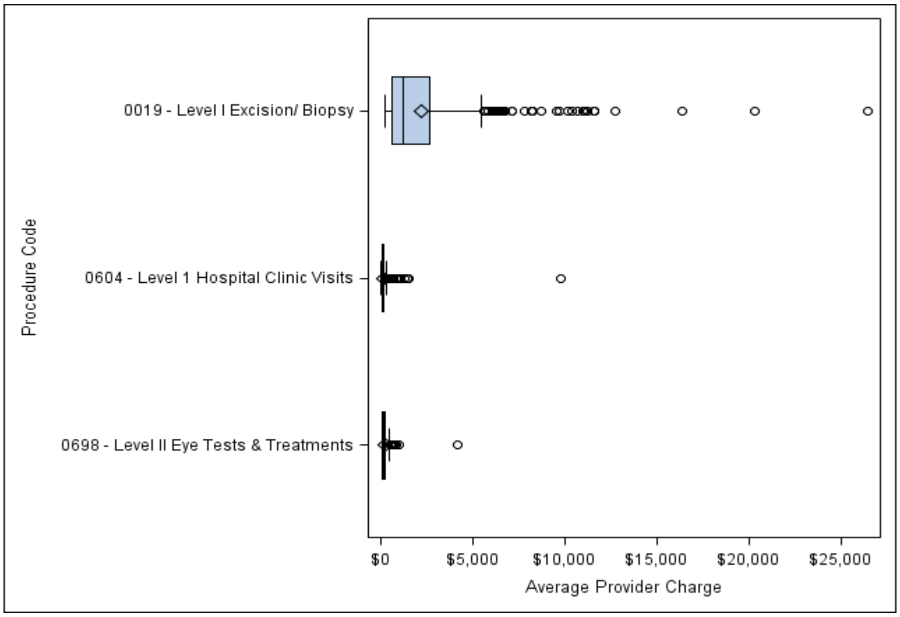

Figure 1: Claimed charges for the 3 procedures with the highest Coefficient of Variation (relative variation).

Notable high cost outliers include Centinela and Whittier Hospital Medical Centers, which are both charging an average of over $20,000 for Level 1 Excisions and Biopsies. Lower Bucks Hospital is charging an average of $9780 for Level I Hospital Visits, and Ronald Reagan UCLA Medical Center is charging an average of $4187 for Level II Eye Tests and Treatments. Further research should be undertaken to understand whether these procedures have inherently variable costs or whether some providers are simply overcharging.

Part One - B and C

The three providers who claim the highest charges for the most number of procedures are Bayonne Hospital Center, Crozer Chester Medical Center, and Stanford Hospital. While it is logical that a large, well-respected hospital like Stanford would account for a substantial number of the highest cost procedures, it is unclear why smaller providers like those noted in Fig. 2a should account for such a large proportion of the highest cost procedures. These providers' pricing practices should be investigated further.

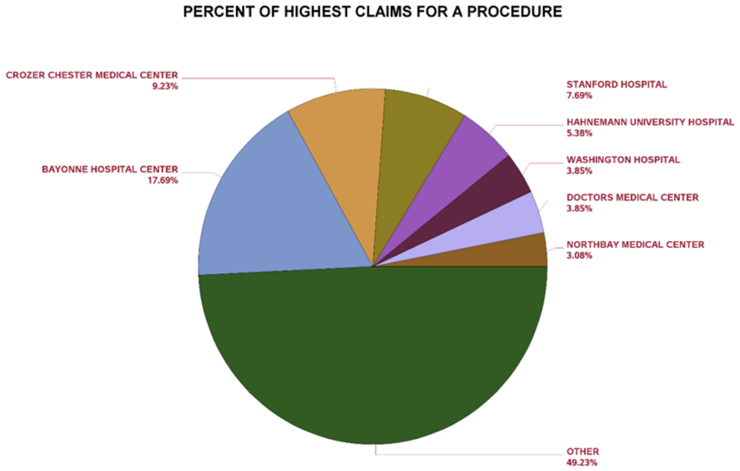

Figure 2a. The percent of procedures for which the noted provider charges the highest amount.

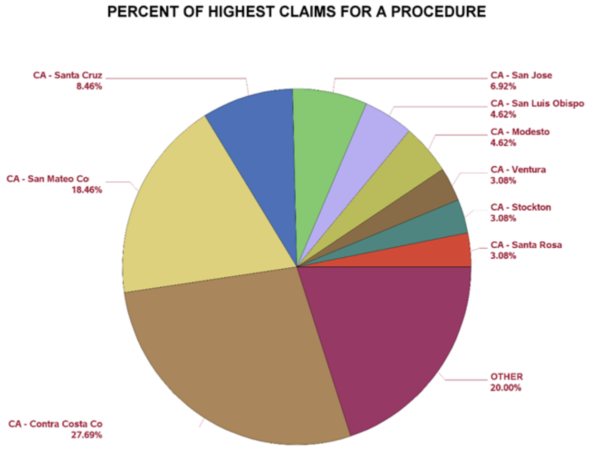

Figure 2b. The percent of procedures for which the noted region has the highest charges.

The three regions in which patients are charged the highest amount for the most procedures are Contra Costa County, San Mateo County and the Santa Cruz region, all in CA. While such geographical clustering could be representative fraud, these findings are more likely indicative of the high cost of living in these areas. As 9 regions in CA account for 80% of the highest cost procedures, research into the high cost of health care in the state could result in considerable savings.

Part One – D

The three providers with the highest number of procedures with the largest difference between their claimed charges to patients and their reimbursement from Medicare are Bayonne Hospital Center, Crozer Chester Medical Center, and Hahnemann University Hospital. Disproportionate differences between claimed charges and Medicare reimbursement may be another indicator of the high cost of living in the suburbs of major east coast cities where these providers are located. However, these providers may warrant more detailed investigations because they are outside the known anomalously high cost regions of California and they all account for a disproportionate number of the absolute highest cost procedures identified in part one - B.

Part Two Total Time Spent: 30 hours

Methods

I used simple feature engineering techniques to generate binary indicators to flag the provider as a university hospital or a region as containing a university hospital and to create interval features for the number of procedures of each level performed by a provider or in a region. Existing outpatient level information was extracted from the procedure codes and new levels were assigned to inpatient procedures based on the presence of chronic conditions. Pearson correlation was measured between all the provided and engineered features and one feature each from a small number of correlated pairs was rejected from further analysis to eliminate redundancy.

I combined two distance-based unsupervised learning approaches to identify points that were the most different from all other points. I applied both approaches to the provider and region summary data. I first calculated the entire Euclidean distance matrix of the given feature space. Using the mean of each feature as the origin of that space, I identified those points farthest from the origin of that space. To supplement these findings, I applied k-means clustering to the same feature space, using a patent-pending method from my own research known as the Aligned Box Criterion (ABC) to estimate the best number of clusters for given feature space. The clustering results allowed me to pinpoint the furthest Euclidean distance outliers that also formed their own cluster far from other clusters. To visualize the combined results of both approaches, I used a deep neural network, known as a stacked denoising autoencoder, to project the newly labeled points from the feature space into a two-dimensional space.[1]

Because the furthest Euclidean distance outliers seemed potentially uninteresting from a fraud detection perspective, I used traditional regression outlier analysis on the provided summary data to identify several other anomalous points.

Testing and Validation

K-means clustering was used to validate the findings from the direct calculation of the full Euclidean distance matrix. A deep neural network was used to project the combined results into a two dimensional space for further exploration and validation. Regression outlier analysis also found the identified Euclidean distance outliers to be leverage points, while the regression outliers with large studentized residuals were found to be points residing at the edges of the larger clusters in the cluster analysis.

Discussion of Results

Part Two – A

The three providers that are the least like all others are the Cleveland Clinic, UCSF Medical Center, and Lahey Clinic Hospital. Figure 3a is a two dimensional projection of the provider feature space which contains Medicare billing information, the number and severity of procedures billed, and an indicator of whether the provider is a university hospital. In both this feature space and the two dimensional projection, the Cleveland Clinic, UCSF Medical Center, and Lahey Clinic Hospital reside in their own cluster that is far from the origin of the space and far from all other clusters. That these points were placed in their own clusters by the k-means method is of special significance given that k-means prefers spherical clusters of a similar size, and that the ABC method was used to estimate the best number of clusters for this analysis. In short, there is statistical support for the hypothesis that these providers are unique.

I also performed a more traditional regression outlier analysis in hopes of identifying lower profile providers who are different in more subtle ways. I used the studentized residuals and leverage points from a regression of claimed charges against Medicare reimbursement and number of procedures billed to locate providers that are charging disproportionately high prices for the amount of Medicare reimbursement they are receiving and the number of procedures they are billing. These potentially suspicious providers are: Bayonne Hospital Center, Doctors Hospital of Manteca, and Delaware County Memorial Hospital. Figure 3b presents these providers as points with high studentized residual values and a low leverage values.

[1] Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." Science 313.5786 (2006): 504-507

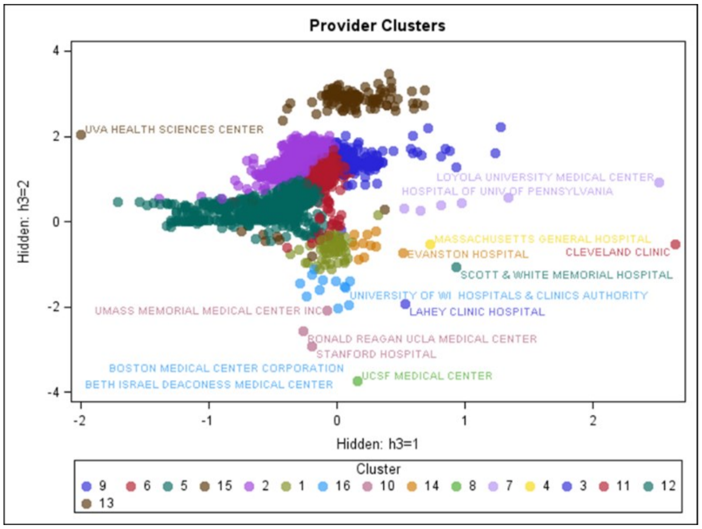

Figure 3a. Provider clusters projected into 2 dimensions with labeled outliers.

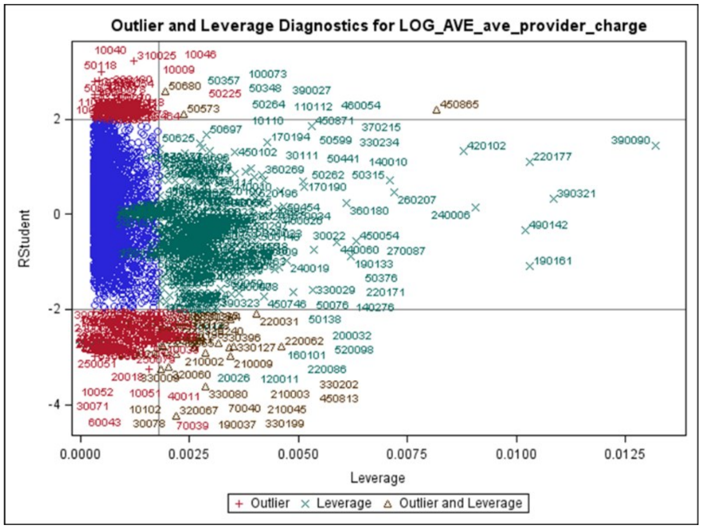

Figure 3b. Providers plotted by studentized residual and leverage.

Part Two – B

Following the same logic and approaches as in subpart A, the regions that are the most different from all other regions are Boston, MA, Cleveland, OH, and Los Angeles, CA. The most suspicious regions are Palm Springs, CA, Hudson, FL, and Tyler, TX.

Part Three Total Time Spent: 25 hours

Methods

Matrix factorization followed by cluster analysis along with A Priori association analysis were used to group unlabeled patients with patients flagged for manual review. I transformed the patient transaction data to represent a spare matrix in dense coordinate list (COO) format. The sparse matrix was decomposed directly from the COO representation into 10 singular value decomposition (SVD) features. These SVD features were merged with numeric encodings of patient demographic data and 1000 k-means clusters were created. Unlabeled patients in clusters with a high proportion of patients selected for manual review were given a non-zero fraud score. The higher the proportion of patients flagged for manual review in a cluster, the higher the unlabeled patients in that cluster were scored for possible fraud.

A Priori association analysis was used to identify frequent sets of procedures amongst the general patient population and the patients flagged for manual review. Frequent sets of procedures within the flagged patient group that were infrequent in the general population were assumed to be evidence of anomalous behavior. An unlabeled patient’s fraud score was incremented for each anomalous set of transactions they participated in. To create a final ranking of the 10,000 most suspicious patients, the fraud scores from both the cluster and association analyses were combined with approximately equal weighting and the patients with the highest overall fraud scores were submitted.

Testing and Validation

Patients who were identified as potentially anomalous by both cluster analysis and association analysis were the most likely to be submitted for additional review. Cluster results were profiled and the clusters containing the highest proportion of manually flagged patients were found to be homogenous. Graph representations of the frequent transactions in the general patient population and the patients flagged for manual review were generated and found to be dissimilar.

Discussion of Results

The six patient clusters with the highest proportions of patients flagged for manual review, and therefore the six most suspicious patient clusters, were found to be homogenous groups composed primarily of higher income females in the 65-74 age range as in Fig. 4. Several dozen additional clusters of anomalous patients were identified and these exhibited varying demographic characteristics.

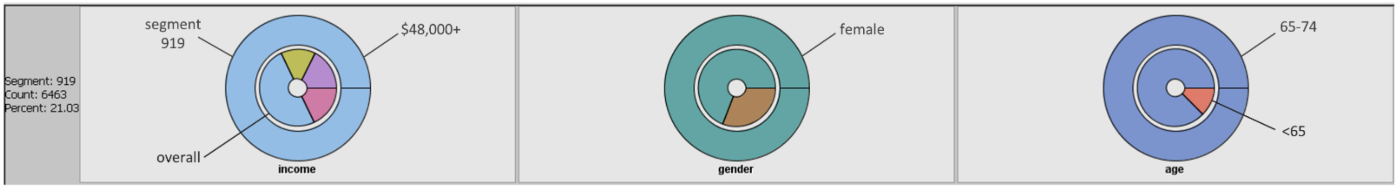

Figure 4: Donut charts representing the overall distribution of patient demographics versus the demographics within a suspicious cluster of patients.

The frequent transactions of the general patient population indicate most patients received one of several most frequent procedures and a small number of other less frequent procedures. Manually flagged patients often received an equal number of many different procedures – a pattern I used to identify possible fraud.

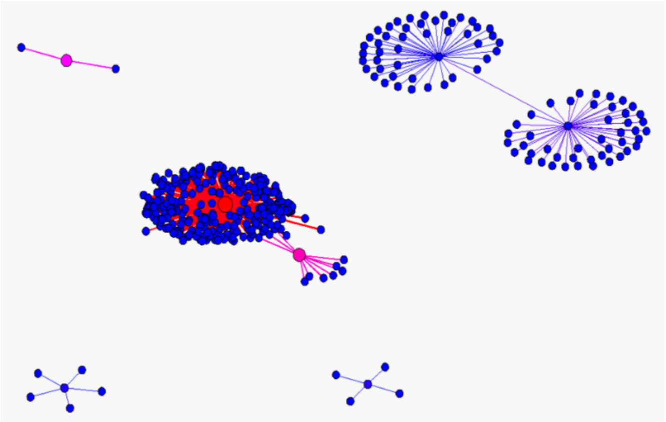

Figure 5a. Constellation plot representing the graph of frequent transactions in the general patient population. Larger node and link size and brighter node and link color represent increasing frequency.

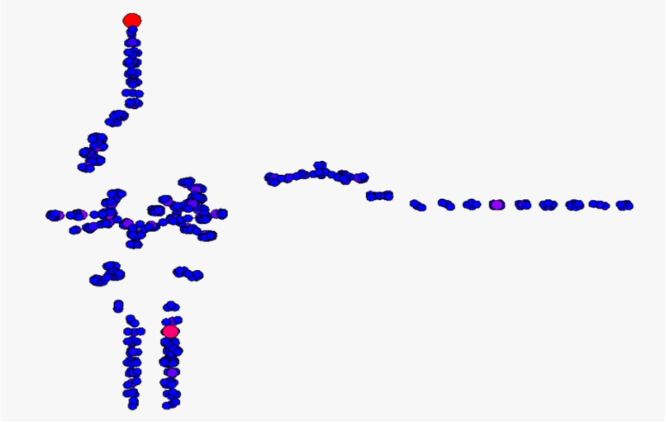

Figure 5b. Constellation plot representing the graph of frequent transactions in the patient population flagged for manual review. Larger node and link size and brighter node and link color represent increasing frequency.