Predicting anomalies in the healthcare system

SUMI CHOUDHURY, RADHIKA NAYAR, APARNA SUNDARAM

(IN ALPHABETICAL ORDER OF LAST NAME)

19 SEPT, 2019

Overview

- Description of the data
- Managing the file size
- Matching the NPIs and avoiding misclassification
- Adjusting for sample imbalance
- Feature engineering, feature selection and diagnostics
- Cluster analysis
- Model building

Data

- Data from:
  - Center for Medicare and Medicaid Services (CMS)
  - DHHS' Office of Inspector Generals' Exclusions Database (LEIE)
- CMS data over 7 years has 56 million records
  - There are multiple records per provider
- 2017 CMS file has data for 1.03 million unique providers
- LEIE file has 73,000 records
  - Mostly unique to indicate only one problem per provider
  - About 100 duplicates indicating multiple problems with a provider

Managing File Size

- CMS file with 56 million records too large for average laptop computing
- Attempts led to crashes and failures
- 2017 CMS file with about 9.5 million records also too large
  - Variables needed to be extracted individually to be checked
- First sample created of 1.2 million records
- Strategy: Sample of 200,000 cases from each of the 7 years
- File size: over 2 GB

Managing File Size

- After additional consultation sample file revised
  - More in line with needs for analysis
- Second sample of about 1.5 million records
- Sampling strategy:
  - Used only the 2017 CMS file
  - Identified the 10 HCPCS codes that had the most submitted charges
  - Accounted for about 20% of all charges submitted
- File size: ~2 GB

Matching NPIs With LEIE File

- NPI is the provider identification number
  - Unique to each provider
- CMS files had complete information on provider NPI
- LEIE files had only 5000 non-missing NPI
  - All other records had missing values
- Inner-join of two files allowed linkage of 137 unique NPIs
  - No clarity on level of misclassification of remaining cases
- Required additional investigation

Checking for Misclassification

- Additional investigation done by partitioning data
  - Facilities/Organizations (< 10,000 records in CMS file)
  - Individuals
- Misclassification checks done on Individual level file
- Strategy: Narrow search to records matching on:
  - Zip code
  - Street address
  - Last name

Checking for Misclassification

- Data cleaning done on zip code and street address
- CMS Zip codes mostly 9 digits and LEIE mostly 5 digits
  - Variation in zip code length in both files for a minority of records
- Street address cleaned to remove stop words
- Both files narrowed down to records common on these variables
- Inner-join on the subset file allowed for additional match on 4 unique NPI
- All other records were truly excluded
  - No misclassification

Adjusting for Sample Imbalance

- Exclusions: 141 unique NPIs were excluded
- Sample size: 1.5 million
  - Unbalanced file on target variable
- Imbalance addressed by weighting
- Weights:
  - 1/ Excluded as a proportion of whole sample (large weight)
  - 1/Non-Excluded as a proportion of whole sample (very small weight)

Feature Engineering

- Multiplied the average charges submitted, amount Medicare paid, amount Medicare allowed by line service count variable
  - This gave us the total charged or paid by HCPCS code by provider
- Took difference between…
- Took ratio between…
- Final file transformed from long to wide
  - Multiple observations per NPI now just one record per NPI
  - Each row associated with HCPCS code now new column

Additional Feature Engineering

- Took natural log of all continuous variables in wide-file
  - Not feasible with difference variable due to negative values
  - Ratios seemed best for modeling
- Categorical variables such as states and provider types condensed
- States re-categorized as regions using Census Bureau classification
  - 4 regions: Northeast, Midwest, South, West
- Provider Type reclassified by level of specialization
  - Basic care/PCP/low specialization
  - Specialists
  - Super-specialists

Diagnostics

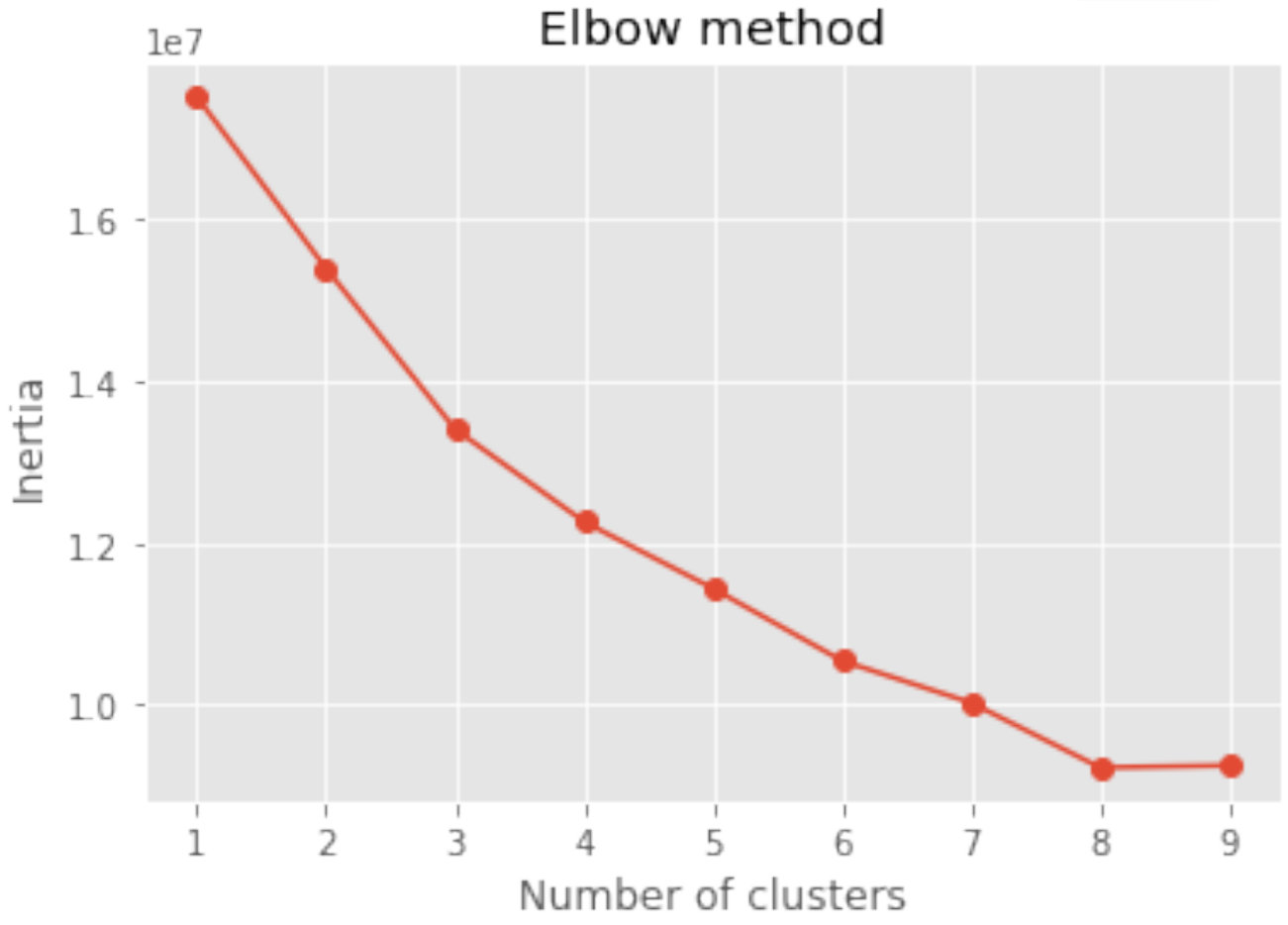
- Correlation between the ratio variables checked
  - Within each set, the ratio variables are almost completely uncorrelated
  - Between the two groups high correlation by HCPCS code (80%-high90%)
  - Argument for selecting only one of the two
- Correlation between the two counts variables checked
  - Within each group, correlation is very low
  - Between the two groups, high correlation by HCPCS code (mostly high 90%)
  - Argument for selecting only one of the two

Cluster Analysis

- Cluster analysis ran on initial seed of 5
- Scree plot indicated that a larger number of clusters is needed
  - Problems with determining exact number due to file size
  - Current scree plot indicates that at least 8 are needed
- Cluster Analysis re-run using k=8
- Clusters output as a new variable in file

Scree Plot

- Indicates that 8 is a good number
- Unable to run plot for higher number of clusters



Diagnostics on Clusters

- Clusters significantly associated with exclusions
- Significantly associated with ratio of paid to submitted charges
  - All HCPCS codes significant except A0427
- Significantly associated with all regions and provider types
- Significantly associated with beneficiary count
  - All HCPCS codes significant except A0427
- Significantly associated with gender

Models – Logit model

- Logit model run with Exclusions/non-Exclusions as target variable
- Variables included in the model on the RHS are:
  - Ratio of paid to submitted charges by HCPCS code (10 vars)
  - Count of unique beneficiary counts by HCPCS code (10 vars)
  - Regions of the country (4 vars)
  - Provider type (3 vars)
  - Gender
  - Cluster
- Analysis was weighted

Logit Model Coefficients

- All variables are significant except:
  - Ratio of paid to submitted for HCPCS code A0427
  - Beneficiary unique count for HCPCS code A0427 and 99214
  - Provider type
- Importantly, the clusters are significant
  - Clusters predict whether or not a provider is excluded