Predicting anomalies in the healthcare system

SUMI CHOUDHURY, RADHIKA NAYAR, APARNA SUNDARAM

(IN ALPHABETICAL ORDER OF LAST NAME)

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
- Project takeaways

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

Sample - Percents Of Selected Columns

Top10_hcpcs	Total_submitted_percent	Total_beneficiary_percent	Total_line_service_percent
99214	6.438335	6.345879	4.194636
99213	4.076711	6.413324	4.001411
99285	3.940294	1.294201	0.485196
66984	2.695674	0.237566	0.317766
99232	2.398989	1.930589	1.936963
99233	1.795099	1.112096	0.935694
A0427	1.720677	0.408946	0.209624
99223	1.482774	1.111840	0.428854
99291	1.425083	0.388740	0.223256
99284	1.187249	0.637510	0.236281

Sample - Top 10 hcpcs description

TOP 10 HCPCS	DESCRIPTION OF TOP 10 HCPCS
99214	Established patient office or other outpatient, visit typically 25 minutes
99213	Established patient office or other outpatient visit, typically 15 minutes
99285	Emergency department visit, problem with significant threat to life or function
66984	Removal of cataract with insertion of lens
99232	Subsequent hospital inpatient care, typically 25 minutes per day
99233	Subsequent hospital inpatient care, typically 35 minutes per day
A0427	Ambulance service, advanced life support, emergency transport, level 1 (als 1 - emergency)
99223	Initial hospital inpatient care, typically 70 minutes per day
99291	Critical care delivery critically ill or injured patient, first 30-74 minutes
99284	Emergency department visit, problem of high severity

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:
  - Total submitted and total std. payment
  - Total allowed and total std. payment
- Took ratio between:
  - Std. pay/submitted
  - Std. pay/allowed
  - Submitted charges/allowed

Feature Engineering:

```
dfcombo['Total_submitted_chrg_amt'] = dfcombo['line_srvc_cnt'] * dfcombo['average_submitted_chrg_amt']
```

```
dfcombo['Total_Medicare_std_payment_amt'] = dfcombo.line_srvc_cnt * dfcombo.average_Medicare_standard_amt
```

```
dfcombo['Total_Medicare_allowed_amt'] = dfcombo.line_srvc_cnt * dfcombo.average_Medicare_allowed_amt
```

```
dfcombo['Net_submit_pay'] = dfcombo.Total_submitted_chrg_amt - dfcombo.Total_Medicare_std_payment_amt
```

```
dfcombo['Net_allow_pay'] = dfcombo['Total_Medicare_allowed_amt'] -  dfcombo['Total_Medicare_std_payment_amt']
```

```
dfcombo['ratio_pay/submit'] = dfcombo.Total_Medicare_std_payment_amt/ dfcombo.Total_submitted_chrg_amt
```

```
dfcombo['ratio_pay/allowed'] = dfcombo.Total_Medicare_std_payment_amt/ dfcombo.Total_Medicare_allowed_amt
```

File Transformation

- Final file transformed from long to wide form:
  - Each row is a unique NPI
  - Shape: (648778, 128)

- 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

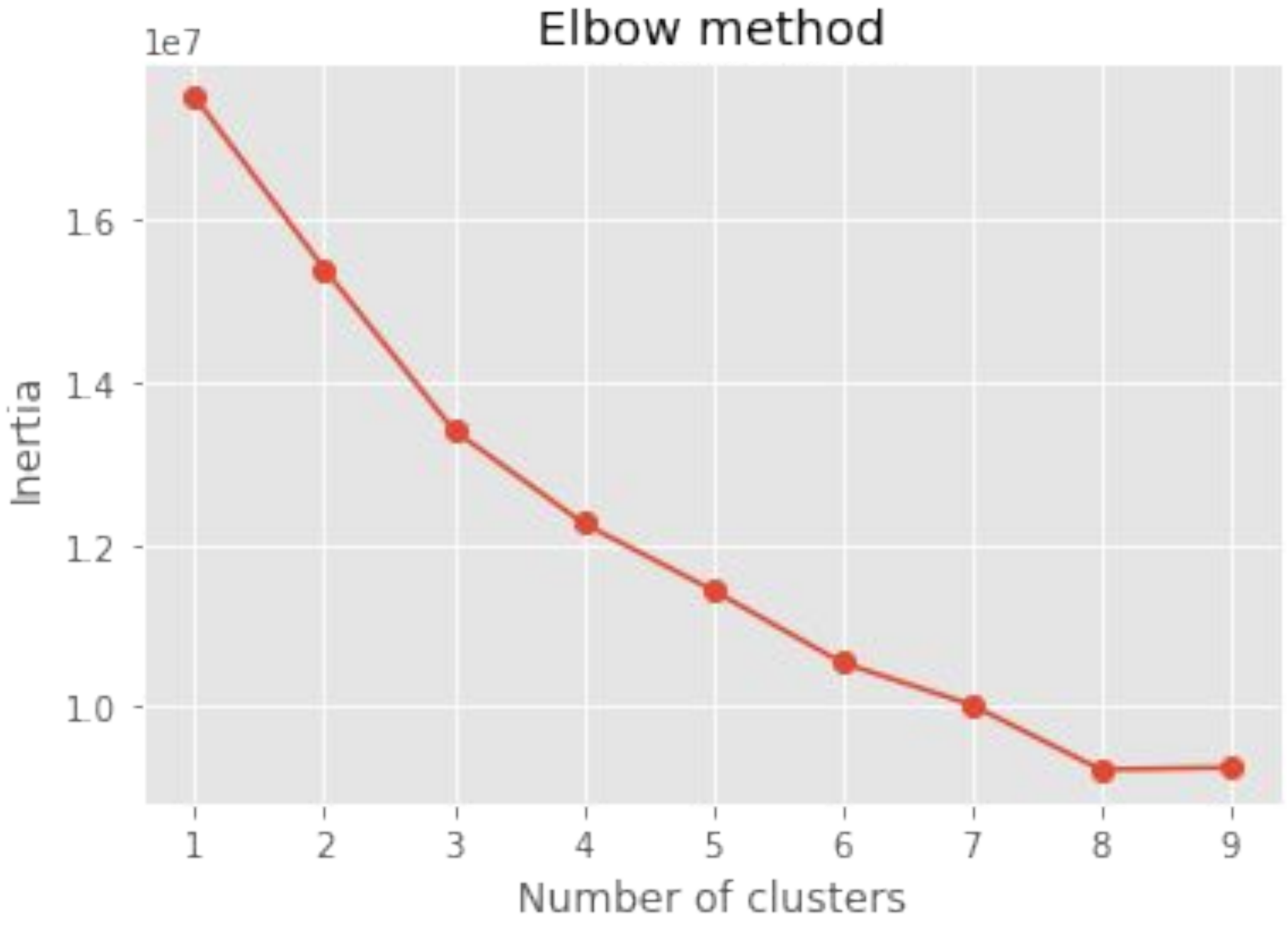
- Correlation between the ratio variables checked
  - Within each set, the ratio variables are almost completely uncorrelated
  - Between the two groups high correlation by HCPCS code (80%-high 90%)
  - Argument for selecting only one of the two
- Correlation between the two counts variables checked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

Excluded in each cluster

- Distribution of Exclusion and Non-Exclusion in 8 different clusters
- The 0s are the non-excluded
- The 1s are the excluded

	1	2	3	4	5	6	7	8
0	72166	84200	44970	93359	138917	52474	84114	78437
1	16	24	5	11	38	14	12	21

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

Logit Model Coefficients

```
Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.973   -0.621   -0.409   -0.219   64.284

Coefficients: (1 not defined because of singularities)
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             2.440e+07  4.038e+07   0.604   0.5457
LPS_66984              -1.384e+00  1.833e-02 -75.461  < 2e-16 ***
LPS_99213               5.909e-02  3.924e-03  15.058  < 2e-16 ***
LPS_99214               1.687e-01  3.811e-03  44.273  < 2e-16 ***
LPS_99223               1.133e-01  6.807e-03  16.640  < 2e-16 ***
LPS_99232              -2.296e-01  6.523e-03 -35.207  < 2e-16 ***
LPS_99233              -4.674e-02  7.477e-03  -6.251 4.09e-10 ***
LPS_99284              -2.074e-01  3.159e-02  -6.566 5.19e-11 ***
LPS_99285               1.730e-01  2.993e-02   5.781 7.44e-09 ***
LPS_99291              -3.461e-01  1.114e-02 -31.068  < 2e-16 ***
LPS_A0427              -2.302e+00  1.746e+00  -1.319   0.1873
northeast1              6.002e-02  1.771e-03  33.901  < 2e-16 ***
midwest1               -3.216e-02  2.028e-03 -15.858  < 2e-16 ***
south1                  7.326e-02  1.869e-03  39.200  < 2e-16 ***
west1                          NA         NA      NA       NA
basic1                 -2.440e+07  4.038e+07  -0.604   0.5457
specialist1            -2.440e+07  4.038e+07  -0.604   0.5457
sup_special            -2.440e+07  4.038e+07  -0.604   0.5457
genderM                 4.157e-01  1.322e-03 314.462  < 2e-16 ***
bene_unique_cnt_66984 -1.072e-03  4.560e-05 -23.508  < 2e-16 ***
bene_unique_cnt_99213 -2.414e-04  4.799e-06 -50.305  < 2e-16 ***
bene_unique_cnt_99214 -9.038e-06  5.127e-06  -1.763   0.0780 .
bene_unique_cnt_99223  8.374e-04  2.075e-05  40.362  < 2e-16 ***
bene_unique_cnt_99232 -1.285e-03  1.498e-05 -85.779  < 2e-16 ***
bene_unique_cnt_99233 -2.431e-04  2.061e-05 -11.797  < 2e-16 ***
bene_unique_cnt_99284 -2.038e-03  4.099e-05 -49.719  < 2e-16 ***
bene_unique_cnt_99285  4.377e-05  1.790e-05   2.446   0.0144 *
bene_unique_cnt_99291  7.247e-04  3.388e-05  21.393  < 2e-16 ***
bene_unique_cnt_A0427  1.573e-03  6.974e-03   0.226   0.8215
cluscol                -4.088e-04  4.526e-04  -0.903   0.3663
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.4040336)

    Null deviance: 325543  on 648777  degrees of freedom
Residual deviance: 262116  on 648749  degrees of freedom
AIC: 1251893

Number of Fisher Scoring iterations: 2
```

Random Forest Models

- Random Forest Models were run on all 648,778 observations in wide-form
  - Variables used in this model are ratio of pay over submitted charges, ratio of submitted over allowed charges, 4 regions: northeast, midwest, south, west, provider type, and gender: Male or Female
- Split into train versus test datasets
  - 60/40 split between train and test
- Set parameters for random forest classifier
  - Number of estimators (trees) = 100, maximum number of features = 10

Random Forest Models - Exclusion Types 1128a3 Feature Importance

	name	score
0	b'ratio_pay.submit_99214'	0.391543
1	b'ratio_pay.submit_99213'	0.248341
2	b'ratio_submitted.allowed_99214'	0.185345
3	b'ratio_submitted.allowed_99213'	0.123514
4	b'specialist'	0.010630
5	b'basic'	0.007406
6	b'gender_M'	0.005235
7	b'south'	0.004973
8	b'midwest'	0.003776
9	b'sup_special'	0.003687
10	b'gender_F'	0.003270
11	b'northeast'	0.003167
12	b'ratio_submitted.allowed_99223'	0.002949
13	b'ratio_pay.submit_99233'	0.001250
14	b'ratio_pay.submit_99232'	0.001247

Random Forest Models - All Exclusion Types Feature Importance

	name	score
0	b'ratio_pay.submit_99213'	0.256740
1	b'ratio_pay.submit_99214'	0.223600
2	b'ratio_submitted.allowed_99213'	0.200846
3	b'ratio_submitted.allowed_99214'	0.165104
4	b'ratio_submitted.allowed_99232'	0.016094
5	b'ratio_submitted.allowed_99284'	0.014030
6	b'ratio_pay.submit_99232'	0.012764
7	b'ratio_pay.submit_99223'	0.012291
8	b'ratio_pay.submit_99233'	0.012167
9	b'ratio_pay.submit_99285'	0.010846
10	b'ratio_submitted.allowed_99233'	0.010083
11	b'ratio_submitted.allowed_99291'	0.009662
12	b'ratio_submitted.allowed_99223'	0.009060
13	b'ratio_pay.submit_99291'	0.008795
14	b'ratio_pay.submit_99284'	0.008356

Project Takeaway

- Working with large datasets and recognizing the limitations
- Creating sub-samples
- Cleaning, coding and manipulating large data
- Understanding the full range of variables in both files
- Merging data and removing misclassification
- Building models on large data

Project Takeaway

- For Random Forest Models:
  - Use One Hot Encoding for all categorical variable levels (do not reduce features such as region or specialty); include more information such as zip codes levels and all provider types
  - Try H2O cloud based service to build a Random Forest model that can accomodate the dummified categorical variables

Project Frustrations

- Routine algorithms either did not run or took forever to run
- Error messages like Error: cannot allocate vector of size 1568.0 Gb
- Machine crashes

THANK YOU

- Please checkout our code on GitHub:
  - https://github.com/aparnasundaram/Medicare_anomaly_detection