PROJECT

PYTHON: HEALTH CARE CASE STUDY

ABSTRACT

Challenge involved detecting anomalies in the United States,Medicare insurance system. Finding anomalous patients, procedures, providers, and regions in the competition's large, complex, and intertwined data sets required industrial-strength tools for data wrangling and machine learning.

SUMMARIES OF PROBLEM, DATA, METHODS

PROBLEM SUMMARY

The Challenge was divided into the following three parts, each of which had specific requirements that pertained to identifying anomalous entities in different aspects of the Medicare system:

- Part 1: Identify providers that overcharge for certain procedures or regions where procedures are too expensive.
- Part 2: Identify the three providers that are least similar to other providers and the three regions that are least similar to other regions.
- Part 3: Identify 10,000 Medicare patients who are involved in anomalous activities.

METHODS SUMMARY

Table shows the wide variety of data pre-processing, analysis, and visualization techniques that were applied complete the tasks as part of the project –

PART	Analytical Techniques	Visualization Techniques
1	Descriptive statistics Straightforward data manipulation	Scatter plots Heat Map
2	Data manipulation K MEANS Clustering	
3	Data manipulation Random Forest K MEANS Clustering	

PART ONE: IDENTIFY PROVIDER AND REGIONS WHERE COSTS ARE HIGH

- The data provided was mostly clean for analysis
- The data was provided in csv format and was imported to python using "read_csv" method of pandas.
- The data frame for inpatient was expressed in terms of 'DRG Definition' and for outpatient; it was expressed in terms of 'APC'

For PART1A - For cost variation calculation

Step 1- Finding Standard deviation and mean for InpatientDRG/OutpatientAPC Step 2 - Coefficient of Variance, the standard deviation divided by the mean

Results:

The 3 Highest cost variation can be seen in Outpatients for following

1. APCS- 0604 - Level 1 Hospital Clinic Visits
2. 0698 - Level II Eye Tests & Treatments
3. 0019 - Level I Excision/ Biopsy

For PART1B - Highest Cost Claims by provider

Step 1 - Finding the max of Average Covered Charges for InpatientDRG/OutpatientAPC

Step 2 - Merging the max value with InpatientDRG/OutpatientAPC dataset

Step 3 – Aggregating the data at provider level to get number of times a provider has charge max value for procedure

Results:

The 3 Highest cost claims can be seen for follwing Providers-

a. BAYONNE HOSPITAL CENTER
b. CROZER CHESTER MEDICAL CENTER
c. STANFORD HOSPITAL

For PART 1C- Highest Cost Claims by Region

Step 1 - Finding the mean of Average Covered Charges for InpatientDRG/OutpatientAPC by DRG/APC Definition and Region

Step2 – Getting the max by region for each DRG/APC Definition

Step3 - Getting number of times a region has charge max value for procedure

Results:

Hospital Referral Region (HRR) Description

1. CA - Contra Costa County 36.0
2. CA - San Mateo County 24.0
3. CA - Santa Cruz 11.

For PART1D- Highest Number of Procedures and Largest Differences between Claims and Reimbursements

Step 1 – Get the claim difference by (Charge – Payment) for each provider and DRG/APC Definition

Step2 – Getting the max by DRG/APC Definition

Step3 – Join the data frame and count how many times a provider has hit MAX.

Results:

We see largest claim difference for highest number of procedures for InpatientDGR for the follwing providers –

- BAYONNE HOSPITAL CENTER, NJ 29
- CROZER CHESTER MEDICAL CENTER, PA 12
- HAHNEMANN UNIVERSITY HOSPITAL, PA 8

PART TWO: IDENTIFY THE LEAST SIMILAR PROVIDERS AND REGIONS

The purpose of this part is to identify the three providers that are least similar to other providers and the three regions that are least similar to other region. Based on the profiling report and cardinality of data following columns will be used for processing

a) Based on the data profiling**:**

Columns to be used to process inpatient data-

'DRG Definition', 'Provider Name', 'Provider State', 'Hospital Referral Region (HRR) Description','Total Discharges', 'Average Covered Charges' and 'Average Total Payments'

Columns to be used to process outpatient data

'APC', 'Provider Name', 'Provider State', 'Hospital Referral Region (HRR) Description','Outpatient Services', 'Average Estimated Submitted Charges' and 'Average Total Payments'

b) Sub-setting & renaming columns so that we have identical columns for Inpatient and Outpatient as:-

- Procedures
- Provider Name
- Provider State
- Region
- Count_of_services
- Charges
- Payment

C) K means Cluster analysis was performed to identify least similar providers and regions.

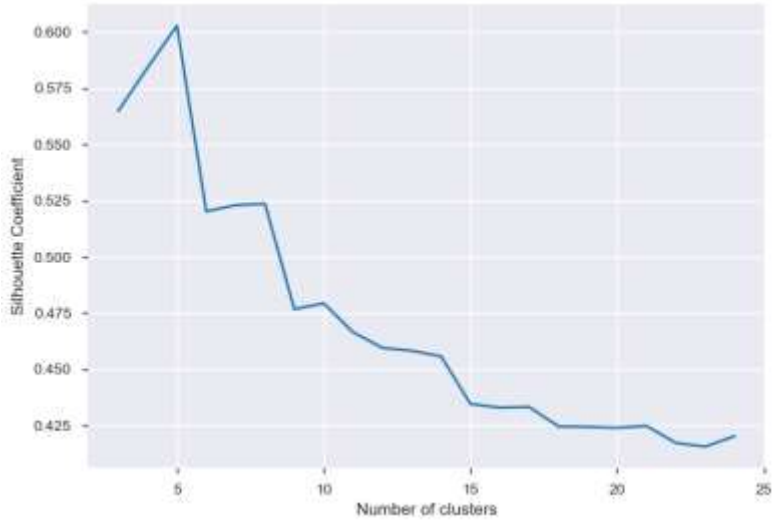
- Data was scaled and then PCA was applied to get optimal components.
- Used. Silhouette Coefficient – optimal cluster.

Result for Provider

```
# calculate SC for K=3 through K=30
k_range = range(3,25)
scores = []
```

```
for k in k_range:
    km = KMeans(n_clusters=k, random_state=1)
    km.fit(reduced_cr)
    scores.append(metrics.silhouette_score(reduced_cr, km.labels_))
scores
```

[0.5648290026091711, 0.5841443695740672, 0.6028124689938463, 0.5201422470203949, 0.5229722201804041, 0.5235640585227276, 0.4767044932966791, 0.4793096967776324, 0.46646153121037204, 0.45947686558792183, 0.4581770881061329, 0.4556950231843945, 0.4345161396390521, 0.4328798200485693, 0.4331564394425185, 0.42450209700050084, 0.42432812434045064, 0.42396499245308006, 0.42472370171037727, 0.41716555075813067, 0.41556465773558676, 0.42020528535615814]



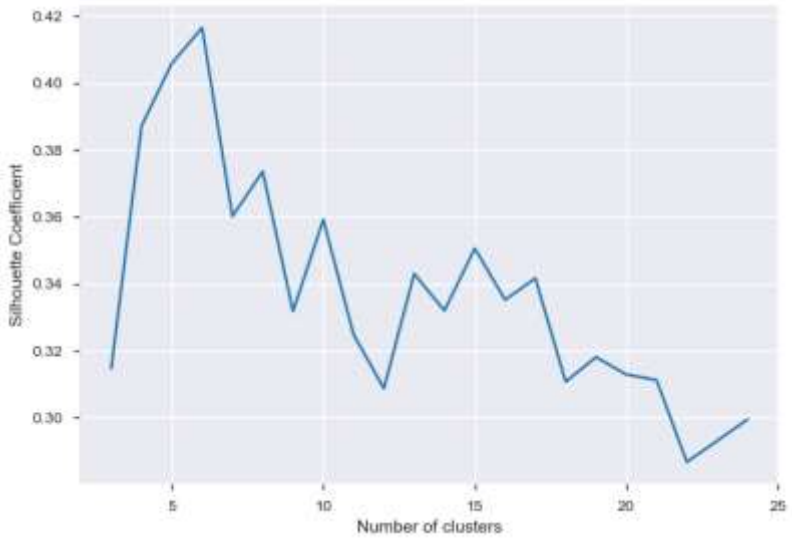
Based on the cluster analysis following Providers stands out as least similar to other providers by K-MEANS Analysis

	Provider Name	Procedur	Provider State	Region	Count_of_services	Charges	Payment	Cluster_23
489	CLEVELAND CLINIC	123	1	1	377234	4653234	1198120	8
2397	SCOTT & WHITE MEMORIAL HOSPITAL	122	1	1	402799	2823332	1061029	8
924	GOOD SAMARITAN HOSPITAL	124	6	9	69164	28175444	6991829	14

Results for Region

```
k_range = range(3, 25)
scores1 = []
for k in k_range:
    kmr = KMeans(n_clusters=k, random_state=1)
    kmr.fit(Medicare_region_scaled)
    scores1.append(metrics.silhouette_score(Medicare_region_scaled, kmr.labels_))
scores1
```

[0.31469461592842424, 0.38723930865855855, 0.4059540045361104, 0.4165834247811949, 0.3602356441277429, 0.3736078516802558, 0.3318186023234452, 0.3592268423387034, 0.3248614405219504, 0.3087469228957531, 0.34302864325217863, 0.3320253561084451, 0.35062408501838627, 0.33524894512822456, 0.3416901481271396, 0.31077277206342585, 0.318112079163691, 0.3129355240744552, 0.3112790703610052, 0.2867793120200481, 0.29308142532774006, 0.29946411961931785]



	Region	Procedures	Provider State	Provider Name	Count_of_services	Charges	Payment	Cluster_24
127	MA - Boston	130	1	41	1494212	61870397.03	30781726	6
21	CA - Los Angeles	130	1	80	505020	251071326	47115949	9
130	MD - Baltimore	100	1	23	94192	25449498.79	23992800	21

PART THREE: IDENTIFY PATIENTS INVOLVED IN ANOMALOUS ACTIVITIES

Generally for anomaly detection, we aim to learn what is most representative of 'normal', and then anything outside of some threshold can be considered an anomaly. We have to identify the cause/pattern for the review patient and based on that the unmarked Patients' needs to be marked.This is a highly imbalanced data; since only 5000 participants were marked for review; while unmarked participant count is 500000

Review_patient_history_samp.csv and Rreview_transaction_coo.csv datasets are the details for reviewed patient. Patient_history_samp.csv and Transaction_coo.csv contained unclassified/unmarked patient.

Step1: As a part of data analysis,first transformed categorical variables to numerical variables using one hot encoding.

Step2: Then created dummies for each *proc* and counted by patient id for both review patient and unlabeled transaction data.

Step3: Added flag *'review_ind'* to both data frame and merge.

Step 4Thus we have 2 datasets; unmarked dataset (*patient dataset*) and marked for review (*review dataset*). The major hurdle was we had 500000 unlabeled and 5000 review labelled patients, which is a highly imbalanced data.So I took a sample of full dataset and added 5000 review labeled patients to create the train dataset.

Step 5. Performed association analysis for getting "REVIEW probability" based on train dataset – I used Random Forest for the purpose

I got 99% Accuracy on train dataset

8634 patient were found having prediction probability > 5 i.e. (predicted label =1).We have our suspicious unmarked patients.

TESTING AND VALIDATION

- For testing the result , used cluster analysis to see if anomalous patients are clustering together.
- I added suspicious un-labeled patients found from random forest and added the review data frame of 5000 patients.
- I used PCA to get optimal components out of long list of columns for better cluster analysis.
- 11740 patients were getting clustered in Cluster 0 for a 5 cluster

CLUSTER_NO.	COUNT OF ID
0	11740
1	40
2	18
3	1299
4	537

ALREADY MARKED REVIEWD	
CLUSTER_NO.	COUNT OF ID
0	4078
1	6
2	18
3	641
4	257

UNMARKED	
CLUSTER_	COUNT OF ID
0	7662
1	34
2	0
3	658
4	280

Recommendation:

1. Unmarked Patient IDs at cluster 0 need to be reviewed with high priority followed by Cluster 3
Cluster 0 have 7662 and Cluster 3 have 658 anomalous patients detected for review.