

5420 Anomaly Detection, Fall 2020

- Harsh Dhanuka, hd2457

5420 Anomaly Detection, Fall 2020 - Harsh Dhanuka, hd2457

Considerations

Heatmap for multi-collinearity

	Total_Discharges	1.0	-0.0	-0.0	-0.0	-0.3	-0.0	0.0	0.0	-0.3	0.0	-0.0	0.0	0.0	-0.0	0.0		1.00
l of To	Average_Covered_Charges	-0.0	1.0	0.8	0.8	0.6	0.3	0.1	0.3	0.2	0.2	0.7	0.1	0.1	0.5	-0.0		
ne and Stat	Average_Total_Payments	-0.0	0.8	1.0	1.0	0.8	0.4	0.2	0.4	0.3	0.1	0.9	0.3	0.3	0.7	0.0	-	0.75
	Average_Medicare_Payments	-0.0	0.8	1.0	1.0	0.8	0.5	0.2	0.2	0.2	0.1	0.9	0.2	0.3	0.6	0.0		
grouped	by Average_Cost_Per_Procedure	-0.3	0.6	0.8	0.8	1.0	0.3	0.1	0.3	0.5	0.0	0.7	0.2	0.1	0.5	-0.0		
	Medicare_%_Paid	-0.0	0.3	0.4	0.5	0.3	1.0	0.2	-0.4	-0.4	0.1	0.4	0.0	0.2	0.3	0.0	- (0.50
	Medicare_%_Paid_State	0.0	0.1	0.2	0.2	0.1	0.2	1.0	-0.1	-0.0	0.3	0.2	0.0	0.0	0.0	0.0		
	Out_of_Pocket_Payment	0.0	0.3	0.4	0.2	0.3	-0.4	-0.1	1.0	0.8	-0.0	0.3	0.3	0.1	0.2	0.0	- (0.25
high mul	Out_of_Pocket_per_discharge	-0.3	0.2	0.3	0.2	0.5	-0.4	-0.0	0.8	1.0	-0.0	0.2	0.2	0.0	0.2	-0.0		
	State_Total	0.0	0.2	0.1	0.1	0.0	0.1	0.3	-0.0	-0.0	1.0	0.1	0.1	0.1	0.0	-0.0		
	Median_Avg_Total_Pyt	-0.0	0.7	0.9	0.9	0.7	0.4	0.2	0.3	0.2	0.1	1.0	-0.0	0.0	0.6	0.0	-	0.00
	Median_Score	0.0	0.1	0.3	0.2	0.2	0.0	0.0	0.3	0.2	0.1	-0.0	1.0	0.9	0.2	0.0		
	Median_Score_by_Provider	0.0	0.1	0.3	0.3	0.1	0.2	0.0	0.1	0.0	0.1	0.0	0.9	1.0	0.2	0.0		
e data	Avg_Payment_by_Median_Income	-0.0	0.5	0.7	0.6	0.5	0.3	0.0	0.2	0.2	0.0	0.6	0.2	0.2	1.0	0.0	-	-0.25
c data.	Total_Disc_by_Pop	0.0	-0.0	0.0	0.0	-0.0	0.0	0.0	0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	1.0		
be using t	he	Irges	arges	nents	nents	edure	Paid	State	ment	arge	Total	Pyt	core	vider	ome	Pop		
train data ł	las	Total_Discha	Average_Covered_Cha	Average_Total_Payr	Average_Medicare_Paym	Average_Cost_Per_Proce	Medicare_%_	Medicare_%_Paid_9	Out_of_Pocket_Pay	Out_of_Pocket_per_disch	State	Median_Avg_Total	Median_S	Median_Score_by_Pro	Payment_by_Median_Inc	Total_Disc_by_		
															5			

1. Revised 2 features

- Grouped the 'Sum total of Total Discharged' by Provider Name, and State
- 'Median Score by Provider' grouped by the Provider Name, and State

2. Drop 7 variables with high multi- d collinearity

3. Split to train_test:

- 75% split, train has 75% of the data.
- Now, for the test data, I will be using the **entire 100% data**, as even the train data has anomalies.

<u>kNN</u> Clustering

Model

Build initial model, and check stability by using the



Scores

The average scores go from -0.5 to 120. So, I make a

subset with scores less than 4.0 to visualize better.



Reasonable Boundaries

I will chose 3 different cut points, which are:

0.0 1.0 5.0

This will result in a 4 cluster analysis.





<u>kNN</u> Clustering

Clusters

4

Check the statistics of the 4 clusters.

Here, I am showing the percentage of data points in each cluster

Percentage of total in each cluster

69.033821 1 27.369454 2 3.439733 3 0.156993 4

Percentage of total y_by_average_cluster



Cluster wise variable <u>**Averages**</u>

Cluster <u>Evaluation</u>

5

Cluster 3 and 4 have less than 5% of the total data points

Out of these two clusters, <u>cluster 4</u> has extremes or high standard deviation from mean for some variables, and hence, I will consider this cluster as suspicious.

Feature-wise cluster EDA will be imperative to justify this claim. On the right, I evaluate a feature 'Median Score'



luster	Total_Discharges	Average_Total_Payments	Medicare_%_Paid	Medicare_%_Paid_State	Out_of_Pocket_Payment	State_Total	Median_Score	Total_Disc_by_Pop
1	-0.221173	-0.283433	0.061537	-0.022363	-0.256388	0.033268	-0.252521	-0.030195
2	0.384266	0.492422	-0.066014	0.078355	0.325327	-0.077728	0.439989	-0.020812
3	1.236714	1.643783	-0.652100	-0.200777	2.240825	-0.039097	1.408806	0.203568
4	3.167817	2.770374	-1.263044	0.572458	6.927906	-0.221593	3.466872	12.445800

conclude that Cluster 4 is highly suspicious.

Scores	y_by_average_cluster	y_by_average_score
y-by-average-score gives us insights about the	1	-0.211456
clusters which are anomalies, as the anomalies might	2	0.282751
have a very high score compared to others.	3	1.697698
higher than all other clusters. So, I can safely	4	13.632964

So, I would pass on the 256 specific entries of the Cluster 4 to the relevant authorities, and call for further investigation on each of the entries, to understand of they are true anomalies. I will provide all the reasoning as I have highlighted above, as to the differences in the means, and walk through the process I have done.



PCA Clustering

Model

Build initial model, and check stability by using the

'Average' aggregate method

Scores

The scores go from -0.1 to 70. So I make a subset of

scores less than 4.0 to visualize better.





0.0

1.0

5.0

Reasonable Boundaries

I will chose 3 different cut points, which are:

This will result in a 4 cluster model.



<u>kNN</u> Clustering

Clusters

4

Check the statistics of the 4 clusters.

Here, I am showing the percentage of data points in each cluster

Percentage of total y_by_average_cluster

94.115230	1
3.715696	2
1.781498	3
0.387576	4

y_by_average_cluster 1 80 Percentage of total 2 3 60 4 40 20 0 1 2 3 4 y_by_average_cluster



Percentage of total in each cluster

Cluster wise variable <u>Averages</u>

Cluster <u>Evaluation</u>

5 Cluster 2, 3 and 4 have less than 5% of the total data points

Out of these three clusters, <u>cluster 4</u> has extremes or high standard deviation from mean for some variables, and hence, I will consider this cluster as suspicious.

Feature-wise cluster EDA will be imperative to justify this claim. On the right, I evaluate a feature 'Out of Pocket Payment'



cluster	Total_Discharges	Average_Total_Payments	Medicare_%_Paid	Medicare_%_Paid_State	Out_of_Pocket_Payment	State_Total	Median_Score	Total_Disc_by_Pop
1	-0.061422	-0.120624	0.026056	-0.004790	-0.112412	-0.005606	-0.080405	-0.024842
2	0.590131	1.522103	-0.208337	0.007671	0.944993	0.062190	1.004088	0.008102
3	1.311465	2.572240	-0.606691	0.187180	2.530000	0.169234	1.571540	0.159379
4	3.229441	2.875316	-1.541189	0.229214	6.608139	-0.012786	2.674932	5.222078

4

ores	y_by_average_cluster	y_by_average_score
y-average-score gives us insights about the	1	-0.150710
sters which are anomalies, as the anomalies might	2	1.368969
re a very high score compared to others.	3	2.935768
ee mai ciusier 4 nas a score annosi 13-14 innes	4	9.372393

Sco

y-b clus hav I se higher than all other clusters. So, I can safely conclude that Cluster 4 is highly suspicious.

So, I would pass on the 638 specific entries of the Cluster 4 to the relevant authorities, and call for further investigation on each of the entries, to understand of they are true anomalies. I will provide all the reasoning as I have highlighted above, as to the differences in the means, and walk through the process I have done.

