Clustering of Medicare 2014 Part D Opioid claims to identify over prescribing providers

Domain Background

Detection and Investigation of Fraud Waste and Abuse in Healthcare using Machine learning. During Fiscal Year (FY) 2016, the Federal Government won or negotiated over $2.5 billion in health care fraud judgments and settlements. Up until recent past almost all healthcare payers have been practicing "Pay" and "Chase" model for Fraud Waste and Abuse. This can be changed using predictive modeling; health Insurance payers have large amounts of claims Data available, upon which we can analyze to build various ML models to identify fraud in near real time when claims are adjudicated.

A simple ICD-10 code for heart failure 150 and "Acute systolic heart failure" 150.21 does make a large difference in the Bill, there are several such examples of Provider Billing codes that are legally allowed to bill but closer examination of historical data reveals the intent of committing Fraud. There is a whole "Billing and Coding" industry with in Healthcare industry which is putting patients in the middle the "Medical Coding" war between Payers and Providers.

This FWA does not end with Provider; it is even committed by beneficiaries or members. Medicaid paid for over 34 million opioid claims in 2012, with 15 percent of Medicaid enrollees having at least one opioid prescription. About 5% received prescriptions from five or more prescribers and about 2% filled them at five or more pharmacies.

The other area of fraud is Durable medical equipment where equipment is ordered but never delivered to the end beneficiary; the equipment is sold again to a retailer. Billing for services, such as periodic maintenance of medical equipment, that never was performed

CMS has developed a Fraud Prevention System in 2010. The Fraud Prevention System helps to identify questionable billing patterns in real time and can review past patterns that may indicate fraud. CMS has used Machine Learning to create predictive models and to identify the FWA.

Problem Statement

Opioid crisis is another issue that has been troubling USA. In this Capstone project I shall perform provider benchmarking on a subset of "Pain Management" Providers who prescribe opioids for Pain relief. I will be performing clustering to identify the providers who are prescribing opioids above the normal levels. This can be achieved by clustering the providers based on the claims submitted; smaller clusters can be used to investigate for over prescription/fraud. The goal is get the list of providers who should be investigated for opioid prescription. Whether a provider should be charged for opioid over prescription shall be decided by reviewing the list produced by clustering.

Datasets and Inputs

I am utilizing publicly available "Medicare Provider Utilization and Payment Data: 2014 Part D Prescriber" CMS data, following is the link for the data.

> https://data.cms.gov/Medicare-Part-D/Medicare-Provider-Utilization-and-Payment-Data-201/465c-49pb

The file has 24121660 (24.1 million) records. I tried to load the csv file using "pd.read_csv", it took forever on my local machine and as well as AWS. I had to Unix commands to scan through the file, apply some filters (grep) and identify a state where the records are less that 1 million. I have tried using many state provider files such as NY, CA, PA the laptop that I have could not manage and the AWS account that I have also could not manage. I have narrowed down to Alabama. So from the above data set I shall be using Alabama providers/claims data for this project. To perform provider benchmarking and identify the providers who do not fit the normal patter of opioid prescription.

Following is the pdf detailing about the columns in the dataset

> https://data.cms.gov/api/views/465c-49pb/files/0931bfc7-1069-4437-961b-e3f43e26ac33?download=true&filename=Part_D_Prescriber_PUF_Methodology_2017-05-25.pdf

Solution Statement

Used the clustering algorithms I shall try labeling the claims data.

Clustering ➔ Finally perform the clustering on the cleaned up and preprocessed data. I will be selecting GaussianMixture, KMeans, and MiniBatchKMeans from sklearn to see which one would be the best. Since we do not know the underlying data structure we shall use silhouette coefficient to find out the best one which will tell us how many clusters are present in the data.

Benchmark model

There might be several of models that have been developed since I am using 2014 CMS data and since CMS itself came up with models in 2010 to identify the fraudulent providers. The office of inspector general of department of Health and Human Services has come up with list of excluded providers, this is called exclusion list. Now please keep in mind that this exclusion list is for all types of frauds/misdiagnosis/malpractices not just Opioid over prescription. Goal is identify the list of providers who shall be further investigated.

OIG exclusion list, list of providers identified as fraudulent

> https://oig.hhs.gov/exclusions/exclusions_list.asp
>
> https://oig.hhs.gov/exclusions/downloadables/UPDATED.csv

Evaluation metrics

Silhouette Coefficient: If the ground truth labels are not known, evaluation must be performed using the model itself. The Silhouette Coefficient (sklearn.metrics.silhouette_score) is an example of such an evaluation, where a higher Silhouette Coefficient score relates to a model with better-defined clusters. The Silhouette Coefficient is defined for each sample and is composed of two scores:

a: The mean distance between a sample and all other points in the same class.
b: The mean distance between a sample and all other points in the next nearest cluster.

The Silhouette Coefficient s for a single sample is then given as:
$s = (b - a)/\max(a, b)$

The Silhouette Coefficient for a set of samples is given as the mean of the Silhouette Coefficient for each sample.

The best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar.

> http://scikit-learn.org/stable/modules/clustering.html#clustering-evaluation

Project design

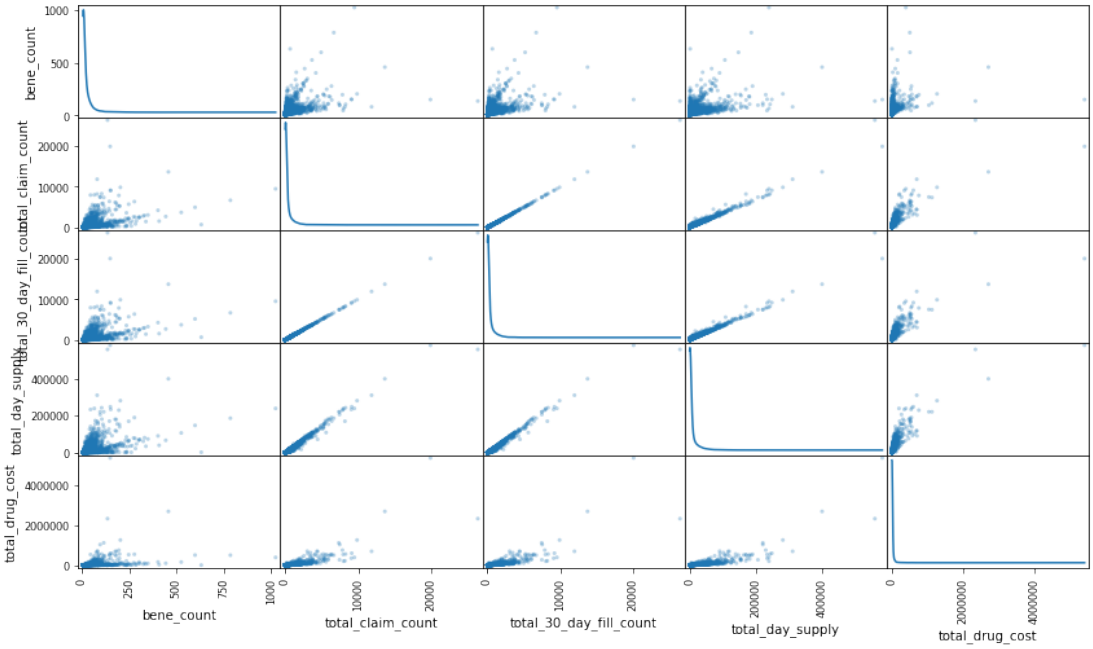
Data clean up ➔ before I go for the clustering there is lot amount of clean up that needs to happen on the data. We do not need all the Medicare Part D claim data; this file consists of prescription claims that are not for opioids also. So let us filter for Opioid claims, then we will have to some data clean such as removing comma from numbers. Convert the columns in proper data types. Determine the columns where there are only flags and remove them. Also please note that Medicare is mostly for 65 and above population, so we do not need the data that is specific to age 65 and above.

For a list of drugs that include opioids, visit

> https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/OpioidDrugList.zip

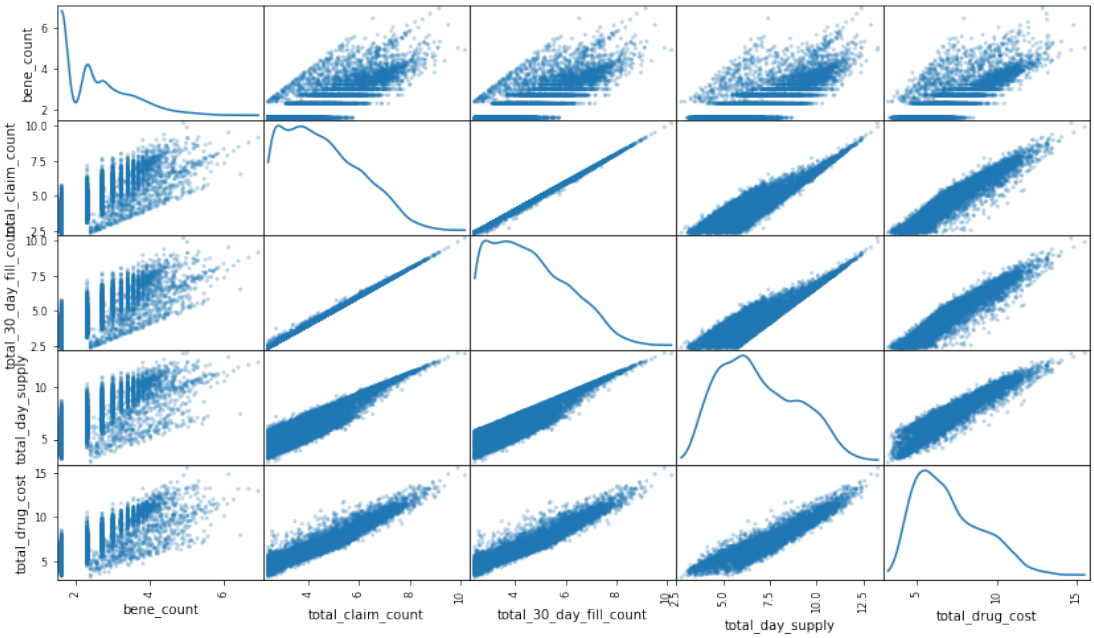
Data exploration ➔ we shall do some data exploration to see whether we have to perform and data preprocessing

Scatter plot before preprocessing



Data Preprocessing ➔ we shall do some data processing such as reducing the variance in data by applying log. We shall also see if we have to perform outlier detection. In this case I would strongly discourage from doing it because we might loose providers who are prescribing above normal prescription levels of opioids. Try and apply some dimensionality reduction using PCA.

Scatter plot after preprocessing



Clustering ➔ Finally perform the clustering on the cleaned up and preprocessed data. I will be selecting GaussianMixture, KMeans, and MiniBatchKMeans from sklearn to see which one would be the best. Since we do not know the underlying data structure we shall use silhouette coefficient to find out the best one which will tell us how many clusters are present in the data.