**Medicare Data Fraud Detection**

Due on: Monday, March 25

Student: Chunhui Zhu

In my project, I am trying to discover how advanced technology in AI and Machine Learning are to help manage the high costs associated with quality healthcare and reduces allocated funds for fraudulent physicians/providers Prat B non-instructional line items for the Medicare fee-for-service (FFS) population.

As I have stated in previous proposal that my data would use two sources websites: 1. 2012-2016 years data collections of the Provider Utilization and Payment Data physician and Other Supplier Public Use File (PUF) from the Centers for Medicare & Medicaid Services (CMS) 2. List of Excluded Individuals and Entities (LEIE) database from Office of Inspector General (OIG).

**The following is my progress in my project.**

I wrote python code to do data mining and mapped fraud data. I have loaded five years (2012-2016) complete the Provider Utilization and Payment Data sets from CMS, and LEIE most updated data set on Feb\_2019 from OIG websites. (The python code can be found in my github: <https://github.com/czhu505/Data698/upload/master/mid-term>.)

The updated LEIE is on monthly based and last updated is in Feb. 2019. It is a complete database containing all exclusions currently in effect. Individuals and entities who have been reinstated are not included in these files. The csv file is 71,284 rows by 18 columns table. In this file, it only contains 5,036 NPI (National Provider Identifier). The NPI is systematically generated when a provider applies from online. It is a unique number for match physician and Another Supplier. I used this list of NPI numbers to map the physician and Other Supplier Public NPI’s from CMS. In this file, the column ‘exclType’ majority are 1128(a)-(c). The 1128(a) rules have five-year minimum periods, and 1128(c) has a 10-year minimum periods. Based on current year 2019, and five-year minimum periods, there are some fraud NPI numbers would missed from those Physician and Other Suppliers whom were detected as Frauds.

So far CMS website has 2012 – 2016 five years Physician and Other Supplier Data sets. The data in the Physician and Other Supplier PUF covers calendar year 2016 and contains 100% final-action physician/supplier Part B non-institutional line items for the Medicare fee-for-service population. Each data set has more than 9 million rows and 26 or 28 columns tables. In the each data set, there are some NPI numbers having multiple records, such as a group which have different business locations. For example, in calendar year 2016 data set, there are 1,000,925 unique NIPs. In my study, I am going to use these nine variables in my analysis, 'npi', 'provider\_type', 'nppes\_provider\_gender','hcpcs\_code','line\_srvc\_cnt', 'bene\_unique\_cnt','bene\_day\_srvc\_cnt', 'average\_Medicare\_allowed\_amt', 'average\_Medicare\_payment\_amt'.

#### I used 5,036 fraud NIP numbers from LEIE to search NIP numbers in Physician and Other Supplier 2012 -2016 Data sets. Here is the finding, there are 1,149 NPI s found in CY\_2012, 906 NPI in CY\_2013, 722 NPI in CY\_2014, 475 NPI in CY\_2015, and 281 NPI in CY\_2016. The Crime Rate between 2012-2016 Physician and Other Supplier is 0.1%, but I believe the number is higher than 0.1% due to the limited time data collection.

The total unique NPIs between CY\_2012 and CY\_2016 is 1,221,295. Total fraud NPI number between CY\_2012 and CY\_2016 is 1,235. Within five years records, there are 247 fraud NPIs having continuous records in these data set.

I used backward method to try to figure the number of new discovered fraud NPI between these five years. Since LEIE holds majority NPI records with minimum 5 years penalty, I assured the NPI once detected as fraud, the NPI won’t exist in next 5 years. In this case, the detected NPIs in 2012, will not appear in 2016. Though the record is less accuracy only in a year base, I only can assure this number is correct.

For the preparation of using machine learning models in prediction fraud NPIs, I extracted the NPIs exclusive between calendar years. For example, there are #281 new found NPIs detected as fraud in 2013 comparing 2012, #281 in 2014, #275 in 2015, #281 in 2016. These data will be separate and been used in training set and test set for machine learning algorithm.

**The following is my next plan to achieve my goal in my final project.**

I will follow the guide and analysis which has discussed in a paper “The Detection of Medicare Fraud Using Machine Learning Methods with Excluded Provider Labels” by Richard A. Bauder and Taghi M. Khoshgofataar. In the paper, they stated the RF algorithm, logistic regression supervised learning models are successful to detect the fraud NPIs in 2012-2015 CMS Data Sets, and used LEIE 2017 Data set for mapping fraud data.

The challenge that I am facing now, is to find the correct method to break down the fraud data set in training and test set. Due to million rows vs a thousand labeled data set, it will easily mass up at the end. And so far, this methodology is lack of discussion in some research paper. And it reminds to my rest of the study while I continuous this project.

## References:

[Medicare Fee-For-Service Provider Utilization & Payment Data Physician and Other Supplier Public Use File: A Methodological Overview] last updated: May 3, 2018, The Centers of Medicare and Medicare Services, Office of Enterprise Data and Analytics <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html>

[Medicare Provider Utilization and Payment Data Physician and Other Supplier PUF: Frequently Asked Questions] last updated: May 4, 2018, The Centers of Medicare and Medicare Services, Office of Enterprise Data and Analytics <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Physician_FAQ.pdf>

[List of Excluded Individuals/Entities (LEIE)] 02-2019 Updated LEIE Database, <https://oig.hhs.gov/exclusions/exclusions_list.asp>;

[The Detection of Medicare Fraud Using Machine Learning Methods with Excluded Provider Labels] by Richard A. Bauder, Taghi M. Khoshgoftaar, College of Engineering & Computer Science, Florida Atlantic University (2018)