Proposed title of the project

Chengming Xie (cx2234), Hongshan Li(hl3353)

Topic : Medical Insurance Fraud Detection

Date: 2/13/2020

Abstract: A brief summary of the project including the purpose, the methods, the goal, and other important things.

Health care is a major industry in the United States (U.S.) and it is important in the lives of many citizens, but unfortunately the continuing rising in medicare care overall experienture and high costs of health-related services leave many patients with limited medical care. The impact of Medicare Fraud is estimated to be between 3% to 10 % for the nation's total healthcare spending, and reached approximately $2 billion including 165 medical professionals in 2018.[1] In order to reduce cost for either insurance companies and the government, our project aims to build a comprehensive and novel fraud detection model to find most suspicious groups of outlying records that belong to the same class could be involved in fraudulent activities. This tool can also offer health care professionals and patients to provide more effective treatment and response for increasing cost of health care. We will not treat medicare claims as a single event but a sequence of event groups by pharmacies and patients. Then compare the similarity between the same group of patients including treatment duration, times of treatments, and drug use to detect anomaly. We applied this method to a set of patient records comprising 26 tables (~60G) from MIMIC-III. to identify suspicious treatments and doctors. Challenges included handling multi-scale datasets, unsupervised outlier detection, and training models for 500 different diseases.

1. Background (Review of Related Literature):

The summary of the related literature, presenting important information and knowledge such as the background of the topic and the current research state.

Health care is a major industry in the U.S. According to CMS.GOV , the costs of healthcare continue to rise, healthcare spending from 2012 to 2014 has increased by 6.7% to reach $3 trillion and Medicare spending accounts for 20% of all health-care spending in the U.S. at about $600 billion. Under current law, national health spending is projected to grow at an average rate

of 5.5 percent per year for 2018-27 and to reach nearly $6.0 trillion by 2027.Health spending is projected to grow 0.8 percentage point faster than Gross Domestic Product (GDP) per year over the 2018-27 period; (CMS.GOV) The impact of healthcare fraud is estimated to be between 3% to 10% of the nation's total healthcare spending. It reached approximately $2 billion and continues to adversely impact the Medicare program and its beneficiaries (NHCAA 2017). Both private insurance companies and the government seek cost-cutting solutions, where the reduction in fraud is one way to help recover costs and reduce overall payments.

Health care fraud is a crime in which someone uses lies, deceptions, or falsehoods when filing a health care claim in an effort to make a profit or to gain some type of benefit. Common fraud schemes involve double-billing or filing duplicate claims for the same service, filing claims for services never provided, billing for services not covered by an insurer's policy, and even providing kickbacks for referrals. Health care fraud can seem like a minor crime, especially when an individual commits an act that seems like it has little impact and doesn't really hurt anyone but it is really a heavy offense in the U.S.

Although medicare insurance fraud is a serious problem in the U.S, there is not a proper, accuracy, time saving and economy way to detect anomaly. The traditional way to uncover fraud is manual relying on internal whistleblowers to inform investigative teams of potential targets. Such processes are tedious, costly and prone to error. Because of the high dimension event sequence data in medicare, people don't have a perfect solution even using big data analytics. It's unlike a simple classification problem in machine learning, we can not treat our records as single independent claim otherwise it will be costly and have a high false positive rate. We have to use novel, advanced algorithms (deep learning etc) to deal with this problem. Therefore, the motivations of our project are: cut cost for medicare, reform the process of investigation, use innovative models to improve accuracy. Since we still don't have a better strategy to mitigate the effect of fraud, there are huge business opportunities for medical insurance fraud detection.

2. Introduction to the Project:

Inform how you are planning to investigate the topic in the project, including the methods to use and the goal to achieve.

The inspiration of this project comes from continuing rising in the medicare expenditure. Especially, there is a significant amount of fraud, waste, and abuse within the Medicare system that costs taxpayers billions of dollars and puts beneficiaries' health and welfare at risk. One way to reduce the medicare cost is to reduce the amount of medical fraud. The manually investigating is inefficient and tedious. Although previous work has shown the effectiveness of constructing machine learning models to automatically detect fraud in medical care, the challenges associated with class-imbalanced big data hinder performance. Moreover, we should treat the claim records as event sequences rather than treat them as a single independent case because recovering money

from single outlier is costly and the high false positive rate will hurt the relationship between insurance providers and medicare professionals. Hence, people still don't get a reliable scheme to identify anomaly in unsupervised medicare data. [1]

The goal of our project is to build an innovative and comprehensive data science tool used to identify suspicious fraud in the medical insurance industry by comparing the similar patient treatments (ex:compare the treatment duration, amount cost of treatment, and drugs use in the treatments between similar patients).(Graphen Medicare intelligence) The tool will help the insurance industry reduce their cost, and also help health care professionals and patients to provide more effective treatment and solutions for responding to increasing the cost of health care. The input is the patient records form MIMIC-III databases. We will train our model for different disease categories independently at the beginning because different diseases have different characteristics. We first use event embedding algorithms such as Skip-n gram and Dynamic Time Warping translating event to vetor to retrieve and align similar patients' treatment records. Secondly, we will use deep learning algorithms such as two recurrent neural networks(RNNS) to predict potential diseases for certain patients based on his/her historical medical record. To detect anomaly treatments, our aim is to detect substantial anomaly linked to multiple claims of the same medical provider. We will first calculate the outlier score in single treatment records according to their distance base scores, and then measure the outlierness of each group of records by weighted rank-based statistics and Weighted rank outlier score (Rob M. Konijn and Wojtek Kowalczyk.) Also, we will use the autoencoder anomaly detection to find outliers. Finally, we will combine results for almost all ICD9 categories among our dataset. The final result will be a user friendly interactive website by Rshiny app that generates top 10 suspicious medical records by patient, and we also try to detect whether substantial fraud linked to multiple claims of the same pharmacy.

3. Introduction to the Dataset:

Introduce the data source, the size, and the background. Also provide some related information such as the methods of getting data and the obstacles.

We require a detailed, comprehensive data set covering most disease categories and corresponding treatments data of each patient. MIMIC-III critical care database is a large, freely-available database comprising 26 tables (~60G) which includes de identified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. MIMIC-III can be accessed on Google BigQuery API so that we can incorporate cloud computing technology and Big Data Analytics such as Spark SQL in our data cleaning process.

The data set has high variety by arising from mixed-type high dimensional features and the combining of comprehensive ICD-9. These data sets are also reliable, as they are provided by reputable hospital resources with transparent quality controls and detailed documentation.

We try to split the dataset based on the International Classification of Disease (ICD-9) and train our event embedding and anomaly detection model on patients who fell into different categories of diseases. The features which will be used to investigate the general fraud detections include the detailed information of different drugs prescribed, the number of prescriptions, the number of days prescribed, the total cost, etc. Other features, such as the provider location or specialty will be combined with the above features to detect the fraud.

4. Plan:

Plans between every milestone and final. Please be specific.

Milestone 1: We were brainstorming and seeking dataset including detailed timestamped diagnoses and treatments provided to a large volume of patients. Based on previous literatures and works, we decided to use the MIMIC-III clinical database as described in Part 3.
We researched related literature on event embedding which can align different patients' timestamped data, RNN algorithm to predict targets' future health conditions and possible treatments. anomaly detection using various methods including deep learning and LOF methods;
In the beginning, we began to fetch data on GCP BigQuery API based on which we generated data insights into entity relationship between 26 tables included in the MIMIC Clinical Database. We aimed to implement big data technology such as Spark SQL/Hive to boost the efficiency on quering data and generate ML-oriented features table to prepare for our future algorithm implementation

Milestone 2: We will begin implementing algorithms in one specific category of disease based on ICD-9, based on previous literatures as well as algorithm framework at Graphen which includes event embedding, Dynamic Time Warping and RNN prediction. Also, we will finish anomaly detection in the similar patients' treatments information to find possible medical fraud (for instance, unnecessary or higher rates of prescription). Our algorithm we will implement in the multi-dimensional "time series" data includes Autoencoder Anomaly Detection, LOF methods, Clustering-based methods, etc.

Milestone 3: We will apply our algorithms and ML techniques to all categories of diseases in which we need to further rationalize our models to tackle the potential problems happening during the training process.

Wrap-up: We will finally realize our visualization on user interface based on demos created by Graphen and improve the interface to include more comprehensive information. Then we can wrap up and begin story-telling on possible business value of our completed work.

5. Challenge:

Although the MIMIC dataset includes detailed timestamped patients' information on treatment and diagnosis, which is different from other dataset, this clinical dataset is limited in MA which does not include multi-center patients' treatment information. In the process of our project, we need to simultaneously figure out possible ways to include more geographically comprehensive patients' treatment information.

Also, models in this project are mainly unsupervised which causes difficulties to evaluate our model performances. Possible evaluations can include further case studies by medical professionals invited to check the rationality of our prediction and anomaly detection. Furthermore, patients may suffer from different categories of diseases and individual ICD-9s may fall into different categories. That is, there may be a mixed effect of different categories of disease on the future health conditions and treatments. Thus, we need to implement more complex algorithms and professional medical knowledge

Reference:

Graphen AI, medical intelligence.

Morris L. Combating fraud in health care: an essential component of any cost containment strategy. Health Aff. 2009;28:1351–6. https://doi.org/10.1377/hlthaff.28.5.1351.

Rob M. Konijn and Wojtek Kowalczyk ：Finding Fraud in Health Insurance Data with Two-Layer Outlier Detection Approach

https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet

Raghavendra Chalapathy, Sanjay Chawla. *Deep Learning For Anomaly Detection: A survey. 2019.*

Shunan Guo, Zhuochen Jin, David Gotz, Fan Du, Hongyuan Zha, and Nan Cao. *Visual Progression Analysis of Event Sequence Data.* 2018.

Zhuochen Jin, Jingshun Yang, Shuyuan Cui, David Gotz, Jimeng Sun, Nan Cao. *Carepre: An Intelligent Clinical Decision Assistance System*. Nov, 2018.

https://oig.hhs.gov/newsroom/media-materials/2018/takedown/

MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available at: http://www.nature.com/articles/sdata201635