COMBATING MEDICAL PROVIDER FRAUD WITH ADVANCED ANALYTICS USING MACHINE LEARNING

April 23, 2018

Written by
Preetesh Shukla
EXL Analytics

lookdeeper@exlservice.com
As per the National Health Care Anti-Fraud Association (NHCCA) financial losses due to health care fraud are in tens of billions of dollars each year. The NHCCA conservatively estimates that 3 percent of all health care spending, or $68 billion, is lost to health care fraud (2015). The Federal Bureau of Investigation (FBI) has estimated fraudulent billings slightly higher and according to their estimates it is between 3 percent and 10 percent of total health care expenditures.

The latest report on questionable claims by the National Insurance Crime Bureau (NICB) states that the number of questionable claims in the year 2015 was ~120,000. All these statistics point to the fact that the amount of healthcare fraud is huge and because of this, insurers and organizations around the world are continuing their fight against health care provider fraud with a variety of initiatives – and degrees of success.

The dollars lost to fraud have direct bearing on the cost and quality of health care. Not only does it contribute to higher insurance premiums, but also every dollar spent on fraudulent claims reduces the amount of money available to improve the quality of care for those incurring legitimate expenses. Preventing fraud requires timely detection. The earlier it is detected, the higher are the chances of averting the insurer’s loss.

Companies fighting medical claim fraud, have a tough task of telling fraudulent claims from honest ones. Although majority of claimants submit honest and accurate claims, a small fraction gives in to the financial incentive and exaggerates the complexity and size of the claim1. In the past insurers just looked at claim information, which made it hard to distinguish fraud and overbilling from the normal variability across providers, specialties and geographic regions. Claim information alone is insufficient, it needs to be combined with the expected inconsistencies for a particular claim along with medical payments, clinical insights and medical logic to detect and prevent fraud. The following are some of the negatives of just looking at claims data.

- Claims databases are incomplete
- Claims are processed by different adjusters causing variability in referring claims
- Looking at historical data does not give a holistic picture, as the providers are also evolving in their tactics

Insurers have stepped up their investment in analytics to focus on fraud investigations and timely detections. Advanced analytics enables targeting fraudulent behavior at early stage and provides useful insights about the source of the fraud. In order to detect fraud, insurers need to focus on the behavioral pattern of the medical provider rather than the act itself, and analytics aids insurers precisely in doing so.

In this whitepaper, we address approaches which make use of data other than the claims data, discuss in detail the fraud detection process, present some typical scenarios of fraud and discuss promising unsupervised machine learning approaches to control medical provider fraud.
Special Investigation Units

Insurance companies have established effective procedures to identify, investigate and deter fraud by deploying SIUs (Special Investigation Units). These SIUs have qualified claim adjusters and special investigators who develop and implement processes to ensure that only legitimate claims are paid. The experience, insight and intuition of these claims personnel have saved insurers millions of dollars in payments over the years.

However, the challenge is that investment required to hire and train these resources and the time spent in analyzing claims is huge. As a result, insurers always faces a resource crunch and only highly suspicious claims get reviewed by SIUs. Only 1 percent of the total claims are processed by SIUs, 99 percent are passed through. Insurers estimate that 10 percent to 25 percent of reported claims contain some element of fraud, therefore despite the presence of SIUs, there is a large amount of loss that insurers incur on account of fraud. The success of an insurance company hinges on selecting the right claims for the SIUs to process; if the company is not able to do so, most of the fraudulent claims would pass through and only a few would get detected, which would result in unnecessary payments and losses.

Therefore, the core of the fraud detection process lies in the identification of a suspicious activity. Conventionally, assessors have done this manually. The manual way of referring claims is a simple approach that tends to process only where the chances of fraud are high. The manual process usually does a poor job at identifying fraudulent cases and getting them examined, as it doesn’t take into account the multiple identities the medical providers may take and is just based on red flags implemented on a case to case basis.

Over the years, Special Investigation Units too have embraced analytics. Analytical approaches increase the scale of investigation and take into account the multiple identities, red flags, historical data and adjusters’ prior experiences to identify the suspicious cases and refer the fraudulent cases to SIUs. Analytics lends a helping hand to SIUs, assisting insurance companies to develop highly consistent claim referral processes and utilize adjusters and investigators in cases which have high likelihood of fraud. This enables insurers to use the expertise of adjusters and investigators optimally in all claims.

Unlike the manual process which is time consuming and involves referrals from individual adjusters based just on review of individual accounts and reference to past incidences, statistical approaches are more efficient. Manual approaches are also susceptible to biases and may result in dollar losses as adjusters with less experience may refer legitimate claims to be fraudulent, while not reporting the illegitimate claims.

SIUs leverage data analytics to identify and tag suspicious claims as fraudulent on the basis of malpractices, some of which are discussed in the following section.

Malpractices-Business

Red Flags

Several providers hospitals, pharmacies and doctors, carry out certain malpractices for their own gains; in this paper we have referred to these malpractices as ‘business red flags’. Business red flags provide a basis for SIUs to compare and rank providers for fraudulent behavior. The most common types of business red flags mentioned in industry literature are2:

- Billing for services that were never rendered – either by using genuine patient information, which at times can be obtained via identity theft, to fabricate entire claims or by padding claims with charges for procedures or services that did not take place.
- Billing for more expensive services or procedures than were actually provided or performed, known as upcoding.
• Performing unnecessary medical services to generate insurance payments.

• Misrepresenting non-covered treatments as medically necessary and covered for obtaining insurance payments – widely seen in cosmetic-surgery schemes, wherein non-covered cosmetic procedures such as "nose jobs" are billed to the patients.

• Falsifying a patient’s diagnosis to justify tests, surgeries or other procedures that aren’t medically necessary.

• Unbundling – billing each step of a procedure as if it were a separate one.

• Billing a patient more than the co-pay amount for services that were prepaid or paid in full by the benefit plan under the terms of a managed care contract.

• Accepting kickbacks for patient referrals and from pharmacies as well.

In addition to the above stated red flags which come from industry literature and knowledge, the following few have been suggested by the EXL team based on the projects we have done:

• Billing for unreasonable number of patients, either by billing huge number of patients in a day or billing for two or more patients simultaneously.

• Prescribing large amounts of compound drugs, which are much costlier than regular, available drugs.

• Prescribing large amounts of narcotics.

• Performing or billing procedures unrelated to the type of injury and diagnosis.

• Misuse of emergency procedure codes, i.e., frequent billing under emergency to get more dollar amount from the insurer.

• Patients traveling larger distances to visit a doctor they know, instead of the one in close proximity of their residence; this is an example of nexus that exists between a doctor and his patient.

These business red flags can be used to formulate rules and to score providers, claims, attorney firms, etc. These red flags have played an important role in fraud detection despite the continuous change and evolution in fraud detection methods.

How Fraudsters Work

One of the common malpractices medical providers follow to defraud is dividing their activities among several IDs. This helps them understake their billed dollars, number of patients and other important information. This practice can be referred to as registered identify fraud. In this type of fraud, the provider’s billed dollars, number of patients and all other vital information is underrepresented and is passed on by adjusters as genuine claims.

The complexity of Medicare and Medicaid programs in terms of the number of providers and payers servicing the population make them especially vulnerable to registered identity fraud. In the year 2014 itself $60 billion of American taxpayer money, or more than 10 percent of Medicare’s total budget, was lost to fraud, waste, abuse and improper payments. Investigation on fraud revealed that there are almost 23,400 potentially fake or bad addresses on Medicare’s list of health care providers. The above statistics are still an under representation.
The complexity of identify fraud is growing and record linkage has evolved too, and in extreme situations it might involve range of linkage scenarios from simple linkage based approaches to complex ongoing linkages. These record linkage techniques along with text mining techniques like Perl Regular Expressions and Natural Language Processing, can not only detect registered identity fraud, but also help identify nexus between medical providers.

However, linking data is just one piece of the puzzle, it needs to be supplemented with data analytics to segregate clusters of fraudulent accounts and forms the core of any medical provider fraud examination and control. Two such machine learning techniques are detailed in the next section.

### Machine Learning Methods & the Limitations of the Manual Approach

The manual approach for fraud detection is based on developing heuristics around fraud indicators. Two approaches are usually taken, first - framing rules that define if the case needs to be sent for investigation, second - creating a checklist with scores for various indicators of fraud. An aggregation of these scores helps determine if the case needs to be sent for investigation. The criteria for determining indicators and the thresholds is tested statistically and periodically recalibrated.

Developing a customized matching algorithms to match Medicaid providers with Medicare providers would involve consolidating similarly named providers into one “family”, so that fraud investigators can analyze a provider more readily. To develop the matching algorithm variables like provider name, address, license number, phone number, specialty, address frequency and last name frequency can be used. Once the record linkage is done, the providers’ combined Medicaid/Medicare activity is reviewed and providers’ records are consolidated into one family eliminating variations due to typos or name-changes, fraud analysts can analyze trends more accurately and the provider may jump out as a clear outlier.

A common solution to the lack of a common identifier is **record linkage**. Record linkage helps curtail registered identity fraud. It helps join data sets based on entities that may or may not share a common identifier, for e.g., database key, national identification number, as may be the case due to differences in record shape, storage location, and/or curator style or preference.

As medical providers may serve both individuals using Medicaid as well as Medicare services of the government. For a correct assessment on fraud, both these services should be considered in tandem and matching Medicaid providers to Medicare providers is of utmost importance.

The savings and efficiency gained from analyzing data can have a huge impact on both an insurer’s ability to mitigate the amount of fraud paid as well as the expense associated with investigating it manually. Numerous statistical methods are being used for detecting fraud since a long time now, for instance outlier analysis, association analysis, decision trees etc. There is no single solution which is applicable in each situation. Statistical techniques like outlier analysis rely on the differentiation of the normal providers from the fraudulent providers,
methods, has not been explored much. In the following sections we discuss two methods to deploy unsupervised learning, first is a type of clustering approach known as Self Organizing Maps (SOM) and second is a type of factor analysis method known as PRIDIT.

a) Self-Organizing Maps (SOM)
as a Clustering Approach

Self-Organizing Maps (SOM) is an unsupervised way of clustering medical provider’s claims data to segregate fraudulent claims from genuine ones. SOM also known as Kohonen Feature Maps, was developed to simulate the way vision systems work in our brain. SOM, is based on a neural network which uses competitive learning techniques to train models. SOM takes in a predefined number of nodes (clusters). The number of clusters vary from situation to situation.

In the case of analyzing medical provider fraud, whenever there is a new provider all nodes compete to get that medical provider assigned to them. Since nodes compete among themselves to acquire the new observations, this is often referred to as “Winner takes all strategy”. The node which will be able to acquire this medical provider will be the one with most similar features and behavior. Also, the entering medical provider modifies the feature distribution of the node and hence we say that the model is self-learning, we don’t need to tell the model beforehand if an observation is in confirmation with the existing state of the node or not.

The principal goal of SOM is to transform the n-dimensional incoming data into one or two dimensional discrete map, and to perform this transformation adaptively in a topologically ordered fashion. For example, in a CMS study, the input dataset could be an n-dimensional table where different columns contain results from different rules and the rows represent an individual claim or provider.
SOM can capture fraud patterns which were not identified in historical data. It also allows us to represent the distribution of n-dimensional features in a 2-D map. In addition, the Kohonen technique creates a network that stores information in such a way that any topological relationships within the input data set are maintained.

This feature of SOM has certain advantages, as the retained topological information between observations facilitate an in-depth study about the statistical features of nodes. We can also do a variable based study across all nodes using this information, for example, for a given node the medical providers can be studied corresponding to a red flag across nodes.

SOM also provides visualization tools which help in visualizing this statistical information and in drawing inference about the suspicious nature of observations in any node. Clustering can tag providers as suspicious or non-suspicious, but all clustering algorithms fail in indicating the severity of fraud, a very important issue that can be tackled using a machine learning approach known as PRIDIT, which is discussed below.

b) PRIDIT - Principal Component Analysis of RIDIT (Relative to an Identified Distribution Integral Transformative) Scores

Adjuster experience is complemented with analytics in PRIDIT. The first step in PRIDIT analysis is to compile a series of indicator variables of fraud - red flags in the case of medical provider fraud. Once rules are formulated based on the red flags they are used to score claims/providers. This score is an ordinal variable which can be used to generate ranks of claims (RIDIT).

Next, each level of the variable is ranked from 1 to n, with 1 being the most likely characteristic to be fraudulent (this is where the adjuster experience is incorporated). The procedure to calculate Beta for this red flag is to calculate the difference, in the percentage of total exposure which has a lower (more fraudulent) ranking and the percentage of total exposure which has a higher (less fraudulent) ranking.

RIDIT score calculation is explained using the below example.

Since RIDIT scores are based on red flags, there would be RIDIT scores corresponding to all decision rules. The RIDIT scores are calculated for all the providers and for all the decision rules, and based on these scores, a principal component analysis on RIDIT scores is carried out to calculate the PRIDIT scores for all the providers.

PRIDIT overcomes the limitation of PCA (Principal Component Analysis) application on ordered categorical variables, as PCA when applied directly to the ordered categorical variables may result in loss of important information. By itself, a PRIDIT score cannot divulge any information about a record’s fraudulent probability. However, PRIDIT scores do provide a way of ranking the providers. Within a dataset, the records with the lowest PRIDIT scores should be viewed as the most likely to be fraudulent. Insurers can apply the results of this analysis.

<table>
<thead>
<tr>
<th>Suspicion Level</th>
<th>Number of observations</th>
<th>Proportion</th>
<th>Proportion below</th>
<th>Proportion above</th>
<th>RIDIT score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>400</td>
<td>.2857</td>
<td>0</td>
<td>.714</td>
<td>-.714</td>
</tr>
<tr>
<td>2</td>
<td>300</td>
<td>.214</td>
<td>.2857</td>
<td>.5</td>
<td>-.2143</td>
</tr>
<tr>
<td>3</td>
<td>700</td>
<td>.5</td>
<td>.5</td>
<td>0</td>
<td>.5</td>
</tr>
</tbody>
</table>
by directing their limited investigative resources towards claims with the lowest PRIDIT scores.

**Conclusion**

While insurers have developed effective fraud departments, the reality is that not all fraud can be identified using manual intervention or statistical approaches. Mechanisms to defraud in healthcare space are constantly evolving. Integrating and mapping the databases along with statistical techniques can go a long way in avoiding registered identity fraud. But again, mapping data to detect red flags such as upcoding, bundling and unbundling of procedure codes, drugs classification, etc. also require a good amount of research and extensive data preparation on a continuous basis, which is very difficult. Determining the right modeling technique to suit a given fraud context in medical provider fraud, can be an equally difficult and challenging task.

Unsupervised machine learning methods present a great alternative in these situations. We explored two unsupervised machine learning approaches in this whitepaper – SOM Clustering and PRIDIT. Both these are useful techniques to capture medical provider fraud, as they work well in the absence of historical data as well as when historical data lacks accuracy regarding what comprises fraud. PRIDIT is useful when applied on ordered variables and generates a continuous score which indicates the severity of fraud, however, it does not work with unordered variables. SOM based clustering on the other hand is able to tackle unordered variables, but does not give a score as output. It only assigns a provider to either fraud or non-fraud categories but fails to indicate the severity of fraud.

Both the approaches suggested in this whitepaper have their own limitations, so an ensemble approach that incorporates business knowledge into machine learning techniques is recommended for evaluating medical provider fraud. By combining SOM based clustering and PRIDIT, carriers can save billions of dollars that are lost every year owing to fraudulent practices. Furthermore, these analytics techniques help providers save time and money spent on investigating and detecting fraud, thus increasing the overall efficiency of their processes.

**References**

EXL (NASDAQ: EXLS) is a leading operations management and analytics company that designs and enables agile, customer-centric operating models to help clients improve their revenue growth and profitability. Our delivery model provides market-leading business outcomes using EXL’s proprietary Business EXLerator Framework®, cutting-edge analytics, digital transformation and domain expertise. At EXL, we look deeper to help companies improve global operations, enhance data-driven insights, increase customer satisfaction, and manage risk and compliance. EXL serves the insurance, healthcare, banking and financial services, utilities, travel, transportation and logistics industries. Headquartered in New York, New York, EXL has more than 27,000 professionals in locations throughout the United States, Europe, Asia (primarily India and Philippines), South America, Australia and South Africa.

© 2018 ExlService Holdings, Inc. All Rights Reserved.
For more information, see www.exlservice.com/legal-disclaimer