BIG-DATA-MEDICARE-FRAUD-DETECTION

Taksh Rastogi
(Business Analytics & Big Data)
(502204181)

I. Abstract

II. People's lives depend on healthcare, which is why it must be reasonably priced. The healthcare sector is a complex system with many moving parts. Rapid growth is being experienced. At the same time, fraud is becoming a serious issue in this sector. Misuse of the medical insurance systems is one of the problems. In the healthcare sector, manual fraud detection is a taxing task. Recently, automated methods for spotting healthcare scams have been developed using machine learning and data mining approaches. In this paper, we attempt to provide a review of healthcare frauds and the methods for spotting them. Several existing studies were looked at in the literature work, with an emphasis on the methods utilised, identifying the significant sources, and the characteristics of the healthcare data. This evaluation has led to the conclusion that future research will focus on sophisticated machine learning techniques and newly discovered sources of healthcare data in an effort to reduce healthcare costs, increase the efficiency of healthcare fraud detection, and improve the quality of healthcare systems. In this paper's analysis of recent studies, machine learning and data mining are used to identify fraud in the healthcare sector. Additional research is required to identify many odd patterns of health insurance system abuse, and more advanced machine learning techniques can be applied to improve results.

Keywords: Machine Learning, Fraud Detection, Healthcare, Review, Anomaly Detection

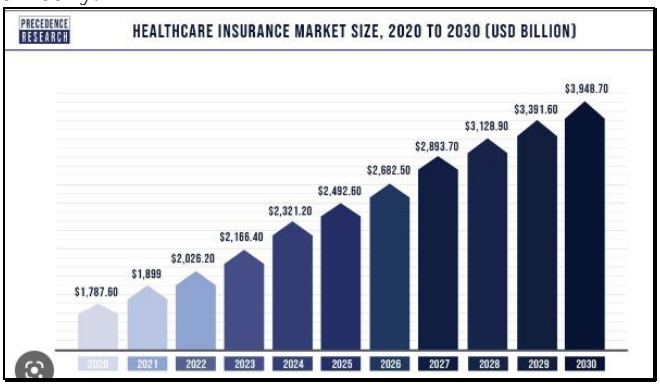
II. Introduction

Access to healthcare has always been a part of people's lives, and this trend will persist. The human body is a sophisticated mechanism. Therefore, it is essential to have medical professionals that have the expertise and training to recognise and treat diseases in diverse body parts. This forces medical professionals to employ a range of therapy modalities for patients with various specialties. The industry's objective is to successfully serve as many patients as possible. Each therapy and service supplied, however, has a price. It is necessary to pay for the time and skills of doctors, drug dealers, and medical staff, as well as a variety of medical amenities. Often, the patients are unable to cover these expenses. Consequently, insurance programmes are used to divide costs among all patients in order to in the healthcare system and to fund the necessary staff and resources. As with any insurance programme, there is a chance for abuse or fraud.

False statements are made in order to profit in white-collar crimes like healthcare fraud.

In addition to its financial impact, fraud has a significant negative influence on the perception of the integrity and usefulness of the data in the healthcare system.

The national health spending increased 4.6% in 2018 to reach 3.6 trillion dollars, according to the Centers for Medicare & Medicaid Services, a division of the Department of Health and Human Services. In terms of the number of claims, this amounted to $11,172 per person. Additionally, according to the National Healthcare Anti-Fraud Association, fraud in the healthcare industry costs the country tens of billions of dollars annually.



The burden of recovery for this significant financial loss is on insurance providers, but patients should have priority. Patients are duped into covering the expense largely in two ways: by paying fictitious co-pays and paying increased insurance premiums. Therefore, it is important to identify the trends in healthcare fraud and implement preventative steps to stop these crimes.

The perception of healthcare fraud as a serious social issue is growing.

Healthcare fraud is undoubtedly a problem for the government, and better methods of detection are needed. To spot healthcare fraud, a lot of effort and in-depth medical knowledge are required.

Healthcare fraud detection has always relied heavily on the knowledge of subject matter specialists, which is inaccurate enough, expensive, and time-consuming. A small number of auditors must manually evaluate and pinpoint the dubious medical insurance claims in order to manually discover healthcare fraud. But more was made possible by contemporary advancements in data mining and machine learning approaches. detection of healthcare fraud that is effective and automated. In

recent years, there has been an increase in interest in mining healthcare data for fraud detection. The many methods for identifying fraudulent behaviours in health insurance claim data are reviewed in this research.

Related work

Healthcare fraud has considerably inflated loss for Healthcare fraud has significantly increased losses for people, businesses, and governments. Combating healthcare fraud has become a crucial issue. As a result, some academics have created techniques for detecting healthcare fraud. Finding, spotting, and reporting frauds as they emerge in the system are the responsibilities of fraud detection systems [1], [2].

Typically, fraud detection is performed in one of two ways. Earlier, manual fraud audit laws were put in place to catch the fraud [3]. The The auditing process calls for in-depth knowledge in that field. These procedures take a long time and are the result of intricate interactions. It requires time-consuming, laborious physical labour. Tools for detecting fraud effectively were created as a result. This highly developed computer-based systems include a wide range of data mining techniques and strategies [1], [3]. Therefore, it is necessary to consider the different types of fraud, healthcare data, and fraud

III. Literature Review

\ For the purpose of identifying healthcare fraud, the majority of researchers use data on healthcare that has been made available by the Centers for Medicare and Medicaid Services (CMS).

Srinivasan et al. [4] proposed an anomaly detection method using Rule-based Data Mining, an unsupervised technique, on the insurance claims data received from Medicare data. Big data has been used to create applications for auditing health insurance claims that look for waste, fraud, and mistakes. These programmes helped private health insurers detect hidden cost overruns that transaction processing systems ignore and were used to spot irregularities in medical insurance claim submissions.

Using medical data from Medicare and Medicaid, Branting et al. combined supervised methods, graph analytics, and decision trees [5]. They offered a technique that involved using network algorithms to graphs made from open-source data sets in order to determine the probability of healthcare fraud.

A study using CMS 2012 data was able to ascertain a physician's practise style by looking at their prior schooling [6]. By providing a regional study together with the national distribution of school procedure payments and charges, they examined medical school fees, procedures, and payments as well as looked for any potential anomalies in the data. In an effort to identify the medical professionals who practise what they do best, the authors examine for connections between physicians' educational backgrounds and the procedures and treatments they use are abusing or ineffectively utilising medical insurance systems.

Using 2012 CMS data, Ko et al. expressly only took into account the field of urology [7]. The diversity in service consumption and payment among urologists is examined by the authors in an effort to estimate savings from an uniform service utilisation.

A machine learning model was developed in a study using the 2013 CMS dataset to identify when doctors submit medical insurance claims that are out of the ordinary [8]. It seeks to ascertain whether and when doctors are acting outside the bounds of their particular specialties, which may be an indication of abuse, fraud, or ignorance of billing practises.

By evaluating precision, recall, and Fscore using 5-fold cross-validation, the model is assessed. The multinomial Naive Bayes algorithm is employed. The results demonstrate that it is possible to successfully utilise machine learning in a novel method to categorise physicians into their respective disciplines using only the procedures they bill for. It identifies physicians who are potentially misusing healthcare insurance systems for further investigation.

III. Data Description

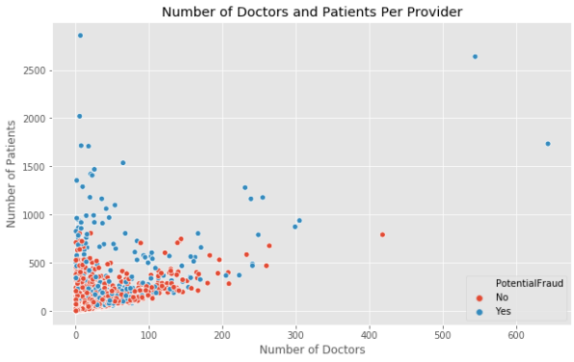
The data was originally split up into eight separate csv files. The dataset with the labels "train beneficiary," "train outpatient," "train inpatient," and "train providers highlighted" contained four of these files. Test beneficiary, Test outpatient, Test inpatient, and Test providers were the other four files that belonged to the unlabeled dataset (providers not flagged as potentially fraudulent). There were 5,410 providers and 558,211 claims in the labelled data. There were 1,353 providers and a total of 135,392 claims in the unlabeled data.

On the labelled dataset, exploratory data analysis, feature engineering, and supervised machine learning were carried out. For unsupervised K-means clustering, just the unlabeled dataset was employed.

Exploratory Data Analysis (EDA)

The beneficiary, inpatient, and outpatient datasets were combined prior to any analysis. Combining the data, we made the decision to first analyse the dataset at the level of the patients and claims because this was also the data's natural format. The following is an example of a question we asked that gave us some of the most important insights:

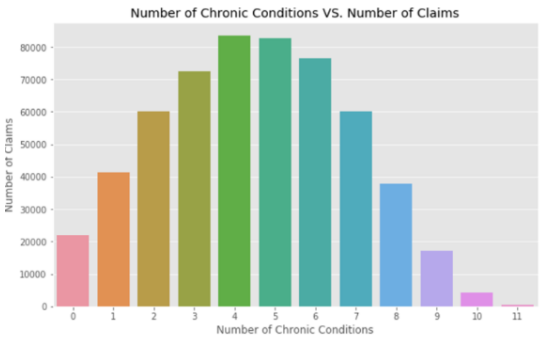
- **Do the number of doctors and patients affect the probability of encountering potentially fraudulent providers?**



The scatter plot above shows that the likelihood of the provider being possibly fraudulent grew as the number of patients, physicians, or both increased. Less instances of possible fraud among providers emerged as the number of patients and physicians declined. This suggests that larger providers, such

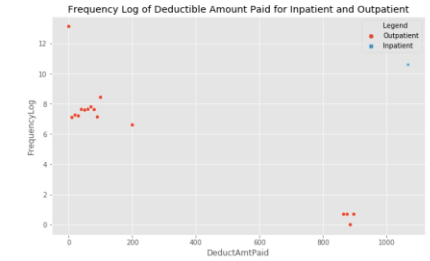
as larger hospitals with wider networks, may be more likely to commit fraud.

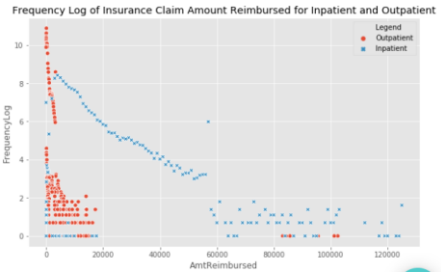
- **Would patients with more chronic conditions have greater number of claims filed in contrast to patients with less chronic conditions?**



Although we anticipated a positive link between a patient's number of chronic conditions and the number of claims made on their behalf, we discovered that patients with 4-6 chronic conditions had the largest number of claims made; the graph depicts this distribution normally.

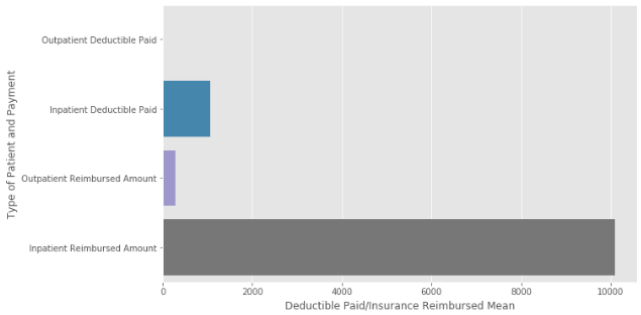
- **How are deductible amounts and insurance reimbursed amounts distributed for inpatients and outpatients?**



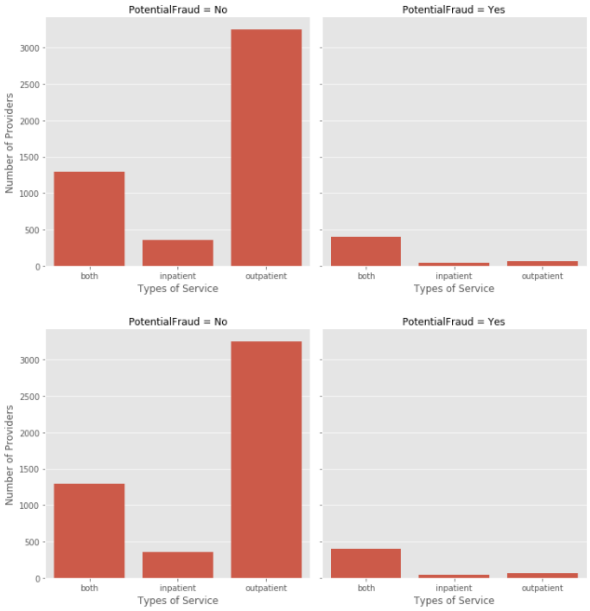


Analysis

The outpatient deductible amount paid is more variable, with a broader variation between $0 and $200, as can be seen in the top graph, whereas the inpatient deductible amount paid is steady at a value of roughly $1100. The most typical value is still 0, though. The bottom graph demonstrates that, with a distribution between zero and twenty thousand dollars, the outpatient insurance claim reimbursement amount similarly tends to be close to zero.
In contrast, the range of values for the inpatient insurance claim reimbursement was much larger and greater, with the greatest sum being reimbursed at about $120,000. This shows that the cost of inpatient services is far higher than the cost of outpatient services. The graph below displays the average expenses for both inpatient and outpatient care..

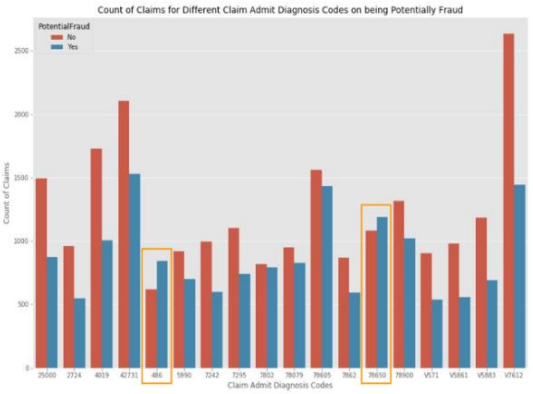


After analyzing the data at the level of the patients and claims, we delved deeper by examining the data from the provider angle:



We discovered that the proportion of non-fraudulent providers delivering only outpatient services was far greater than the proportion providing either only inpatient services or both inpatient and outpatient services. However, compared to providers who only offer inpatient or outpatient care, there are significantly more fraudulent providers who offer both types of services. This further suggests that larger providers are more likely to be dishonest.

- **Are the total counts of claims for different claim admit diagnosis codes greater for potentially fraudulent or non-fraudulent providers?**



As can be seen in the graph above, non-fraudulent providers surprisingly had the larger counts of claim admit diagnosis codes with exceptions to two codes: 486 and 78650. Thus, further research and analysis should be done on these two codes.

.

Data on Feature Engineering

The creation of a workable data frame, which needed a significant amount of effort and imagination, was one of our most difficult jobs for this project. As was previously noted, three datasets of features at the level of patients and claims were supplied to us, although the target was at the level of the providers who were labelled as possibly being involved in fraud or not. As a result, in order to develop a new data set based on the providers, we had to combine and modify the inpatient, outpatient, and beneficiary data. 79 characteristics were present in the combined datasets. The categorised breakdown of the feature is shown in the chart below.

Features' Categories		
Days Admitted	Financial	Age
Race	Type of Service	Claims
States	Counties	Chronic Conditions
Diagnosis Codes	Procedure Codes	Gender
Number of Patients	Number of Doctors	Attending / Operating Physicians

Examples

Here are some examples of how we combined or created new features from the aforementioned categories to modify the original features:

[1] Age: based on the claim start date and the patient's birthdate, age was computed for patients; age was then mapped to providers by figuring out the typical age of patients serviced by those providers.

[2] States: counted the states in which the providers were active Counties: counted the counties in which the providers were active Chronic Conditions: a total of 12 chronic conditions, including Alzheimer's disease and ischemic heart disease, were listed. For each provider, the number of patients with each condition was examined.

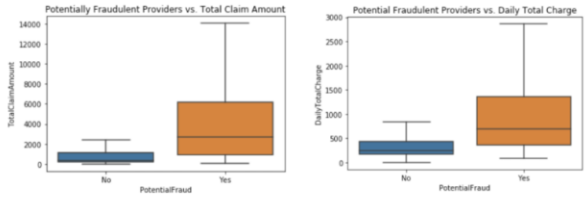
Whether the provider provided inpatient, outpatient, or both services is indicated by a dummified or label-encoded variable referred to as the service type.

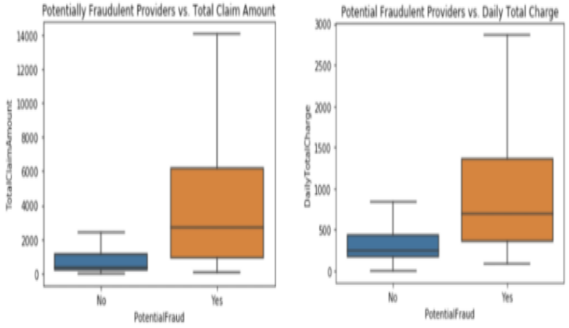
In-Depth Explanation

Now, we will provide more in-depth explanations for a few of our more important engineered features beginning with the features in the financial category.

Outpatient deductible values were concentrated around $0, as was shown in the EDA section, whereas inpatient deductible values were fixed at $1068. For the amount of insurance reimbursement, there was, however, a little more differentiation. We discovered that fraudulent providers received a $20 greater median outpatient insurance reimbursement than non-fraudulent providers. For fraudulent providers compared to non-fraudulent providers, the inpatient median was around $1,000 more.

We chose to develop additional features integrating this data with the





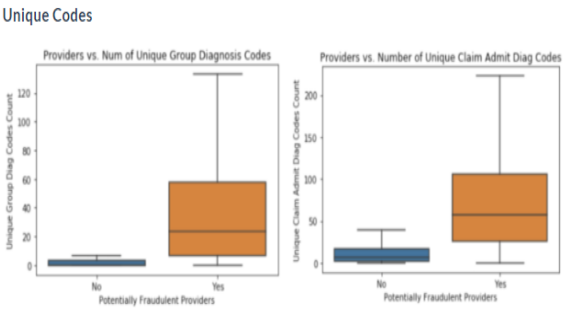
By extracting the tweets from Twitter API and the tweets were organized into the word cloud to analyze what words have been frequently used by Twitter users and also what is the emotion behind these words. It can be seen from the figure that words like Insurance, Fraud, healthcare, explanation, liability and crime were frequently used by the users in the context of the research topic.

Findings

After integrating the features, we can more clearly distinguish between financial features that are fraudulent and those that are not. The total per claim median on the left plot is roughly $340 for providers who don't commit fraud and $2700 for those who do. The overall number of claims for fraudulent providers is typically $2500 higher than the total amount of claims for legitimate suppliers. We found that, on average, fraudulent providers charge $470 more per day than non-fraudulent providers (see the right plot), which is the distribution of the daily total charge.

The findings in relation to the financial data are illuminating and clear. In order to spread the fraudulent behaviour and avoid detection, it makes more sense to conceal the fraud within the whole claim than to overcharge in a single area where there is a more obvious set pattern.

The "Unique Group Diagnosis Codes Count" and "Number of Unique Claim Admit Diagnosis Codes" are two additional remarkable characteristics that we designed. Patients are categorised into different categories using group diagnosis codes based on related diagnoses and costs. The first diagnosis at admission is specified by claim admit diagnosis codes. The unique group diagnostic codes and unique claim admit diagnosis codes used in provider claims were counted to derive the individual features.



When compared to non-fraudulent providers, fraudulent providers utilise, on average, 38 more distinct group diagnosis codes than they do, with a median of 23 for the former. In comparison to non-fraudulent providers, fraudulent providers employ 67 more distinct claim admission diagnosis codes on average, with a median of 59 compared to 8 for non-fraudulent providers.

Referring to the EDA once more, we saw that fraudulent providers did not submit more claims overall or per code. Instead, we discover that a relevant indicator in this case is the total number of unique codes used. This connects to networks and service offerings in general. The majority of the providers who have been labelled as fraudulent are more advanced and have larger networks.in bigger hospitals and are operating within both inpatient and outpatient.

Therefore, it is unquestionably true that fraudulent suppliers will employ a greater number of unique codes. This was a highly intriguing discovery, and a further examination of the number of distinct group diagnosis codes will be provided in a subsequent section.

Once we had our initial dataset, we ran Extra Trees Classifier to determine the relevance of the features and Lasso Regression to determine which features should be eliminated. Our top five crucial features for spotting dishonest providers were "Number of Unique Group Diagnosis Codes", "Number of Unique Claim Admit Diagnosis Codes", "Service Type", "Total Claim Amount", and "Daily Total Charge".

Penalized Logistic Regression

We then wanted to test the validity/strength of our features, so we performed penalized logistic regression.

Features	Train Accuracy Score	Test Accuracy Score
• Number of Duplicated Beneficiary IDs • Patients with 12 Chronic Conditions	0.65	0.63
• Number of Duplicated Beneficiary IDs • Patients with 12 Chronic Conditions • Total Claim Amount	0.76	0.76
• Number of Duplicated Beneficiary IDs • Patients with 12 Chronic Conditions • Total Claim Amount • NumUniqGroupDiagCode	0.85	0.85

[3] First, using the fraudulent flag as the target, we ran the penalised logistic regression on the two features that performed the worst (the number of duplicate beneficiary IDs and patients with 12 chronic diseases). With a train accuracy score of 0.65 and a test score of 0.63, this model did not perform as well as was predicted.

[4] The performance dramatically improved once we included one of our best features, total claim amount, with train and test accuracy scores at 0.76. The number of distinct group diagnosis codes was added as our final feature, and the train and test accuracy scores were raised to 0.85. Due to the growth of the accuracy ratings, we were thus reassured of the power of our features and that they created reliable models. We do not currently have near-perfect accuracy scores, which was also predicted, but the train and test scores were similar, indicating that overfitting was not an issue.

EDA, feature engineering, and modelling iterations later, we entered our final machine learning models with

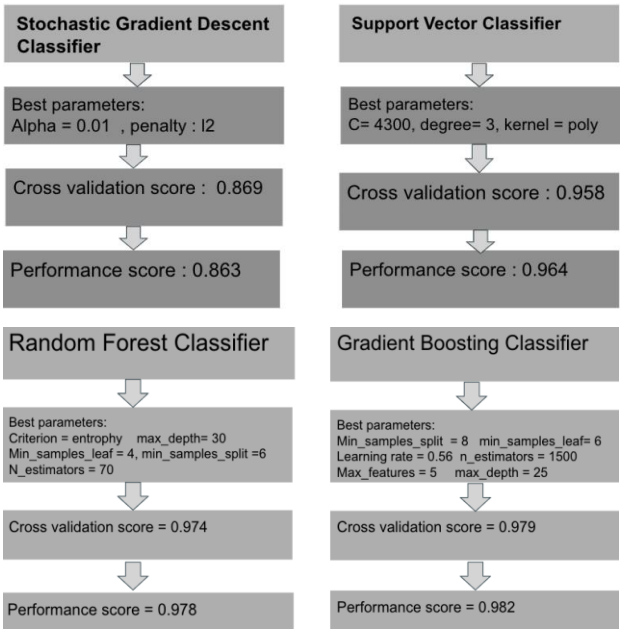
IV. PROCEDURE, & METHODS(ALGORITHMS)

Machine Learning Models

[5] We have to address the dataset imbalance before fitting the machine learning models. Only 10% of the sample was made up of fraudulent providers. We increased the sampling of the minority class to solve this issue. We added a random sample of students from the minority class to our initial data set.

[6] Classifiers that are linear and nonlinear

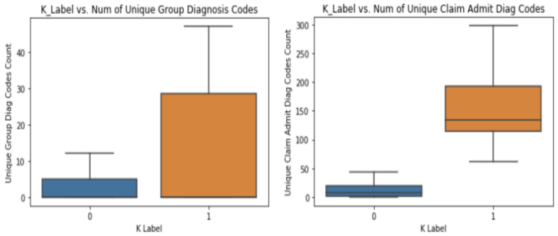
Then, we created both linear and non-linear classifiers, comparing the accuracy ratings of each. Using Scikit-GridSearchCV, learn's we adjusted the hyperparameters for each model to enhance the outcomes.



The stochastic gradient descent, a linear classifier, was outperformed by the non-linear models. The Random Forest and Gradient Boosting classifiers had better results.

K-Means

On the unlabeled (test) dataset, we also used K-Means clustering and added those K-labels. Then, we conducted some analysis on this data to determine the key characteristics from the earlier (labelled)data.

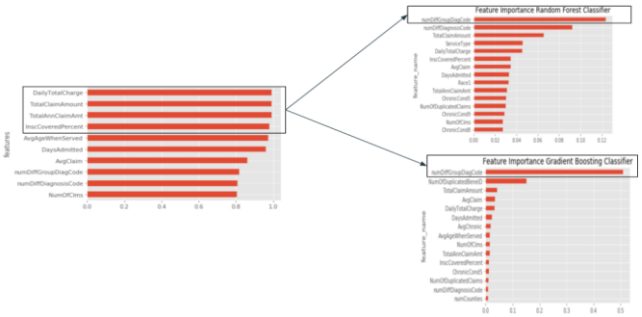


[7] The two groups designated 0 and 1 inside the particular group diagnosis codes and claim admit diagnostic codes features were clearly distinguished, as was the case in our research. The minority class in this dataset is the class with the label 1, and the minority class in the labelled dataset was the fraudulent class.

[8] The key lesson from this is that there is differentiation in higher dimensions and that there are two distinct portions that are unmistakably divided by kmeans. This is helpful in real life where we are not given labels in advance since it helps us to comprehend that the data has structure.

Final Analysis

We conducted a unary classification to check the accuracy of individual features and compared them to the most important features from the Random Forest and Gradient Boosting classifiers.



[9] The three categories on the left with the greatest accuracy ratings were Daily Total Charge, Total Claim Amount, and Total Annual Claim Amount. However, the Random Forest and Gradient Boosting classifiers on the right both demonstrate that the most crucial attribute was the number of various group diagnosis codes.

[10] Before we go back to the topic of the number of different group diagnosis codes, let's first have a better understanding of what a group diagnostic code is. The Diagnosis-Related Group Code (DRG) is a system for categorising patients into specific groups. A payment weight is applied to each DRG.

For instance, if a doctor simply notes "appendicitis" before conducting an appendectomy, the lowest or neutral DRG category will be used. A higher DRG category will be used if "acute appendicitis" is listed as the ailment. A discrepancy in these areas could result in cost increases of up to threefold. This illustrates how the power of the "DailyTotalCharge" feature and other connected features in the financial category are impacted by the "numDiffGroupcode" feature.

Where might a potential abnormality originate?

Upcoding

Instead of the accurate diagnosis, which would have resulted in a lower DRG, a doctor purposefully entered a more serious diagnosis, increasing the DRG.

The group diagnosis code was upgraded by a medical coder to make a patient appear to be in a more serious condition on the claim.

Unbundling

A DRG payment typically covers all costs related to an inpatient stay from the time of admission to the time of discharge. Unbundling is the process of invoicing a number of separate codes for a collection of operations when really only one all-inclusive code should have been used.

Recommendations

[11] Because there are so many codes and claims sent every day, it is challenging to identify fraudulent claims. Clinics and hospitals with a large patient population and a wide network of physicians are examples of providers who are deemed dishonest.

By developing profiles of patients with chronic diseases, we reasoned that a market basket analysis and network concept map (shown below) could be helpful. The inconsistencies of the group diagnosis codes utilised may then be found using the profiles.

V. CONCLUSION

O The following characteristics were the most crucial for identifying dishonest providers: Unique Group Diagnosis Codes, Unique Claim Admit Diagnosis Codes, Total Claim Amount, and Service Type.For future work, we would like to tune the hyperparameters of the K-means model to affirm whether our label presumptions are accurate. Using market basket analysis, we would also like to examine fraudulent suppliers in greater detail and develop new features based on the correlations.

VI. REFERENCES

[1] Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. Journal of Network and Computer Applications, 68, 90-113.

[2] Behdad, Mohammad, et al. "Nature-inspired techniques in the context of fraud detection." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42.6 (2012): 1273- 1290

[3] Konasani, Venkatareddy, Mukul Biswas, and Praveen Krishnan Koleth. "Healthcare fraud management using big data analytics." An Unpublished Report by Trendwise Analytics, Bangalore, India (2012).

[4] Srinivasan, Uma, and Bavani Arunasalam. "Leveraging big data analytics to reduce healthcare costs." IT professional 15, no. 6 (2013): 21-28.

[5] Branting, L. Karl, Flo Reeder, Jeffrey Gold, and Timothy Champney. "Graph analytics for healthcare fraud risk estimation." In Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on, pp. 845-851. IEEE, 2016..

[6] Feldman, Keith, and Nitesh V. Chawla. "Does medical school training relate to practice? Evidence from big data." Big data3, no. 2 (2015): 103-113.

[7] Ko, Joan S., Heather Chalfin, Bruce J. Trock, Zhaoyong Feng, Elizabeth Humphreys, Sung-Woo Park, H. Ballentine Carter, Kevin D. Frick, and Misop Han. "Variability in Medicare utilization and payment among urologists." Urology 85, no. 5 (2015): 1045- 1051.

[8] Bauder, Richard A., Taghi M. Khoshgoftaar, Aaron Richter, and Matthew Herland. "Predicting medical provider specialties to detect anomalous insurance claims." In Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on, pp. 784- 790. IEEE,20