Name: Gouri Eknath Sutar

Email address: gouri.sutar8@gmail.com

Contact number: +91 9623441705

Anydesk address: **814 361 811**

Years of Work Experience: **5**

Date: **28th Jan 2023**

Self Case Study -1: Healthcare Provider Fraud Detection Analysis

1. **What Is a Healthcare Provider?**

   A healthcare provider is a person or entity that provides medical care or treatment. Healthcare providers include doctors, nurse practitioners, midwives, radiologists, labs, hospitals, urgent care clinics, medical supply companies, and other professionals, facilities, and businesses that provide such services.

2. **Who Are Healthcare Providers?**
   - The physical therapist that helps you to recover from your knee injury.
   - The home healthcare company that provides your visiting nurse.
   - The durable medical equipment company that provides your.
   - home oxygen or wheelchair.
   - Your pharmacy.
   - The laboratory that draws and processes your blood tests.
   - The imaging facility that does your mammograms, X-rays, and magnetic resonance imaging (MRI) scans.

- The speech pathologist (speech therapist) that works with you to make sure you can swallow food safely after a stroke.
- The outpatient surgery clinic where you had your colonoscopy done.
- The specialty laboratory that does your DNA test.
- The urgent care centre or walk-in clinic in your neighbourhood shopping centre.
- The hospital where you receive inpatient (or in some cases, outpatient) care.
- The emergency department staff that stabilise you in the event of an accident or serious illness.

3. **What is Healthcare fraud ?**

   Health care fraud is not a victimless crime. It affects everyone, individuals and businesses alike and causes tens of billions of dollars in losses each year. It can raise health insurance premiums, expose you to unnecessary medical procedures, and increase taxes.

   Healthcare fraud is an organised crime which involves peers of providers, physicians, beneficiaries acting together to make fraud claims. As per the U.S. legislation, an insurance company should pay a legitimate healthcare claim within 30 days. So, there is very little time to properly investigate this. Insurance companies are the most vulnerable institutions impacted due to these bad practices. As per the Government, the total Medicare spending increased exponentially due to frauds in Medicare claims.

Overview

A. Provider Fraud is one of the biggest problems facing Medicare. According to the government, the total Medicare spending increased exponentially due to frauds in Medicare claims. Healthcare fraud is an organised crime which involves peers of providers, physicians, beneficiaries acting together to make fraud claims.

B. Healthcare fraud and abuse take many forms. Some of the most common types of frauds by providers are:

   a) Billing for services that were not provided.

   b) Duplicate submission of a claim for the same service.

   c) Misrepresenting the service provided.

   d) Charging for a more complex or expensive service than was actually provided.

   e) Billing for a covered service when the service actually provided was not covered.

C. The datasets consists of -

   a. Train.csv and test.csv - This file consists of Provider Data , Inpatient Data, Outpatient Data, Beneficiary Details Data and various categorical features and also the PotentialFraud which we have to predict.

   b. Provider Data

      This data provides insights of Provider Id , PotentialFraud mark on that ids. we have to predict PotentialFraud for future data.

   c. Inpatient Data

      This data provides insights about the claims filed for those patients who are admitted in the hospitals. It also provides additional details like their admission and discharge dates and admit d diagnosis code.

   d. Outpatient Data

This data provides details about the claims filed for those patients who visit hospitals and not admitted in it.

e. Beneficiary Details Data

This data contains beneficiary KYC details like DOB, DOD, Gender, Race, health conditions (Chronic disease if any), State, Country they belong to etc.

D. The goal of this project is to " predict the potentially fraudulent providers " based on the claims filed by them.along with this, we will also discover important variables helpful in detecting the behaviour of potentially fraud providers. This is basically a classification problem.

Research-Papers/Solutions/Architectures/Kernels

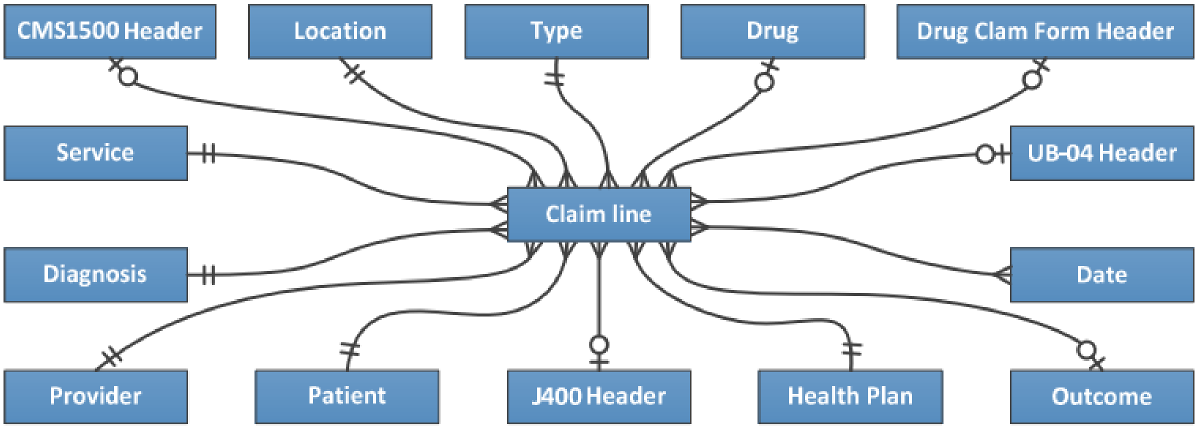
1. **https://drive.google.com/file/d/1DdTpjcmeJ9b5d29UTrEdFqhll86f-dBC/view?usp=sharing**
    a. Observations -
        i. In the US healthcare system, it is estimated that fraud, waste, and abuse result in a loss of approximately $700 billion. Due to its distributed management model, limited cross-program communications, and difficult-to-track patient population of low-income adults, their children, and individuals with certain disabilities.
        ii. Medicaid has been a particularly vulnerable target for fraud in recent years. Beyond the transaction level, one must examine the data for effective fraud detection. The

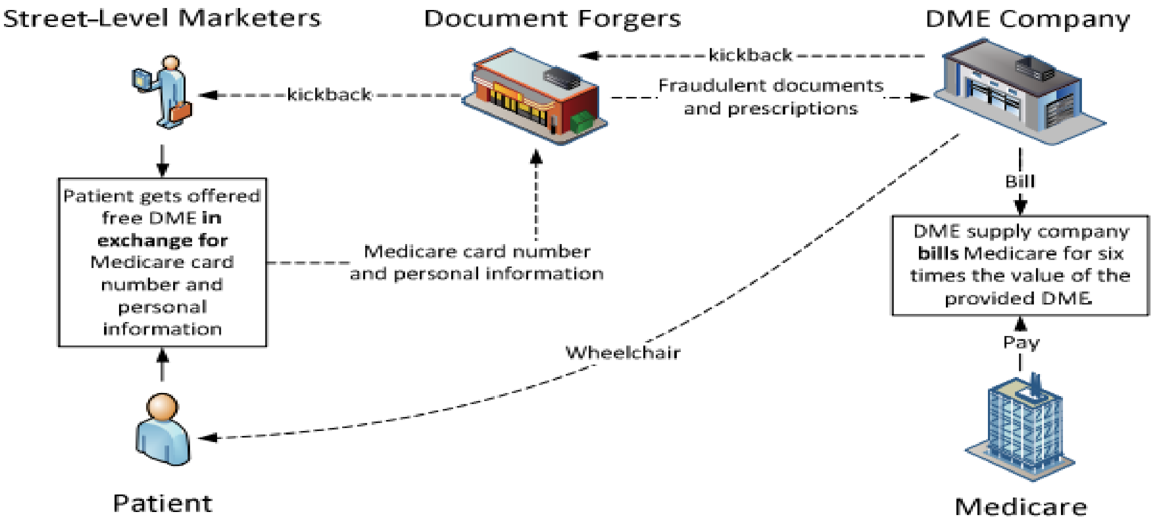
purpose of this paper is to develop a Medicaid multidimensional schema, as well as to provide a collection of multidimensional data models and analysis techniques that assist in predicting the likelihood of fraudulent activities.

b. Takeaways -
    i. Fraud is the intentional deception or misrepresentation of something that an individual knows to be false or does not believe to be true and makes, knowing that the deception could result in some unauthorised benefit to himself or herself or some other person.
    ii. Providers, beneficiaries (patients), and insurance companies are the three main actors involved in healthcare fraud. Since providers are the ones who start the process of billing insurers, they quickly become the centre of fraudulent schemes. When a provider joins Medicaid, they agree to the state's reimbursement rates and send claims directly to the state or managed care organisation for payment.
    iii. Before requesting Medicaid reimbursement, the patient pays the provider's bill if the provider does not participate in Medicaid. After processing the claim, the agency or insurer sends the patient an explanation of benefits outlining the paid-for services, their codes, and their costs.
    iv. This scheme includes multiple types of fraud, including phantom billing, medically unnecessary services to detect this and similar schemes, one could compare physician service profiles with their peers using a data model that

might have indicated this issue was the high cost of prescriptions.

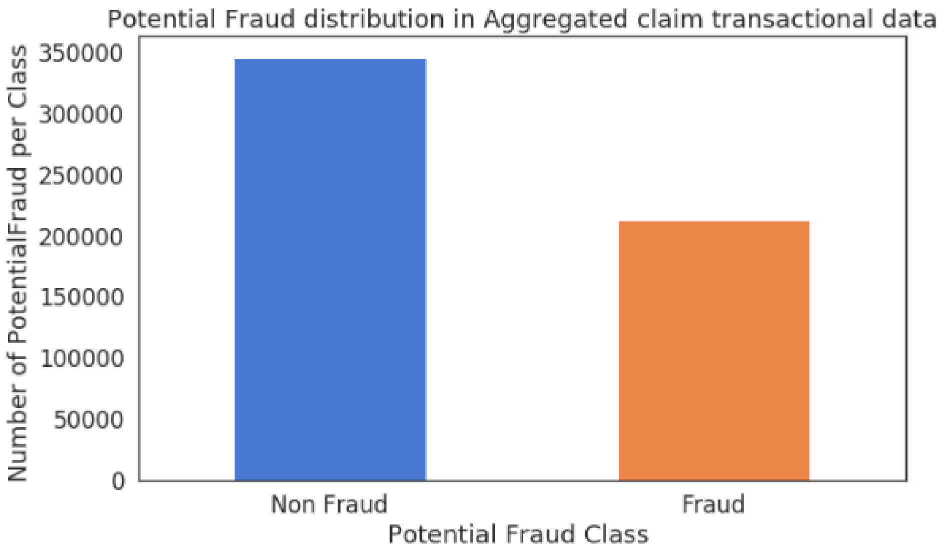
v. Identifying these patients as a suspect patient group reveals the clinic and possibly additional providers that may be involved in this scheme.



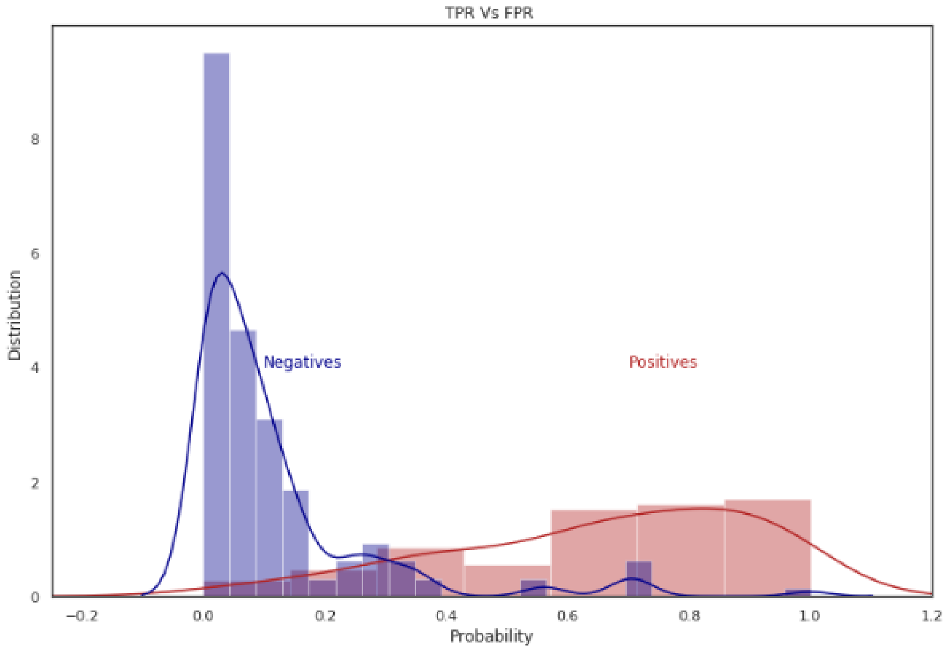

Street-Level Marketers
Document Forgers
DME Company
kickback
kickback
Fraudulent documents and prescriptions
Bill
Patient gets offered free DME in exchange for Medicare card number and personal information
Medicare card number and personal information
DME supply company bills Medicare for six times the value of the provided DME.
Pay
Wheelchair
Patient
Medicare

2. https://www.kaggle.com/code/abhi739/healthcare-fraud2

This is the kaggle solution that I went through.



This dataset is extremely unbalanced. There are 10% fraud providers and 90% Non fraud providers.



After a model has been constructed, probability thresholding can be used if the algorithm returns probabilities and score. You basically shift the classification threshold trade-off level that is appropriate. Typically, this can be improved by creating an evaluation metric curve. This poses a limitation. The accuracy of predicting the other class will also decrease if the cutoff is changed.

a. Observations -
    - test data does not contain all levels of variables in comparison to train data, so adding Train data to Test data will help you get good average scores for new features in Test data. Therefore, in order to evaluate results, our strategy will be to append test data to train data, generate new average features, and use only test data.
    - When compared to outpatient data, the number of claims is lower for inpatient data.
    - Even though there are fewer claims in inpatient data, there is a higher percentage of fraudulent activity in inpatient data 57% than in outpatient data 36%. This is because the amount that is reimbursed per claim for inpatient care is significantly higher than the amount that is reimbursed per claim for outpatient care 35 times the earlier calculation.
    - The scatter plot reveals that the likelihood of a fraudulent transaction is high when a patient is younger than 60 years old and has a claim period of more than 20 years.
    - There are a lot of everyday transactions that are categorised as frauds. One can set a threshold to make a tradeoff between fraud class predictions and non-fraud class predictions based on a business decision. We will be

able to better understand providers' fraudulent behaviour and perform better at detecting new fraud patterns by adding more data and training over time.

Takeaways:

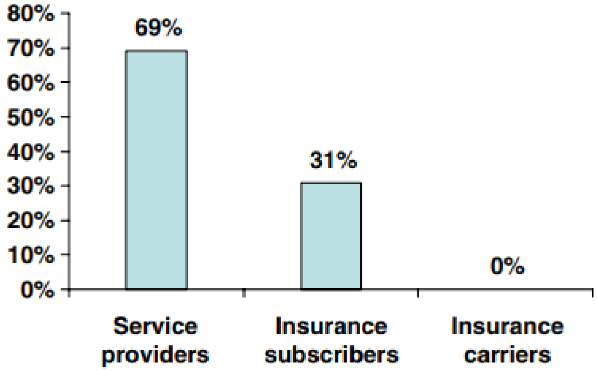
- In This solution they used to train data variables from test data so it will have the effect of data leakage so it will not perform well in real world data.
- Need to go with each feature correlation with available csv files to get a better relation to predict fraud with best feature results. For that we can use the relation plots to compare relation scores with that.

3. **https://cpb-us-w2.wpmucdn.com/sites.gatech.edu/dist/4/216/files/2015/09/p70-Statistical-Methods-for-Health-Care-Fraud-Detection.pdf**

   Abstract:

   The summary, categorization, and comparison of statistical methods used to detect health care fraud, the discussion of the most crucial steps in data preprocessing, and the identification of the primary sources and characteristics of the data used to detect fraud. The results of this survey are used to talk about the areas of existing research that haven't been covered enough or to suggest new areas for research.

   The development of scalable, accurate, and rapid fraud detection algorithms and methods is the overarching goal of this field of study. Dependent on the availability of labels in the training dataset, statistical methods for detecting health care fraud typically fall into two categories: supervised and unsupervised. This section will first go over the most frequently used methods in each class before talking about how the two classes work together.

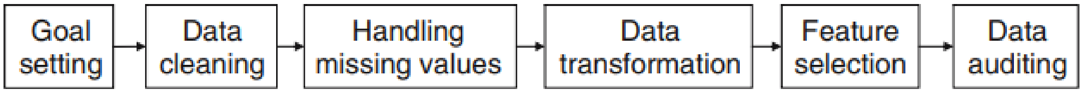


1 Percentages of papers on detecting three types of fraud

Setting a goal:

The majority of current methods aim to distinguish between fraudulent and legitimate cases, but this is only the intermediate goal of fraud detection. The ultimate objective is to locate and eliminate the factors that lead to fraud so that it can be avoided in the future.

This step aims to identify and prioritise the kinds of fraud that should be the focus of detection. Experts in the domain as well as statistical analysts are needed for this step to be successful. Domain experts know which kinds of fraud are most likely to occur or cause the most financial damage, so they prioritise finding these kinds of fraud first.



From a data analysis standpoint, statistical analysts provide expert opinions to determine whether the data used for fraud detection meets the requirements for effective statistical analysis, such as sample size and data quality.

Health care data: sources, characteristics,and preprocessing

The primary source of raw data for the detection of health care fraud is the insurance industry, which includes both public health departments and private insurance companies.

The US Health Care Financing Administration is one of the major government health departments that have been mentioned in the literature. The majority of the raw data, regardless of where they come from, are insurance claims.

A service provider and an insurance subscriber are both involved in a claim for insurance. There are two characteristics of the claim data.

1. They have a large number of characteristics that can be used to identify the kinds of fraud committed by the service providers and insurance customers in question.

2. The service provider and insurance subscriber involved in each claim typically have their own distinct identifiers. It is possible to obtain a global view of a service provider's behaviours over time and across various insurance subscribers, as well as a global view of an insurance subscriber's behaviours over time and across various service providers, by linking various claims using the unique identifiers.

Future Enhancement:

To improve the determination of health insurance companies in the detection system for fraud. Subsystems in their system are used to identify fraud committed by service providers, insurance customers, and their employers. To prevent overfitting, this system employed a method known as "early stopping." When training a Neural Network, this method makes use of two distinct datasets: When the network begins to overfit the data, one is used to update the weights and biases, and the other is used to stop training.

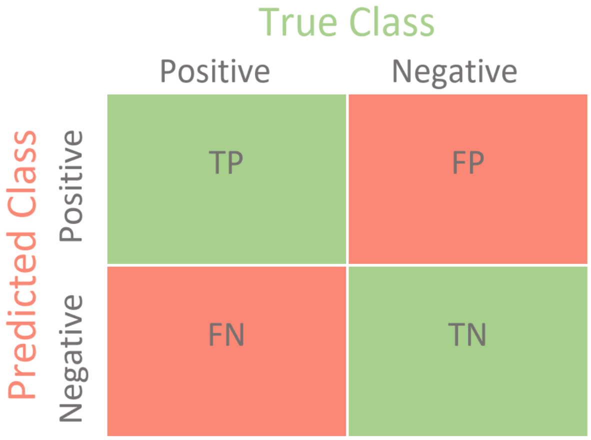
Performance metric:

Since there are few cases of healthcare fraud in the dataset, "accuracy" will not be the appropriate metric. Plotting the confusion matrix will be an important first step. The misclassification, False Positive (FP) and False Negative (FN), must then be examined. FN indicates that the cases the model predicted are genuine but are actually fraudulent. FP indicates that the cases the model finds are fraudulent, but in reality, they are legitimate.

Therefore, the performance indicators are:

a) Matrix of Confusion:

From the confusion matrix, we can train a model to determine whether a provider is fraudulent or not.The table is where the counts for TP, FP, TN, and FN will be plotted. The model's performance can be seen from this table. We can predict the model is



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

b) F1 Score:

F1 score can be interpreted as a measure of overall model performance from 0 to 1, where 1 is the best. To be more specific, F1 score can be interpreted as the model's balanced ability to both capture positive cases (recall) and be accurate with the cases it does capture (precision). It is the harmonic equivalent of recall and precision.

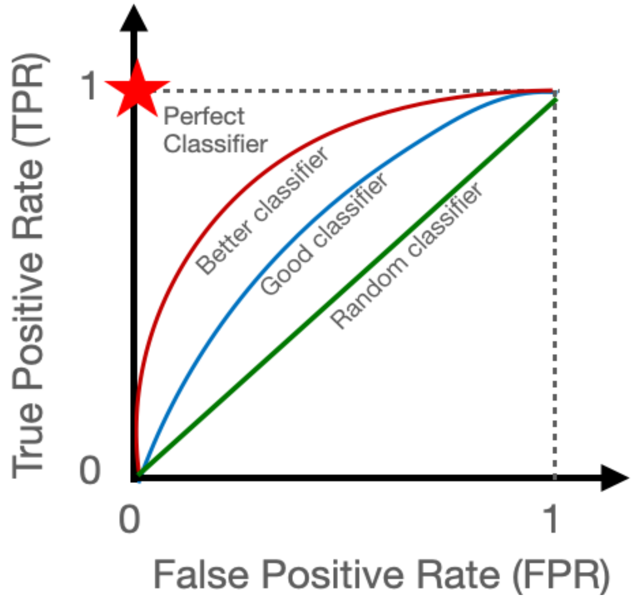
Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

It is the formula for the F1 score. The F1 score, which includes both Precision and Recall, will be the appropriate metric for this issue.

c) **AUC Score:**

"AUC" stands for "Receiver Operating Characteristics" (ROC). For various thresholds, ROC plots True Positive Rate (TPR) in relation to False Positive Rate (FPR). The predicted probability score's ranking is what determines the area under the curve, not absolute values.



The accuracy evaluated using the AUC. The test's accuracy increases as the ROC curve approaches the graph's upper left corner, where the sensitivity and false positive rates equal 0 and the specificity equals 1. As a result, the ideal ROC curve has an AUC of 1.0.

d) The FPR and FNR:

We need to check the False Positive Rate (FPR) and False Negative Rate (FNR) separately because the cost of misclassification is very high. The FPR should be as low as possible to get a more relevant result of the model.

First Cut Approach

Based on the research and readings that I have done. I will follow the below steps -

1. In this dataset we have to clean that data and then combine the different aspect as we know here 4 different files of the train dataset i.e. Provider, Inpatient, outpatient, beneficiary, which contain procedures, Chronic Conditions etc.
2. This is a binary classification problem where we have to identify the fraud data.
3. Feature engineering is one of the most important aspects of this Kaggle problem. So I will look to come up with various aggregated features based on procedure data. Features like Chronic Conditions done by a particular beneficiary holder, Reimbursement Amount, Diagnosis Group Code and many more other features.
4. As the dataset in healthcare fraud is highly imbalanced fraud cases , ‘accuracy’ won’t be the proper metric. An important initial step will be

to plot the confusion matrix. we need to check the misclassification of data.

5. The model creation part was seriously impacting the Accuracy and F1 score. So we need to come up with architecture which will help us in training different models on top of that. Finally we can stack the predictions using logistic Regression and best fit models like Decision Tree, SVM, KNN, etc.
6. For hyperparameter tuning, Optimization seems like a nice strategy as it will help us to come up with the best hyperparameters.