Name: Shiva Gupta

Email address: Shivagupta2706@gmail.com

Contact number: +91-8299407616

Anydesk address: 125 304 535

Years of Work Experience: 7 years

Date: 16th Aug 2022

Self Case Study -1: ***HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS***

"After you have completed the document, please submit it in the classroom in the pdf format."

Please check this video before you get started:
https://www.youtube.com/watch?time_continue=1&v=LBGU1_JO3kg

Overview Section:

Introduction :

Fraud is **wrongful or criminal deception intended to result in fnancial or personal gain**. Nowadays, fraud is very common in any area like tax fraud, credit card fraud e.t.c.. It is also very common in the Healthcare area. Health care fraud is a type of fraud involving the use of the health care system by an individual, medical provider, or insurance company in a deceitful manner in order to proft from it. Nowadays, Insurance institutions face many false claims and duplicate claims. And the time to approve and verify the claims is very less, around 15-30 days.So it is not possible to check each and every claim manually. In many cases, Healthcare providers fll the details in

insurance form and fle the claims at the place of other benefciaries which is very hard to fgure out that is legitimate or fraudulent without checking ground truth. So here we will analyse and detect “Healthcare Provider Fraud” based on given data.

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.

Ref: HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS | Kaggle

Business Problem:

Many providers make false claims at the place of other benefciaries so Insurance institutions are more vulnerable in healthcare fraud due to this reason, insurance companies increased their insurance premium and as a result healthcare is becoming costly data by day.

Here, our task is to minimize the false claims by providers and to do so, we will have to provide a solution which will predict the potential fraudulent providers based on previous claims fled by them. Along with it, we will focus on discovering the important variables which help to detect the behaviour of

potential fraudulent provers. And using this solution, we will be able to make the right decision for the false claims which will save the money from false claims and prevent the Insurance company from fnancial loss.

ML Formulation:

For this business problem, given data is labelled data and the target variable has discrete values hence we can easily formulate this problem as a supervised classifcation problem.

The discrete values of the target variable are “Fraudulent” and “Legitimate”. So Binary classifcation models will be built based on claims (fraudulent or legitimate) by providers to predict whether a provider is potentially fraudulent or not.

Business constraints:

1. Model should be interpretable which helps to tell the reason behind predicting any query point as Fraudulent or Legitimate. Which will be helpful for non-technical people or business people.
2. Misclassifcation cost will be high because if the model predicts fraudulent where it is actually “Legitimate” then we will lose customer trust and if the model predicts “Legitimate” where it is actually “Fraudulent” then Insurance companies have to bear that claim amount. So we have to focus more on false positives(FP) and false negatives (FN) while training models.
3. The insurer will have enough time to settle the claim amount to providers. So there is no restriction for low latency.

Data set column analysis:

We are considering Inpatient claims, Outpatient claims and Benefciary details of each provider for this problem. The dataset is downloaded from Kaggle where data is splitted already in the Train and Test dataset.

Source data Link: HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS | Kaggle

Here, we will do column analysis of the Train dataset for Inpatient claims, Outpatient claims and Benefciary details of providers.

a. Inpatient Data (Train/Test data):

This data provides insights about the claims fled 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. In this dataset, there are 30 variables and a lot of information is missing in it.

Columns	Type	Description
BeneID	object	Beneficiary unique id
ClaimID	object	Claim unique id
ClaimStartDt	object	Claim start data (format: yyyy-mm-dd)
ClaimEndDt	object	Claim end data (format: yyyy-mm-dd)
Provider	object	Provider unique id
InscClaimAmtReimbursed	int64	reimbursed insurance claim amount
AttendingPhysician	object	Physician who attended the patient
OperatingPhysician	object	Physician who operated the patient

OtherPhysician	object	Physician who was present in the treatment of patient rather than attending and operation physician
AdmissionDt	object	date when patient admitted to hospital
ClmAdmitDiagnosisCode	object	code of initial diagnosis when patient admitted with initial symptoms
DeductibleAmtPaid	float64	the amount that a policy holder has to pay before the insurance company starts paying up
DischargeDt	object	date when patient discharge from hospital
DiagnosisGroupCode	object	group code of diagnosis which was performed on patient
ClmDiagnosisCode_1	object	Code of first diagnosis of patient which was added as first diagnosis of claim by provider
ClmDiagnosisCode_2	object	Code of second diagnosis of patient which was added as second diagnosis of claim by provider
ClmDiagnosisCode_3	object	Code of third diagnosis of patient which was added as third diagnosis of claim by provider
ClmDiagnosisCode_4	object	Code of 4th diagnosis if any
ClmDiagnosisCode_5	object	Code of 5th diagnosis if any
ClmDiagnosisCode_6	object	Code of 6th diagnosis if any
ClmDiagnosisCode_7	object	Code of 7th diagnosis if any
ClmDiagnosisCode_8	object	Code of 8th diagnosis if any
ClmDiagnosisCode_9	object	Code of 9th diagnosis if any
ClmDiagnosisCode_10	object	Code of 10th diagnosis if any
ClmProcedureCode_1	float64	Code of 1st procedure that patient undergo for treatment
ClmProcedureCode_2	float64	Code of 2st procedure that patient undergo for treatment
ClmProcedureCode_3	float64	Code of 3st procedure that patient undergo for treatment
ClmProcedureCode_4	float64	Code of 4st procedure that patient undergo for treatment
ClmProcedureCode_5	float64	Code of 5st procedure that patient undergo for treatment

ClmProcedureCode_6	float64	Code of 6st procedure that patient undergo for treatment

b. Outpatient Data (Train/Test data):

This data provides details about the claims fled for those patients who visit hospitals and not admitted in it. In this dataset, there are 27 variables and a lot of information is missing in it.

Columns	Type	Description
BeneID	object	Beneficiary unique id
ClaimID	object	Claimfeciary unique id
ClaimStartDt	object	Claim start data (format: yyyy-mm-dd)
ClaimEndDt	object	Claim end data (format: yyyy-mm-dd)
Provider	object	Provider unique id
InscClaimAmtReimbursed	int64	reimbursed insurance claim amount
AttendingPhysician	object	Physician who attended the patient
OperatingPhysician	object	Physician who operated the patient
OtherPhysician	object	Physician who was present in the treatment of patient rather than attending and operation physician
ClmDiagnosisCode_1	object	Code of first diagnosis of patient which was added as first diagnosis of claim by provider
ClmDiagnosisCode_2	object	Code of second diagnosis of patient which was added as second diagnosis of claim by provider
ClmDiagnosisCode_3	object	Code of third diagnosis of patient which was added as third diagnosis of claim by provider
ClmDiagnosisCode_4	object	Code of 4th diagnosis if any

ClmDiagnosisCode_5	object	Code of 5th diagnosis if any
ClmDiagnosisCode_6	object	Code of 6th diagnosis if any
ClmDiagnosisCode_7	object	Code of 7th diagnosis if any
ClmDiagnosisCode_8	object	Code of 8th diagnosis if any
ClmDiagnosisCode_9	object	Code of 9th diagnosis if any
ClmDiagnosisCode_10	object	Code of 10th diagnosis if any
ClmProcedureCode_1	float64	Code of 1st procedure that patient undergo for treatment
ClmProcedureCode_2	float64	Code of 2st procedure that patient undergo for treatment
ClmProcedureCode_3	float64	Code of 3st procedure that patient undergo for treatment
ClmProcedureCode_4	float64	Code of 4st procedure that patient undergo for treatment
ClmProcedureCode_5	float64	Code of 5st procedure that patient undergo for treatment
ClmProcedureCode_6	float64	Code of 6st procedure that patient undergo for treatment
DeductibleAmtPaid	int64	the amount that a policyholder has to pay before the insurance company starts paying up
ClmAdmitDiagnosisCode	object	code of initial diagnosis when patient admitted with initial symptoms

c. Benefciary Details (Train/Test data):

This data contains benefciary KYC details like health conditions,the region they belong to etc.

Columns	Type	Description
BeneID	object	Beneficiary unique id
DOB	object	Date of birth of Beneficiary
DOD	object	Date of death of beneficiary. Value will be null when beneficiary is alive
Gender	int64	Gender of beneficiary
Race	int64	Race(Human category) of beneficiary
RenalDiseaseIndicator	object	Indicator of Kidney disease
State	int64	State of Beneficiary
County	int64	Country of beneficiary
NoOfMonths_PartACov	int64	There are 4 parts of Medicare (A,B,C, and D) Number of months of Part A (Inpatient/ hospital) coverage
NoOfMonths_PartBCov	int64	Number of months of part B (Outpatient/Medical) Coverage
ChronicCond_Alzheimer	int64	Chronic condition of Alzheimer (1: Yes and 2:No)
ChronicCond_Heartfailure	int64	Chronic condition of Heart failure (1: Yes and 2:No)
ChronicCond_KidneyDisease	int64	Chronic condition of Kidney Disease (1: Yes and 2:No)
ChronicCond_Cancer	int64	Chronic condition of Cancer (1: Yes and 2:No)
ChronicCond_ObstrPulmonary	int64	Chronic condition of obstructive pulmonary disease (1: Yes and 2:No)
ChronicCond_Depression	int64	Chronic condition of Depression (1: Yes and 2:No)
ChronicCond_Diabetes	int64	Chronic condition of Diabetes (1: Yes and 2:No)
ChronicCond_IschemicHeart	int64	Chronic condition of Ischemic heart disease (1: Yes and 2:No)
ChronicCond_Osteoporasis	int64	Chronic condition of Osteoporosis (1: Yes and 2:No)

ChronicCond_rheumatoidarthritis	int64	Chronic condition of rheumatoid arthritis (1: Yes and 2:No)
ChronicCond_stroke	int64	Chronic condition of Stroke (1: Yes and 2:No)
IPAnnualReimbursementAmt	int64	Total reimbursement amount as inpatient annually
IPAnnualDeductibleAmt	int64	Total amount paid before claim as inpatient annually
OPAnnualReimbursementAmt	int64	Total reimbursement amount as outpatient annually
OPAnnualDeductibleAmt	int64	Total amount paid before claim as outpatient annually

d. Labeled data (Train data):

Train labeled dataset consists of Provider Id and the provider is potentially fraudulent or not.

Performance metric:

Chance of imbalance data is high in fraud detection that is why the following metric will be useful for evaluation.

1. F1-score
2. Confusion Matrix
3. FPR and FNR
4. log-loss

But, for a balanced dataset, we can also try AUC score.

Research-Papers/Solutions/Architectures/Kernels

Mention the urls of existing research-papers/solutions/kernels on your problem statement and in your own words write a detailed summary for each one of them. If needed you can include images or explain with your own diagrams. it is mandatory to write a brief description about that paper. Without understanding of the resource please don't mention it

1. (PDF) Fraud detection in health insurance using data mining techniques (researchgate.net)

This paper is written by **M.E. Student (Vipula Rawate)** under **Associate professor G Anuradha**.

In order to detect and avoid the fraud, data mining techniques are applied. This includes some domain knowledge of the health care system and its fraudulent behaviors, analysis of the characteristics of healthcare insurance data. Data mining which is divided into two learning techniques which is supervised and unsupervised learning, supervised and unsupervised is employed to detect fraudulent claims. But, since each of the above techniques has its own set of advantages and disadvantages, by combining the advantages of both the techniques, an advanced hybrid approach for detecting fraudulent claims in the healthcare insurance industry is proposed.

Using an unsupervised learning model, a data point with a new category of disease can be clustered in a particular group but supervised may not be

able to detect that new data point. And if the claim is duplicate with a diferent date then it can easily detect using a supervised learning model.

Final cut solution in this paper,use Hybrid framework wherein Evolving clustering method(ECM) followed by classifcation (SVM) to detect the fraudulent claims.

Block Diagram of Approach:

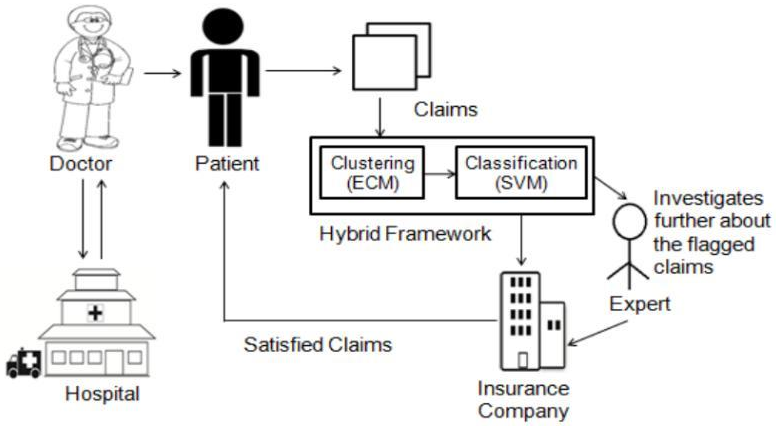


Fig. 8. Hybrid Model

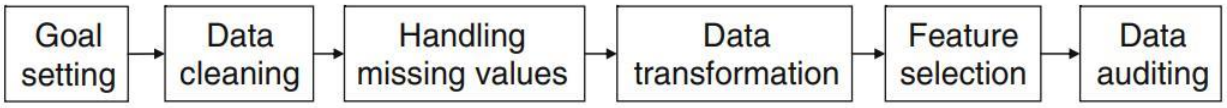
Observation: Using simply supervised or unsupervised learning will not be a good idea for this problem so hybrid models will work better for this problem.

The reason, we have seen above in the paper, so we can take advantage of unsupervised and supervised learning by using hybrid architecture.

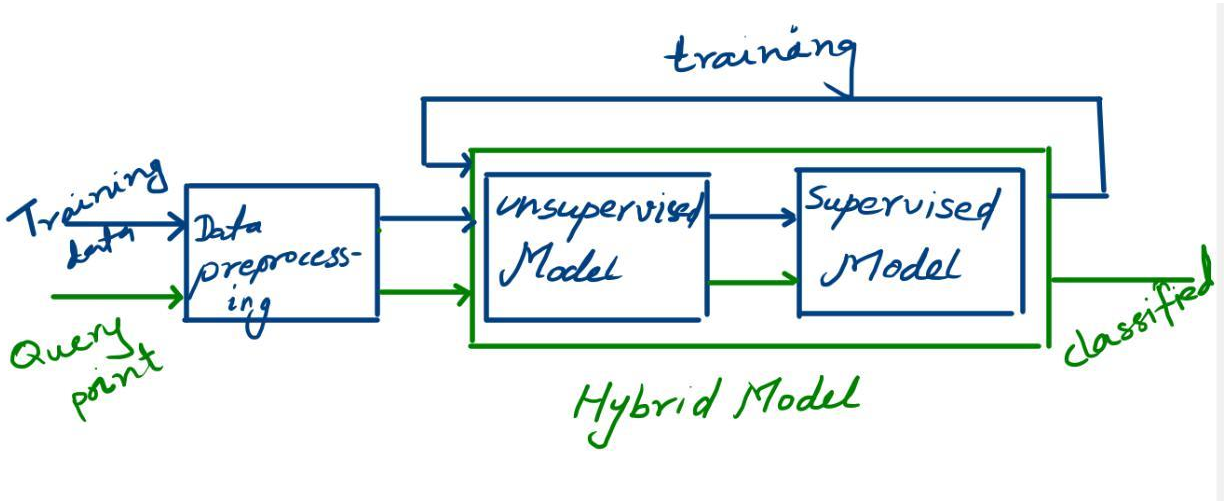
2. [(PDF) A survey on statistical methods for health care fraud detection (researchgate.net)]

The paper **"A survey on statistical methods for health care fraud detection**" is written by **Jing Li, Kuei-Ying Huang , Jionghua Jin & Jianjun Shi.** This paper aims to provide a comprehensive survey of the statistical methods applied to health care fraud detection, with focuses on classifying fraudulent behaviors, identifying the major sources and characteristics of the

data based on which fraud detection has been conducted, discussing the key steps in data preprocessing, as well as summarizing, categorizing, and comparing statistical fraud detection methods. The data preprocessing steps follow:



After data preprocessing, they have tried with multiple supervised models(Decision Tree) and **hybrid model** (unsupervised model followed by Supervised learning model). In this paper, they have explained data preprocessing briefy as well as the statistical modeling part for healthcare fraud. So in this paper, the Hybrid model is also being used to detect query points that are “Fraudulent” or “Legitimate” .



: Architecture (Black Box)

Observation:

Preprocessing of data is a technique which is used to transform raw data into useful and efcient data so that models can learn better and predict

better. In this paper,preprocessing steps were in focus like data cleaning, handling missing value, data transformation and feature selection.

For Imputation, They tried hot-deck Imputation and regression Imputation. And for feature selection they tried with top 30-40 features and which would be selected using feature importance.

3. Improving Fraud and Abuse Detection in General Physician Claims: A Data Mining Study (nih.gov)

To improve fraud and abuse detection in General Physician claims, we should work on the feature engineering part. So in this paper, Authors explain how they created features for each physician, standardized the value using Z scores, used a hierarchical clustering method to segment physicians, and then identifed clusters of physicians that were suspected of abuse and fraud.

In the Feature engineering section, they have introduced 9 features for physician's abuse and 5 features for fraudulent behaviour.

Features of medical resource abuse:

- Percentage of the patients that they were visited more than once in a month
- The average of the prescrip drug items in a claim
- The average cost of a drug prescription claim
- The ratio of the fve expensive antibiotic prescriptions to all physician claims
- The ratio of injection prescription to all physician claim
- The ratio of total injection prescription to all physician claim
- The ratio of total prescript antibiotic to all physician claims
- The ratio of injected antibiotic to physician claim
- The ratio of injected corticosteroid prescription to all physician claim

Features of fraudulent behaviors:

- Percentage of reduplicative patients
- Percentage of reduplicative patients-pharmacy
- Percentage of reduplicative patients-pharmacy in a month
- The average cost of a drug prescription claim
- The ratio of claims referred to a high-cost pharmacy

Observation:

By reading this paper, I fgured out how feature engineering is a game changing idea for this detection problem. Everyone should make a habit to create at least a few new features if possible before training a model.

4. Medicare fraud detection using neural networks | Journal of Big Data | Full Text (springeropen.com)

In this paper, researchers mainly focused on neural networks and solutions for imbalanced data. The chance of imbalanced data is high for fraud detection so to get rid of this problem we can use “Over-sampling” or “Under-sampling” so in this paper, they tried random over sampling (ROS) , random under-sampling (RUS) and also a hybrid technique (ROS-RUS).

And they really got a better auc score with a balanced dataset. Even ROS-RUS maximizes efciency with a 4× speedup in training time.

This study compares six deep learning methods for addressing class imbalance and assesses the importance of identifying optimal decision thresholds when training data is imbalanced. And they are referring to data which is provided by CMS.and the baseline neural network architecture is used for this classifcation problem. Its hyperparameters are discovered through a random search procedure that evaluates models on a validation set. The number of hidden layers, the number of neurons per layer, and

regularization techniques are the primary focus of hyperparameter tuning in this study. They compare the efectiveness of each hyperparameter with the ROC AUC score and loss scores of the models and visualize results across 100 epochs. And they got 0.85 AUC score with the best model.

Observation:

This paper mainly focused on two things, frst the technique for getting rid from imbalanced data. So they use over-sampling, under-sampling and hybrid methods. And second neural network architecture and hypertuning of models.

5. (PDF) Application of Clustering Methods to Health Insurance Fraud Detection (researchgate.net)

   Insurance companies generally get help by human inspections and heuristic rules to discover abuse and fraud. First, it is impossible to discover all abuse and health care fraud by manual inspection over a lot of claims. Second, new types of health care fraud. The aim of this paper is to fnd out suspicious abuses and frauds from massive datasets . In order to achieve this goal, this paper applies two clustering methods, SAS EM and CLUTO , to a large real-life health care database and compares the performances of these two methods. In this study, SAS Enterprise Miner and CLUTO, are used for clustering analysis. Experimental results indicate that CLUTO is faster than SAS EM while SAS EM provides more useful clusters than CLUTO.

   And the data processing and storage of this data was conducted within the insurance company's infrastructure in compliance with privacy regulations.

   Coming to understanding the clustering result, assessing cluster quality is the most challenging step in clustering analysis. Diferent

clustering methods may generate very diferent clusters for the same dataset. Researchers have proposed various quality measures for clustering and most of them can be grouped into two categories: external quality measure and internal quality measures .

Observation:

This paper focused on two clustering techniques, which are SAS EM and CLUTO where experimental results indicate that CLUTO is faster than SAS EM while SAS EM provides more useful clusters than CLUTO.

Data preprocessing is also important so for it, they had conducted preprocessing of data within the insurance company's infrastructure in compliance with privacy regulations.

First Cut Approach

Explain in steps about how you want to approach this problem and the initial experiments that you want to do. ***(MINIMUM 200 words) ***

***When you are doing the basic EDA and building the First Cut Approach you should not refer any blogs or papers ***

After reading the papers for this problem,I will go below the frst cut approach.

1. **Data acquisition:** data is available on kaggle so we can use that data.
2. **Data Preprocessing:** we have the dataset in hand, next we will do data preprocessing and prepare the data for data analysis.

a. For Imbalanced data, we will try SMOTE technique,class_weight and hybrid technique (ROS-RUS)
b. Data cleaning, handling missing values,feature engineering and data transformation will be performed as preprocessing.

3. Perform Exploratory data analysis
    a. Univariate data analysis
        i. **Categorical variable**
            1. Counter plot
            2. Stacked plot
            3. Barplot
        ii. **Numerical variable**
            1. PDF (Histogram)
            2. CDF
    b. Bivariate data analysis
        i. Scatter plot
        ii. Joint plot
    c. Multivariate data analysis
        i. Pair plot
    d. Check multicollinearity in data
    e. Check distribution of data (left skewed, right skewed and Normal distribution)

    Note: we will also answer hypotheses based on EDA.

4. **Data splitting**: will split dataset into Train, Test and validation data by using **stratifed** technique and keeping random_state = some value
5. Transform the data to feed the model
    a. **Categorical variable** : apply one-hot encoding to convert into numeric vector.

   b. **Numerical variable** : apply column standardization for all distance based algorithms.
6. **Modeling:**
   a. Create a random model and get the loss of this model. And will make sure all further models have to perform better than this model.
   b. Will explore all models which were mentioned in the paper for this data set like SVM, XGBoost, Neural network and hybrid model. And hyper tune using Grid or Random search cross validation with log loss metric.
   c. Will plot performance of models to understand model behaviours.
7. **Model Validation:** Will compare which model is working really well on this dataset by performance metric (F1-score, confusion metric and FPR and FNR) and make that model as the fnal model.
8. **Deployment:** will train the fnal model with best parameters and pickle the model. And will try to deploy the model on heroku using Flask framework.

Notes when you build your fnal notebook:

1. You should not train any model either it can be a ML model or DL model or Countvectorizer or even simple StandardScalar
2. You should not read train data fles

3. The function1 takes only one argument “X” (a single data points i.e 1*d feature) and the inside the function you will preprocess data point similar to the process you did while you featurize your train data
    a. Ex: consider you are doing taxi demand prediction case study (problem defnition: given a time and location predict the number of pickups that can happen)
    b. so in your fnal notebook, you need to pass only those two values
    c. def fnal(X):
        preprocess data i.e data cleaning, flling missing values
        etc compute features based on this X use pre trained
        model
        return predicted outputs
    fnal([time, location])
    d. in the instructions, we have mentioned two functions one with original values and one without it
    e. fnal([time, location]) # in this function you need to return the predictions, no need to compute the metric
    f. fnal(set of [time, location] values, corresponding Y values) # when you pass the Y values, we can compute the error metric(Y, y_predict)
4. After you have preprocessed the data point you will featurize it, with the help of trained vectorizers or methods you have followed for your train data
5. Assume this function is like you are productionizing the best model you have built, you need to measure the time for predicting and report the time. Make sure you keep the time as low as possible
6. Check this live session: https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/4148/ hands-on-live-session-deploy-an-ml-model-using-apis-on-aws/5/module-5-feature-engine ering-productionization-and-deployment-of-ml-models