……………………………………………...……………………………………………...……………………………………………...……………………………………………

**Name:** G. Ganesh Reddy

**Email address:** ganeshreddybhavani@gmail.com

**Contact number:** 9182888417

**Anydesk address:** 389 531 466

**Years of Work Experience**: 1 year 2 months

**Date:** 16/3/22

……………………………………………...……………………………………………...……………………………………………...……………………………………………

**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**

\*\*\* Write an overview of the case study that you are working on. ***(MINIMUM 200 words)*** \*\*\*

**Introduction:**

Fraud can be defined as a dishonest act committed by an individual or a group of people with the knowledge and for the financial benefits. Healthcare fraud can be defined as misrepresenting information, concealing information or deceiving a person or an entity in order to get benefitted financially. In Health care the fraud involves the health care system by an individual, Physicians, Doctors, Healthcare providers and the insurance companies.

In our case, we will analyze all the details of the healthcare provider and conclude whether legitimate or not. If the providers fill all the details on behalf of the beneficiaries and makes a claim to get benefitted then it is considered as a Fraud. Health care fraud is one of the biggest problems in healthcare domain across world. In USA, the insurance company should clear all the compensations within 30 days of claim. So, there is less time to investigate carefully and also the claims are increasing rapidly so it’s hard to investigate all the claims manually. So, its wise adopt a computerized technique that automatically investigate through the beneficiary details and suggest whether a claim is Fraud or not.

**ML Formulation and Business problem:**

From references we understand, according to the survey, it is estimated that over 15% of claims are fraud and insurance companies in USA incur losses over 30 billion USD annually. And in India insurance companies incur approximately 600-800 crores annually.

From given dataset our objective is the predict whether the provider is fraud or not. We need to obtain a probability score of the provider fraudulent activity, by analyzing the beneficiary details and reasons why healthcare provider is fraud. So that we can prevent insurance companies from incurring financial losses.

**Business constraints:**

1.Cost of misclassification is very high, where if we predict legitimate provider as fraud (False Positive) it costs for further investigation and also a matter of companies’ reputation. If we predict fraud provider as a legitimate provider (False Negative) then we will end up with huge financial losses.

2. No strict latency requirements.

3. Feature interpretability is highly important- As Insurance company should justify the fraudulent activity of the provider and need to set up manual investigation if needed.

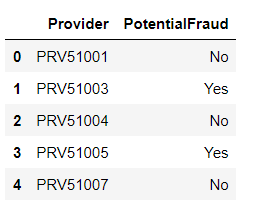
Mapping to ML:

We need to build a binary classification algorithm based the details filled by the provider, inpatient data, outpatient data and beneficiary data to predict whether the health care provider is fraud or not.

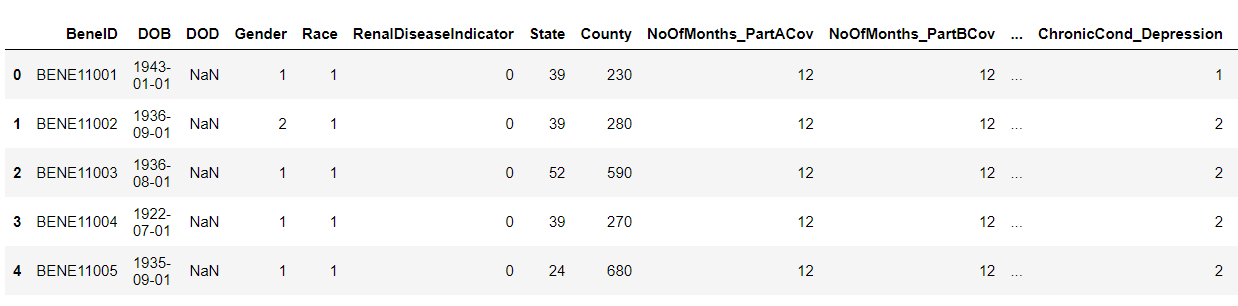
**Data Overview:**

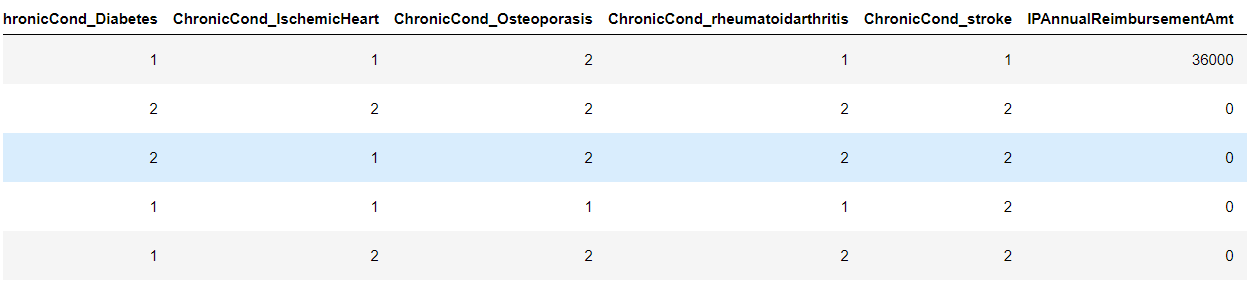
We have train and test data with 4 datasets each. Which are

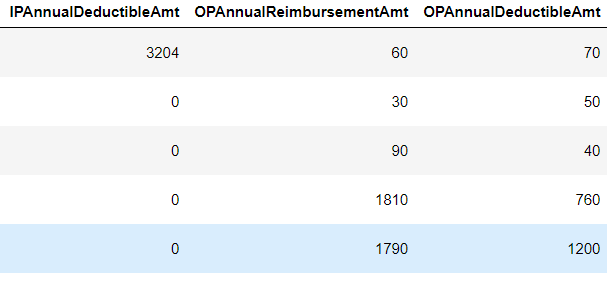
**Provider data** - ProviderID, Potential Fraud.



**Beneficiary data-**

****

****

****

1. BeneID: It contains the unique id of the beneficiary.

2. DOB: It contains the Date of Birth of the beneficiary.

3. DOD: It contains the Date of Death of the beneficiary if the beneficiary id dead else null.

4. Gender, Race, State, Country: It contains the Gender, Race, State, Country of the beneficiary.

5. RenalDiseaseIndicator: It contains if the patient has existing kidney disease.

6. ChronicCond\_\*: The columns started with “ChronicCond\_” indicates if the patient has existing that particular disease. Which also indicates the risk score of that patient.

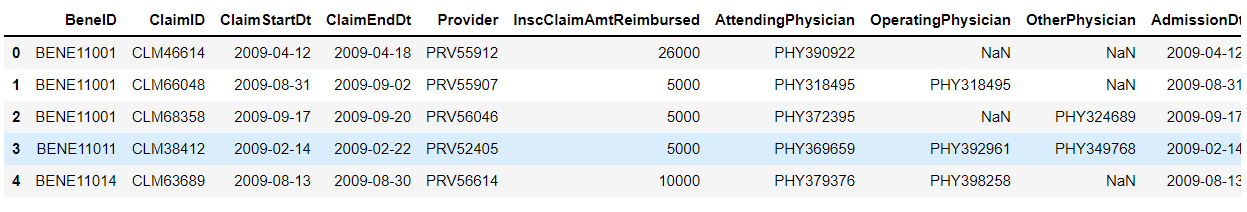
7. IPAnnualReimbursementAmt: It consists of the maximum reimbursement amount for hospitalization annually.

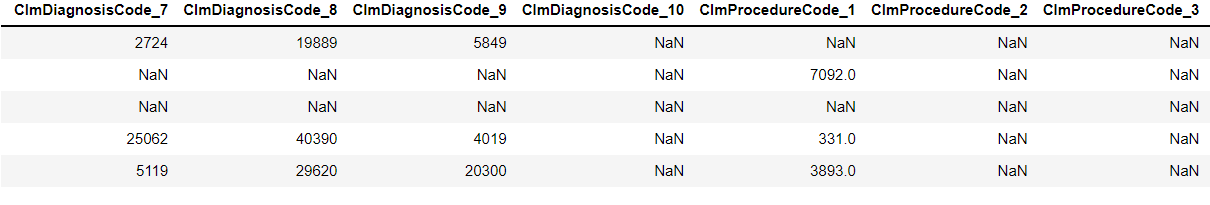
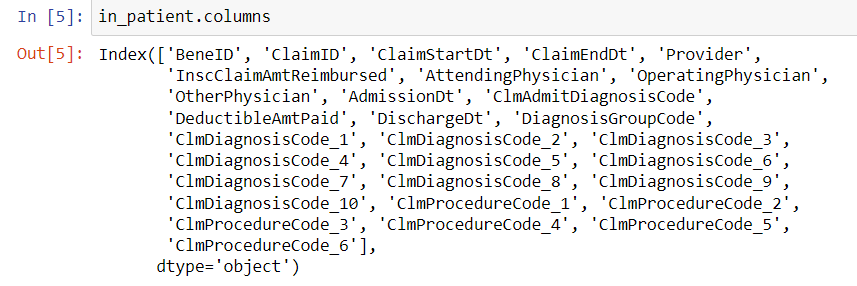
8. IPAnnualDeductibleAmt: It consists of a premium paid by the patient for hospitalization annually.

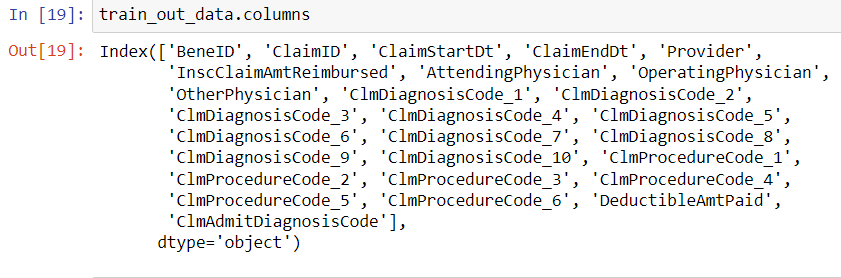
9. OPAnnualReimbursementAmt: It consists of the maximum reimbursement amount for outpatient visits annually.

10. OPAnnualDeductibleAmt: It consists of a premium paid by the patient for outpatient visits annually.

**Inpatient data -**

****

**  Outpatient data** –



1. BeneID: It contains the unique id of each beneficiary i.e., patients.

2. ClaimID: It contains the unique id of the claim submitted by the provider.

3. ClaimStartDt: It contains the date when the claim started in yyyy-mm-dd format.

4. ClaimEndDt: It contains the date when the claim ended in yyyy-mm-dd format.

5. Provider: It contains the unique id of the provider.

6. InscClaimAmtReimbursed: It contains the amount reimbursed for that particular claim.

7. AttendingPhysician: It contains the id of the Physician who attended the patient.

8. OperatingPhysician: It contains the id of the Physician who operated on the patient.

9. OtherPhysician: It contains the id of the Physician other than AttendingPhysician and OperatingPhysician who treated the patient.

10. ClmDiagnosisCode: It contains codes of the diagnosis performed by the provider on the patient for that claim.

11. ClmProcedureCode: It contains the codes of the procedures of the patient for treatment for that particular claim.

12. DeductibleAmtPaid: It consists of the amount by the patient. That is equal to Total\_claim\_amount - Reimbursed\_amount.

13. AdmissionDt: It contains the date on which the patient was admitted into the hospital in yyyy-mm-dd format.

14. DischargeDt: It contains the date on which the patient was discharged from the hospital in yyyy-mm-dd format.

15. DiagnosisGroupCode: It contains a group code for the diagnosis done on the patient.

**PERFORMANCE METRICS:**

As our health care dataset is highly imbalanced, accuracy score is not a good metric to measure performance. And cost of misclassification is high we opt for following metrics.

* 1. Confusion matrix
  2. F1 score – Harmonic mean of precision and recall.
  3. AUC score – Area under the curve, close to 1 better the model performance.
  4. FPR, FNR - cost of misclassification is high we need to check out these two metrics carefully, should be low for better model.

# 

# **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\*\*\*

**URL1.**

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

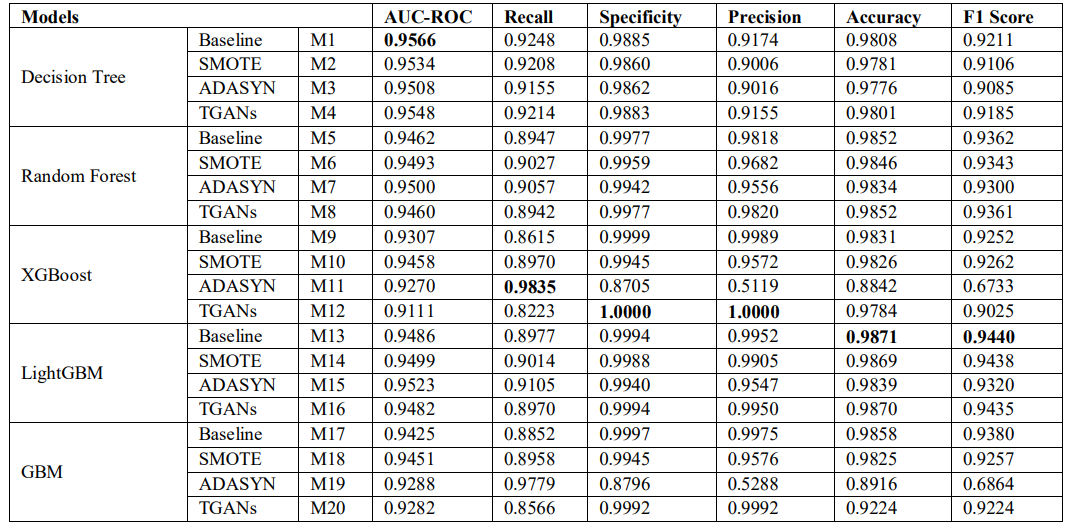
1. **Observations**
2. Classification of fraudulent activities by various categories like service providers like doctors, physicians, ambulance services, insurance companies.
3. Overview of the data by merging all the datasets based on primary and foreign keys. The user can get overview of service provider behaviour with basic analysis.
4. Missing value handling techniques hot-deck imputation and regression imputation.
5. Hot-deck imputation- it tries to fill incomplete values with the most similar complete information cases.
6. Regressor imputation- A regression model is fitted for each variable with missing values with other variables without missing values as input variables.
7. Perform correlation checks to reduce redundant features.
8. Decision trees are more interpretable in this problem but the only problem is overfitting, for that ensembling techniques are used.

1. **Takeaways**
   1. Handling missing values with hot-deck imputation and regressor imputation.
   2. Correlation checks to delete redundant features.
   3. Feature engineering techniques for features likes average claim amount reimbursed per provider, total no of claims per provider.
   4. Feature engineering is most important to come up with new features.

**URL2.**

[**http://ijettjournal.org/archive/ijett-v69i3p216**](http://ijettjournal.org/archive/ijett-v69i3p216)

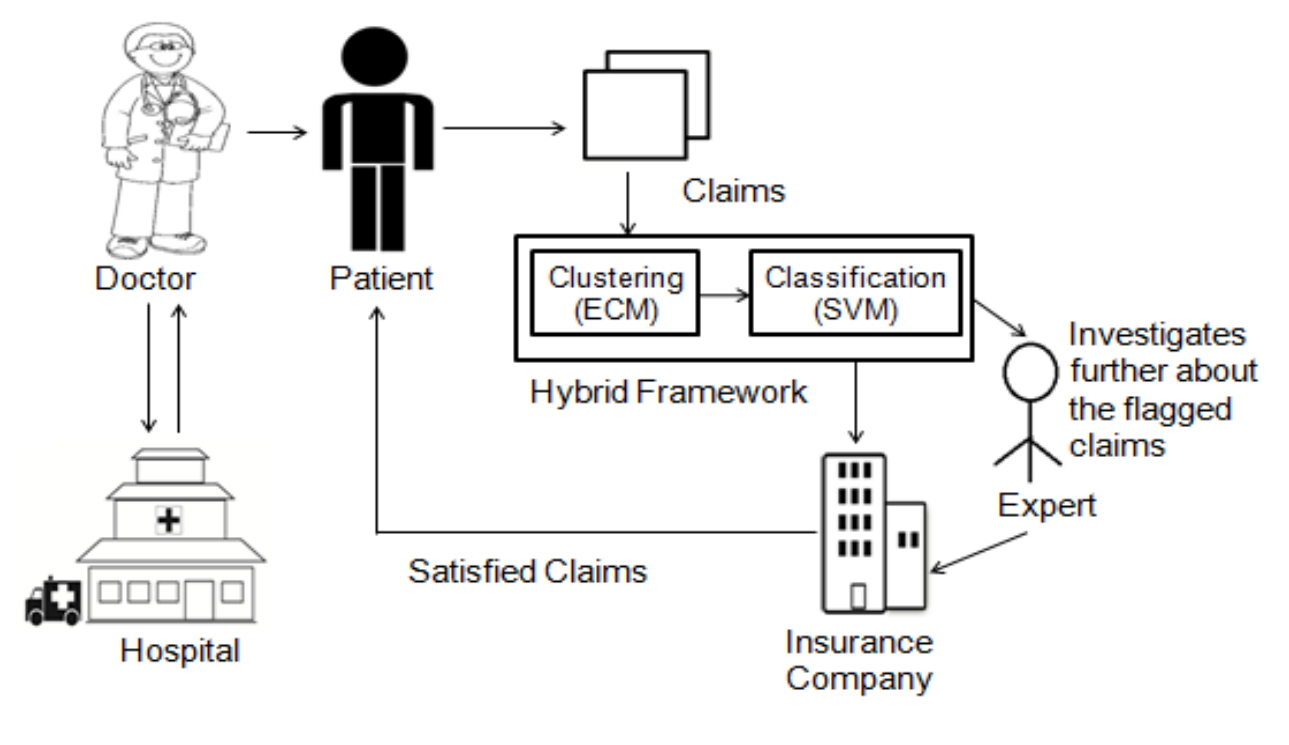
1. **Observations**
2. The dataset if from Ayushman Bharat (PM-JAY), the world’s largest health insurance scheme and is highly imbalance. As we have less fraudulent cases comparative to non-fraudulent cases.
3. For handling data imbalance, the user came up with Synthetic Minority Oversampling Technique SMOTE, Adaptive Synthetic Sampling approach ADASYN, Tabular Adversarial Generative Networks TGANs.
4. The user followed comparative modelling on various models using all the data imbalance handling techniques mentioned above.
5. SMOTE can be understood as new minority samples being synthesized between the two real samples in the dataset set. ADASYN can be performed by considering density distribution of the minority samples. TGANs is a neural network-based generative model which is trained to learn the distribution of the minority samples.
6. The performance analysis of all models is listed below.



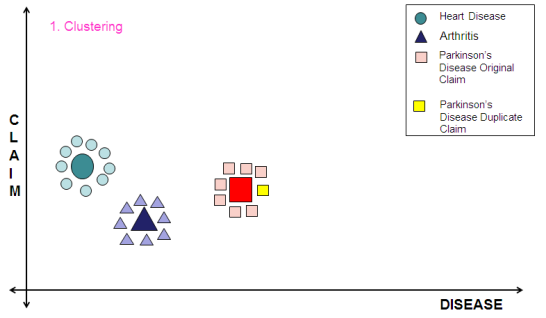
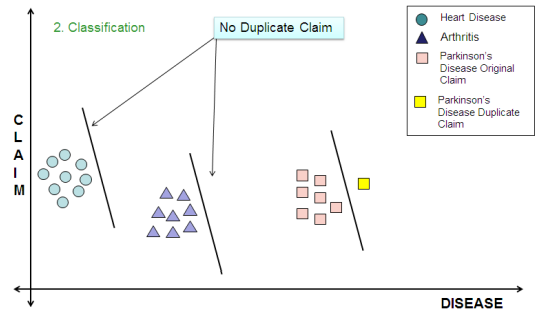
1. Among all the models XGBoost has performed well and achieved 100% specificity and precision while using TGANs as oversampling technique.
2. The user had removed some of the missing features which are mostly empty and cannot contribute much to the model.
3. **Takeaways**
4. Handling data imbalance techniques such as SMOTE, ADYSN, TGANs.
5. XGBoost algorithm performs well comparative to other models.
6. Key performance metrics used.

**URL3:** [**https://www.researchgate.net/publication/282538462\_Fraud\_detection\_in\_health\_insurance\_using\_data\_mining\_techniques**](https://www.researchgate.net/publication/282538462_Fraud_detection_in_health_insurance_using_data_mining_techniques)

1. **Observations:**
2. The user tried both supervised and unsupervised techniques on the data and concluded both have some advantages and disadvantages. Where supervised technique fails to classify new data which is not similar train data and can correctly classify duplicated data. Where clustering techniques fails to group duplicated claims but it can group datapoints which are not similar to train data as a different cluster.
3. The novel approach proposed in this paper is Evolving Clustering Method (ECM) for clustering because the data is dynamic and new data is generated continuously and Support Vector Machine (SVM) for classification.



1. In this approach, first the claims are clustered based on the diseases/treatment/procedural codes followed and then classified using support vector machines.

1. **Takeaways**
2. A hybrid model, combination of clustering technique and support vector machines.
3. In oder to cluster, need to come up with the features like common diseased claims/ common procedural codes followed claims/ common physicians attended for patient for treatment.
4. Need to check the duplicated claims, where date or some other column is modified and reamaining details are kept same and raised for claim.

**URL4**

[**https://www.annalsofrscb.ro/index.php/journal/article/download/2409/2028/4484**](https://www.annalsofrscb.ro/index.php/journal/article/download/2409/2028/4484)

1. **Observations**
2. In order to label the raw data, the user have used clustering techniques such as k means clustering. Based on hyperparameter tuning the best number of clusters are formed.
3. The user measured the certainty of a fraudulent activity as propensity score.
4. The user came up with Gini coefficient to measure the model performance, it can be interpreted as

**GINI COEFFICIENT:**

GC = 2 \*AUC – 1

Where GC-Gini coefficient

AUC- Area under curve

1. The perfect model will have Gini coefficient as 1.
2. **Youden’sJ Statistic:**

J = Sensitivity\* + Specificity\* – 1

J = TPR + (1 – FPR) – 1 = TPR – FPR

The user chooses the Youden’sJ statistic as threshold value to predict the target.

1. Here the cost function is the cost lost due to fraud and cost spend for investigation.
2. Various models are tried on the data such as linear regression, Random Forest, XGBoost.

Random Forest model is rejected due to overfitting and XGBoost has performed well for imbalanced data.

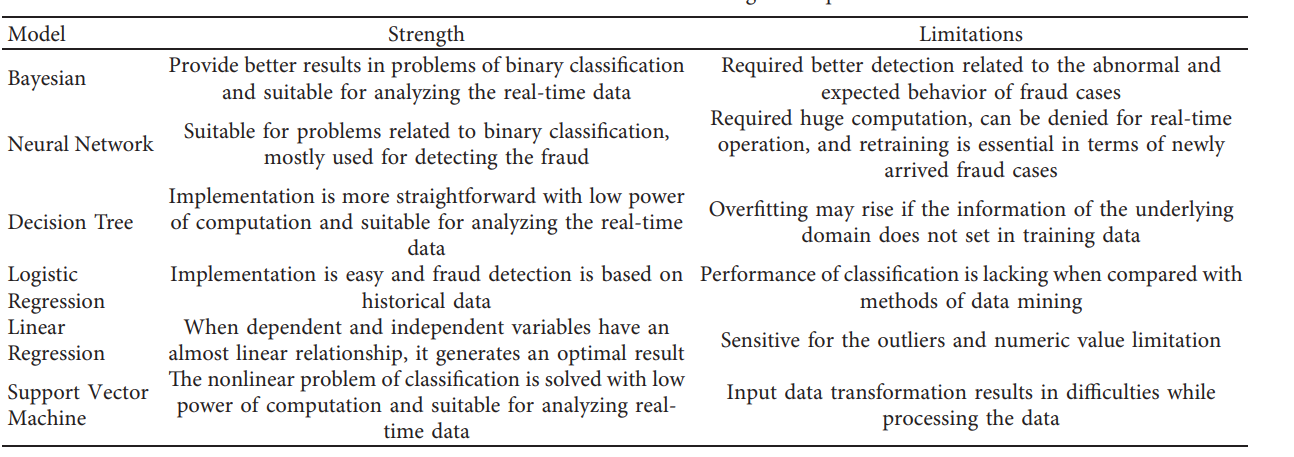
1. In production. The model is refreshed at frequent intervals with the new patterns of the fraudulent activities.

1. **Takeaways**
2. Performance metric should be chosen by considering the cost function of the problem
3. Random forest overfits the model and XGBoost performs well for imbalanced data. Still the performance of models varies based on the data patterns.
4. Retrain the model in production to deal with new patterns.

**URL5**

[**https://www.hindawi.com/journals/scn/2021/9293877/**](https://www.hindawi.com/journals/scn/2021/9293877/)

1. **Observations**
2. The user followed comparative analysis of each model and listed the pros and cons of each model.



1. The data source is UCI machine learning repository, which contains information of payment transactions data through both online and offline (cash and credit/debit cards).
2. **Takeaways**
3. As seen earlier, Random Forest model performs well comparative to simple classification models.
4. This methodology states sequential CNN model has not shown good performance than Random Forest. But generally, we expect sequential models to perform well.

**URL6:**

[**https://www.theseattledataguy.com/healthcare-fraud-detection-with-python/#page-content**](https://www.theseattledataguy.com/healthcare-fraud-detection-with-python/#page-content)

1. **Observations**
2. The user focussed more on exploring data, and find out some key features which can help to find the target better.
3. Some key features are age, the distribution of ages is similar in both fraud and non-fraud cases.
4. Second is duration of the patient admitted
5. Total money spend for fraud claims and non-fraud claims and per claim cost. This feature is very important as the cost function is the totally how we are going to benefit by the machine learning model.
6. **Takeaways**
7. Feature engineering and Data analysis is most important to come up new features.
8. It helps to start the problem with a clear observation.
9. Some of the key features as age, money spend on claims, money spend on per claims and more.

# 

# **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 \*\*\*

Based on the research papers and blogs I followed, I came up with the approach

1. Fraudulent data will be highly imbalanced. So, I will do oversampling with one of the techniques (SMOTE, ADASYN, TGAN) discussed in the above research paper.
2. Data analysis on each column of data and see the distribution using plotting techniques.
3. Inpatient and outpatient datasets have most columns in common, I will merge the using joins.
4. To handle missing values, I follow either hot-deck imputation or regressor imputation as discussed in one of the papers.
5. Feature engineering is most important, so I will analyse and calculate features like age, duration of admitted, some features checking physicians attended, procedural codes followed, diagnosed codes followed.
6. Loss function is the amount we lost due to fraud claims, to minimize that I will try to calculate the per claim amount, average amount claimed, total claims per month and more different features.
7. Based on the observations on various papers and blog we can conclude ensemble models like Random Forest, XGBoost models performs well.
8. I will do comparative analysis on models with all the features available and the only important features that contribute most to the model.
9. Hyper parameter tuning is performed to get the better models with good parameters to optimize the loss function.
10. Performance metrics are F1 score and auc where the data is imbalanced so, accuracy will not be a good metric to evaluate.

**Notes when you build your final 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 files
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
   1. Ex: consider you are doing taxi demand prediction case study (problem definition: given a time and location predict the number of pickups that can happen)
   2. so in your final notebook, you need to pass only those two values
   3. def final(X):

preprocess data i.e data cleaning, filling missing values etc

compute features based on this X

use pre trained model

return predicted outputs

final([time, location])

* 1. in the instructions, we have mentioned two functions one with original values and one without it
  2. final([time, location]) # in this function you need to return the predictions, no need to compute the metric
  3. final(set of [time, location] values, corresponding Y values) # when you pass the Y values, we can compute the error metric(Y, y\_predict)

1. 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
2. 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
3. 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-engineering-productionization-and-deployment-of-ml-models>