**Guidelines for PGPDSE FT Capstone Project – Interim Report**

**Industry Review**

Health care has become one of the strongest contributors to the economy lately. The expenditure of healthcare has stood up to 1.2% of the GDP and is to see a rise in the following years. It has a potential market size of up to 8.6 trillion by 2022. The industry is an integration of sectors within the economic system that provides goods and services to treat patients. It includes the generation and commercialization of goods and services lending themselves to maintaining and re-establishing health of which providing insurance would come under the economic wing of the sector which is a widespread phenomenon on rising cases of health issues in respect with the happenings around.

 As there is a growth of leaps and bounds in the economy of this sector it also leads to fraud in case of insurance claims which cause a spike in the expenditure exponentially due to false insurance claims which seems to be a major problem.

 The insurance agencies who are involved in providing policies incur losses due to it.

**Literature Survey - Publications, Application**

**Publications:**

In recent years, people have made a series of new research in this field. R. Ikono et al. (2019) reviewed 88 articles from journal articles, conference minutes, and books based on the research question’s relevance. The results of this review indicate that traditional fraud detection methods were difficult to be implemented in the healthcare system, as new fraud patterns continue to evolve to circumvent fraud detection method.

Globally ,the healthcare industry rates AI and IoT as the most disruptive technologies.In india,most private healthcare providers have adopted foundational technologies such as HIS,ERP,appointment booking,RFID asset,tracking etc.

Healthcare has become a major expenditure since 1980.Both the size of the healthcare sector and the enormous volume of money involved make it attractive fraud target.According to the office of Management and Budget ,in 2010,about 9% ,or around 47.9 billion of Medicare expenditure was lost due to fraud .therefore effective fraud detection is important for reducing the cost of healthcare system.

Detecting healthcare fraud and abuse,however, needs intensive medical knowledge .Many health insurance  systems rely on human exports to manually review insurance claims and identify suspicious ones.This results in both system development and claim reviewing being time-consuming ,especially for the large national insurance programs .

In recent years,systems for processing electronic claims have been increasingly implemented to automatically perform audits and reviews of claims data.These systems are designed for identifying areas requiring special attention such as erroneous or incomplete data input ,duplicate claims and medically non-covered services.Although these systems may be used detection capabilities are usually limited since the detection mainly relies on pre-defined simple rules specified by domain experts.

**Applications:**

Human beings always search for methods, tools, or techniques that reduce the human effort for performing a certain task efficiently. In Machine Learning, algorithms are designed in such a way that they try to learn by themselves using past experience. After learning from the past experience, the algorithms become quite capable of reacting and responding to conditions for which they are not explicitly programmed. So, Machine Learning helps a lot when it comes to fraud detection. It tries to identify hidden patterns that help in detecting fraud which has not been previously recognized. Also, its computation is fast as compared to the traditional rule-based approaches.

There is no exact rule which can clearly distinguish the abnormality of medical insurance transactions. Moreover, the number of abnormal records is tiny compared to the massive number of regular treatment records. For those two reasons above, the relatively small dataset of labeled abnormal records limit the algorithm accuracy.

Due to the influence of various concomitant diseases, patient characteristics, doctor preferences, and additional noise factors in medical treatment records, the situation is complicated, making the anomaly challenging to find out .

Because intentional deception fraudsters often use multiple methods to conceal their fraudulent behaviors behind enormous usual transaction data, traditional means based on rules are challenging to find fraudsters and hard to cover the updated fraud behaviors.

For those reasons listed above, the real-world healthcare scenario is so complex that many reasonable behaviors seem abnormal, and hence the abnormal detection system in the healthcare domain is known as hard to develop and apply. In order to get rid of this dilemma, we tried to use a machine learning method to detect medical fraud cases.

**Dataset and Domain**

**Data Dictionary**

About the data set (Insurance claim data)

The dataset contains information about the insurance claims, beneficiaries, and providers that may contribute to the potential fraudulent claims.

Attribute information:

1. **BeneID-** Beneficiary ID.
2. **ClaimID -** Claim ID .
3. **ClaimStartDt -** The date when the claim was submitted.
4. **ClaimEndDt-** The date when reimbursement was done.
5. **Provider -** Person/Company that provides a healthcare service.
6. **InscClaimAmtReimbursed-** Expense that was paid, and are covered for under policy (Reimbursed).
7. **AttendingPhysician-** A physician who has treated or currently treating, a person seeking insurance.
8. **OperatingPhysician-** A physician who performs a procedure/surgery on the person seeking insurance
9. **OtherPhysician**  - Other physicians involved

10. **AdmissionDt**      - The date on which the beneficiary was admitted in healthcare facility

11**. ClmAdmitDiagnosisCode**         -

12  **DeductibleAmtPaid**   - The amount paid for the diagnosis/procedure by the beneficiary

13 **DischargeDt** - The date on which the beneficiary was discharged from healthcare facility

14  **DiagnosisGroupCode**     - Code indicating the healthcare diagnosis package

15 **ClmDiagnosisCode\_1 to 10**     - Code indicating different diagnostic tests

24  **ClmProcedureCode\_1 to 6**       - Code indicating different medical procedures

30**PotentialFraud**         - Indicates if this is a potential fraudulent claim

31  **DOB**   - Date of Birth of Beneficiary

32  **DOD** - Date of Death of deceased Beneficiary

33  **Gender**   - Gender of the beneficiary

34  **Race**  - Racial background of the beneficiary

35  **RenalDiseaseIndicator**  - Indicates if the beneficiary has been affected with Renal disease

36**State**  - State in which the beneficiary is residing

37  **County**  - County to which the beneficiary belongs

38  **NoOfMonths\_PartACov**                   -

39**NoOfMonths\_PartBCov**                     -

 40  **ChronicCond\_Alzheimer** - Indicates if the beneficiary has been affected with Chronic Alzheimer’s

41  **ChronicCond\_Heartfailure**  - Indicates if the beneficiary has been affected with Chronic Heart condition

42  **ChronicCond\_KidneyDisease**  - Indicates if the beneficiary has been affected with Chronic Kidney Disease

43  **ChronicCond\_Cancer** - Indicates if the beneficiary has been affected with Chronic Cancer

44  **ChronicCond\_ObstrPulmonary**  - Indicates if the beneficiary has been affected with Chronic Obstruction of Pulmonary

 45  **ChronicCond\_Depression** - Indicates if the beneficiary has been affected with Chronic Depression

 46  **ChronicCond\_Diabetes** - Indicates if the beneficiary has been affected with Chronic Diabetes

47  **ChronicCond\_IschemicHeart**- Indicates if the beneficiary has been affected with Chronic Ischemic Heart

 48  **ChronicCond\_Osteoporasis** - Indicates if the beneficiary has been affected with Chronic Osteoporosis

49  **ChronicCond\_rheumatoidarthritis** - Indicates if the beneficiary has been affected with Chronic rheumatoid arthritis

 50  **ChronicCond\_stroke** - Indicates if the beneficiary has been affected with Chronic stroke

 51  **IPAnnualReimbursementAmt**  - Amount reimbursed for the inpatient beneficiary annually

 52  **IPAnnualDeductibleAmt**   - Amount deducted from the inpatient beneficiary annually

 53  **OPAnnualReimbursementAmt** - Amount reimbursed for the outpatient beneficiary annually

 54  **OPAnnualDeductibleAmt**  - Amount deducted from the outpatient beneficiary annually

**Variable categorization (count of numeric and categorical)**

The number of numerical variables are 7 and the number of date and time variables are 6 and the number of categorical variables are 32.

There are a total of 55 features in the dataset.

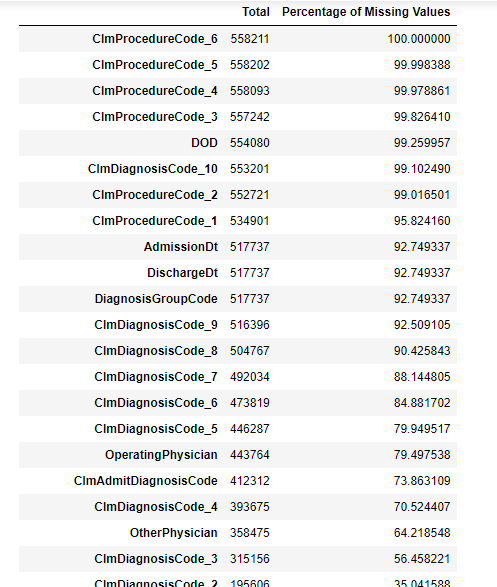
**Pre Processing Data Analysis (count of missing/ null values, redundant columns, etc.)**

There are no redundant columns as such in the dataset.

The categorical columns tend to have a higher percentage of null values . But on further observation few features turned out to be null due to the presence of values of its kind on other features. For example-if a claim procedure is done the other claim procedure has to be marked null which was treated and converted into zeros or NA for easier understanding and processing of the data .

Top five columns with maximum percentage of **null values** are

* Claim procedure codes- which tend to have a null percentage of at least 90 on which particularly claim procedure 6 has a null value percentage of 100 which was dropped and other 5 values were treated.
* DOD-It is a feature which describes or records the date of expiry of the patients who has made a claim even though there is not a very visible effect of it with the dependent variable of potential fraud it is not dropped for now.
* Claim Diagnosis code-This basically describes the procedure that was undergone and the claim of charges made in regards with it. It has ten sub categories. So a patient falling into multiple categories is a very rare occurrence and contains at an average of over 90 percent of null values which were treated.
* Dates(admission and discharge)- The records of admission and discharge dates are missing at an average of 92 percent which are significantly insignificant when compared to claim start date and claim end date.
* Operating physicians- It is over 79 percent as such physicians are not required for every patient being admitted to the hospital and it was treated accordingly.



**Business Problem Understanding**

* Health insurance fraud can be explained as a situation where an insured or medical service provider furnishes fraud, false or misleading information to the insurer with the intention to obtain unfair benefits from a policy for the policy holder or service providing source.
* Such fraud leads to serious losses for the insurance service providers but it could also result in impacting the health insurance advantage for genuine customers. Thus the objective is to develop a state where it becomes difficult to commit fraud and get away with it. This is in the best interest of insuring the public at large, the Insurance Companies as also the society.
* 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.

**Business Objective:**

Thus the objective is to  analyze and explore a dataset for medical costs in health insurance. And to derive insights on important variables helpful in detecting the behavior of potential fraud providers.

**Project Outcomes:**

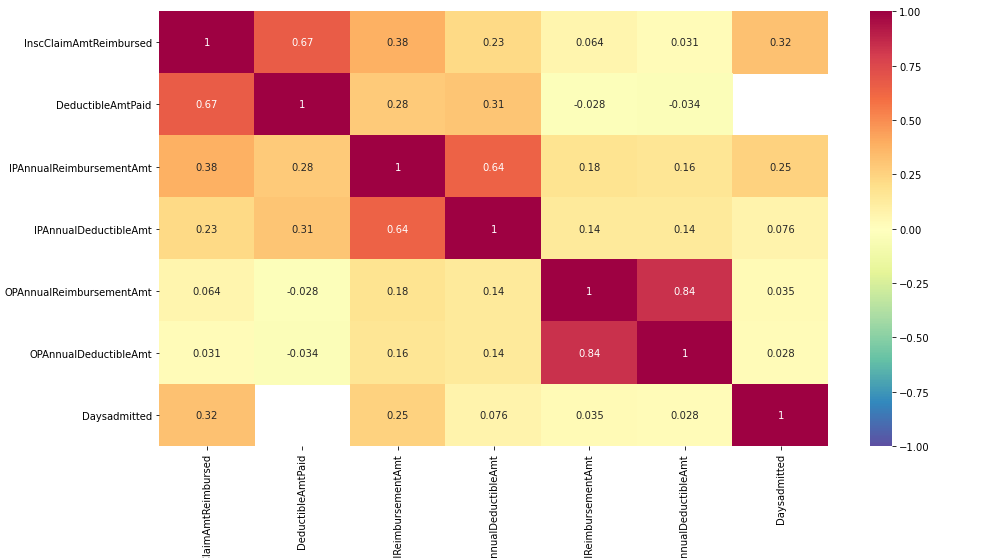
It is to basically determine the fraudulent happening across various factors of an insurance claim and trace a pattern on it in respect to the Beneficiary ID, physician handling the patient and other factors involved in it through which a ML model is formed with help of clustering algorithm.

**Data Exploration (EDA)**

We have discussed the presence of null values already and we have treated it accordingly as well . We have already discussed on the number of categorical and numerical features of the dataset(48 and 7).

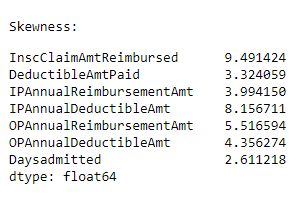
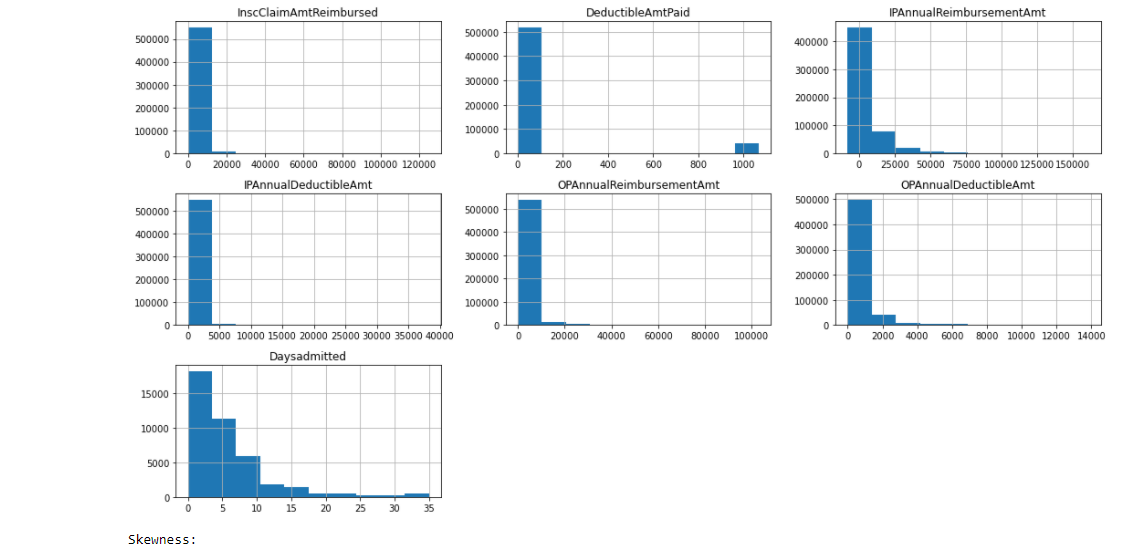
**Correlation of Numerical Features:**

There is high correlation between OP Annual Deductible Amount and OP Annual Reimbursement Amount and the same can be said for IP Annual Deductible Amount and IP Annual Reimbursement Amount.This explains the fact that the payment and repayment missionary seems to be actually in line with one another. So the other factors of physicians and patients teaming up and claiming ambiguous claim procedure codes which would be discussed later.

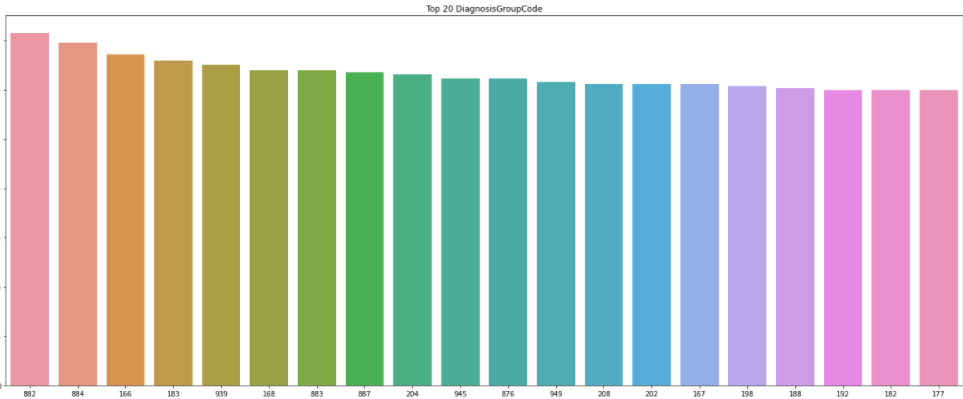


**Distribution of variables**

The distribution of numerical variables can be drawn with the help of skewness. The variables seem to be extremely positively skewed. The reason for such skewness is the nature of the variables. The uneven or skewed distribution does not denote a bad dataset. But, the idea of skewness is shared for referring to the distribution of features.



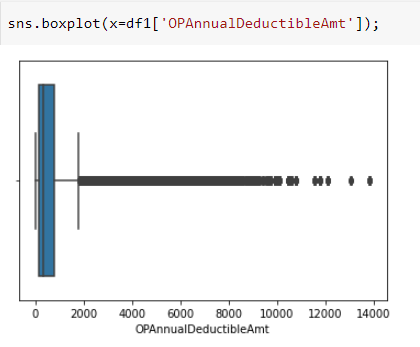
The categorical features basically consist of claim procedures and respective features which have distribution below.

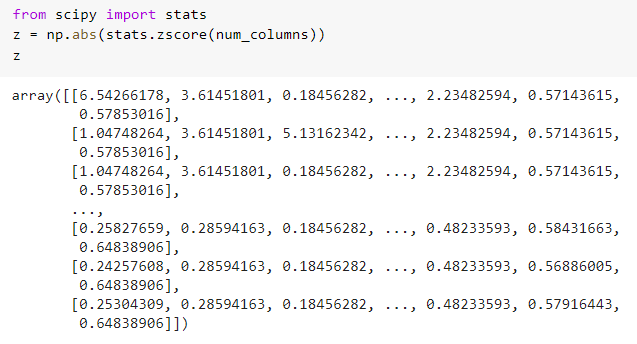


**Presence of outliers and its treatment:**

**Feature Engineering:**

As per discussions with our mentor and on analysis of data we could conclude that transformation and scaling of data would not add value to the model thus skipping it and proceeding to feature engineering.





**Addition of Features to the dataset:**

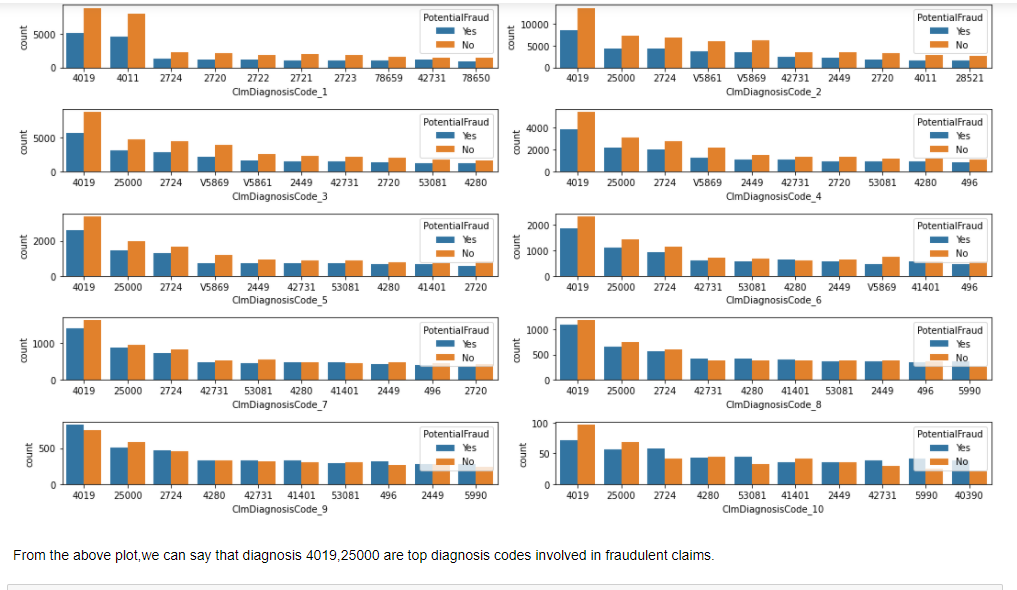
There are three more features added to the dataset which are as follows.

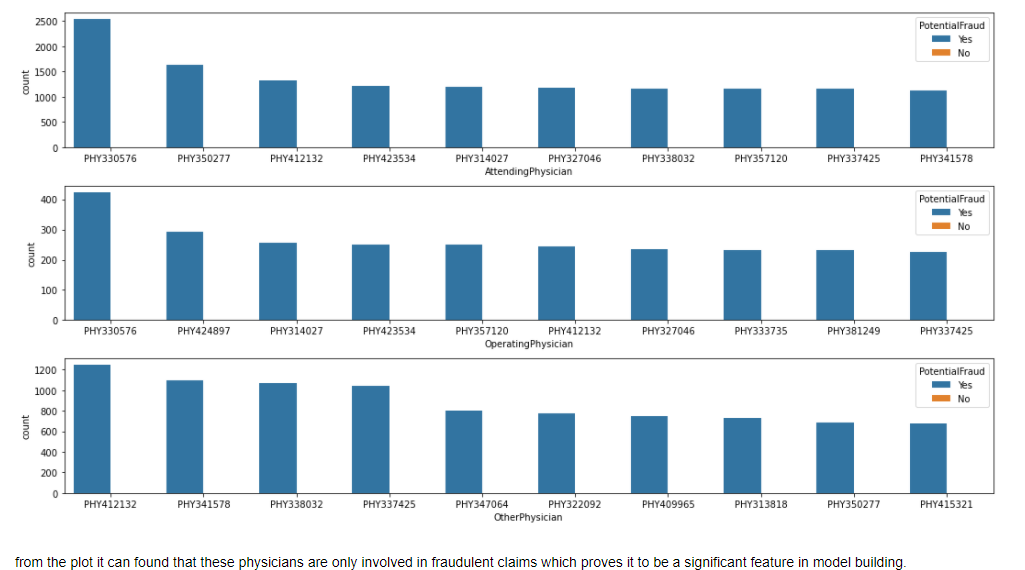
Weather dead- Is the patient alive or dead.

Age- age of the patients

**Feature Selection:**

The key features to be selected  would include beneficiary ID, claim procedure ID ,physicians involved and claim diagnosis are to be considered and the below plots prove them.





**Outliers:**

**Assumptions:**

  Check for the assumptions to be satisfied for each of the models in

* Regression – SLR, Multiple Linear Regression, Logistic Regression
* Classification – Decision Tree, Random Forest, SVM, Bagged and boosted models
* Clustering – PCA (multi-collinearity), K-Means (presence of outliers, scaling, conversion to numerical etc.)