SYNOPSIS

FRAUD DETECTION IN HEALTHCARE INSURANCE

OVERVIEW:

Health Care Industry:

Healthcare industry primarily focuses on the maintenance and progression of health which involves the diagnosis, treatment and prevention of diseases in humans.

There are several types of treatment procedures that are carried out for patients in different specialties. Every treatment has a price to bear, which could raise uncertainties to many patients on affording it. In order to remediate that situation, Health insurance schemes come on page to actively cover the expenses to a greater extent.

As the industry is expanding at an expeditious pace, the fraudulent are happening at wide verse and stays a critical problem to articulate.

Types of Fraud in HealthCare [1]:

1. **Fraud by the Service Provider:**
    - Billing for the medical services that are not actually performed
    - Billing for each stage of a medical procedure as a separate treatment
    - Billing for expensive medical services than the one actually performed
    - Performing unnecessary medical services
2. **Fraud by Insurance:**
    - Filing claim for medical services which was not actually received
    - Using another person's insurance coverage
    - Non-disclosure of Pre-Existing Diseases and manufacturing diagnosis reports to justify tests, examinations and surgeries to prove claim worthiness
    - Misrepresenting treatments that are not covered as medically necessary
    - Duplicate submission of a claim for the same service

NOTE : Any fraudulent activity can have more than one party may be involved (a patient, a physician and insurance company)

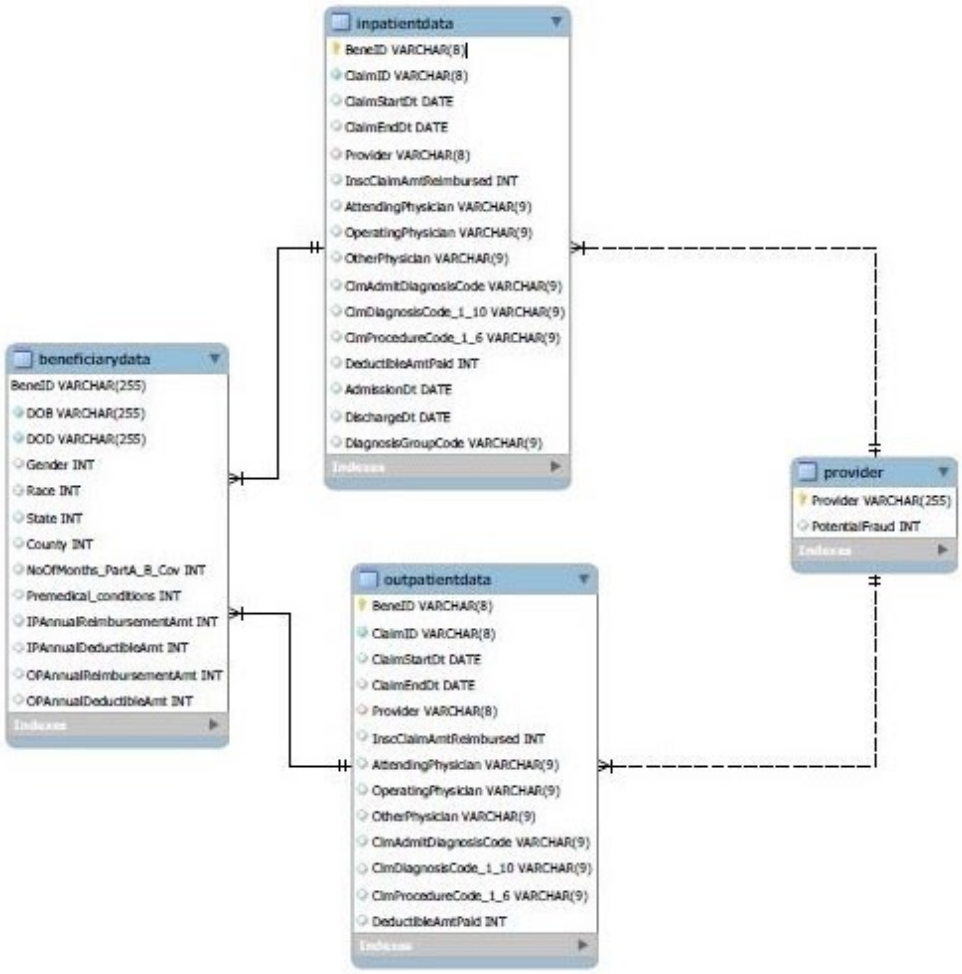
The US has two major national health insurance programs namely Medicare and Medicaid. In 2016, the US spent $3.4 trillion on health care expenditures. The National Health Care Anti-Fraud Association estimates conservatively that health care fraud equals 3 percent of the nation's total health care expenditures. According to the percentages above, fraud, waste and abuse would range from $102 billion to $340 billion [2].

PROJECT OBJECTIVE: The goal of the project is to predict the potentially fraudulent providers based on the claims filed by them, which would aid in saving an enormous amount of money from fraudsters and help those in real need.

DATASET SOURCE: The source of the dataset is from Kaggle which is used to analyse the Insurance Claim Fraud for Medicare [3].

DATASET DESCRIPTION:

There are four files in the dataset with several features including the target variable that indicates whether the particular provider's claim is fraud or not.



There are four files available in the dataset that are to be merged.

1.InpatientData 2.OutpatientData 3.BeneficiaryData 4.ProviderData.

The InpatientData and OutpatientData will first be merged using the common columns followed by the merge of the BeneficiaryData and then the merge of the ProviderData.

After merging, the raw data contains a total of 5,58,211 rows and 55 columns.

DATASET DICTIONARY:

1. Inpatient Data : It provides information about the claims filed for the patients who are admitted in the hospitals. This file has an exact of 40,474 rows and 30 columns.

2. Outpatient Data : It provides information about the claims filed for the patients who visit hospitals but not admitted in the hospital.This file has an exact of about 5,17,737 rows and 27 columns.

S.No	Column Name	Column Description
1	BeneID	Unique ID of the beneficiary
2	ClaimID	Unique ID for each claim submitted
3	Provider	Unique ID for the Insurance Provider
4	ClaimStartDt	Start Date for the Claim
5	ClaimEndDt	End Date for the Claim
6	AdmissionDt	Admission Date of the Patient in Hospital
7	DischargeDt	Discharge Date of the Patient from Hospital
8	AttendingPhysician	Unique ID of the Main Physician who attended the patient
9	OperatingPhysician	Unique ID of the Operating Physician who attended the patient
10	OtherPhysician	Unique ID of the Other Physician who attended the patient
11	InscClaimAmtReimbursed	Money settled to beneficiary based on the Insurance Premium
12	DeductibleAmtPaid	Amount to be paid by the Beneficiary from their own resources for Inpatient
13	DiagnosisGroupCode	Unique Diagnosis Code as per the ICD Code [4]
14	ClmAdmitDiagnosisCode	Unique Diagnosis Admit Code for the Claim given at the time of the Beneficiary's admission to the hospital
15	ClmDiagnosisCode_1–10	Diagnosis Codes for that are performed on the patients by providers which is based on the ICD Codes
16	ClmProcedureCode_1-6	Procedure Codes that patients undergo based on the ICD Codes

3. Beneficiary Details Data : It provides the KYC details of the patient found in the Inpatient Data and Outpatient Data. This file has an exact of 1,38,556 rows and 25 columns.

S.No	Column Name	Column Description
1	BeneID	Beneficiary ID that is registered with the Insurance Provider
2	DOB	Date of Birth of the Beneficiary
3	DOD	Date of Death of the Beneficiary
4	Gender	Gender of the Beneficiary
5	Race	Race of the Beneficiary
6	RenalDiseaseIndicator	Whether the beneficiary has renal disease or not
7	State	State Code for the registered members
8	County	County Code for the registered members
9	NoOfMonths_{PartA/B}Cov	No of months of coverage (Part A and Part B) covered by Medicare
10	Premedical conditions	There are some columns such as RenalDiseaseIndicator,ChronicCond_Depression,ChronicCond_Diabetes etc to indicate if the member has any prior medical condition.
11	IPAnnualReimbursementAmt	In patient Annual reimbursement amount
12	IPAnnualDeductibleAmt	Amount to be paid by the Beneficiary from their own resources for Inpatient
13	OPAnnualReimbursementAmt	Outpatient Annual reimbursement amount
14	OPAnnualDeductibleAmt	Amount to be paid by the Beneficiary from their own resources for Outpatient

4. Provider Data : It provides the details of the Provider ID and the Potential Fraud Claim. This file has about 5,410 rows and 2 columns.

S.No	Column Name	Column Description
1	Provider	Unique ID of the healthcare providers
2	PotentialFraud	Indicates if the claim is fraud or not

Inference from the Dataset:

To predict the Potential Fraud, the target variable is identified as PotentialFraud (found in the Provider data). It is a binary classification problem where the output will be 1 if it is a counterfeited claim.

The probabilistic outputs are essential to calculate as they may help determine the chances of capturing the fraudulents.

APPROACH:

The first step is to examine the data at the elementary stage using the various concepts of Exploratory Data Analysis (EDA).
The EDA procedure would consists of the following two parts:

1. **Descriptive Analysis:** It consists of Univariate, Bivariate and Multivariate Analysis. It also involves finding the relation of the variables with the target variable.
2. **Inferential Analysis:** Based on the Descriptive analysis, the deductions will be verified with statistical tests.

MODEL BUILDING:

Since the project objective has been identified as a classification problem, the analysis will start with the Logistic Regression Model (Basic Classification Model).

After building the basic model, more Complex Classification Models such as Decision Tree, Random Forest and Boosting Techniques will be built to identify the best model (by tuning the hyperparameters) to find the potential fraud provider.

MODEL EVALUATION METRICS:

Every model's (basic or complex) performance can be identified using Metrics like:

- Precision : It is a good measure to determine when the cost of False Positive is high.
- Recall : It is the best metric to select the best model when there is a high cost associated with False Negative.
- F1 Score : It seeks a balance between the Precision and Recall
- Accuracy : It is the ratio of the number of correct predictions to the total number of input samples.

CONCLUSION: We will draw a model which would predict and analyze the potential fraud provider with maximum effectiveness.

REFERENCES:

[1]Shivani S. Waghade, Prof. Aarti M. Karandikar (2018) 'International Journal of Applied Engineering Research' in ISSN 0973-4562 Volume 13, Number 6 pp. 4175-4178

[2]https://www.sas.com/en_hk/insights/articles/risk-fraud/medicaid-benefit-fraud.html

[3]https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis

[4]https://en.wikipedia.org/wiki/International_Classification_of_Diseases