Project Summary

Batch details	DSE Chennai August 2020
Team members	Krithikaa Madhumitha Lakshmi Sruthi B Monica Chandrasekar Samyuktha Mahesh M Surya Narayanan
Domain of Project	Finance & Risk Analytics
Proposed project title	Fraud Detection In Healthcare Insurance
Group Number	1 (One)
Mentor Name	Mr.Animesh Tiwari

Acknowledgement

We would like to thank our mentor Mr. Animesh Tiwari for providing his valuable guidance and suggestions over the course of our Project Work. We also thank him for his continuous encouragement and interest towards our Project work.

We are extremely grateful to all our teaching and non-teaching staff members of GREAT LEARNING, who showed keen interest and inquired us about our development and progress.

We greatly admire and acknowledge the constant support we received from our friends and team members for all the effort and hard work that they have put into completing this project.

TABLE OF CONTENTS

TABLE OF CONTENTS

1. INDUSTRY REVIEW

1.1 BACKGROUND RESEARCH:

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. Every country has its unique healthcare system.

A country's healthcare quality is measured on the following criteria: [1]

1. Expenditure : It is the measure on how much the country spend on healthcare
2. Quality : It is the measure of the quality of the healthcare
3. Availability : It is the availability of the healthcare for the population
4. Population's Health : It is the measure of how effective it is on the population
5. Upfront Cost : It is the measure of how much people pay upfront for healthcare

As the industry is proliferating at an expeditious pace, 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.

Health Insurance Schema in US:

The US healthcare sector accounts for an outsized share of spending relative to health outcomes, which leads to several uncertainties around politically driven reform of the industry.

Most of the Americans are covered by both public and private health insurance. About 70% of the insured Americans are covered by private insurance plans through their employers, and 30% are covered by government-funded programs. The government insurance programs include Medicare, Medicaid, and the Children's Health Insurance programs.

People end up getting bankrupt by just paying the bills for the medical expenses. Statistics show that Medical bills cause a bankruptcy for every 30 seconds which has led to around 6,50,000 people go bankrupt every year.

Medicare:

Medicare is a national health insurance program in the United States, begun in 1966 under the Social Security Administration and now administered by the Centers for Medicare and Medicaid Services (CMS). Medicare is the federal health insurance program for People who are 65 or older, Certain younger people with disabilities, People with End-Stage Renal Disease (permanent kidney failure requiring dialysis or a transplant, sometimes called ESRD). [2]

There are four parts of Medicare: [3]

- Part A provides inpatient/hospital coverage.
- Part B provides outpatient/medical coverage.
- Part C offers an alternate way to receive your Medicare benefits
- Part D provides prescription drug coverage

Medicare Part A: It covers Inpatient care in a hospital, Skilled nursing facility care, Nursing home care (inpatient care in a skilled nursing facility that's not custodial or long-term care), Hospital care, Home health care.

Medicare Part B: It covers the Services or supplies that are needed to diagnose or treat the medical condition and that meet accepted standards of medical practice. Health care to prevent illness (like the flu) or detect it at an early stage, when treatment is most likely to work best.

Medicare Part C : Medicare Part C is not a separate benefit. Part C is the part of Medicare law that allows private health insurance companies to provide Medicare benefits. These Medicare private health plans, such as HMOs and PPOs, contract with the federal government and are known as Medicare Advantage Plans.

Medicare Part D: Medicare's prescription drug benefit is the part of Medicare that provides outpatient drug coverage. Part D is provided only through private insurance companies that have contracts with the federal government, it is never provided directly by the government.

Projected Health Care Expenditures:

In 2016, the US spent $3.4 trillion on health care expenditures [4]. The National Health Care Anti-Fraud Association estimates conservatively that health care fraud ranges from 3 to 10 percent of the nation's total health care expenditures. According to the above percentages, fraud, waste and abuse would range from $102 billion to $340 billion.

Experts project that health care expenditures will soar as high as $5.5 trillion by 2025 . Tracing back to 2009, it is reported that Medicare fraud alone is $100 billion annually [5].

Fraud, Waste and Abuse in the Healthcare Industry: [6]

- Fraud is a certainly intended scheme to defraud the program to gain money or property owned by the program.
- Waste is overutilization of services that results in futile costs in the program.
- Abuse determines the payment of services for the intentionally misrepresented facts.

As fraud in the industry remains a critical problem to articulate, our focus is on Fraud Detection.

Types of Fraud in HealthCare [7]:

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 service

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 may have more than one party involved (ie... either a patient, a physician or an insurance company).

News Articles on Fraud in Medicare:

"The Personal Toll of Whistle-Blowing", Sheelah Kolhatkar (2019) [8]

In 2009, Darren Sewell (Vice-President of Freedom Health Insurance Company) felt perplexed over the issue that though the entire world was facing recession, healthcare was thriving of its surplus demand in service. He focussed on investigating this in his own firm and found that his firm is also a part of the syndicate of fraud activities. He found numerous discrepancies which the common traits are misinterpretation of the details, wrong billing and more.

Darren Sewell then met with a FBI officer Ed Ortega and went undercover to obtain the evidence . A suit was filed against this medical fraud activity later by Sewell, which is known as "the whistleblower suit". On behalf of the suit, most of the medical fraud came into the purview of the organization.

"35 Audits Find Medicare Advantage Plans Overbilling CMS", Public Payers News (2016) [9]

Medicare Advantage Plan puts the patient into two baskets which is based on the individual's risk score. Higher risk score would be given to a patient who has higher healthcare costs whereas the lower risk score would be given to a patient who has lower healthcare costs. Medicare Advantage Plan gives a higher percentage of coverage for patients with higher risk score.

Fraudulent activities which have been identified at this stage involves manipulation of the risk score. Out of 37 audits , 35 audits claimed to have false risk scores and only 2 were found to be legit. 35 of these audits did not possess patient's documents that indicated higher risk scores.

The amount of money overpaid for those false risk scores were in millions while the other genuine 2 claims were only in hundreds and thousands. The audit takes multiple years for inspection. CMS spent $117M to recover $14M of overpayment. $30M money has been spent for auditing medicare advantage plans.

"CMS Considers AI and Value-Based Care Fraud Prevention Strategies", Public Payers News (2019) [10]

Five pillars that were established to prevent Fraudulent claims:

1. Taking action against current fraudulent activities
2. Establishing frameworks for prevention of fraud from reoccurring
3. Providing protection against upcoming risks
4. Using advanced analytics and artificial intelligence to detect
5. Setting guidelines to prevent providers from fraud activities and hence reducing administrative burden.

1.2 TOPIC SURVEY IN BRIEF:

Problem Understanding:

From a business point of view, it can be very helpful if one has an idea which Provider is a Fraud. This will help businesses in saving enormous amounts of money from frauds which would aid those who are in real need.

Current Solution to the Problem: [9] [10]

A rule based system is followed, where unusual large transactions or repetitive transactions at typical locations are observed. A legacy software is used to identify the transaction which cannot process the real time data which is critical for digital space.

Proposed Solution to the Problem: [9] [10]

ML-based fraud detection will be considered for these solutions, with effective use of data analytics and related business insights using statistical, predictive, and cognitive models to detect frauds.

1.3 CRITICAL ASSESSMENT OF TOPIC SURVEY: [11]

The U.S. Department of Health and Human Services (HHS) projected that fraud and abuse account for between 3 to 15 percent of annual expenditures for healthcare in the United States which ranges from $100–170 billion.

To help combat fraud and abuse, the federal government's False Claims Act (FCA) would sue the violators for terrible damages with a penalty fine of $5,500–11,000 per false claim. While it is difficult to stop an individual who intentionally commits fraud, there are certain external and internal systems and processes that can be implemented to better detect fraud and abuse and to deter future fraud and abuse.

Data mining techniques have allowed payers to use more sophisticated techniques such as data mining, reporting, and rules engines for fraud and abuse detection.

2. LITERATURE SURVEY

Jing Li , Kuei-Ying Huang , Jionghua Jin , Jianjun Shi (Dec, 2007) "A Survey on statistical methods for health care fraud detection".

The Healthcare System in the United States is quite expensive. To assist people, a federal government health program, Medicare was put into action. But as the cost increased , an increase in the Fraud,abuse and waste started increasing as well.

Three kinds of Frauds are mentioned in this paper : Service Providers Fraud, Insurance Subscribers Fraud and Insurance Carriers Fraud.

The paper proceeds on how one should approach the fraud detection dataset. The following steps such as : Goal Setting , Data Cleaning, Missing Value Treatment, Data Transformation, Feature Selection and Data Auditing

Once all these steps are done then statistical models for detections are built. There are three types for model building: Supervised Learning , Unsupervised Learning and Semi-Supervised Learning. Post model building the performance evaluation has to be done.

Senthilkumar, Bharatendara K, Amruta A Meshram, Angappa Gunasekaran, Chandrakumarmangalam (July, 2018) "Big Data in Healthcare Management: A Review of Literature"

This paper reviews the definition, process, and use of big data in healthcare management. Major steps of big data management in the healthcare industry are data acquisition, storage of data, managing the data, analysis on data and data visualization.

The most challenging parts for big data in healthcare are data privacy, data leakage, data security, efficient handling of large volumes of medical imaging data.

The authors suggest few new data visualization tools to the healthcare analyst to make effective decision making. Big data has a great perspective to progress healthcare management and transform the healthcare industry to the next level.

Peter Travaille, Roland M. Müller, Dallas Thornton, Jos Van Hillegersberg (June, 2011) "Electronic Fraud Detection in the U.S. Medicaid Healthcare Program"

This paper analyzes the applicability of existing electronic fraud detection techniques in similar industries to the US Medicaid program.

Given the fact that Medicaid is the payer of last resort and receives little feedback from the actual beneficiary of paid healthcare services. The dependence on electronic fraud detection is significantly greater than in similar studied industries.

As learned from the credit card industry, the clear benefits of supervised learning techniques should be weighed against the costs of streamlining data acquisition and closing the feedback loop from adjudicated claims to labeled claims data

While it is not a technology problem, an in-depth assessment should evaluate the effects of Medicaid policy changes such as increasing Medicaid provider enrollment standards, delaying payment to allow for more claim review time, or providing incentives to report fraudulent activity found on EOBs.

Shivani S. Waghade, Aarti M. Karandikar (2018) "A Comprehensive Study of Healthcare Fraud Detection based on Machine Learning"

Healthcare is an essential in people's lives and it must be affordable. It is expanding at an expeditious pace. At the same time, fraud in this industry is turning into a critical problem. One of the issues is the misuse of the medical insurance systems.

Recently, machine learning and data mining techniques are used for automatically detecting healthcare frauds. Supervised, unsupervised and semi-supervised learning are the three categories of Machine learning approaches.

In most of the cases, semi-supervised learning approaches are used by many researchers. After reviewing different studies on healthcare fraud detection, it can be concluded that frauds or abuse that occur in health insurance systems can be of different unusual patterns which can be achieved by taking the correlations between different entities of healthcare data can be taken into account.

3. METHODOLOGY

Cross Industry Standard Process for Data Mining (CRISP-DM) methodology is followed. CRISP-DM provides the complete blueprint for conducting a Data Mining project. It provides a uniform framework for experience documentation. [12]

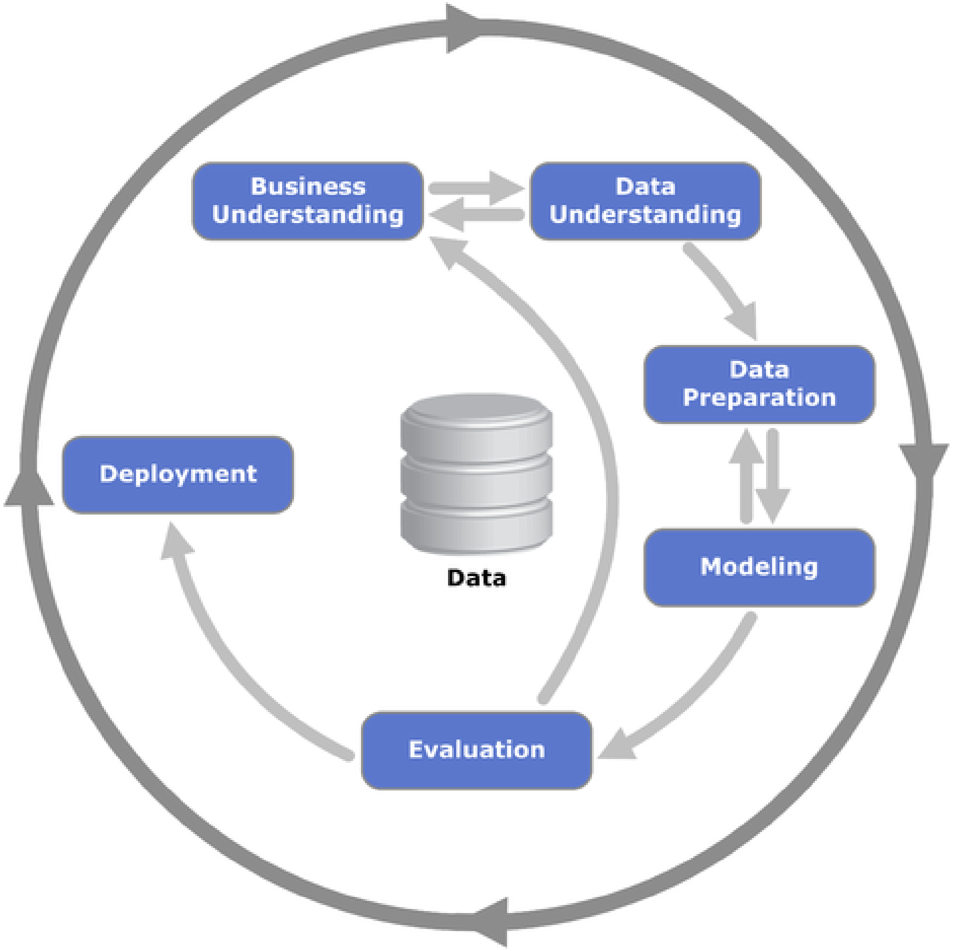


Fig 3.1 : Methodology, CRISP - DM (Fig Source : Wikipedia)

The sequence of the phases is not strict and moving back and forth between different phases as it is always required. It is essential to have a standard process as it aids to project planning and management. Published in 1999 to standardize data mining processes across industries, it has since become the most common methodology for data mining, analytics, and data science projects. The fact behind the success of CRISP-DM is that it is an industry, tool, and application neutral.

It breaks down the Life Cycle of a Data Mining project into the following phases,

Phase 1 : Business Understanding

Understanding the business requirement plays a paramount importance. The primary step to find out the business objective. In order to achieve that task, we may need profound knowledge on comprehending the outline of the issue which relates to the business objective.

After understanding the requirements, the following step is to lay down the aspects which would be intended by the respective Data-Scientist until the project is accomplished and the desired results are delivered.

Phase 2 : Data Understanding

Data understanding goes seamless with respect to the business understanding stage where the data collection, description, exploration and validation take place in this stage.

At this phase, Exploratory Data Analysis will take place in order to identify the significant variables influencing the business prediction.

Phase 3 : Data Preparation

The third phase is Data Preparation which is also called as Data Wrangling or Data Munging Phase, which involves the development of the final dataset amicable for the modelling. the final dataset for modelling. It covers all activities to construct the final dataset from the initial raw data. In this phase Data Cleaning, Imputation and Feature Engineering are performed

Phase 4 : Modelling

Data modelling is proceeded after the data preparation stage which involves selecting the ideal modelling technique, test design will be generated and the model will be built. Several models will be built based on tuning iterations until the best model is found

Phase 5 : Evaluation

Evaluation is the stage where the selected model technique is evaluated based on the right fit. The steps executed to construct the model to ascertain it properly achieves the business objectives.

Phase 6 : Deployment

Deployment is the final stage of the process where monitoring and maintenance takes place from the best evaluated model. However, this project does not have the requirement to execute the deployment.

4. DATASET AND DOMAIN

4.1 DATASET SOURCE:

The data used in this project was retrieved from Kaggle which is used to analyse the Insurance Claim Fraud for Medicare[13].

4.2 DATASET OVERVIEW:

The data in its original form consisted of eight different csv files. Four of these files belong to the dataset labeled with potential fraud column: train beneficiary, train outpatient, train inpatient, and train providers flagged. The other four files belong to the unlabeled dataset (without potential fraud column) : test beneficiary, test outpatient, test inpatient, and test providers.

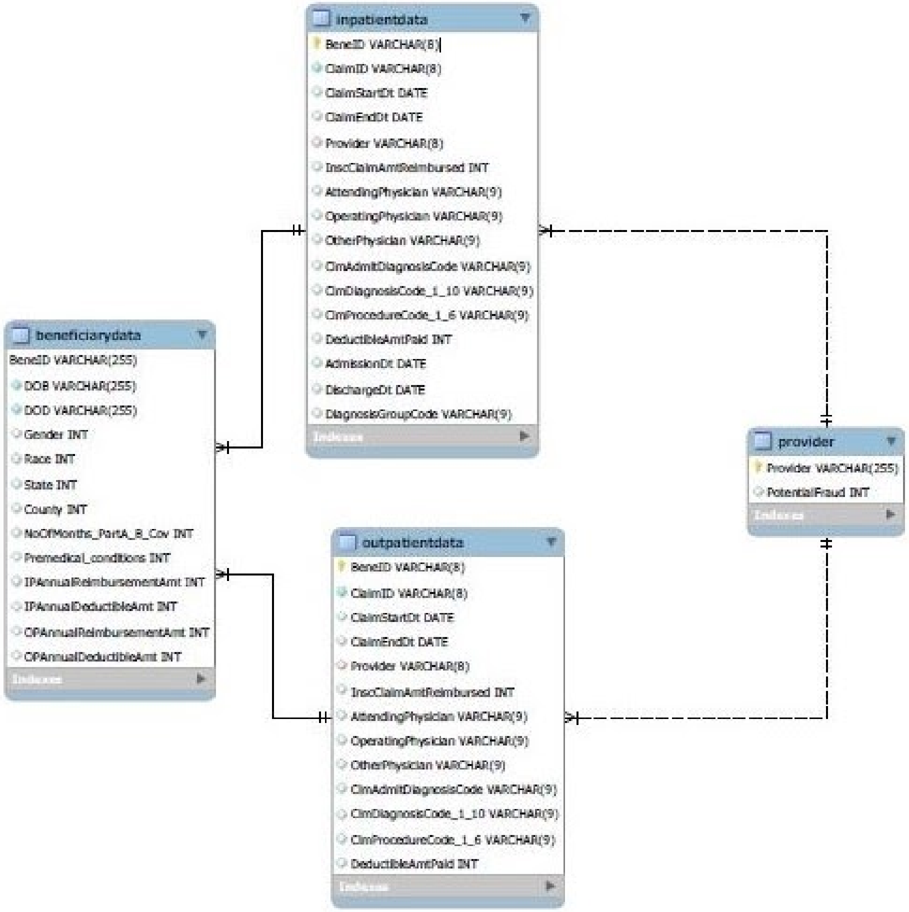

inpatientdata
BeneID VARCHAR(8)
ClaimID VARCHAR(8)
ClaimStartDt DATE
ClaimEndDt DATE
Provider VARCHAR(8)
InscClaimAmtReimbursed INT
AttendingPhysician VARCHAR(9)
OperatingPhysician VARCHAR(9)
OtherPhysician VARCHAR(9)
ClmAdmitDiagnosisCode VARCHAR(9)
ClmDiagnosisCode_1_10 VARCHAR(9)
ClmProcedureCode_1_6 VARCHAR(9)
DeductibleAmtPaid INT
AdmissionDt DATE
DischargeDt DATE
DiagnosisGroupCode VARCHAR(9)
Indexes
beneficiarydata
BeneID VARCHAR(255)
DOB VARCHAR(255)
DOD VARCHAR(255)
Gender INT
Race INT
State INT
County INT
NoOfMonths_PartA_B_Cov INT
Premedical_conditions INT
IPAnnualReimbursementAmt INT
IPAnnualDeductibleAmt INT
OPAnnualReimbursementAmt INT
OPAnnualDeductibleAmt INT
Indexes
provider
Provider VARCHAR(255)
PotentialFraud INT
Indexes
outpatientdata
BeneID VARCHAR(8)
ClaimID VARCHAR(8)
ClaimStartDt DATE
ClaimEndDt DATE
Provider VARCHAR(8)
InscClaimAmtReimbursed INT
AttendingPhysician VARCHAR(9)
OperatingPhysician VARCHAR(9)
OtherPhysician VARCHAR(9)
ClmAdmitDiagnosisCode VARCHAR(9)
ClmDiagnosisCode_1_10 VARCHAR(9)
ClmProcedureCode_1_6 VARCHAR(9)
DeductibleAmtPaid INT
Indexes

Fig 4.2: ER Diagram for Dataset

4.3 DATA DICTIONARY:

Data Definition: [14]

- Inpatient: A patient who receives Medical treatment being admitted in the hospital.
- Outpatient: A patient who receives Medical treatment without being admitted in the hospital.
- Diagnosis: Diagnosis is the process of determining which disease or condition explains a person's symptoms and signs. Based on the Diagnosis of a patient a physician will be assigned to the patient.
- Physician: A physician is a doctor of medicine. Each patient is assigned to a physician based on their Diagnosis. For instance a person who has been diagnosed with Heart Failure will be assigned to a Cardiologist and cannot be assigned to a Neurologist.
- Diagnosis Code: Diagnosis code is used as a tool to identify and group the diseases or symptoms of the patient. It is in conjunction with ICD-9 (International Classification of Diseases Ninth Revision). Diagnosis Code 4019 indicates Unspecified essential hypertension, 1970 indicates Secondary malignant neoplasm of lung. [15]
- Procedure Code: Procedure code is used to identify what procedure was done to or given to a patient. It is in conjunction with ICD-9-CM (International Classification of Diseases Ninth Revision Clinical Modification). Procedure Code 7092 indicates Scar conditions and fibrosis of skin, 331 indicates Incision of lung.
- Pre-medical conditions: The existing medical conditions that any person has are pre-medical conditions (Comorbidities).
- Beneficiary: A beneficiary is a person whom the government must cover certain healthcare costs including premium deductibles and copays.
- Claim: An insurance claim is a formal request to an insurance company asking for a payment based on terms of insurance policy. Claim will be made by the Provider for the Beneficiary.
- Provider: A provider is a person who is incharge of filing and claiming the insurance for the Beneficiary.
- Reimbursed Amount: Reimbursed Amount is the amount compensating the out of pocket expenses by the provider to the beneficiary. Health insurance companies often face reimbursement due to urgent medical procedures that need to be performed.

- Deductible Amount: Deductible Amount is the amount that the Beneficiary has to pay as part of a claim whenever it arises and the rest of the amount is paid by the insurance.

Data Dictionary:

1. Beneficiary Details Data: It provides the KYC details of both the Inpatients and Outpatients. It also contains the details of the pre-medical conditions of the patient. The corresponding file contains 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	State	State Code for the states in U.S.
7	County	County Code for the counties in U.S.
8	NoOfMonths_{PartA/B}Cov	No of months of Medicare Part A and Medicare Part B insurance is covered for the Beneficiary
9	Premedical conditions	Columns such as RenalDiseaseIndicator, ChronicCond_Depression,ChronicCond_Diabetes indicate if the member has any prior medical condition.
10	IPAnnualReimbursementAmt	Inpatient's Annual Reimbursement Amount
11	IPAnnualDeductibleAmt	Inpatient's Annual Deductible Amount
12	OPAnnualReimbursementAmt	Outpatient's Annual reimbursement Amount
13	OPAnnualDeductibleAmt	Outpatient's Annual Deductible Amount

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

3. Outpatient Data : It provides information about the claims filed for the patients who visit hospitals but not admitted in the hospital. The corresponding file contains an exact of 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 Reimbursed to the Beneficiary for the particular claim
12	DeductibleAmtPaid	The amount to be paid by the Beneficiary before the health insurance pays anything
13	DiagnosisGroupCode	A system which classifies hospital cases according to certain groups
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 assigned to patients on the basis of the Disease diagnosed (follows ICDCode-9)
16	ClmProcedureCode_1-6	Procedure Codes assigned to patients on the basis of the procedure performed (follows ICDCode-9)

4. Provider Data : It provides details of the Provider ID and the Potential Fraud Claims. The corresponding file contains an exact of 5410 rows and 2 columns.

S.No	Column Name	Column Description
1	Provider	Unique ID of the insurance 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 getting fraud.

4.4 VARIABLE CATEGORIZATION:

On analysing the 56 columns in the dataset we have identified:

Numerical Features	6
Categorical Features	8
Object Features	42

4.5 PROBLEM STATEMENT:

The dataset summarizes a set of heterogeneous features about the Insurance Claim Fraud in Medicare in 2008 and 2009. The goal of the project is to save enormous amounts of money from frauds which would aid those who are in real need.

4.6 TOOLS USED:

- Programming Language: Python
- Visualizations: Python and Tableau Desktop

5. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis refers to the process of performing initial investigations on data so as to discover patterns, spot anomalies, test hypotheses and to check assumptions with the help of summary statistics and graphical representations.

Since numerous features are present in the dataset , certain features need to be dropped to before feature engineering and exploratory data analysis privy to the missing value treatment which helps to gain more insights.

5.1 FEATURE ENGINEERING:

S.No	Column Name	Column Description
1	IsAdmitted	If the patient is from the Inpatient file, it is marked as 1 and if the patient is from the Outpatient file, it is marked as 0.
2	Claim Duration	It is difference between the Claim Start Date and Claim End Date
3	Claimed_Age	It is the difference Claim Start Date and Date of Birth
4	Admit Duration	It is the difference between the Discharge Date and Admission Date
5	Annual_reimbursement_amount	It is the sum of IPAnnualReimbursementAmt and OPAnnualReimbursementAmt
6	Annual_deductible_amount	It is the sum of IPAnnualDeductibleAmt and OPAnnualDeductibleAmt
7	Claim_Amount_For_Bene_Per_Claim	It is the sum of InscClaimAmtReimbursed and DeductibleAmtPaid
8	Avg_Claim_Amt_Per_Bene	It is the average of the Claim Amount for each Beneficiary
9	Avg_Reimbursed_Amt_per_bene	It is the average of the InscClaimAmtReimbursed for each Beneficiary
10	Total_Claims_Beneficiary	It calculates the Total Claim made by each Beneficiary
11	Total_Provider_Claims_ Covered	It calculates the Total Claims covered by each Provider
12	Total_Unique_Physicians	It calculates the Total Unique Physicians for each Beneficiary

S.No	Column Name	Column Description
13	Total_Physicians	It calculates the Total Physicians for each Beneficiary
14	Is_Dead	It is created based on DOD column (Self Comparison property of Nan)
15	DeadActions	It is to check if any action was taken after the Death of the Person.
16	Total_Diag_per_bene	Total Diagnosis for each Beneficiary
17	Total_procedure_per_bene	Total Procedures undergone by each Beneficiary
18	Fraudulent Providers for each Beneficiary	Total Number of Fraudulent Providers for each Beneficiary
19	Beneficiary for each Fraudulent Providers	Total Number of Beneficiaries for each Fraudulent Provider

Post performing the Feature Engineering, we observed 558211 number of rows and 72 columns.

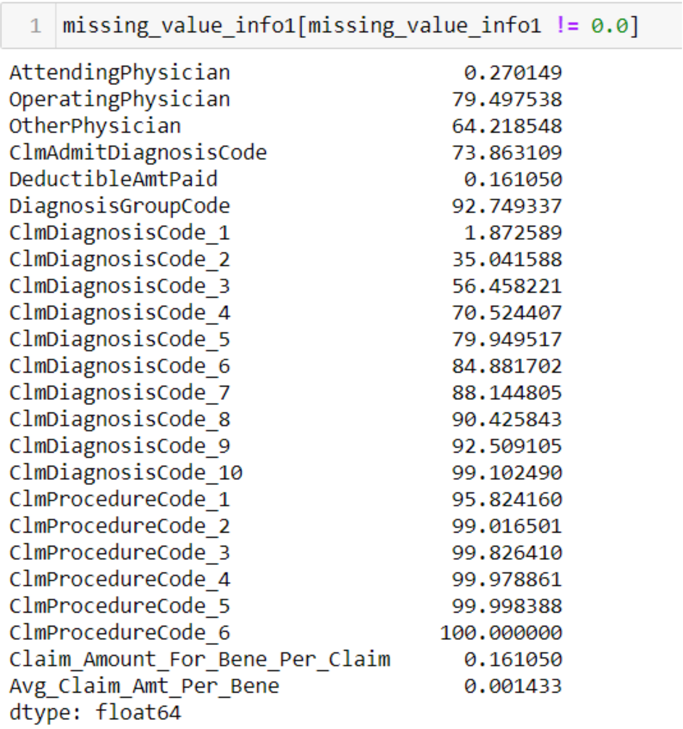
5.2 MISSING VALUE ANALYSIS:

Missing values in data is a common phenomenon in real world problems. Knowing how to handle missing values effectively is a required step to reduce bias and to produce powerful models. In the case of multivariate analysis, if there is a larger number of missing values, then it can be better to drop those cases (rather than do imputation) and replace them. On the other hand, in univariate analysis, imputation can decrease the amount of bias in the data, if the values are missing at random.

Below table provides the percentage of the missing values in each column.

Missing Values have been treated in three ways.

1. Dropping the features which have more than 40% missing values
2. Filling the columns with either single mean or median
3. Filling the columns with either grouped mean or median.



```
missing_value_info1[missing_value_info1 != 0.0]
```

```
AttendingPhysician                 0.270149
OperatingPhysician                79.497538
OtherPhysician                    64.218548
ClmAdmitDiagnosisCode             73.863109
DeductibleAmtPaid                  0.161050
DiagnosisGroupCode                92.749337
ClmDiagnosisCode_1                 1.872589
ClmDiagnosisCode_2                35.041588
ClmDiagnosisCode_3                56.458221
ClmDiagnosisCode_4                70.524407
ClmDiagnosisCode_5                79.949517
ClmDiagnosisCode_6                84.881702
ClmDiagnosisCode_7                88.144805
ClmDiagnosisCode_8                90.425843
ClmDiagnosisCode_9                92.509105
ClmDiagnosisCode_10               99.102490
ClmProcedureCode_1                95.824160
ClmProcedureCode_2                99.016501
ClmProcedureCode_3                99.826410
ClmProcedureCode_4                99.978861
ClmProcedureCode_5                99.998388
ClmProcedureCode_6               100.000000
Claim_Amount_For_Bene_Per_Claim    0.161050
Avg_Claim_Amt_Per_Bene             0.001433
dtype: float64
```

Fig 5.2.1 : Missing Values Before Treatment

Treating the Missing Value Columns:

1. **DeductibleAmtPaid** - The missing values for this feature is filled by taking the average of deductibleAmtPaid per beneficiary. Replacing with 0 will not work here as a beneficiary will have the same amount of Deductible Amount through the contract period with the Insurer.

2. **Claim_Amount_For_Bene_Per_Claim :** This feature is filled by taking the average of Claim_Amount_For_Bene_Per_Claim per beneficiary. Replacing with 0 will not work here as a beneficiary will have the different amount for each Claim Amount through the contract period with the Insurer.

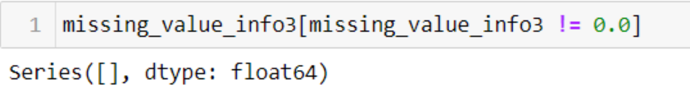
3. **Avg_Claim_Amt_Per_Bene:** This feature is filled by taking the average of the Claim_Amount_For_Bene_Per_Claim per beneficiary. Replacing with 0 will not work here as the Claim Amount for each beneficiary will be different.

4. **AttendingPhysician**: It would be unjust to fill this column by considering the entire dataset as one single group. But in reality this dataset can be segregated into two groups. 'yes_fraud' group that will have all the data pertaining to those providers who are fraud, and 'no_fraud' group that will have data pertaining to

those providers who are not fraud. AttendingPhysican is filled in such a fashion that by taking the mode of each group separately.

5. **ClmDiagnosisCode_1** : This feature gets filled by taking the mode of the ‘yes_fraud’ and ‘no_fraud’ group separately like a similar way of filling done in AttendingPhysician.

6. **ClmDiagnosisCode_2** : This feature gets filled by taking the mode of the ‘yes_fraud’ group separately and ‘no_fraud’ group separately just similar ways of filling done in AttendingPhysician.



```
missing_value_info3[missing_value_info3 != 0.0]
```

```
Series([], dtype: float64)
```

Fig 5.2.2 : Missing Values After Treatment

5.3 OUTLIER TREATMENT:

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

Outliers can be of two types: Univariate and Multivariate

- Univariate outliers are outliers in a 1-dimensional space.
- Multivariate outliers are outliers in an n-dimensional space.

Types : Data Entry Errors, Measurement Error, Experimental Error, Intentional Outlier, Data Processing Error, Sampling error and Natural Outlier.

Most of the ways to deal with outliers are similar to the methods of missing values like deleting observations, transforming them, binning them, treating them as a separate group, imputing values and other statistical methods like winsorization.

There are 18 numerical columns in this dataset. Boxplot is used to visualize these numerical columns to check for the presence of outliers.

From these 18 numerical columns , it can be seen that 13 are continuous columns and the rest are discrete numerical columns.

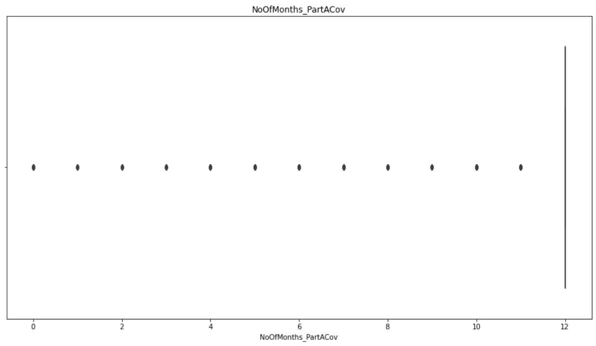


Fig 5.3.1 : Discrete Column

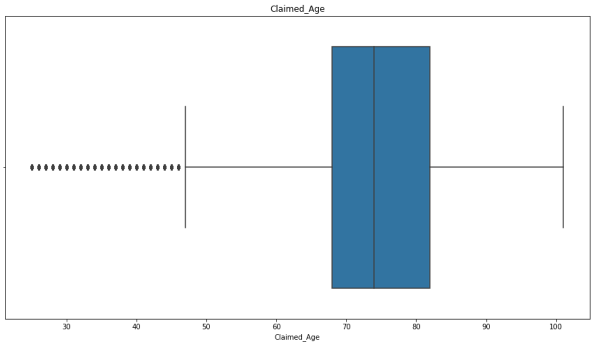


Fig 5.3.2 : Continuous Column Before Outlier Treatment

13 columns requires outlier treatments, which are

1. InsClaimAmtReimbursed
2. Claimed_Age
3. Annual_reimbursement_amount
4. Annual_deductible_amount
5. Claim_Amount_For_Bene_Per_Claim
6. Avg_Claim_Amt_Per_Bene
7. Avg_Reimbursed_Amt_per_bene
8. Total_Claims_Beneficiary
9. Total_Provider_Claims_Covered
10. Total_Unique_Physicians
11. Total_Physicians
12. Total_Diag_per_bene
13. Total_procedure_per_bene

Winsorization technique has been used for treating the outliers.

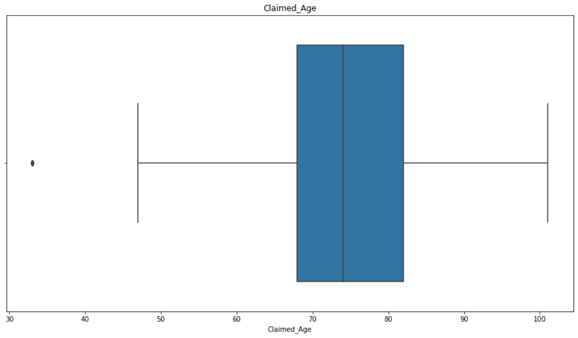


Fig 5.3.3 : Continuous Column After Outlier Treatment

After treating the columns, the outliers are either completely removed or reduced. The outliers that still exist are the true outliers that cannot be treated.

5.4 EXPLORATORY DATA ANALYSIS:

Univariate analysis is the simplest form of analyzing data. It doesn't deal with causes or relationships. It's major purpose is to describe the data, analyze and derive the structured pattern for it.

Bivariate analysis means the analysis of bivariate data. It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values. It usually involves the variables X and Y.

Multivariate analysis is used to study more complex sets of data. Multivariate analysis can reduce the likelihood of Type I errors.

5.4.1 Distribution of Target Variable:

Before plotting the distribution, we need to encode it such that the target variable satisfies 1 for Yes and 0 for No.

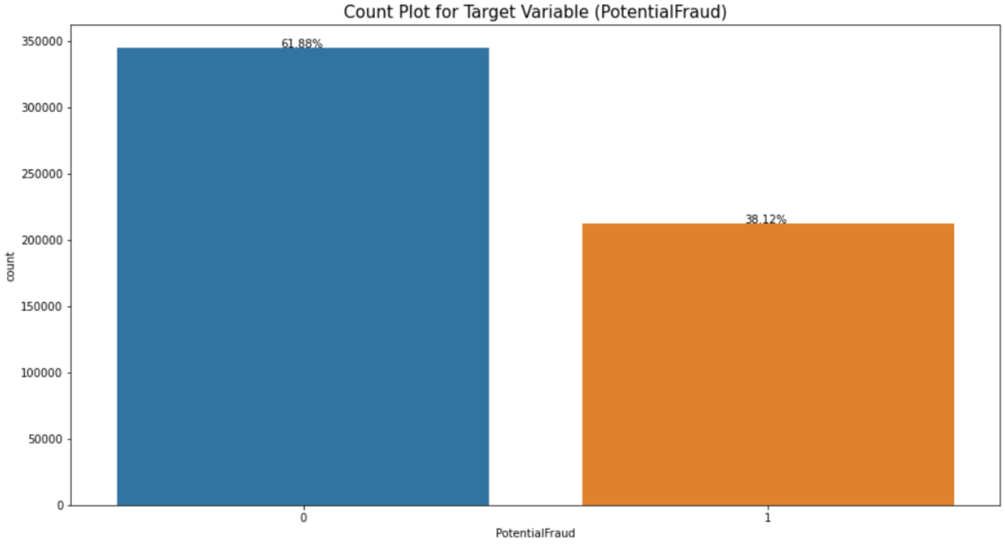


Fig 5.4.1 : Distribution of Target Variable

From the above distribution, it is noticed that the dataset is quite balanced.

5.4.2 Top Fraudulent Providers for each Beneficiary:

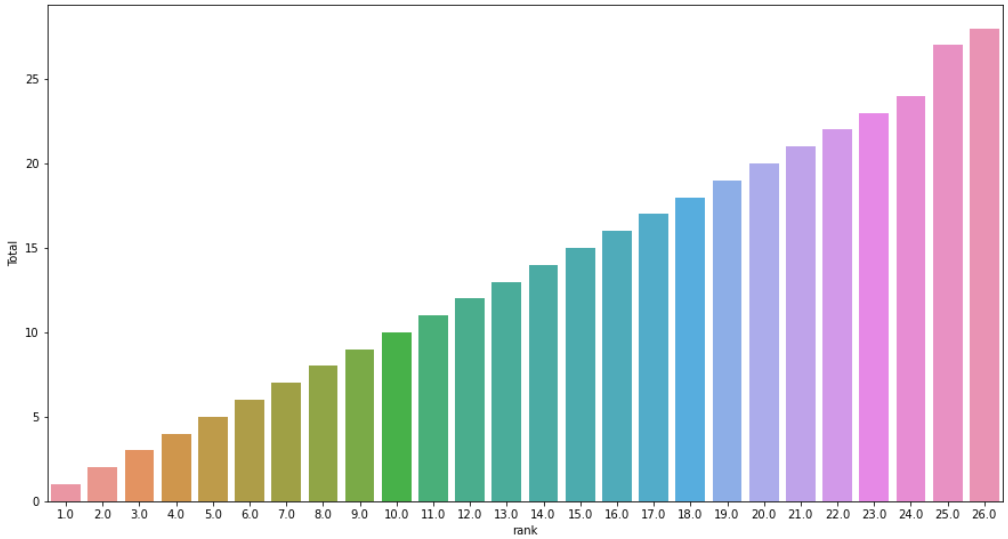


Fig 5.4.2 : Top Fraudulent Providers for each Beneficiary

From the above plot, we can infer that the Provider ID with Rank 26 has the highest Number of frauds for the Beneficiaries.

5.4.3 Top Beneficiary for each Fraudulent Providers:

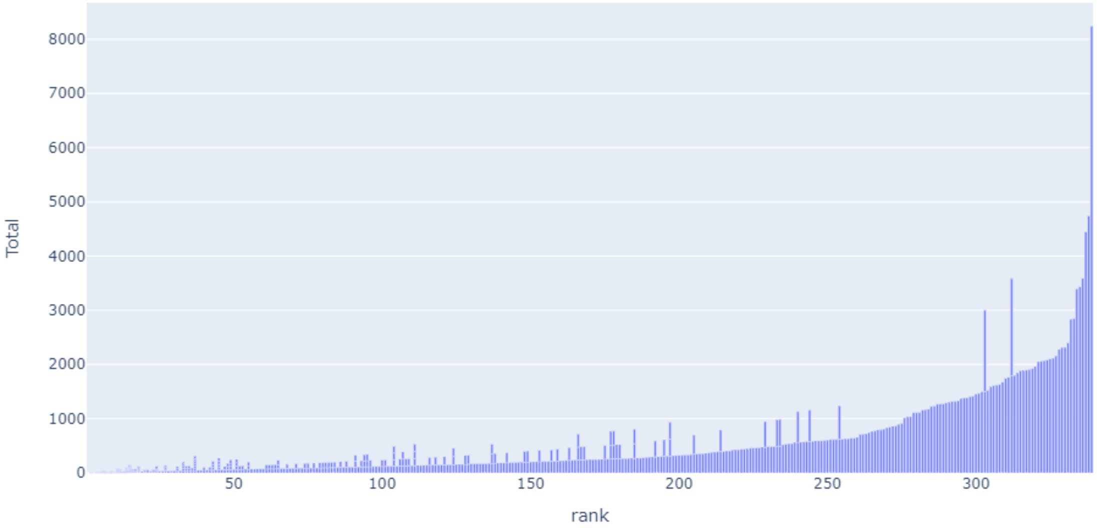


Fig 5.4.3 : Top Fraudulent Providers for each Beneficiary

From the above plot it is inferred that, Beneficiary with Rank 339 has the highest Number of Frauds to the Providers.

5.4.4 Admit Duration vs Claim Details:

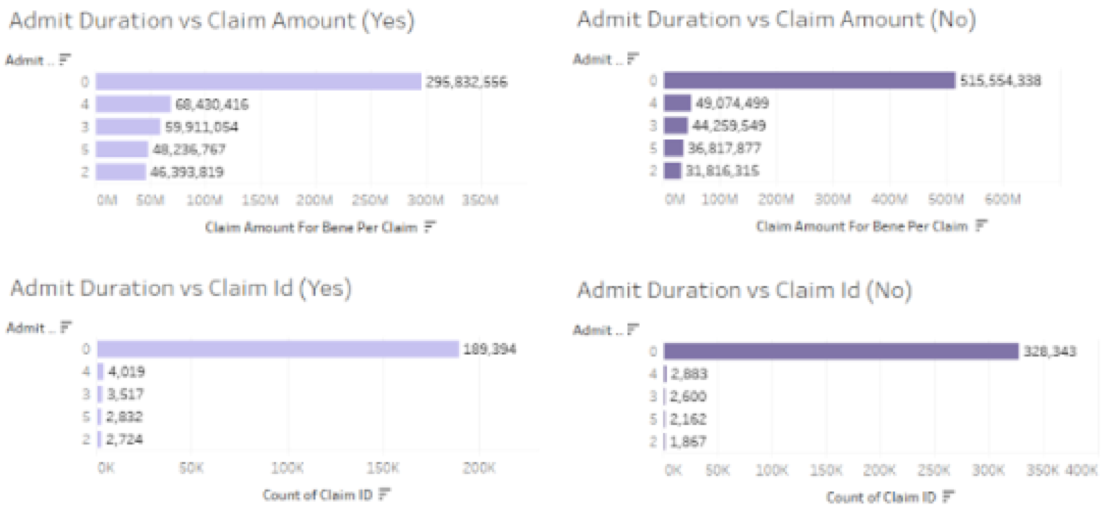


Fig 5.4.4 : Admit Duration vs Claim Details

Based on the above plots on Admit Duration indicate that the majority of the Fraud Claim has happened without a single day admission in the hospital. The maximum Fraud Claim amount is a little over 2.9M. It can also be seen that there are more ClaimID registers in that period as well.

5.4.5 Provider vs Total Provider Claims:

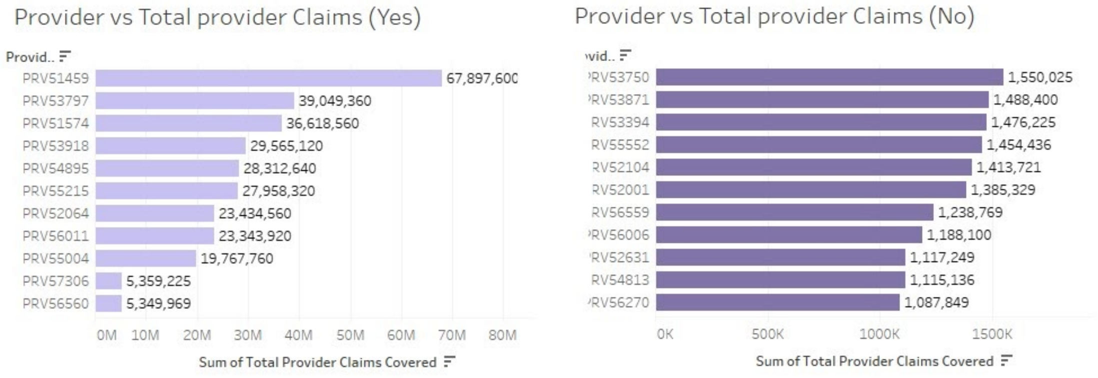


Fig 5.4.5 : Admit Duration vs Claim Details

From the above plots, we can infer that a fraudulent provider's total of claims is 60 times higher than non - fraudulent provider's total claims.

5.4.6 Admit Duration Vs Deductible Amount

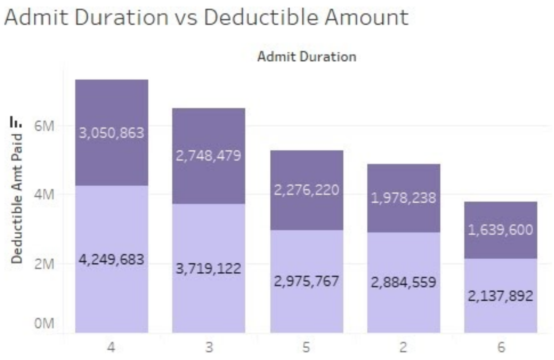


Fig 5.4.6 : Admit Duration Vs Deductible Amount

From the graph based on the Admit Duration vs Deductible Amount, it is noted that there is a small difference between Deductible Amount paid between the Fraudulent and Non-fraudulent claims. Even Though the difference is quite small, this small difference can make a huge impact on the Amount that can be saved for the Insurance Company

5.4.7 Admit Duration Vs Total Provider Claims

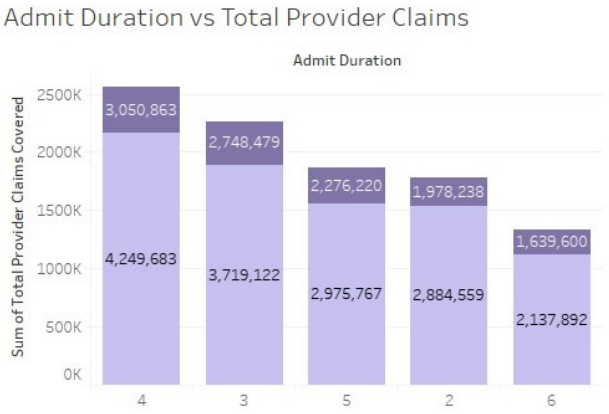


Fig 5.4.7 : Admit Duration Vs Total Provider Claims

From the graph based on the Admit Duration vs Total Provider Claims, it is noted that there is a small difference between Claims raised by the Fraudulent and Non-fraudulent claims. Even Though the difference is quite small, this small difference can make a huge impact on the Amount that can be saved for the Insurance Company.

5.4.8 IsAdmitted Vs Potential Fraud

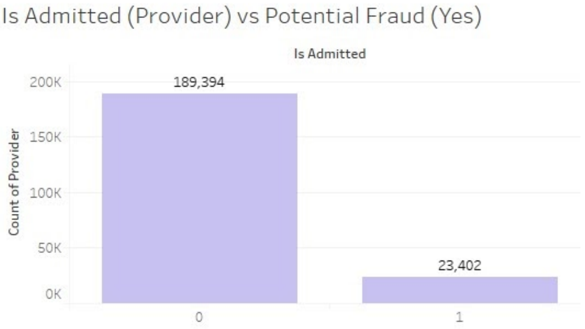


Fig 5.4.8 : Is Admitted vs Potential Fraud

From this graph, we can infer that the Number of Fraudulent Providers is high incase of IsAdmitted as 0.

5.4.9 Analysis based on the Claimed Age

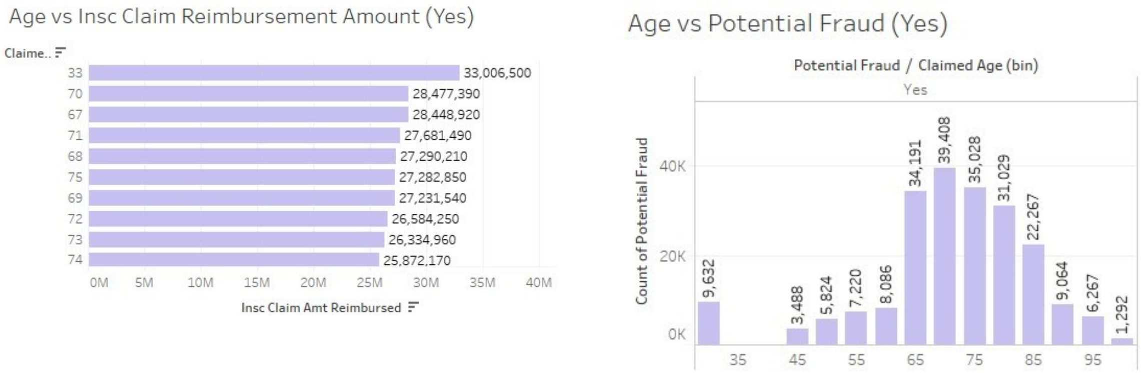


Fig 5.4.9 : Analysis based on the Claimed Age of the Beneficiary

From this graph, we can infer that more of the potential fraud cases are for the patients at the age of 33. As per the Medicare if a person is below 65 he is the policy due to disabilities. This shows the facts that the claims made at the age of 33 are fraudulent as the person with actual disabilities may not have received his support system.

5.4.10 Analysis based on the Attending Physicians

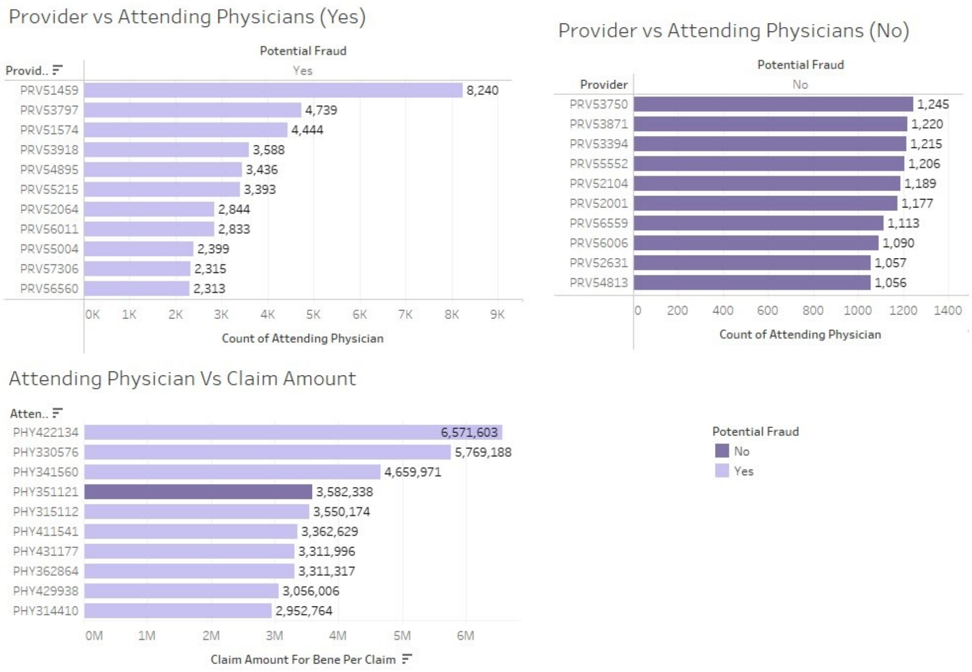


Fig : 5.4.10 Analysis based on the Attending Physicians

The above graphs show that there is a clear relationship between the Attending Physicians, the Provider and the Claim Amount. We can infer that the Claims Made by the Fraudulent Provider and Attending Physician are more compared to the Claims Made by the Non-Fraudulent Provider and Attending Physicians. This goes in hand with the Fraudulent Physician in the Claim Amount.

5.4.10 Age vs Claimed Amount

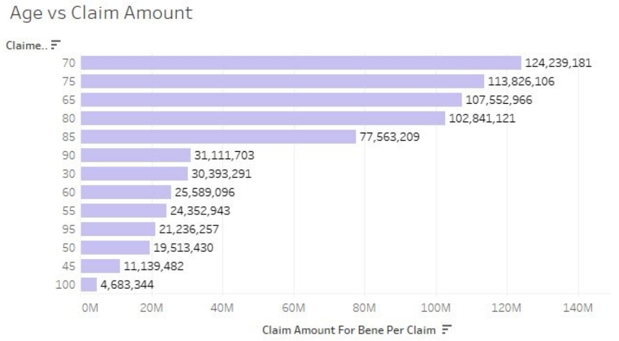


Fig 5.4.10 : Age vs Claimed Amount

From the graph we can infer that most of the fraudulent claims have occurred when the Beneficiary is above 65. This gives the possibility that the Claim Form was filled by the Provider and he has given false information to the Insurance company in order to receive a claim.

5.4.11 Correlation Matrix:

Correlation is a measure that determines how two features are related with each other. Correlation matrix is essential before and after any feature transformation, feature engineering and feature selections.

The heatmap in the picture below shows how independent features are correlated with each other. The scale on the right side indicates the level for the correlation. Anything close to 1 is highly correlated. In other words, those two features represent the same insights. By design, a feature when checked for a correlation with itself will always be 1.

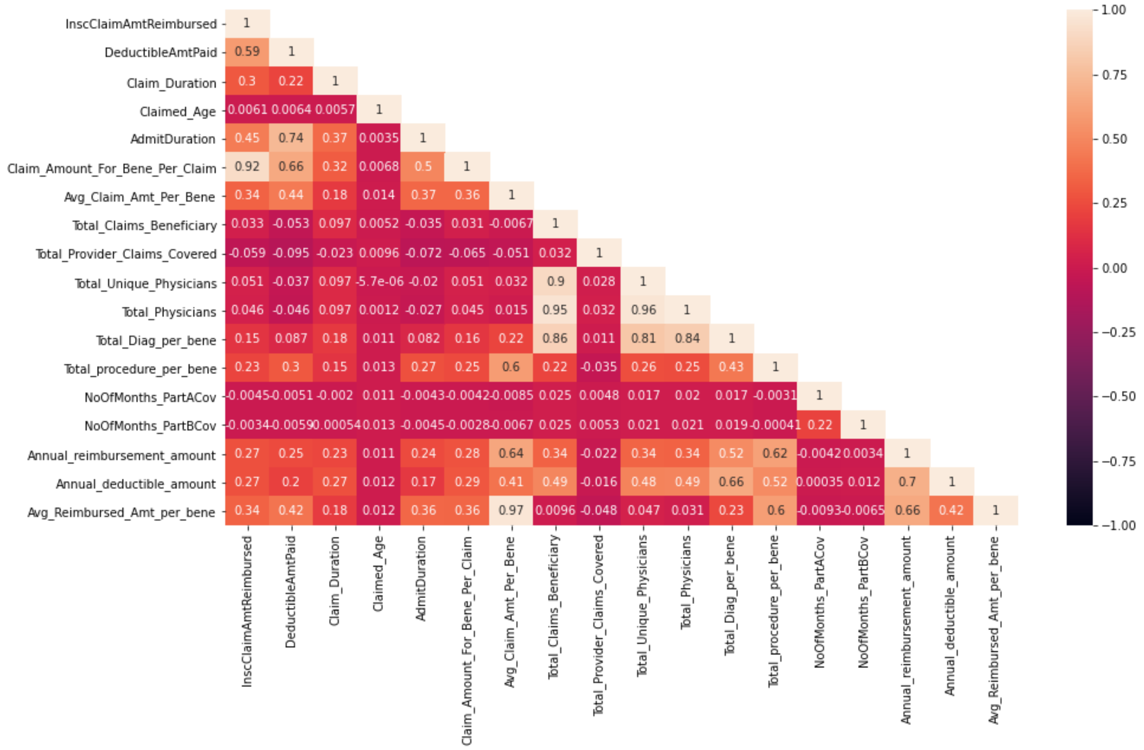


Fig 5.4.11 : Correlation Matrix

The corresponding features have strong correlation as their correlation value is greater than 0.7 Total_Physicians, Total_Unique_Physicians, Total_Diag_per_bene, Claim_Amount_For_Bene_Per_Claim and Avg_Reimbursed_Amt_per_bene.

This can be seen along the diagonal of the heatmap when all the features correlated with itself give the value of 1 with white background color. But it can also be seen that there are other features that are highly correlated with each other as well. This gives an indication for the presence of multicollinearity.

5.5 HANDLING THE CATEGORICAL VARIABLES:

This dataset has a combination of Numerical and Categorical Variables. The Categorical Variables need to be handled such that the model handles both the data types.

1. The **premedical conditions** fall under the classification of either 0 or 1.

   Different methods of encoding such as Label Encoding or One Hot Encoding were tried, but integrity of the data was lost. So the ideal way to encode the variables was replacing the 2 as 0.

   The pre-medical conditions variables are RenalDiseaseIndicator, ChronicCond_Alzheimer, ChronicCond_Heartfailure, ChronicCond_KidneyDisease, ChronicCond_Cancer, ChronicCond_ObstrPulmonary, ChronicCond_Depression, ChronicCond_Diabetes, ChronicCond_IschemicHeart, ChronicCond_Osteoporasis, ChronicCond_rheumatoidarthritis and ChronicCond_stroke.

2. **Attending Physician**: The attending physician is encoded by taking the list of the top 25 attending physicians. If the attending physician is in the top 25 list, it is marked as 1 else it is marked as 0.

3. **ClmDiagnosisCode_1**: The claim diagnosis code 1 is encoded by taking the list of the top 25 claim diagnosis code. If the claim diagnosis code 1 is in the top 25 list, it is marked as 1 else it is marked as 0.

4. **ClmDiagnosisCode_2**: The claim diagnosis code 2 is encoded by taking the list of the top 25 claim diagnosis code. If the claim diagnosis code 1 is in the top 25 list, it is marked as 1 else it is marked as 0.

The columns ClaimStartDt, ClaimEndDt, AdmissionDt, DischargeDt, DOB, DOD, IPAnnualReimbursementAmt, OPAnnualReimbursementAmt, IPAnnualDeductibleAmt, OPAnnualDeductibleAmt, DeadActions, AttendingPhysician, ClmDiagnosisCode_1, and ClmDiagnosisCode_2 are dropped as they are redundant.

6. STATISTICAL SIGNIFICANCE OF VARIABLES

Statistical significance tests are designed to address the problem and quantify the likelihood of the samples of skill scores being observed in given the assumption that they were drawn from the same distribution. Statistical learning theory deals with the problem of finding a predictive function based on data.

It is used to compare the two features to build the models to predict the output. In our case the target column is Potential Fraud which is categorical in nature.

If Null Hypothesis is selected, the feature that we used for the test is not significant to build the model, because there is no use to predict the potential fraud. If it satisfies Alternate Hypothesis it is useful in predicting the potential fraud.

The Statistical tests are performed after Handling the Categorical Variables as each categorical variable has many values. Since the Data is not normal for the Numerical Columns, MannWhitneyU test is performed and for the Categorical Columns Chi Square Contingency test is performed.

The following features were tested with Chi Square Contingency and the Alternate Hypothesis has been proven:

- State and Potential Fraud
- County and Potential Fraud
- Race and Potential Fraud
- Renal Disease Indicator and Potential Fraud

The following features were tested with MannWhitneyU and the Alternate Hypothesis has been proven:

- Total Provider Claims Covered and Potential Fraud
- Total Unique Physicians and Potential Fraud
- Claimed Age and Potential Fraud
- InscClaimAmtReimbursed and Potential Fraud
- AdmitDuration and Potential Fraud

7. MODEL BUILDING PREREQUISITES

7.1 SCALING THE DATA:

The data is scaled using standard scaler, standard scaler converts the values to Z-scores by subtracting the values from the mean and dividing it by the standard deviation.

7.2 BUILDING THE MODEL DATA:

Before building the model data, One hot encoding technique will be used to handle the Race column as categorical variable .

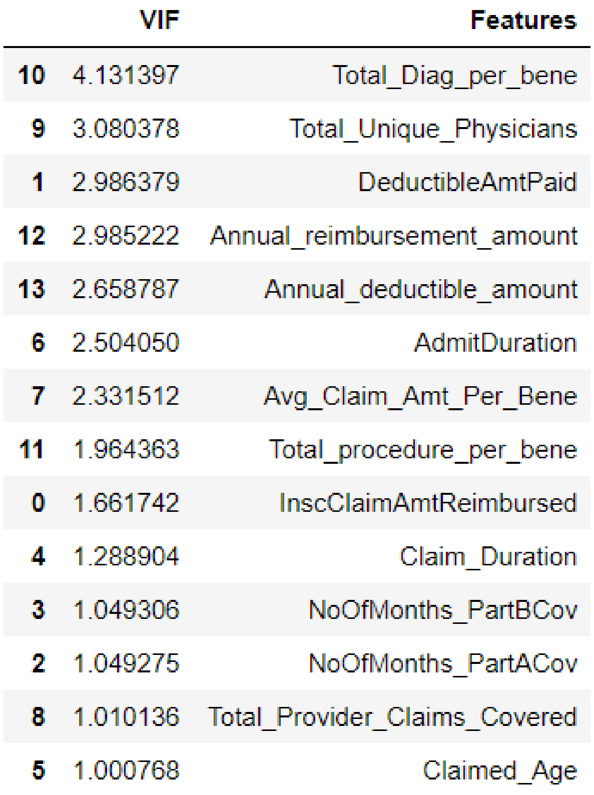
The columns BeneID, Provider and Claim ID will be dropped as they are unique and cannot be handled in encoding.

The model data is built by concatenating the scaled numerical columns along with encoded categorical columns and with target variable to form the model data. Now there are 40 columns and 5,58,211 rows.

7.3 ASSUMPTIONS FOR LOGISTIC REGRESSION:

Multicollinearity is a situation where two or more predictors are highly linearly related. In general, an absolute correlation coefficient of > 0.7 among two or more predictors indicates the presence of multicollinearity. Precise estimate of coefficients gets affected due to the presence of multicollinearity in the regression model.Therefore multicollinearity has to be treated. This is done by calculating Variance Inflation Factor and dropping the predictors whose VIF values are above 5.

After treating Mulit-collinearity:



	VIF	Features
10	4.131397	Total_Diag_per_bene
9	3.080378	Total_Unique_Physicians
1	2.986379	DeductibleAmtPaid
12	2.985222	Annual_reimbursement_amount
13	2.658787	Annual_deductible_amount
6	2.504050	AdmitDuration
7	2.331512	Avg_Claim_Amt_Per_Bene
11	1.964363	Total_procedure_per_bene
0	1.661742	InscClaimAmtReimbursed
4	1.288904	Claim_Duration
3	1.049306	NoOfMonths_PartBCov
2	1.049275	NoOfMonths_PartACov
8	1.010136	Total_Provider_Claims_Covered
5	1.000768	Claimed_Age

Fig 8.2 : After treating Multicollinearity

7.4 SELECTING THE SAMPLE FOR MODEL DATA

Since the dataset is large, a sample of the dataset will be chosen to proceed with the model building and evaluation.

After several trial and error methods, the sample dataset of size 0.30 was chosen as it represents the whole population data represents .The method used to select the sample is train test split method with strafiy out parameter is given such that the proportion of the target variable is maintained during the selection of the sample.

The sample population contains 167464 rows and 36 columns. Since the Data is not normal, MannWhitneyU test is performed to check the distribution of the Data statistically.

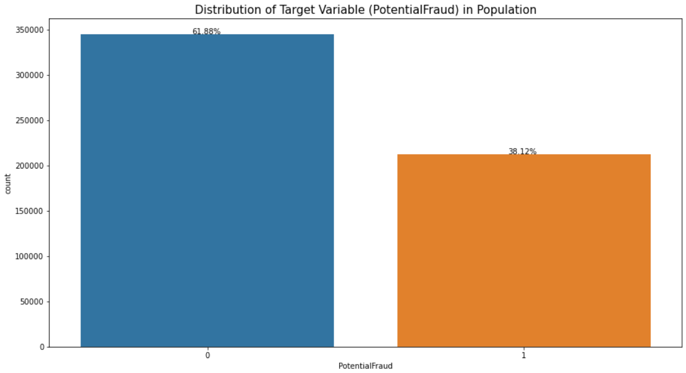


Fig 8.3.1 Distribution of Target Variable in Population

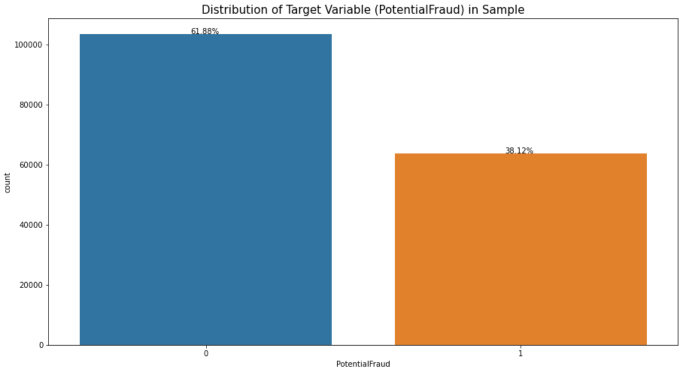


Fig 8.3.2 Distribution of Target Variable in Sample

Model Evaluation:

Let us identify the False Positive, False Negative, True Positive and True Negative

True Positive : A fraud claim is correctly marked as fraud

True Negative : A non-fraudulent claim is correctly marked as non-fraud

False Positive : A non-fraudulent claim is marked as fraud claim

False Negative : A fraud claim is marked as non-fraud claim

7.5 TRAIN TEST SPLIT

The train-test split is a technique for evaluating the performance for any supervised learning algorithm. The data is split into dependent feature Y (target) and independent features X.

The dataset is then divided into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model, instead it is used to predict the expected values. This second dataset is referred to as the test dataset.

The data is then split into training and testing sets in order to avoid data leakage. For our model performance evaluation 70:30 split is done where 70% of the data is used for training the model and 30% of the data is used for testing the model.

The objective is to estimate the performance of the machine learning model on new data, data not used to train the model.

8. MODELS FOR CLASSIFICATION[16]

Hyperparameters tuning : Hyperparameter tuning is the process of tuning the parameters to get good performance for the model. Machine learning algorithms never learn these parameters as they are defined by the programmer. Hyperparameter tuning aims to find such parameters where the performance of the model is highest or where the model performance is best and the error rate is least.

GridSearchCV: Grid Search is an effective method for adjusting the parameters in supervised learning and improving the generalization performance of a model. Grid Search CV tries all possible combinations of the parameters of interest and find the best ones.

8.1 LOGISTIC REGRESSION:

Logistic Regression is a Machine Learning algorithm which is used for the classification problems. It is a predictive analysis algorithm based on the concept of probability. It is built as a base model for the Target Variable falls under the Binary Classification. The optimal model hyper parameters are chosen using argmax. The tuned model is fitted and the model is evaluated.

Being the sparse matrix dataset, Logistic Regression model here gave the threshold value as 1. The threshold value cannot be 1 as the range for a logistic regression is between 0 and 1. Sparse matrix works well in tree-based models.

8.2. NAIVE BAYES:

Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. Naive Bayes Classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. The model is fitted and the model is evaluated.

8.3 DECISION TREE CLASSIFIER:

Decision Tree Classifier is a non-linear model that belongs to the family of supervised machine learning algorithms. It can be used for both a classification problem as well as a regression problem. The goal of the algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

The optimal model hyper parameters are chosen using GridSearchCV. The tuned model is fitted and the model is evaluated.

8.4 RANDOM FOREST CLASSIFIER:

Random Forest is a versatile ensemble learning method capable of performing both regression and classification tasks. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The optimal model hyper parameters are chosen using GridSearchCV. The tuned model is fitted and the model is evaluated.

8.5 XGBOOST CLASSIFIER:

XGBoost is an algorithm that has recently been dominating applied machine learning. XGBoost (eXtreme Gradient Boosting) is an advanced implementation of gradient boosting algorithm. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. This approach supports both regression and classification predictive modeling problems.

The optimal model hyper parameters are chosen using GridSearchCV. The tuned model is fitted and the model is evaluated.

8.6 MLP CLASSIFIER:

MLPClassifier stands for Multi-layer Perceptron classifier which in the name itself connects to a Neural Network. Unlike other classification algorithms such as Support Vectors or Naive Bayes Classifier, MLPClassifier relies on an underlying Neural Network to perform the task of classification.

The optimal model hyper parameters are chosen using GridSearchCV. The tuned model is fitted and the model is evaluated.

8.7 IDENTIFICATION OF BEST MODEL

As per our dataset, we used various classification models in order to find the best result in terms of accuracy and prediction. We found that the XGBoost classification model is providing better results in comparison to other models such as Random Forest, Decision Tree, MLP.

Accuracy Score for Train Dataset after Hyperparameter Tuning:

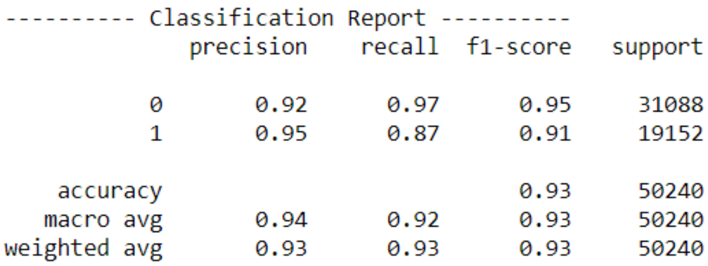
- Logistic Regression : 0.82
- Naive Bayes (without hyperparameter tuning) : 0.39
- Decision Tree Classifier : 0.88
- Random Forest Classifier : 0.84
- XGBoost Classifier: 0.98
- MLP Classifier: 0.82

Accuracy Score for Test Dataset after Hyperparameter Tuning:

- Logistic Regression : 0.82
- Naive Bayes (without hyperparameter tuning) : 0.39
- Decision Tree Classifier : 0.88
- Random Forest Classifier : 0.84
- XGBoost Classifier: 0.93
- MLP Classifier: 0.82

In order to verify XG Boost as our best model, we used the Classification Report to evaluate the performance of the model. Using the evaluation techniques it is inferenced that XGBoost is performing better than above mentioned models.

Classification Report for XGBoost Classifier:

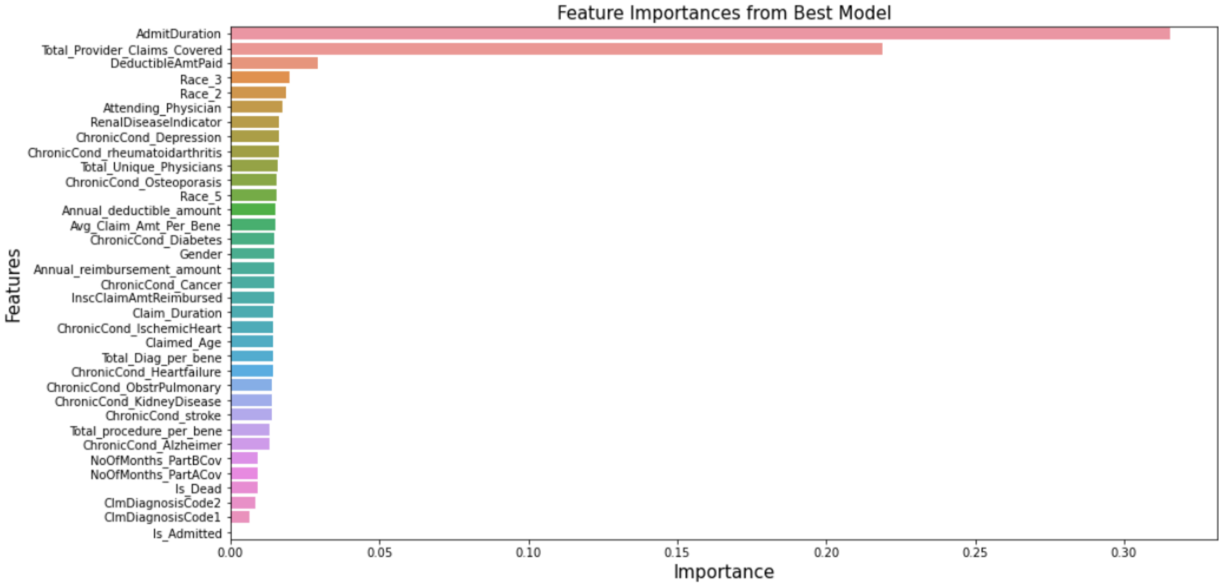


```
---------- Classification Report ----------
              precision    recall  f1-score   support

           0       0.92      0.97      0.95     31088
           1       0.95      0.87      0.91     19152

    accuracy                           0.93     50240
   macro avg       0.94      0.92      0.93     50240
weighted avg       0.93      0.93      0.93     50240
```

Fig 8.7 : Classification Report for XGBoost Classifier

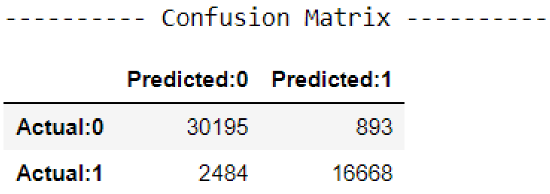
8.8 IDENTIFICATION OF IMPORTANT FEATURES

As XGBoost is identified as the best model, the Important Features from the Model have been identified. The key features identified are AdmitDuration, Total_Provider_Claims_Covers, DeductibleAmtPaid, Race_3, Race_2 and Attending Physician.



9. BUSINESS CASE

Based on the XGBoost, the confusion matrix is as below:



```
---------- Confusion Matrix ----------
```

	Predicted:0	Predicted:1
Actual:0	30195	893
Actual:1	2484	16668

Fig 9 : Confusion Matrix XGBoost Classifier

No of Fraudulent Claim	19152
Total Data points	50240
19152 are fraud out of 50240	0.3812101911
Total Loss calculates is	$751,255,150
Each fraud costs	$14953.32703
Model precision is 95%	18194
Savings	$272060832
Risk of losing the non fraud customers	$4465000

10. CONCLUSION

Key Takeaway Points:

1. The Xgboost model was able to save $272060832.
2. It has been observed that the age group of 65 and above and the age group of 35 face a lot of fraudulent providers.
3. Collaboration of physicians with providers was found.
4. Outpatient had a lot of fraudulent providers.
5. Lot of fraudulent activities take place when a provider claims money for a wheelchair but the beneficiary does not receive that wheelchair.
6. Assignment of Wrong diagnosis code leads to registering the patient in a high risk category resulting in losing more money by Medicare.
7. Proper checkpoints have to be established inorder to prevent false registration of patients' names to receive the claim amount.
8. Proper checkpoints have to be established inorder to prevent wrong upcoding of patients' diagnosis code,procedure code and other disease indicators.
9. Immediate checking of the patients health history has to be made to speed up the fraud prevention act.

Future Work:

During the initial 2000's, the health care management is redundant on sharing the patient's claim and health related information as it comes under the purview of classified information. As a matter of fact, that clause has paved the way to the occurrence of fraudulent activities by the imposters which has led to loss of several million dollars from the citizen's insurance amount.

Due to these restrictions, our dataset is structured in a way that the details of the attending physician, operating physician, other physician, diagnosis code and information, procedure codes, presence of disease, beneficiary details and hospital are not disclosed. Furthermore, the categorical details like state name, country, race and gender are encrypted to numerical data which is insufficient to make more predictions.
Likewise, the demographics of the beneficiaries (i.e....age, salary) are not available with the dataset. These imbalances in the dataset with insufficient details caused hurdles in modelling and predicting the business case.

However, during 2019 the health care has made several amendments to restrain the money fraud from the claim amount by making the data transparent and disclosed to all in order to access and claim for the incidents.With respect to the amendments, we could make the prediction with ease and also aids us in modelling technique to deliver qualitative results with better significance in results.

Limitations and Risks:

1. Diagnosis Code,Procedure Code are unknown
2. Beneficiary details are unknown
3. Physician Details are unknown
4. Disease Confirmation is not available
5. State and County are unknown
6. These are the risks with which the model was built.

If more information was available then more investigations could have been made.

11. REFERENCES

1. https://www.youtube.com/watch?v=DublqkOSBBA
2. https://www.medicare.gov/what-medicare-covers/
3. https://www.medicareinteractive.org/get-answers/medicare-basics/medicare-coverage-overview/
4. https://www.sas.com/en_us/insights/articles/risk-fraud/medicaid-benefit-fraud.html
5. https://stonebridgebp.com/library/uncategorized/how-healthcare-fraud-affects-us-all/
6. https://www.essencehealthcare.com/fraud-waste-abuse/
7. 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
8. https://www.newyorker.com/magazine/2019/02/04/the-personal-toll-of-whistle-blowing
9. https://healthpayerintelligence.com/news/35-audits-find-medicare-advantage-plans-overbilling-cms
10. https://healthpayerintelligence.com/news/cms-considers-ai-and-value-based-care-fraud-prevention-strategies
11. https://fin.plaid.com/articles/algorithmic-and-rules-based-fraud-models/
12. https://www.opusconsulting.com/rule-based-vs-machine-learning-effective-fraud-prevention/
13. https://perspectives.ahima.org/healthcare-fraud-and-abuse/
14. https://paginas.fe.up.pt/~ec/files_0405/slides/02%20CRISP.pdf
15. https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis
16. https://searchhealthit.techtarget.com/
17. https://en.wikipedia.org/wiki/List_of_ICD-9_codes
18. https://www.analyticsvidhya.com/blog/