ANALYSIS OF FRAUD DETECTION IN HEALTH INSURANCE SECTOR

A REPORT SUBMITTED TO

SAVITRIBAI PHULE PUNE UNIVERSITY

TOWARDS PARTIAL FULFILLMENT OF DEGREE

OF

MASTER OF SCIENCE (M. SC.)

IN STATISTICS

IN THE FACULTY OF SCIENCE AND TECHNOLOGY

SUBMITTED BY

MONIKA(2134)

SWAPNIL MORKHADE(2135)

GAYATRI WADGHULE(2153)

UNDER THE GUIDENCE OF

DR. MOHAN KALE

DEPARTMENT OF STATISTICS

AND

CENTRE FOR ADVANCED STUDIES IN STATISTICS,

SAVITRIBAI PHULE PUNE UNIVERSITY,

PUNE-411007,

INDIA.

MAY, 2023

Certificate of the Guide

This is to certify that, the following students of M.Sc. Statistics,

(1) MONIKA (2134)

(2) SWAPNIL MORKHADE (2135)

(3) GAYATRI WADGHULE (2153)

have successfully completed their project titled **ANALYSIS OF FRAUD DETECTION IN HEALTH INSURANCE SECTOR** under the guidance of **DR. MOHAN KALE** and have submitted this project report on **20th MAY 2023** as a part of the course ST-402, towards partial fulfillment of requirements for degree of M.Sc. Statistics in Savitribai Phule Pune University in academic year 2022-2023.

DR. MOHAN KALE
(Project Guide)

Prof. T. V. Ramanathan
(Head of the Department)

Acknowledgments

We would like to express our sincere gratitude to all those who have supported us in completing our project on "Analysis of fraud detection in health insurance sector".

First and foremost, we would like to thank our project guide Dr. Mohan Kale sir, for their valuable guidance and constant support throughout this project. Their insights and suggestions have been immensely helpful in shaping our understanding of the topic and ensuring that our work meets the required standards.

We are also thankful to Mr. Abhijeet Gosavi,VP CitiBank, who suggested us to work on this topic with the necessary resources and data to carry out this project successfully. Their contribution was vital in helping us gather relevant information and conduct meaningful analysis.

Furthermore, We would like to acknowledge the support and cooperation of Dr. Mukund Ramtirthkar and the other staff members at Department of Statistics, Savitribai Phule Pune University, who provided us with their time and expertise in answering our queries and sharing their experiences on the topic.

Finally, We would like to extend our appreciation to friends for their unwavering encouragement and motivation during the course of this project. Their support has been a constant source of inspiration for us.

Thank you all for your invaluable contributions to this project.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Summary

1.1 Introduction

During Covid-19 many people faced health related issues but they did not had any health insurance covering medical expenditures.Hence, it was difficult for them to manage the high expenditure. As a result people have now realised the importance of insurance coverage to fulfill non-planned medical expenditures. In view of this there has been a rapid growth in the insurance industry, which inturn lead to considerably increase in the number of insurance claims. As health is everyone's first priority making this sector fraud free in all aspects is absolutely essential. In it's report "Insurance Business" indian insurance market is growing at a rate of 14.5 % .

According to the 2019 report of National Health Care Anti-Fraud Association on healthcare fraud detection, the total losses in 2018 was 56,19,84,09,510 INR (USD 679.18 million) which is expected to reach 2,10,17,10,30,000 INR (USD 2.54 billion) by 2024. Healthcare fraud is an organized crime which involves peers of providers, physicians, beneficiaries acting together to make fraud claims.

1.2 Motivation

1.2.1 Aim

- To Build a binary classification model based on the claims filed by the provider along with Inpatient data, Outpatient data, Beneficiary details to predict whether the provider is potentially fraudulent or not.
- To calculate the probability score for fraudulent claims and insurance company can accept or deny the claim or set up an investigation on that provider.
- To Find out the important features which are the reasons behind the potentially fraudulent providers. Such as if the claim amount is high for a patient whose risk score is low, then it is suspicious.

1.2.2 Objectives

Healthcare fraud can occur in many forms, here ,we have focused on Provider's Fraud.

- The goal of this project is to predict the potentially fraudulent providers based on the claims filed by them.
- To discover important variables helpful in detecting the behaviour of potentially fraud providers.
- To Study fraudulent patterns in the provider's claims to understand the future behaviour of providers
- Financial loss is a great concern, but also protecting the healthcare system so that they can provide quality and safe care to legitimate patients.

1.3 Dataset

1.3.1 Data Description

Source of the data: Data are collected from references of research paper published by A. Jenita Mary and S. P. Angelin Claret **Analytical study on fraud detection in healthcare insurance claim data using machine learning classifiers**.
https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis
Data collection method: Secondary Data
Introduction to the dataset:
1) Inpatient Data: This data provides insights about the claims filed for those patients who are admitted in the hospitals. It also provides additional details like their admission and discharge dates and admit d diagnosis code.
2) Outpatient Data: This data provides details about the claims filed for those patients who visit hospitals and not admitted in it.
3) Beneficiary Data: This data contains beneficiary KYC details like health conditions,regioregion they belong to etc.
Description of the variables: There are total 55 variables with more than 7,00,000 observations. Variables their Unique value count and Number of Missing observation in that variable are shown in table below.

Variable	Unique	Missing Obs	Type
State	52	0	Nominal
RenalDiseaseIndicator	2	0	Nominal
Race	4	0	Nominal
Provider	6763	0	Nominal
PotentialFraud	3	135392	Nominal
OtherPhysician	57493	445235	Nominal
OperatingPhysician	43654	551963	Nominal
OPAnnualReimbursementAmt	2084	0	Count
OPAnnualDeductibleAmt	792	0	Count
BeneID	148072	0	Nominal
DOB	900	0	Interval
DOD	13	688432	Interval
Gender	2	0	Nominal
ChronicCond Alzheimer	2	0	Ordinal
ChronicCond Heartfailure	2	0	Ordinal
ChronicCond KidneyDisease	2	0	Ordinal
ChronicCond Cancer	2	0	Ordinal
ChronicCond ObstrPulmonary	2	0	Ordinal
ChronicCond Depression	2	0	Ordinal
ChronicCond Diabetes	2	0	Ordinal
ChronicCond IschemicHeart	2	0	Ordinal
AdmissionDt	400	643578	Interval
DischargeDt	366	643578	Interval
DiagnosisGroupCode	739	643578	Nominal

Table 1.1: Dataset Description

1.4 Definitions and mathematical preliminaries

1.4.1 Markov Process

A Markov model is a stochastic method for randomly changing systems that possess the Markov property. This means that, at any given time, the next state is only dependent on the current state and is independent of anything in the past.

Definition 1.4.1 (Markov Property:)**.** If the probabilities for the future values of a process are dependent only on the present, the process has the Markov property. Mathematically, for a process with time set {1, 2, 3, ...} and a discrete state space:

$$P\left(X_n = x_n \middle| X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, X_{n-3} = x_{n-3}, ..., X_1 = x_1\right) = P\left(X_n = x_n \middle| X_{n-1} = x_{n-1}\right)$$

whenever conditional probabilities on both sides are well defined.

1.4.2 Random Forest

This is an ensemble learning method that constructs a large number of decision trees on bootstrapped training samples, and then combines their predictions during inference.But when building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from full set of predictors. The predictions of individual trees are combined through majority voting for classification, or averaging for regression.

1.4.3 Support Vector Machines

These models are supervised learning models used for classification and regression. It is an extension of support vector classifier that results from enlarging the feature space in a specific way, using kernels. The goal of the SVM is to find the best boundary, called the maximum margin hyperplane, that separates the different classes

1.4.4 Decision Tree

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

1.4.5 Naive Bayes

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

1.4.6 Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

1.4.7 Neural Network

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

1.4.8 Binary Classification Neural Network

A binary classification neural network is a type of artificial neural network used for solving binary classification problems. In binary classification, the goal is to categorize inputs into one of two classes or categories.

1.4.9 Recurrent Neural Network

A recurrent neural network (RNN) is a type of artificial neural network that is specifically designed to handle sequential data by incorporating feedback connections.

1.4.10 Confusion Matrix:

The confusion matrix is a performance measurement tool in machine learning and classification problems. It is a matrix that provides a detailed breakdown of the predictions made by a classifier, comparing them to the actual ground truth values.

Actual	**Predicted**	
	Fraudulent	**Non-Fraudulent**
Fraudulent	True Positives (TP)	False Negatives (FN)
Non-Fraudulent	False Positives (FP)	True Negatives (TN)

True Positives (TP): The number of instances that are correctly predicted as positive (e.g., correctly predicted as fraudulent in the case of health insurance fraud detection).
False Positives (FP): The number of instances that are incorrectly predicted as positive (e.g., non-fraudulent instances predicted as fraudulent).
False Negatives (FN): The number of instances that are incorrectly predicted as negative (e.g., fraudulent instances predicted as non-fraudulent).
True Negatives (TN): The number of instances that are correctly predicted as negative (e.g., correctly predicted as non-fraudulent).

1.4.11 Sensitivity:

Sensitivity, also known as True Positive Rate (TPR) or Recall, measures the proportion of actual positive cases correctly identified by a classifier.
In our case sensitivity is measure of how well the model is detecting a actual fraud provider as potentially fraud.
Formula:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

1.4.12 Specificity:

Specificity measures the proportion of actual negative cases correctly identified by a classifier. In our case sensitivity is measure of how well the model is detecting a actual non-fraud provider as potentially non-fraud.
In health insurance fraud detection Specificity is more important than sensitivity. because our model should not show any actual non fraud provider as fraud. since punishment to non-criminal is more dangerous than not punishing a criminal.

Formula:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

1.4.13 F1 Score:

The F1 score is a measure of a classifier's accuracy that considers both precision and recall. It provides a single metric that balances the trade-off between precision and recall.

Formula:

$$\text{F1 Score} = \frac{2 \times (Precision \times Recall)}{Precision + Recall}$$

1.4.14 Precision:

Precision represents the proportion of true positive cases among the predicted positive cases. It measures the classifier's ability to correctly identify positive cases. In the context of health insurance fraud detection, precision is a performance metric that indicates the proportion of correctly identified fraudulent cases among all the predicted fraudulent cases.

Formula:

$$\text{Precision} = \frac{TP}{TP + FP}$$

1.4.15 Accuracy:

Accuracy measures the overall correctness of the classifier by calculating the proportion of correctly classified cases (both positive and negative) among all cases.

Formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

1.4.16 Receiver Operating Characteristic (ROC) Curve:

The ROC curve is a graphical representation of the performance of a binary classifier at various classification thresholds. It illustrates the trade-off between the True Positive Rate (Sensitivity) and the False Positive Rate (1 - Specificity) as the decision threshold is varied. The ROC curve is created by plotting the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) at different classification thresholds. The area under the ROC curve (AUC-ROC) is commonly used as a measure of the classifier's performance, where a higher AUC-ROC indicates better performance.

These formulas and definitions are commonly used in machine learning classification models to evaluate the performance of binary classifiers.

1.5 Literature Review

This literature review aims to explore the existing literature on health insurance fraud detection using machine learning and data analytics techniques, with a particular focus on the effectiveness and limitations of these methods. By synthesizing and analyzing the available literature, this review aims to provide insights into the current state of the art in health insurance fraud detection and identify promising directions for future research.

1.5.1 Markov model with machine learning integration for fraud detection in health insurance

By-Gupta RY, Mudigonda SS, Baruah PK and Kandala PK (2021) Markov model with machine learning integration for fraud detec tion in health insurance. arXiv preprint arXiv:2102.10978

Introduction:

The work proposes a model for fraud detection based on Markovian theory/Model. The dataset used for testing is on health insurance data.

Data Description:

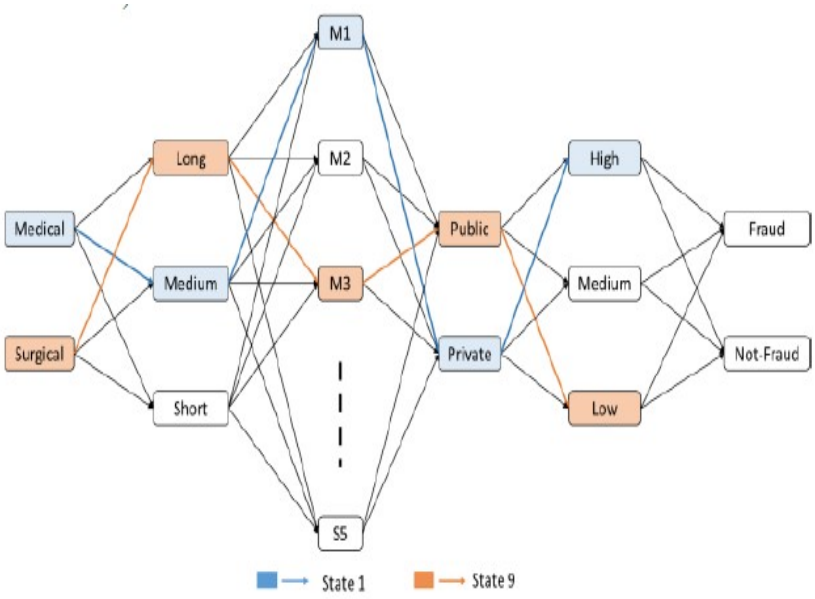
Claims and policy data for One policy year 20th August 2019 which contains 26 features.

Status	Count	% of total
Fraud	38,082	9.95%
Not-fraud	344,505	90.05%
Grand Total	382,587	100%

Methodology:

Each of the features in the dataset was categorized to into groups based on quantiles such that the groups has equal number of claims in it. They have considered five most significant features such as Benefit Type,No of Days Stayed,Primary Diagnosis Code,Hospital Type,Net Amt and after doing categorization of features,the total number of states thus formed was 1,188. Label the states to each claim are shown below.

Benefit Type	No of Days Stayed	Primary Diagnosis Code	Hospital Type	Net Amt	States
MEDICAL	medium	M1	Private	high	1
MEDICAL	medium	M1	Private	medium	2
MEDICAL	short	M1	Private	high	3
MEDICAL	medium	M1	Public	medium	4
MEDICAL	long	M1	Public	medium	5
MEDICAL	long	M1	Private	high	6
MEDICAL	medium	M1	Public	medium	4
MEDICAL	long	M3	Private	high	7
MEDICAL	medium	M1	Private	medium	2
MEDICAL	long	M1	Private	high	6
…	…	…	…	…	…
SURGICAL	long	S5	Private	high	8



The model is fitted to calculated the probability of claim being fraudulent or not. each of the states has the probability of it being fraudulent or not

For ex: Probability for state 1 being fraudulent is calculated as P(Claim=Fraud /State=1)

and P(Claim= Not Fraud /State=1) =1- P(Claim=Fraud /State=1)

Similarly one can find the respective probabilities for all the states To Improve the performance of Markov model with Gradient Boosting Method (GBM) three additional features are taken into consideration, viz. Medical Service Provider ID , Hospital District and Amount paid to Hospital. The data are divided into training and test dataset in the ratio 70:30. Gradient Boosting Method is used for building a fraud detection model.we will try for 80:20 and 85:15 dividing the data into training and test and observe the results. For the GBM modelling a total of 300 trees were used. The maximum depth of each tree (i.e., the highest level of variable interactions allowed) was kept as 5. Learning rate or step-size reduction was kept at 0.1. In addition to the usual fit, a 10-fold cross-validation was performed.

Results:

Measure	Markov Model	Markov Model with GBM
Sensitivity	0.5940	0.8486
Specificity	0.9795	0.9846
Precision	0.7638	0.8607
Accuracy	0.9407	0.9710
F1-Score	0.6683	0.8546

Table 1.2: Model Performance Metrics

The Markov based model gave the accuracy of 94.07 % with F1-score at 0.6683. F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.
However, the improved Markov model performed much better in comparison with the accuracy of 97.10 % and F1-score of 0.8546. It was observed that the improved Markov model gave much lower false positives compared to Markov model.

Sensitivity of the model 0.59 meaning that 59 % of the fraud cases were correctly identified. And a specificity of 0.97 means that 97 % of the non-fraud cases were correctly identified. The accuracy of the model is 0.94, meaning that 94 % of the labels were correctly identified by the model. The F1 score is a popular performance measure for classification and often preferred over, for example, accuracy when data is unbalanced(i.e. percentage of fraud and non fraud observations in the data are significantly different). Our dataset is also unbalanced since there are 38,082 fraudulent cases out of total 3,82,587 The Model proposed has area under the curve(AUC) 0.8424 and with GBM it improves to 0.9926 The proposed model shows a significant improvement when a boosting technique is used.

1.5.2 Analytical study on fraud detection in healthcare insurance claim data using machine learning classifers

-By A. J. Mary and S. P. Angelin Claret

Published Online: 30 November 2022

Introduction:

The prediction of health insurance fraud has become an active research topic. The anonymous activities prevailing in insurance claims have affected the financial growth of health insurance companies. In this research paper k-means clustering, Support Vector Machines (SVM), and Naive Bayes (NB) have been used to analyze healthcare provider fraud detection

Tools Used For The Analysis:

For the evaluation of the ML algorithum such as SVM,Naive bayes and K-mean clustering, Weka(Waikato Environment for Knowledge Analysis), a Statistical ML tool used and for data cleaning MS-EXCEL has been used.

Weka: It is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Data Description:

The dataset for further analysis has been taken from the Kaggle website with the name "Healthcare provider fraud detection analysis". The dataset contains three datasets: beneficiary, inpatient(IPD), and outpatient(OPD).The collected dataset is classified into three groups approved, denied, and, cancelled. This research paper considers approved and denied claims as non-fraud and fraud.

Methodology:

a)Data Cleaning- The records that hold missing values and The beneficiary dataset are eliminated in this study. So, that's why only 5011 records are under study out of 14780862 records. The finalized attributes/variables considered for the study are: Claim ID; Claim start date; Claim end date; Provider ID; Attending physician; Claim amount; Diagnosis code 1 to 10 and the status.

b)Performance Of Classifiers-

In the dataset claim with approved status is labeled as 'fraud'.

The results for the classifiers,namely, SVM,NB, and K-means under 10-cross validation and 66% split are mentioned below in table-

Classification Model	Vallidation	Precision	Recall	specificity	Accuracy
k-mean clustering	10- folds	91.3 %	92.4%	92.1%	92.3%
	66%split	90%	92.7%	91.8%	92.2%
Naive Bayes Class	10- folds	90%	97.7%	92.2%	94.6%
	66% split	92%	98%	97.8%	95.6%
SVM	10- folds	98.2%	96.9%	98.4%	96.8%
	66% split	97.6%	97.4%	93.8%	96.8%

Table 1.3: Result of The Classification Model

Performance of The Proposed Classifiers-

The performances of the three classifiers are evaluated by estimating the false positive rate.It is the proportion of the false positive samples to the total positive samples. It is otherwise termed as 'missed alarm rate'.

False Positive Rate of all models are less than 5% which proves the efficiency of all classifiers on insurance claim datasets.But SVM provides us with the FPR of 1.3% which means the proportion of incorrect predictions is very low as compared to other classifires ,also, sensitivity & specificity are high for SVM hence, SVM works better than other two classifiers.

Results:

Classification Model	Validation	False Positive Rate
k-mean clustering	10- folds	4.13%
	66% split	4.13%
Naive bayes classifier	10- folds	4.3%
	66% split	3.4%
SVM	10- folds	1.3%
	66% split	1.11%

Table 1.4: False Positive Rate analysis of proposed classifiers

The obtained results display that all three classifiers have given considerable good results. But, the accuracy given by the SVM is the highest. It is recommended that in the future, multi-criteria-based decision support systems will be explored.

1.5.3 Effective Fraud Detection in Healthcare Domain using Popular Classification Modelling Techniques

-By Sheffali Suri, Deepa V Jose.

International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 ,Issue-11, (September 2019)

Introduction:

In Healthcare,increasing healthcare costs along with the hike in fraud cases have made it difficult for people to approach these services when required. To avoid such situations, we must understand and identify such illegal acts and prepare our systems to combat such cases. Thus, there is a need to have a powerful mechanism to detect and avoid fraudulent activities.

In this paper they have employed Logistic regression and Random forest modelling approaches. Also, these approaches are evaluated and compared to come up with the most optimal idea for detection of maximum frauds from real-time datasets.

Methodology:

Here in this section, details related to datasets, existing algorithms and the performance metrics have been discussed.

a)Data Source-The datasets used here has been taken from Kaggle ,categorized as Inpatient claims, Outpatient claims and Beneficiary details of each provider. The data source also contains datasets named Train and Test data set stating whether the claim filled is fraud or not.

b)Data Cleaning-Replacing missing value by using central tendency statistics of each of the columns or by relevant constant, the next step involves feature selection.Age is picked out as an important feature as the frequency of fraud claims being filed is high for age ranging from 30-70 years when compared to 1-30 and 70+.Based on the EDA, 26 features are selected from the test dataset for further modeling and evaluation.

c)Model Construction-

1)Logistic Regression: There exist two classes i.e. Binomial logistic regression which can be coded as 1 (yes, fraud claim) and 0(no, non-fraud claim). The dependent variable is classified under these two classes on the basis of independent variables.

2)Random Forest: It is an ML algorithm which involves a large number of decision trees operating together.The class with maximum majority becomes the final predicted class for the data point.

c)Model Evaluation-Here, Confusion matrix performance is used for measuring the performance of the models.

Based on Confusion Matrix, Accuracy, Sensitivity ,Specificity are mea-

Actual	**Predcited Non-fraud claim**	**Predicted Fraud Claim**
Non Fraud Claim	True Negative(TN)	False Positive(FP)
Fraud Claim	False Negative(FN)	True Positive(TP)

Table 1.5: Confusion Matrix

sured for evaluation.

Results:

Evaluation	**Logistic regression**	**Random Forest**
Accuracy(train)	0.92	0.88
Accuracy	0.91	0.87
Sensitivity	0.67	0.81
Specificity	0.93	0.87
Kappa Value	0.54	0.47
AUC	0.80	0.84
F1 score(train)	0.647	0.60
F1 score(validation)	0.591	0.54

Table 1.6: Comparative Analysis between Logistic Regression and Random Forest

From this table, It is quite evident to consider Logistic Regression as a better classification model in fraud detection when compared to Random Forest. The implementation of Logistic Regression points towards the linearity between variables where as Random forest is able to handle large sets of High dimensional data.

Future Work:

The central idea of these approaches is to come up with suspicious data from big data sets.For future work, the same approaches can be tested on data from various other sources along with different sampling techniques. Also, an approach towards the Development of Self-evolving fraud detection methods can be proposed as well.

1.6 Chapterwise summary

- In chapter 1 , we have explained about out problem which we are working on , dataset description, mathematical definitions, literature review of research papers and chapterwise summary of each chapter;
- For chapter 2 , we have done explorartory data analysis(EDA) and there interpretations;
- In chapter 3,we have explained about data cleaning procedure,feature creation and selection plus model fitting by using various approaches;
- In chapter 4,Conclusions are made for the overall business problem;
- In chapter 5,we have discussed about scope and limitations of our study.

■

Chapter 2

Exploratory Data Analysis

2.1 Distribution of the class label

We are plotting the below plot to check the distribution of the class label.

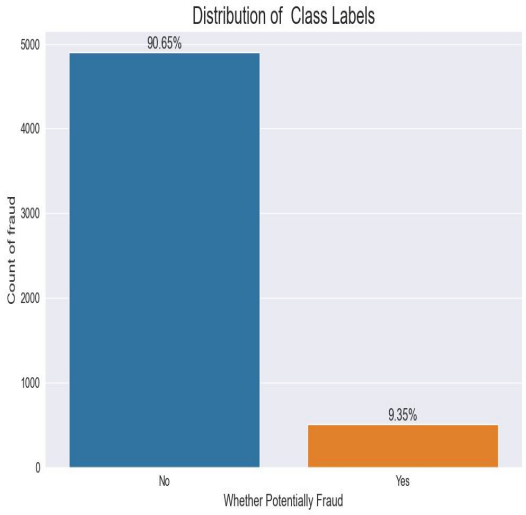


Figure 2.1: **Plot of the class label**

Observation: This is an highly imbalanced dataset. There are 10% fraudulent providers and 90% non-fraudulent providers.

2.2 Analysis of Beneficiary data

Distribution of Gender in Beneficiary Data

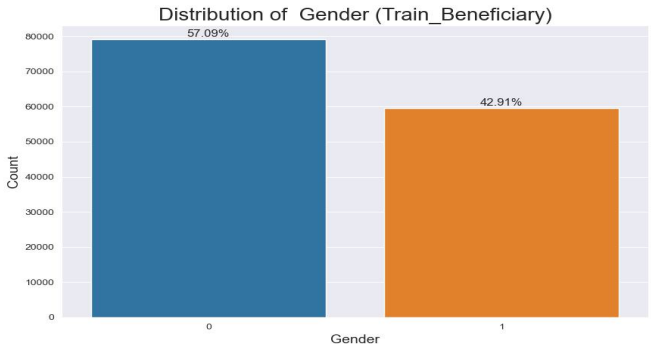


Figure 2.2: **Plot of Gender**

Observation: The ratio of genders in beneficiary data is Gender 0 : Gender 1 = 57% : 43% from this we can conclude that data is not gender biased.

Plotting top 20 States in terms of beneficiary count

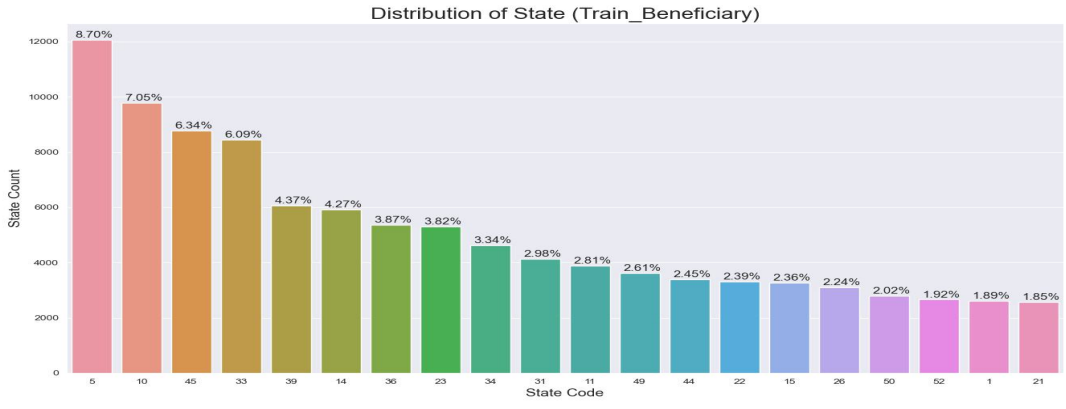


Figure 2.3: **Top 20 States**

Observation: States with code 5,10,45,33 and 39 are the top 5 states & 8.7% of the beneficiaries belongs to state 5.

Plotting top 20 Countries in terms of beneficiary count

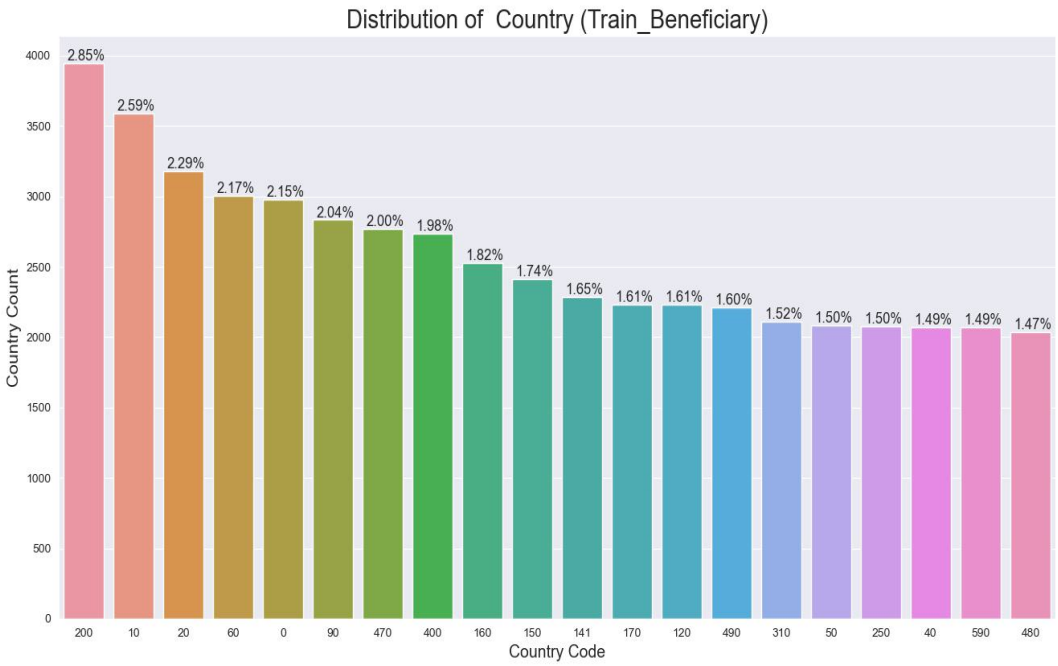


Figure 2.4: **Top 20 Countries**

Observation:

1)Countries with code 200, 10, 20, 60 and 0 are the top 5 countries.

2)2.85% of the beneficiaries belongs to country with code 200.

Plotting countplot of Race in the beneficiary data:

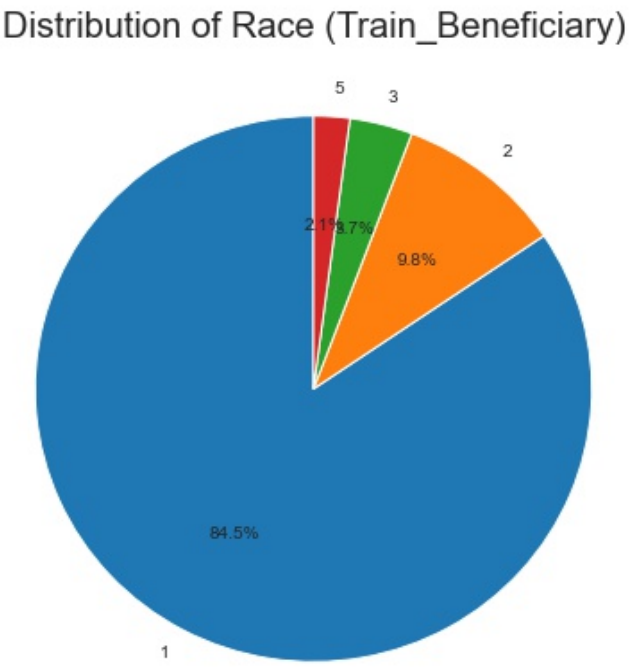


Figure 2.5: **Countplot of Race**

Observation:

1)Race 1 is the most in terms of beneficiary count.

2)85% beneficiaries belongs to race 1.

3)There is no race 4 in the dataset.

Plot the distribution of the Patient Risk Score in the beneficiary data to check overall health conditions of all the beneficiaries.

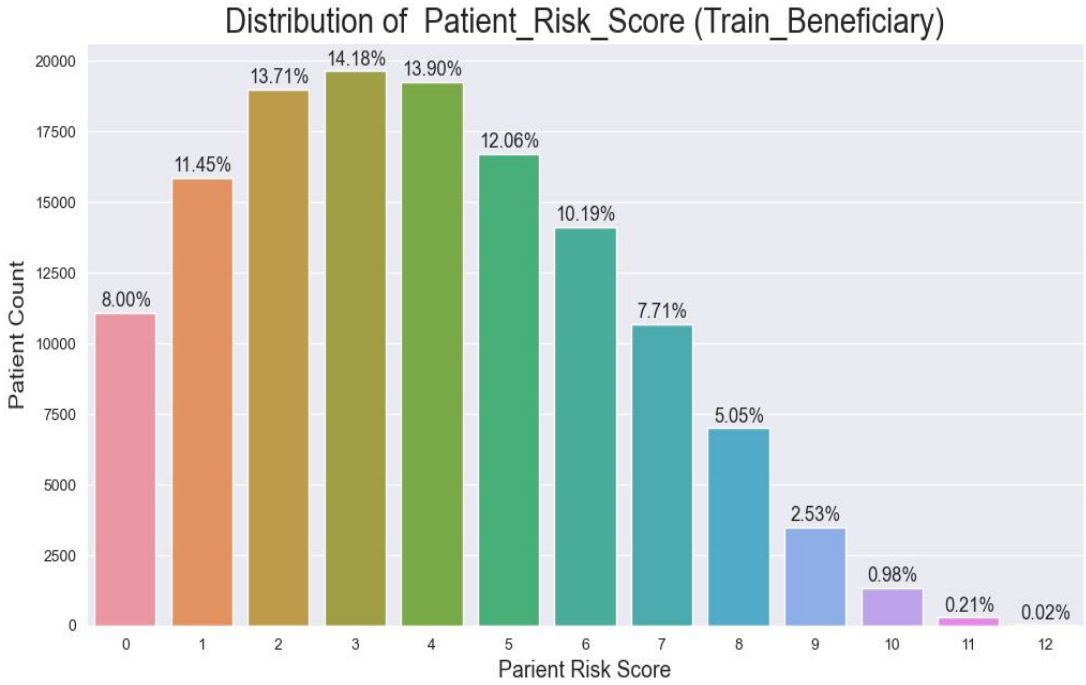


Figure 2.6: **Plot of the Patient Risk Score**

Observation:

1)The distribution of patient risk score is right tailed.The majority of the patients count is located towards the left side of the graph.The mean or average value of the data tends to be larger than the median value.

2)Most of the patients with risk score 2, 3, 4, 5.

3)Very few patients are there with risk score 9, 10, 11, 12.

Plot of Distribution of IP Annual Reimbursement Amount:

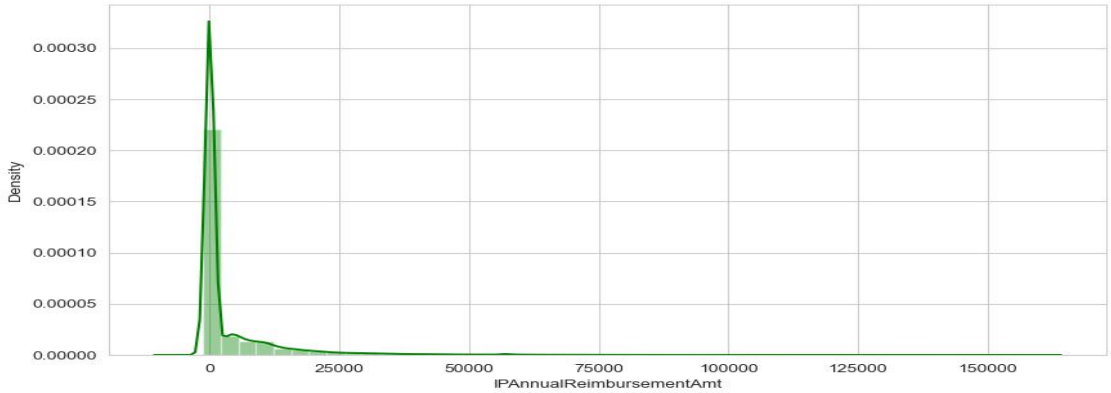


Figure 2.7: **Plot of IP Annual Reimbursement Amount**

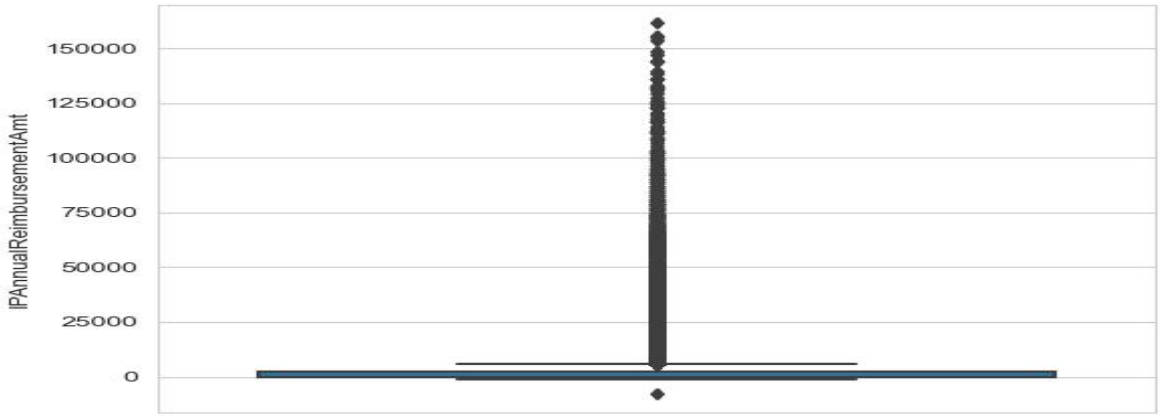


Figure 2.8: **Boxplot of IP Annual Reimbursement Amount**

Observation:

1)25th and 50th percentile of annual reimbursement amount is zero.

2)75th of annual reimbursement amount is 2800.

3)There may be some outliers as 100th percentile is 161410.

4)Total annual reimbursement amount is 507162970.

Distribution of Age in Beneficiary Data:

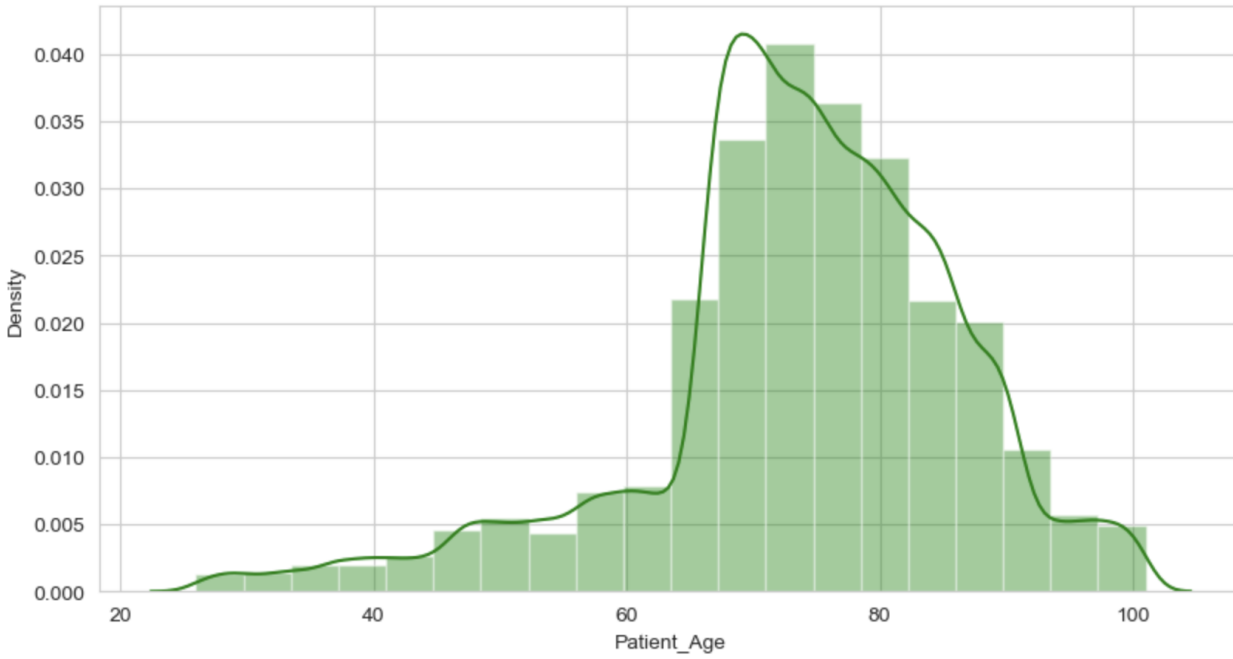


Figure 2.9: **Countplot of Age**

Observation:

1)There are very less number of patients in the age group 20-40

2)Most of the patients are with age group 60-90

3)Number of patients are less in the age group 90-100.

2.3 Analysis of Inpatient Data

Distribution of Attending Physician in Inpatient Data

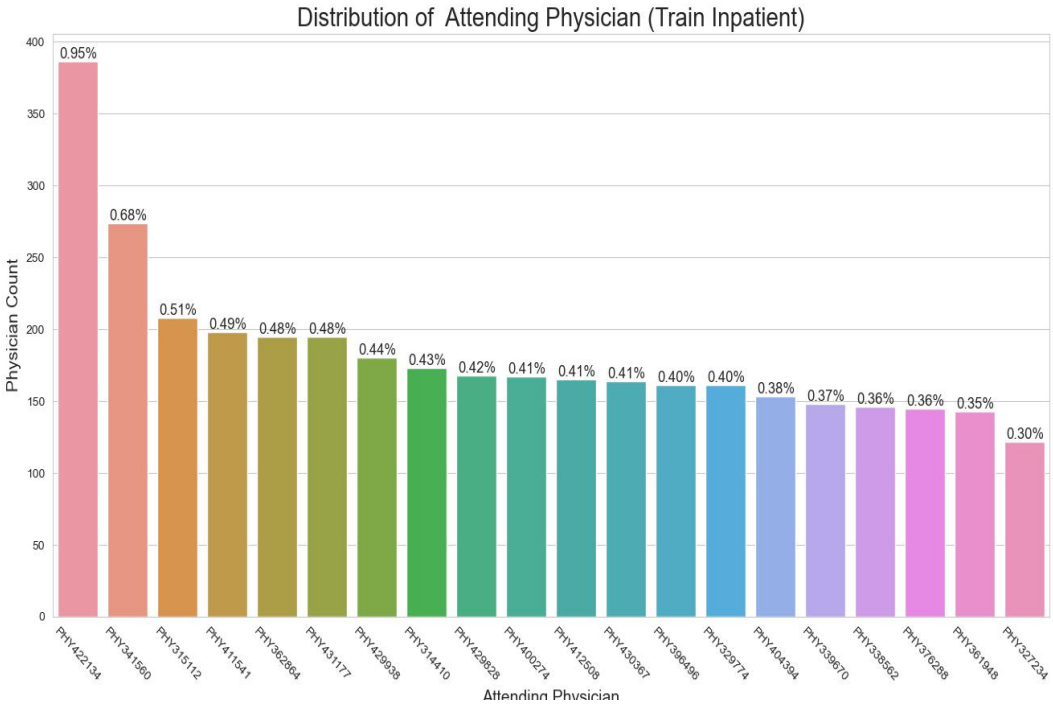


Figure 2.10: **Plot of Attending Physician**

Observation:

1)PHY422134, PHY341560, PHY315112, PHY411541, PHY431177 are the top 5 attending physicians in terms of number of patients visit.

2)PHY422134 treated 1% of the total patients.

Distribution of starting claim year in Inpatient data

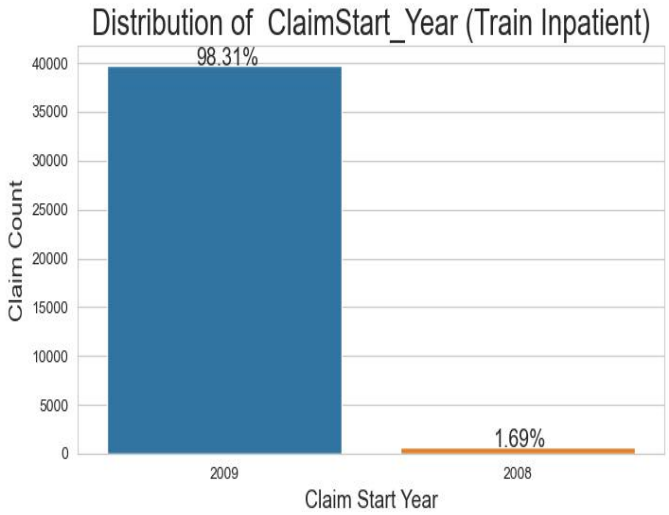


Figure 2.11: **Distribution of starting claim year**

Observation: For 98.3% of the inpatients claim started in 2009.Only for 1.7% inpatients claim started in 2008

Monthwise Distribution of starting claim year in Inpatient data

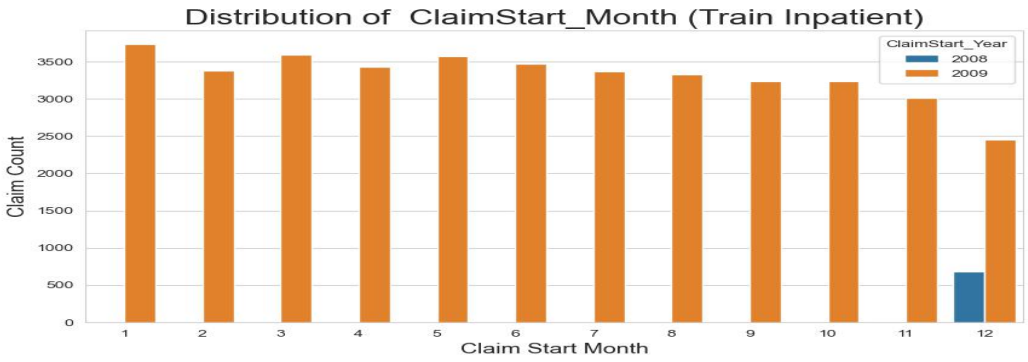


Figure 2.12: **Mothwise distribution of starting claim year**

Monthwise Distribution of starting claim year in Inpatient data

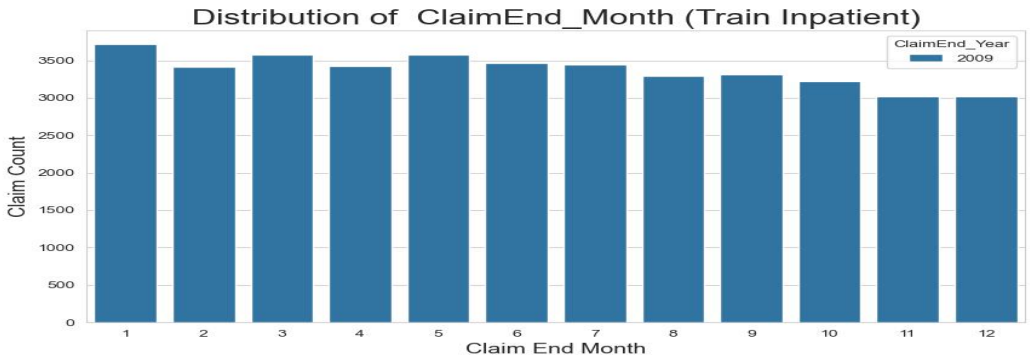


Figure 2.13: **Distribution of ending claim month**

Observation:

1)Claims are almost equally distributed (unifromly distributed)accross all the months for the year 2009.Claims are observed only in December for the year 2008. That means the data collected from Dec2008 to Dec2009.

2)All the claims ended in 2009, distributed accross all the months.

Distribution for Insurance Claim Amount Reimbursed

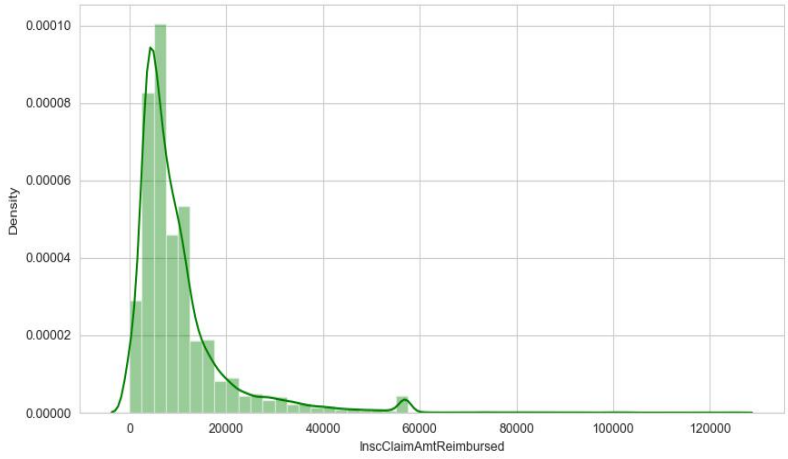


Figure 2.14: **Plot of insurance claim amount reimbursed**

Observation:

Total Insurance Claim Amount Reimbursed for inpatient is 40474.For very few claims Insurance Claim Amount Reimbursed are very high indicating to suspicious activities.

Similarly we can do the analysis of Outpatient data.Now, we'll be moving forward to draw insights from our final dataset that we have merged from four datasets.

2.4 Analysis of Final dataset

Distribution for Inpatient Outpatient in Final dataset

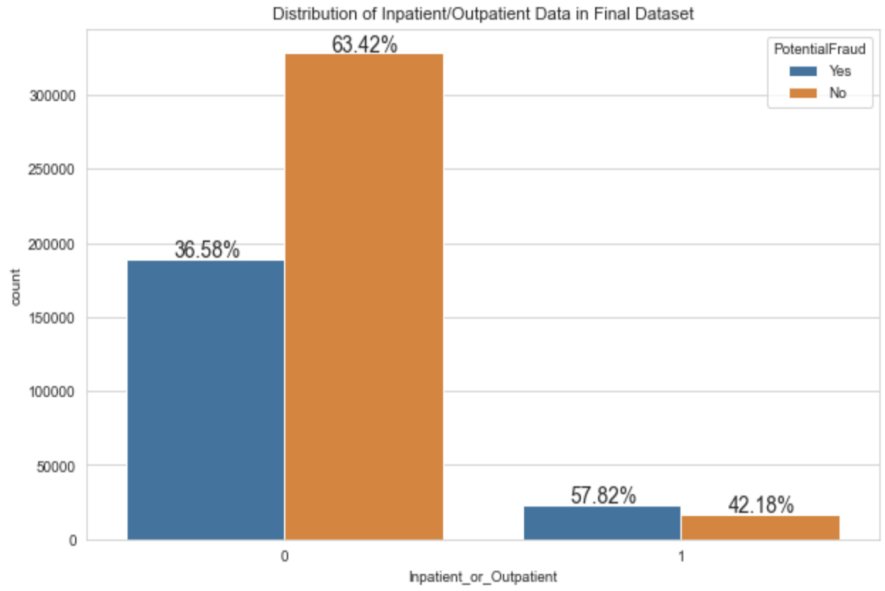


Figure 2.15: **Plot for Inpatient Outpatient in Final dataset**

Observation:

The number of claims are less for inpatient data compared to outpatient data. Even though the claims are less in inpatient data, percentage of fraudulent activity is more in inpatient data(57.8%) whereas it is 36.5% in outpatient

data. This is beacuse per claim reimbursement amount for inpatient is much higher(35 times calculated earlier) than the per claim reimbursement amount of outpatient.

A simple reasoning to these could be the fact that outpatients are not admitted, hence they didn't committed frauds.On the other hand inpatients are admitted hence they might want to retrieve more money.

Distribution for Race in Final dataset

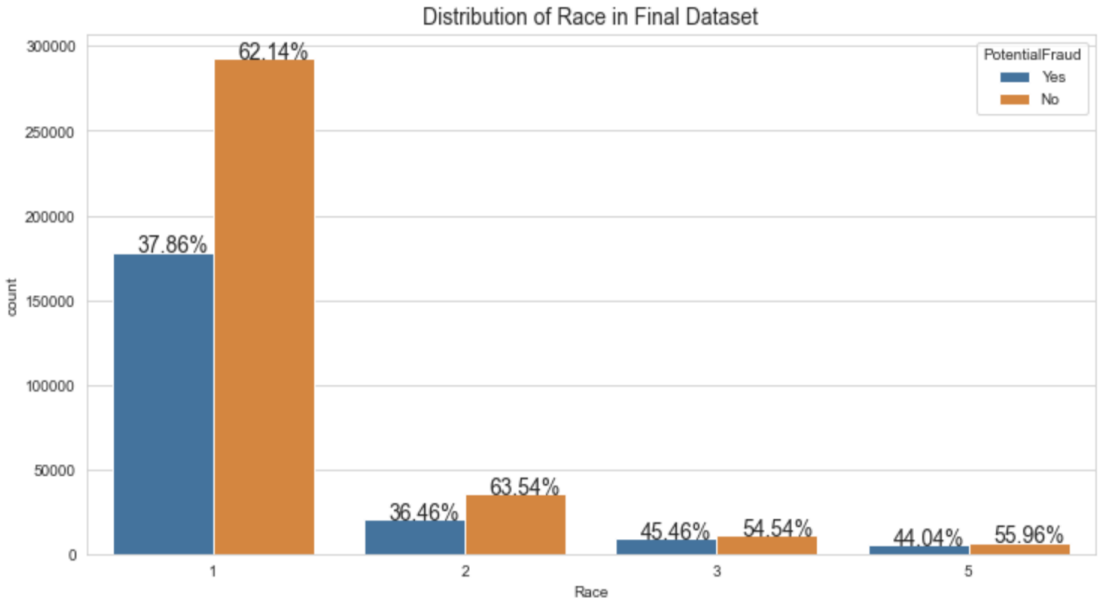


Figure 2.16: **Plot of Race in Final dataset**

Observation:

1)Total number of transactions are more for Race 1, 37.8% are fraudulent out of them.

2)The ratio of fraudulent transaction is most for Race 3 (45.5%)

3)So, race is an important feature in fraud detection.

Histogram of Insurance Claim Amount Reimbursed in final Dataset

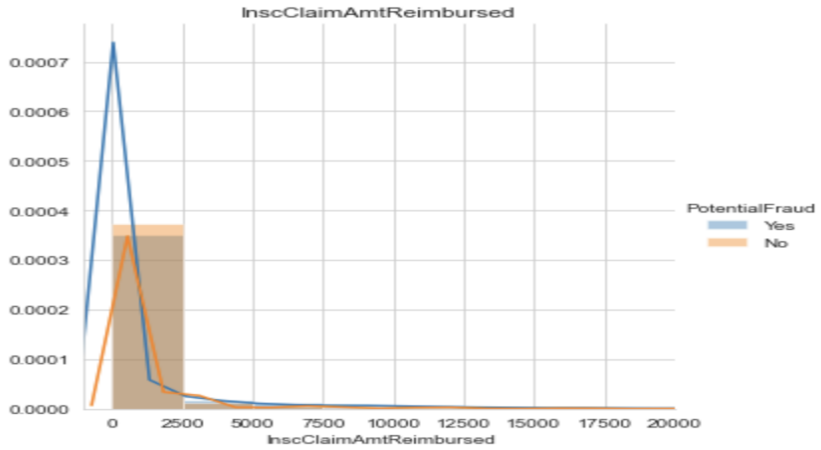


Figure 2.17: **Histogram of Insurance Claim Amount Reimbursed**

Observation: From the histogram we can observe that when the claim amount is less, the number of fraud claims are much higher compared to legitimate claims.

Boxplot of Insurance Claim Amount Reimbursed in final Dataset

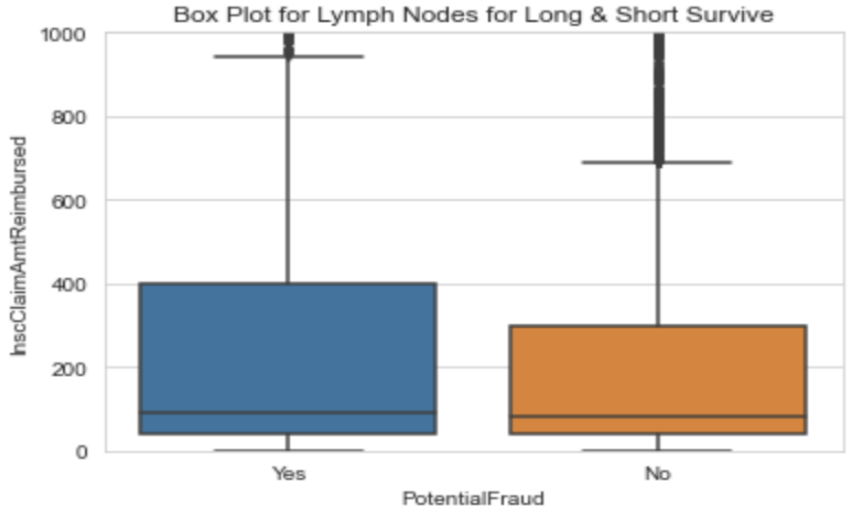


Figure 2.18: **Boxplot of Insurance Claim Amount Reimbursed**

Observation: 25th,50th percentiles are very less for claim amount reimbursed. 75th percentile for fraud claims is higher than non-fraud claims.

Patient Age vs Claim Period

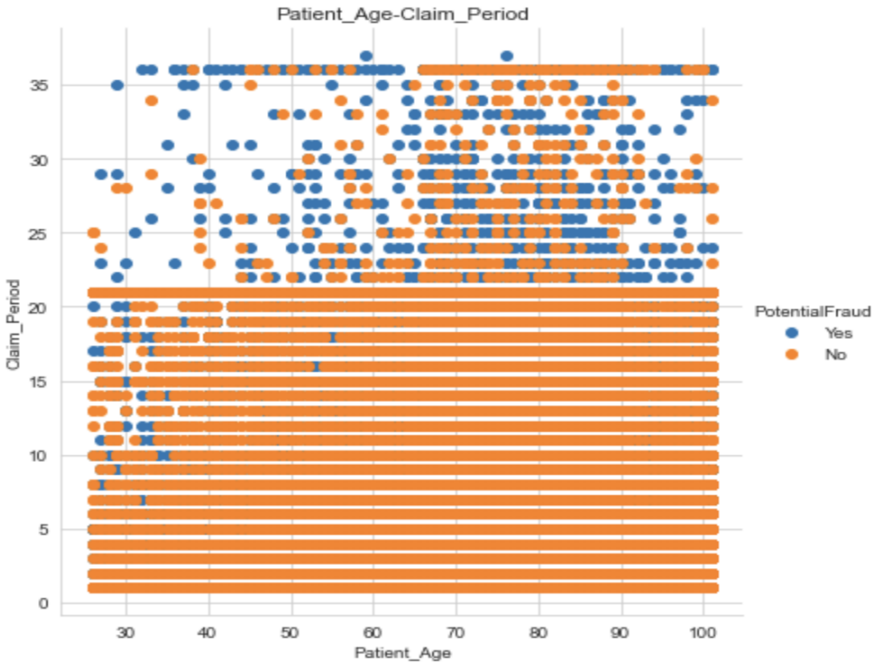


Figure 2.19: **Scatter Plot Patient Age vs Claim Period**

Observation:

From the scatter plot we can see that when patient's age <60 years and claim period more than 20 days, the probability of the transaction is fraudulent is high.

Patient Age vs Insurance claim amount reimbursed

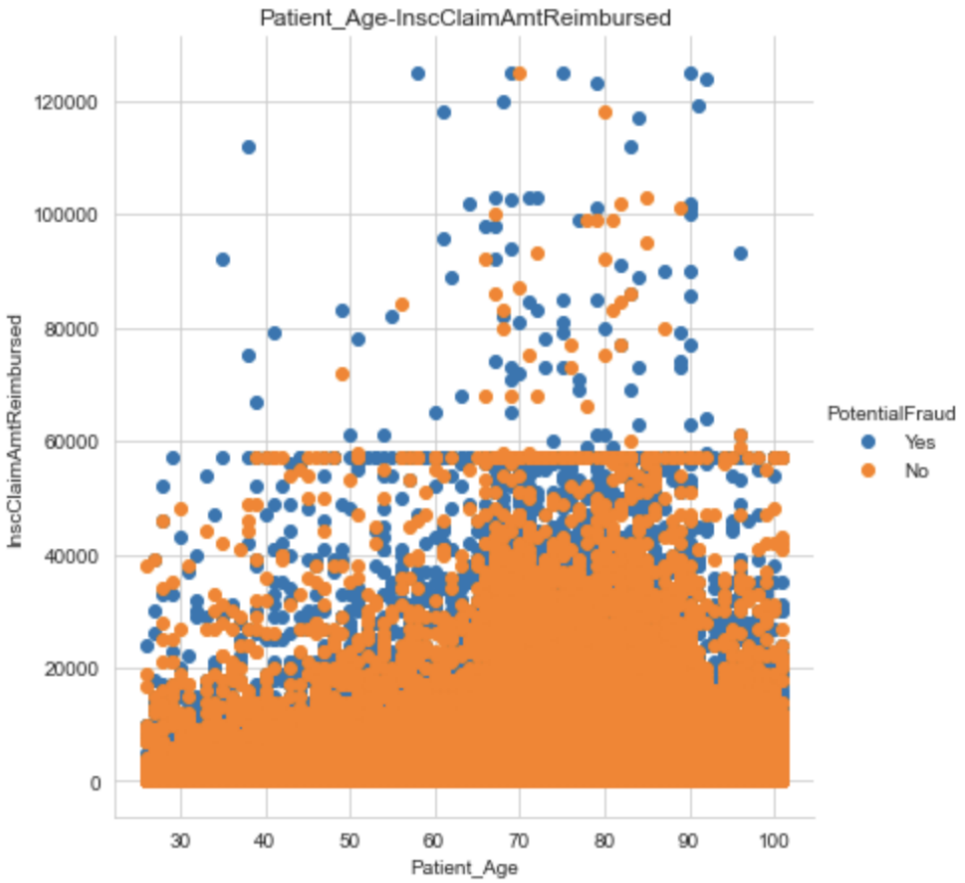


Figure 2.20: **Scatter Plot: Patient Age vs Insurance claim amount reimbursed**

Observation:

From the Scatter Plot of Patient Age vs InscClaimAmtReimbursed we can observe that if patient's age<60 years and claim amount>60000 it tends to be a fraudulent transaction. If the patient's age>88 yrs and claim amount>60000 the probability to be fraudulent is high.

Distribution of Total claims filed by attending physicians

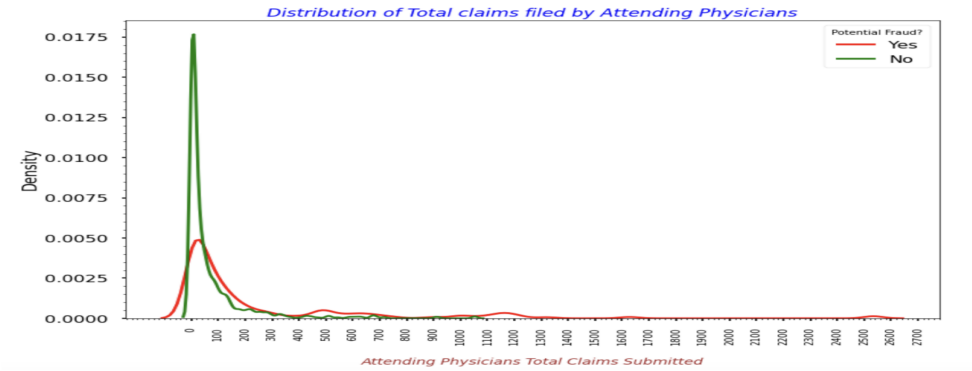


Figure 2.21: **Plot of Total claims filed by attending physicians**

Observation:

The above KDE suggests that the newly added feature Attending Physician-total-claims may be useful in segregating the potentially fraud and non-fraudulent cases.

For example, we can say that if total claims filed by a Attending Physician is greater than 500 then chances of being fraudulent are high.

Distribution of Providers interaction with attending physicians

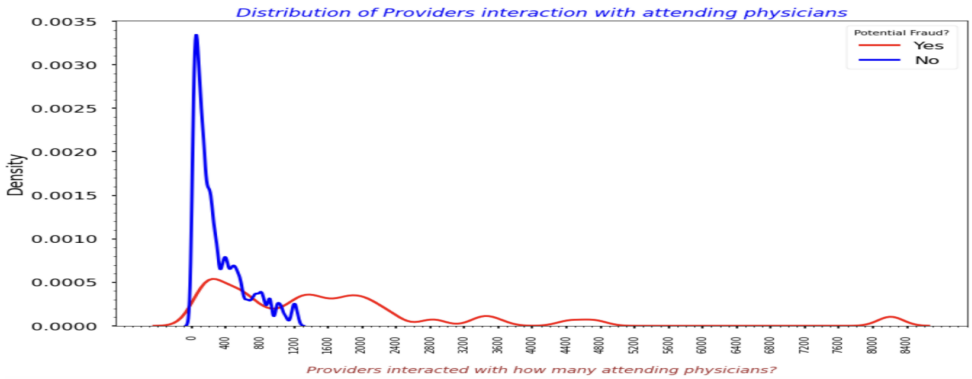


Figure 2.22: **Plot of Providers interaction with attending physicians**

Observation: The above KDE plot are quite interesting as we can see that if Provider-Attending Phy is high then chances of fraud is quite high

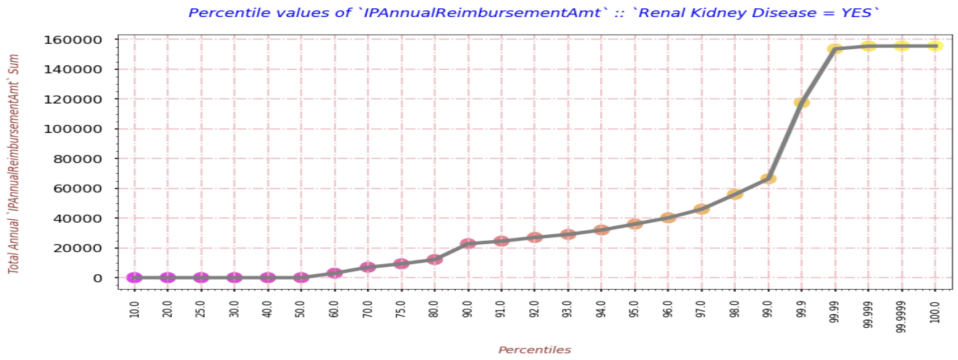


Figure 2.23: **Percentile values of IPAnnualreimbursement amount to Renal kidney disease indicator**

Observation: The above graph shows us that some of the reimbursements paid by the PAYER are very high as compared to the rest of the records. This can be a potential sign of fraudulent cases because generally the criminals file some forge cases with exponentially high amounts.

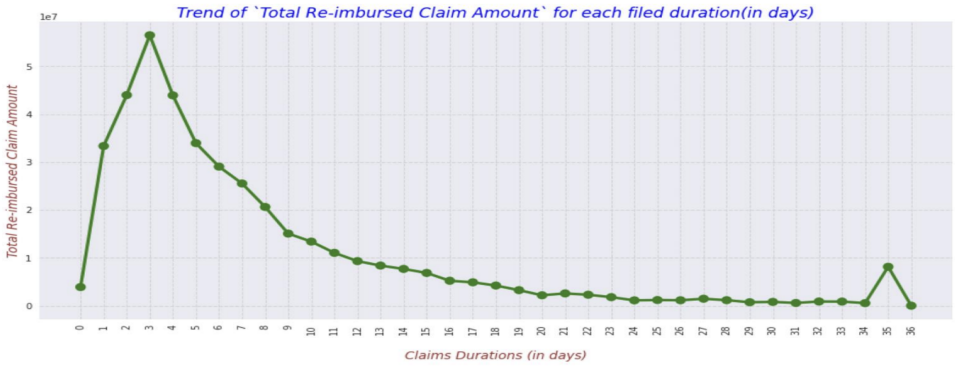


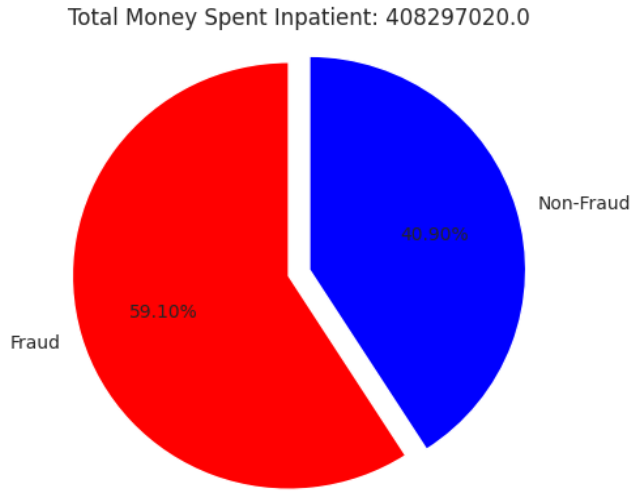
Figure 2.24: **Trend of Total Re-imbursed Amount vs claim period**

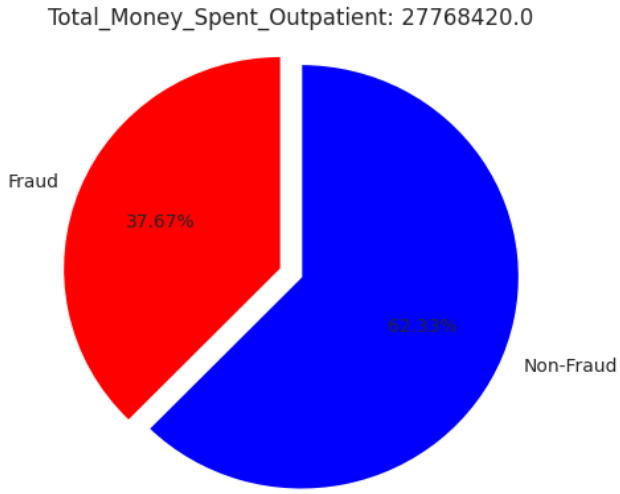
Observation:

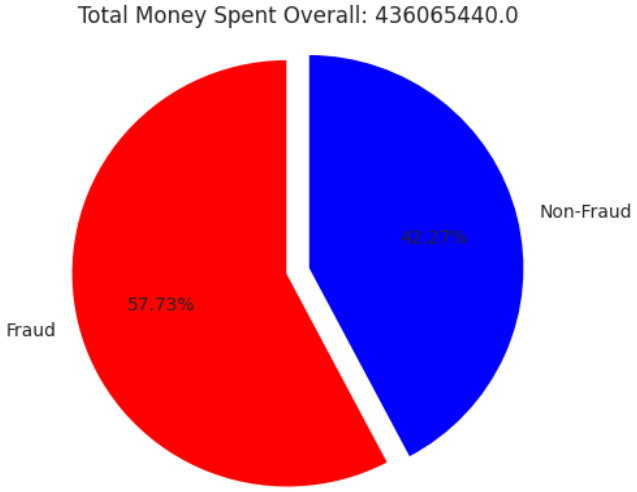
1)The above graph tells us that the Total Re-imbursed Amount is the highest for 3 days claims

2)And, for claims with durations from 12 to 34 the total re-imbursed amount is very less, however, for 35 days duration ,we can witness a clear spike that can be a potential sign of fraudulent.

2.5 Analysis Based on Insurance Amount Reimbursed







	Inpatient	Outpatient	Total
Total Money Spent	408297020.0	27768420.0	436065440.0
Average Reimbursed Amount per Patient	10088	282	-
Total Money Spent in Fraud	241288510.0	10460410.0	251748920.0
Percentage of Money Spent in Fraud	59.10%	37.67%	57.73%

Table 2.1: Healthcare Fraud Statistics

Chapter 3

Analysis of Data

3.1 Data Cleaning

Data cleaning, also known as data scrubbing, refers to the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in a dataset. It is an essential step in data preparation and analysis, as it ensures that the data is accurate, reliable, and suitable for further analysis or modeling.

3.1.1 Merging of Datasets

We have 4 different datasets, which are interconnected by foreign keys. We need to merge them using the foreign keys to get a overall dataset. Below is a brief overview of the dataset.

- Merge Inpatient and Outpatient data based on common columns.
- Merge beneficiary details with inpatient and outpatient data on BeneID.
- Merge provider details with previously merged data on ProviderID.

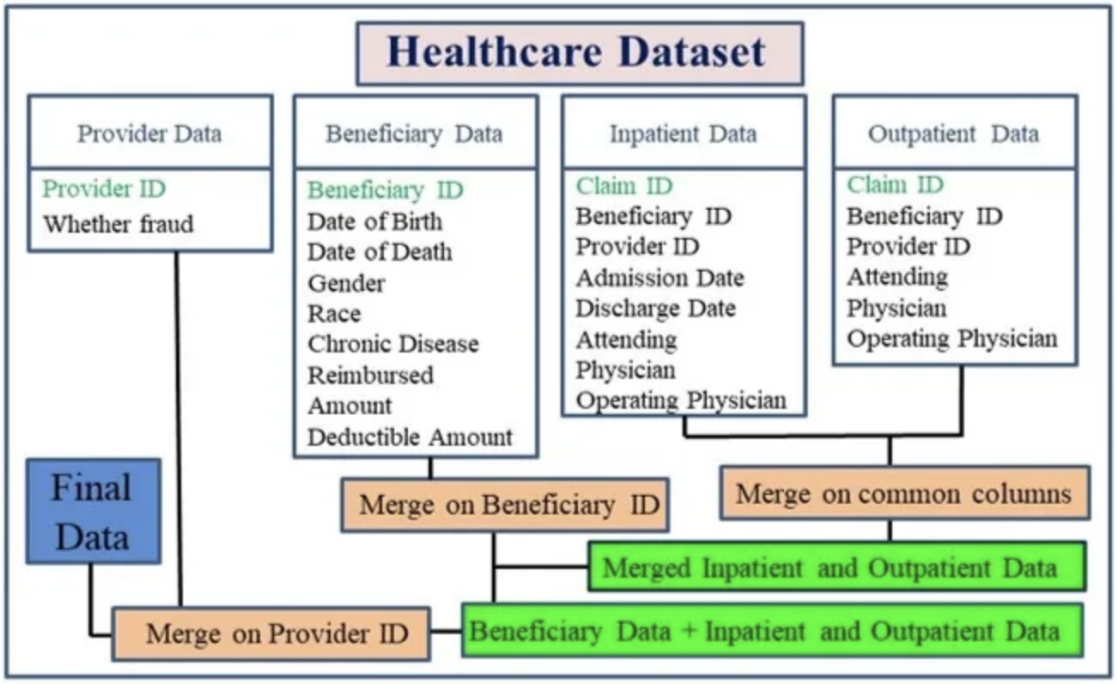


Figure 3.1: **Final Dataset**

3.1.2 Missing Value Imputation

Missing value imputation is a crucial step in data cleaning for healthcare fraud detection, as missing data can significantly affect the accuracy and effectiveness of the analysis. Here are some commonly used techniques for handling missing values in the context of healthcare fraud detection:

Complete Case Analysis: In this approach, records with missing values are simply removed from the dataset. While this method is straightforward, it may result in a loss of information if the missing data is not randomly distributed. It is suitable when the amount of missing data is minimal.

Mean/Median/Mode Imputation: This technique involves replacing missing values with the mean, median, or mode of the available data for that variable. It is a simple approach that works well for numerical or categorical

variables . However, it does not capture the uncertainty associated with the imputed values.

First, we have done missing value imputation on the basis of variables by using appropriate techniques such as imputing with mode if the variable is categorical , mean if it is continuous,median if there is presence of outliers etc.

3.1.3 Feature Creation

Creating new features, also known as feature engineering, is a crucial step in healthcare fraud detection data analysis. By deriving new features from existing ones, we tend to capture additional information and patterns that may improve the performance of fraud detection models.
We are using the **'groupby'** operation and aggregating with mean or other aggregation functions. It allows us to summarize and extract information from groups or categories within final dataset.
1)**Provider-Level Features**: We have a dataset of healthcare claims with information on providers, patients, and costs,so ,we created features at the provider level by grouping the data by providers and calculating aggregate statistics such as Providers fill and submit the claim they are mainly associated with the fraudulent activity. So, we have group by the provider and take the mean of reimbursed, deducted, etc. If the average claim amount or claim period is high for a provider, this is suspicious.
2)**Beneficiary-Level Features**: Beneficiaries also associated with fraudulent activity. So, group by the data-frame by Beneficiary Id and take mean. If the average claim amount is high for a beneficiary then this is suspicious.
3)**Physician-Level Features**:Physicians are also associated with fraudulent activity. So, group by AttendingPhysician, OperatingPhysician, and OtherPhysician and take mean. High amounts for a physician are suspicious.

3.1.4 Removal of Insignificant Variables

Datasets may contain identifiers or metadata columns that are not directly related to the analysis. For example, unique identification numbers, timestamps, or other administrative information might not be relevant to fraud detection and hence, can be removed to reduce the computational burden and improve the efficiency of analysis.
we have removed the following columns.

```
remove_columns=['BeneID', 'ClaimID', 'ClaimStartDt','ClaimEndDt','AttendingPhysician','OperatingPhysician', 'OtherPhysician'
                'ClmDiagnosisCode_1','ClmDiagnosisCode_2', 'ClmDiagnosisCode_3', 'ClmDiagnosisCode_4','ClmDiagnosisCode_5',
                'ClmDiagnosisCode_6', 'ClmDiagnosisCode_7','ClmDiagnosisCode_8', 'ClmDiagnosisCode_9', 'ClmDiagnosisCode_10'
                'ClmProcedureCode_1', 'ClmProcedureCode_2', 'ClmProcedureCode_3','ClmProcedureCode_4', 'ClmProcedureCode_5',
                'ClmProcedureCode_6','ClmAdmitDiagnosisCode', 'AdmissionDt','ClaimStart_Year', 'ClaimStart_Year', 'ClaimStar
                'ClaimEnd_Year', 'ClaimEnd_Month', 'Admission_Year', 'Admission_Month', 'Discharge_Year', 'Discharge_Month',
                'DischargeDt', 'DiagnosisGroupCode','DOB', 'DOD','Birth_Year', 'Birth_Month','State', 'County']
```

And then convert the type of Gender and Race to categorical and do **one-hot encoding**.After that standardize the data using **StandardScaler**. Finally, dataset is ready for model fitting.

Here,response variable is 'Potential fraud' and independent variables are such as- 'Age','Per Provider insurance claim amount reimbursed' etc.

3.2 Model Fitting

Now our dataset is ready. We need to try different models and validate the performance of every model. Based on the performance of the validation data we need to pick the best one for deployment.
Below are the different approaches that we have followed in this project.

3.2.1 Approach 1:Supervised Machine Learning Models

- Split the data into Train and Validation data and make it in the four given ratios 80:20, 75:25, and 65:35 respectively to check which splitting criterion gives the best results.
- Use Logistic Regression, Random forest, Decision Tree, Support Vector Classifier, and Naive Bayes for all these 3 datasets. Pick the best model based on the performance score.

80:20 Results

Model	Accuracy	Sensitivity	Specificity	F1 score	Rank
Logistic Regression	0.855	0.7763	0.9032	0.8027	4
Random Forest	0.895	0.75	0.9838	0.8444	1
Decision Trees	0.83	0.7368	0.8870	0.7671	5
SVM	0.855	0.7631	0.9112	0.8	3
Naive Bayes	0.87	0.7236	0.9596	0.8088	2

Table 3.1: 80:20 results

These metrics suggest that the random forest model performs well in healthcare fraud detection. An accuracy of 0.895 indicates that the random forest model is able to correctly predict the outcome for 89.5% of the cases in the dataset. This means that it is accurate in identifying both fraud and non-fraud cases around 89.5% of the time.

A specificity of 0.9838 indicates that the random forest model has a high level of specificity. **It correctly identifies 98.38% of the non-fraud cases out of all the actual non-fraud cases in the dataset.** This means that

the model is very good at distinguishing non-fraud cases from the total set of negative cases.

Given a specificity of 0.9838, we can subtract it from 1 to find the false positive rate: FPR = 1 - specificity = 1 - 0.9838 = 0.0162

Therefore, the false positive rate is 0.0162 or 1.62% in this case. This means that out of all the non-fraud cases, the random forest model incorrectly classifies around 1.62% as fraud (false positives). The remaining 98.38% are correctly classified as non-fraud.

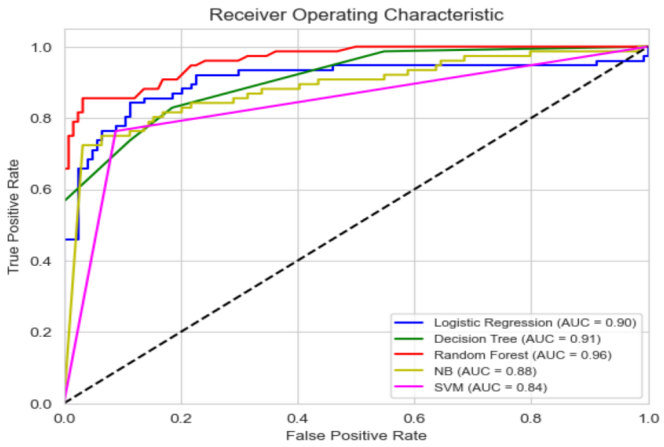


Figure 3.2: **ROC and AUC for 80:20 split**

Observation:

Random forest has the highest Area Under Curve(AUC)of 0.96 indicating that the model is highly effective in identifying and flagging potential fraudulent activities, making it a valuable tool for fraud detection and prevention.

75:25 Results

Model	Accuracy	Sensitivity	Specificity	F1 score	Rank
Logistic Regression	0.864	0.7789	0.9161	0.8131	3
Random Forest	0.892	0.7684	0.9677	0.8439	1
Decision Trees	0.84	0.7578	0.8903	0.7826	5
SVM	0.86	0.8	0.8967	0.8128	4
Naive Bayes	0.852	0.7052	0.9419	0.7836	2

Table 3.2: 75:25 results

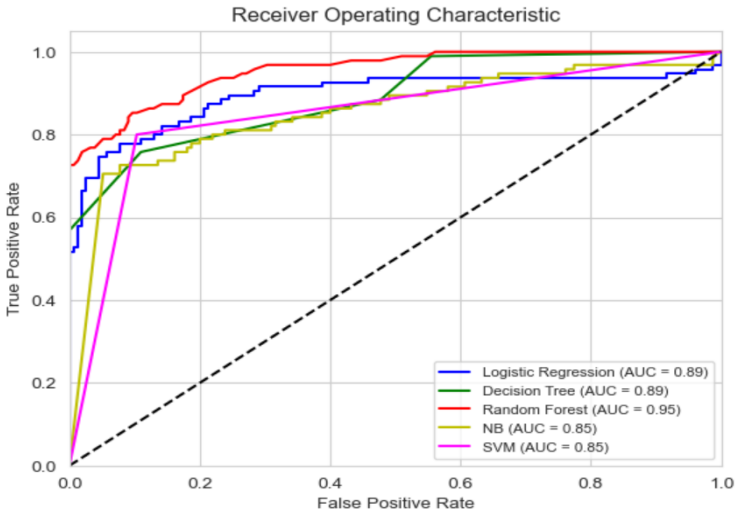


Figure 3.3: **ROC and AUC for 75:25 split**

Observation:

The results are similar to 80:20 split criterion with random forest as the best classifier.

65:35 Results

Model	Accuracy	Sensitivity	Specificity	F1 score	Rank
Logistic Regression	0.84	0.7368	0.9032	0.7777	5
Random Forest	0.8914	0.7443	0.9815	0.8389	1
Decision Trees	0.8428	0.7443	0.9032	0.7826	4
SVM	0.8514	0.7443	0.9170	0.792	3
Naive Bayes	0.8314	0.6466	0.9447	0.7445	2

Table 3.3: 65:35 results

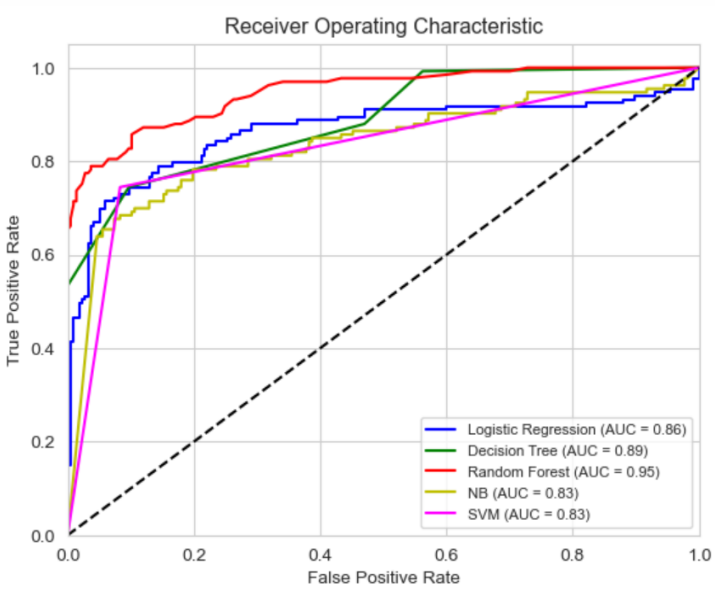


Figure 3.4: **ROC and AUC for 65:35 split**

Observation:

The results are similar to 80:20 split criterion with random forest as the best classifier.

Comparative Analysis

Sampling Ratio	Best Model	Accuracy	Sensitivity	Specificity	F1 Score
80:20	Random Forest	0.895	0.75	0.9838	0.8444
75:25	Random Forest	0.892	0.7684	0.9677	0.8439
65:35	Random Forest	0.8914	0.7443	0.9815	0.8389

Table 3.4: Selection of Best Model

Observation:

Since,**type 2 error** are considered to be more severe than type 1 error so,we will focus on minimizing Type 2 errors .

Type 2 Error (False Negative): The model fails to detect a fraudulent claim and incorrectly labels it as non-fraudulent.

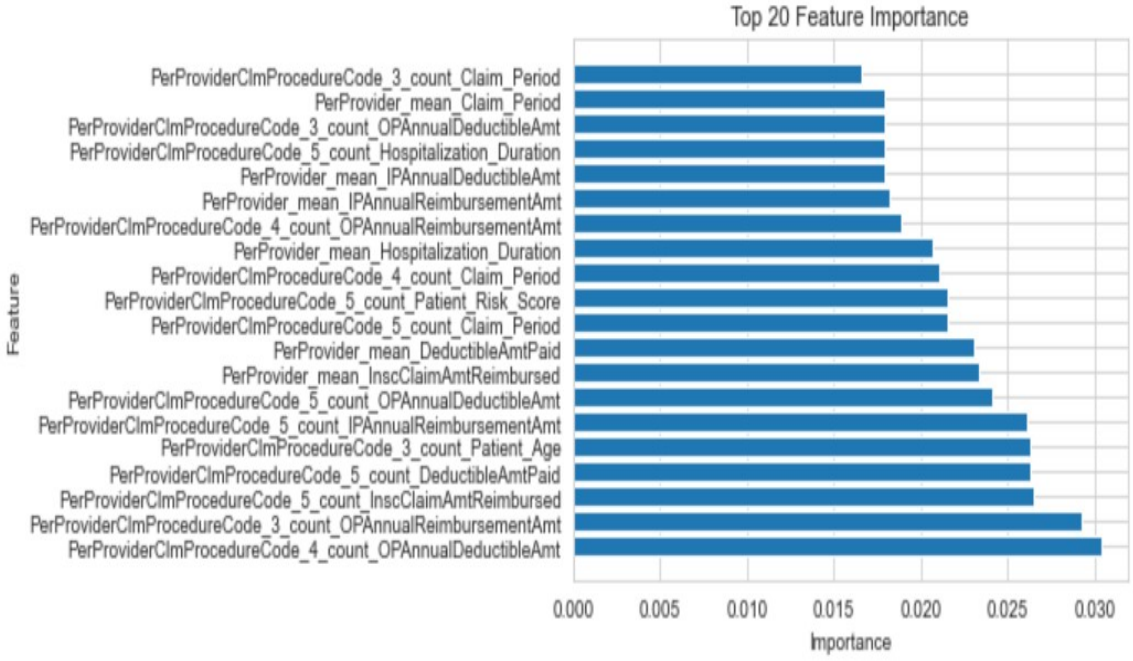
So,taking in view of above point we will consider the model with highest specificity and good accuracy.Here, Our objective is correctly identifying fraud and non-fraud claims and minimising type 2 errors.Therefore from above table random forest is working as the best model with accuracy of 89.5% and specificty of 98.38% for 80:20 split criterion.

3.2.2 Approach 2: LDA and QDA

Variable Selection

Variable selection is done using random forest as we have mixed type of variables,Lasso and Ridge regression are primarily used for regularizing and improving the predictive performance of models, rather than for feature selection in mixed-type of variables.

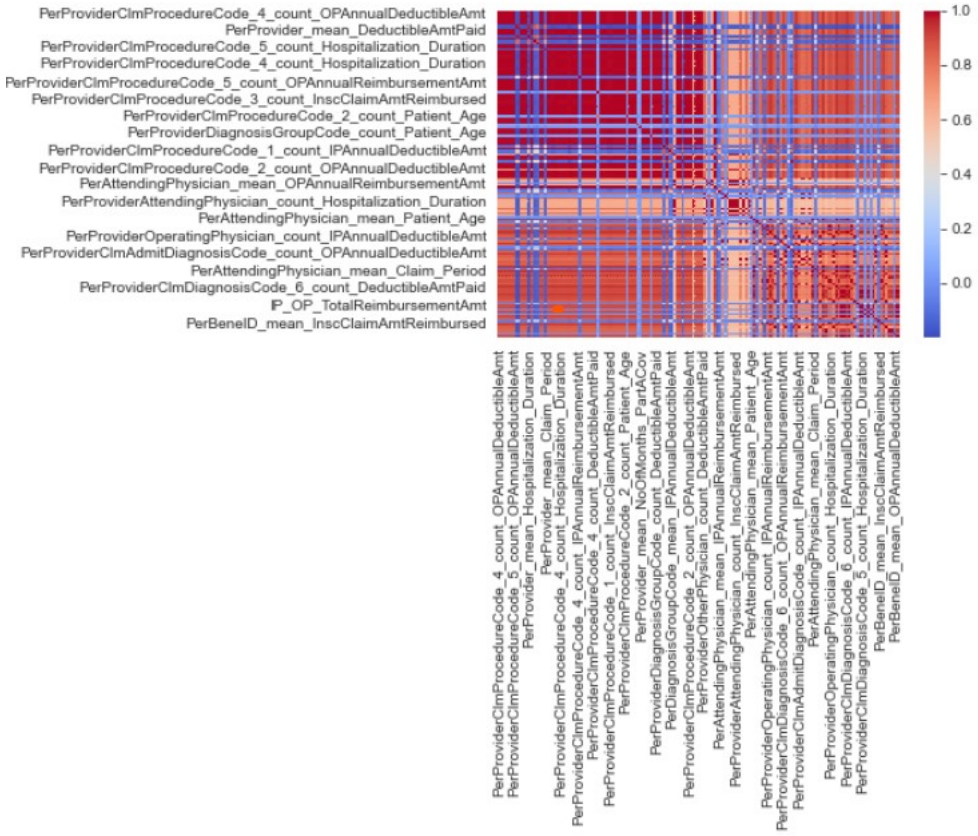
By using random forest model we have selected top 20 features by variable importance. the following plot shows variables in ascending order of their importance.We have chosen random forest for feature selection because it can deal with both categorical and numerical data and our dataset have both type of variables.



We have analysed the distribution of variables whose feature importance is greater than 0.001 we get 146 such variables. Now fit the model on 146 selected features. calculate the correlation matrix of selected features. we get 146 by 146 correlation matrix. Find the highly correlated features then remove

highly correlated features from the selected features create a new dataframe with only the remaining selected features.

Plot shows Heatmap of the previously calculated correlation matrix.



Use the StandardScaler preprocessing tool for data preprocessing and then split the data into train and test.

Linear Discriminant Analysis(LDA)

Linear Discriminant Analysis (LDA) is a statistical technique commonly used for dimensionality reduction and classification problems. While LDA is not typically used for health insurance fraud detection, it can be adapted and applied to identify patterns and anomalies in health insurance claims data that indicate potential fraudulent activities.

Methodology: 1) In the context of LDA, a set of features (independent variables) that are most informative for distinguishing between fraudulent and non-fraudulent claims need to be selected. To train a supervised classification model using LDA, a training dataset is required.

2) An LDA model can be trained using the labeled dataset. LDA aims to find a linear combination of features that maximizes the separation between the fraudulent and non-fraudulent classes. It achieves this by modeling the distribution of each class and estimating the class-specific means and covariance matrices.

Dimensionality Reduction: One of the key benefits of LDA is its ability to reduce the dimensionality of the feature space while preserving the discriminatory information between classes. LDA achieves this by projecting the high-dimensional data onto a lower-dimensional subspace that maximizes the separation between the classes. This reduction in dimensionality can help visualize the data and improve the efficiency and performance of subsequent classification tasks.

Model Evaluation and Validation: The performance of the LDA needs to be evaluated and validated. Techniques like 10-fold cross-validation can be employed to assess the model's generalization capability and avoid overfitting.

Cross-validation scores: 0.8125 , 0.805, 0.8125, 0.835, 0.80125, 0.80375, 0.82625, 0.82875, 0.83875, 0.8025

Mean cross-validation score: 0.81662

LDA demonstrates a relatively high accuracy of 0.8265, indicating that it correctly classifies 82.65% of the health insurance claims as either fraudulent or non-fraudulent. The precision of 0.9237 suggests that when the model predicts a claim as fraudulent, it is accurate 92.37% of the time. However, the recall (also known as sensitivity) of 0.5910 indicates that the model only identifies

59.10% of the actual fraudulent claims. The F1 score, which combines precision and recall, is 0.7208, indicating a balance between the two metrics. The high specificity of 0.9702 suggests that the model is effective at correctly classifying non-fraudulent claims.

Quadratic Discriminant Analysis(QDA)

Quadratic Discriminant Analysis (QDA) is a statistical technique used for dimensionality reduction and classification problems, similar to LDA. While QDA is not typically employed for health insurance fraud detection, it can be adapted and applied to identify patterns and anomalies in health insurance claims data that indicate potential fraudulent activities.

Methodology (QDA): A QDA model can be trained using the labeled dataset. QDA assumes that each class follows a multivariate normal distribution and estimates class-specific means and covariance matrices. It aims to find a quadratic decision boundary that best separates the fraudulent and non-fraudulent classes.

Model Evaluation and Validation: Techniques like 10 fold cross-validation can be employed to assess the model's generalization capability and prevent overfitting.

Cross-validation scores: 0.78, 0.785 , 0.80125, 0.78125, 0.78375, 0.7675, 0.7825 , 0.8025, 0.81, 0.77625

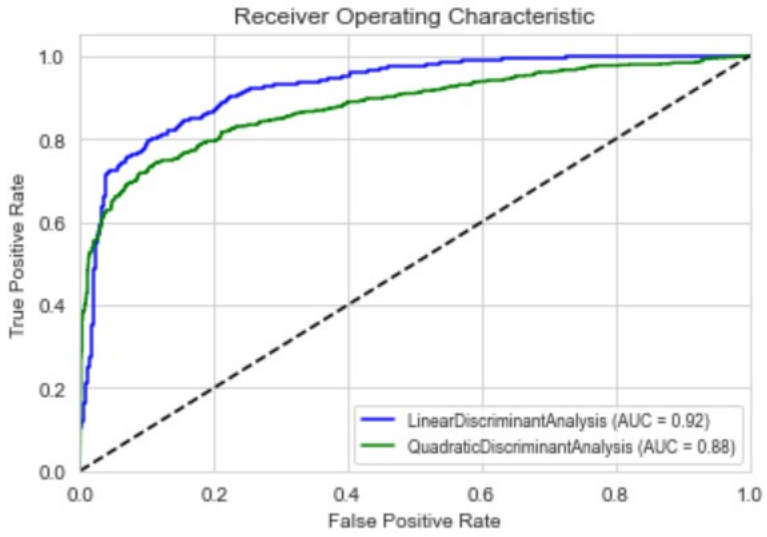
Mean cross-validation score: 0.787

QDA demonstrates a slightly lower accuracy of 0.8045 compared to LDA. The precision remains at 0.7208, indicating that when the model predicts a claim as fraudulent, it is accurate 72.08% of the time. However, the recall (sensitivity) improves to 0.7902, indicating that the model identifies a higher proportion of actual fraudulent claims (79.02%). The F1 score of 0.7539 sug-

gests a balance between precision and recall. The specificity of 0.8132 indicates that the model is effective at correctly classifying non-fraudulent claims, but it is slightly lower than LDA.

Metric	LDA	QDA
Accuracy	0.8265	0.8045
Precision	0.9237	0.7208
Recall	0.5910	0.7902
F1 Score	0.7208	0.7539
Sensitivity	0.5910	0.7902
Specificity	0.9702	0.8132

Table 3.5: Performance Metrics for LDA and QDA



Conclusions: Overall, LDA exhibits higher accuracy and specificity, suggesting that it performs better at identifying non-fraudulent claims accurately. On the other hand, QDA has a higher recall, indicating that it identifies a larger proportion of actual fraudulent claims. We will prefer LDA over QDA due to computational ease since there is no significant difference in performance metrics of both the model.

3.2.3 Approach 3: Deep Learning Models

Deep learning is a subfield of machine learning that focuses on training deep neural networks with multiple layers. By training a neural network on historical data that includes both fraudulent and non-fraudulent cases, the network can learn to recognize patterns and relationships that differentiate between fraudulent and non-fraudulent claims.

Here, we have applied two types of Neural Networks Binary Classifiction Neural Networks and Recurrent Neural Networks (RNNs) on final dataset.

Binary Classifiction Neural Networks : In our analysis, Health insurance fraud detection typically involves classifying claims into two categories: fraudulent or legitimate. A binary classification neural network is specifically designed to handle this type of problem by learning to distinguish between two classes. It is suited to model the decision-making process of identifying fraudulent claims.

Recurrent Neural Networks (RNNs): RNNs have a hidden state that allows them to remember information from previous claims in the sequence. This memory enables the network to learn and detect patterns that may span across multiple claims, such as recurring fraudulent behaviors or suspicious billing sequences.

Results:

Model	Accuracy	Sensitivity	Specificity	F1 score
BNN	0.8625	0.7898	0.9076	0.8148
RNN	0.617	0.75	0.8523	0.78

Table 3.6: Neural Network Results

Here,BNN works better than RNN with specificity of 90.76% which is quite good in detecting non-fraud claims correctly.
Moreover,accuracy of BNN which is 86.25% is also higher then RNN.

3.2.4 Approach 4

Approach 4 deals with Fitting a Markov Model on the health insurance fraud detection dataset.

Firstly we had selected some important variables by fitting random forest on Final_Dataset_Train_FE dataset and selected six most important variables by variable importance.

Variable
InscClaimAmtReimbursed
DeductibleAmtPaid
Hospitalization_Duration
Claim_Period
ExtraClaimDays
Inpatient_or_Outpatient

We have fitted the Markov model on these variables but we get very low accuracy for fitted Markov model . So to boost the performance of the model we selected three variables from variable importance by random forest and taken two more variable by domain knowledge which are more likely to have prediction power for potential fraud.

Markov Model Using Naive Bayes

We have selected the five variables which are most important for fitting the Markov model.

Variable
Patient_Age
InscClaimAmtReimbursed
Hospitalization_Duration
PerProviderClmProcedureCode_4_count_OPAnnualDeductibleAmt
Patient_Risk_Score

Each of the feature in the dataset used for this purpose was categorized into groups based on quantiles, such that the groups had equal number of claims in it. If the variable is already categorical then we kept it as it is. The sequence of values taken in the feature was labelled into states.For e.g.for a datset with three features where each of the feature can take take three values the total number of possible states would be 27. similarly in this work we have considered the five most significant variables and after doing the categorization of these features the states are formed.

Each of the claim has a class label as fraud or not-fraud. Now based on the data a model has been fitted to determine the probability of claim being fraud or not. Each of the state has the probability of it being feaudulent or not.
P(PotentialFraud=1 /State=1) is calculated.

and
P(PotentialFraud=0/State=1) =1- P(PotentialFraud=1 /State=1) similarly we can find the respective probabilities for all the remaining states.

we can calculate this probabilities by two methods.
1.By Naive Bayes Algorithm (By Bayes Theorem)
2. By One-Step Transition Probability matrix.

For this the dataset was divided into train and test in the ratio 80:20.

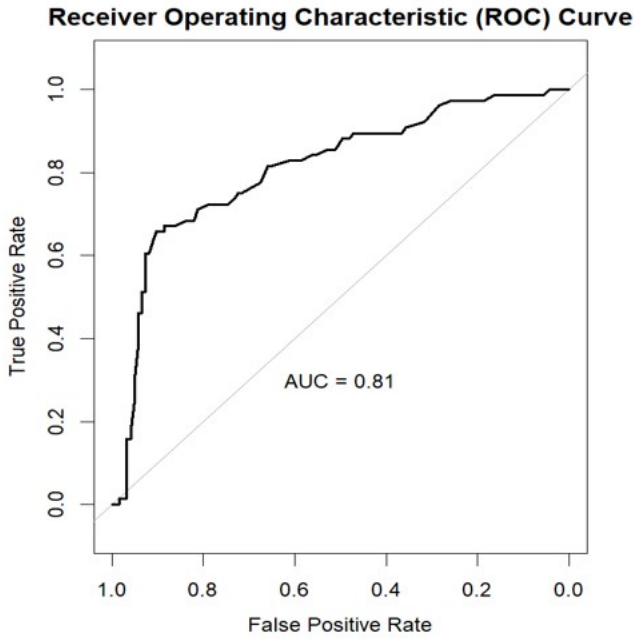
Using the method explained above the probability of each state being fraud is derived for the train dataset. The model was then tested on the test dataset. The confusion matrix below summarized the model.

		Y_pred	
		0	**1**
Y	**0**	570	48
	1	138	244

Table 3.7: Confusion Matrix for Markov Model Using Naive Bayes

It can be Observed that sensitivity of Markov model with Naive Bayes is 0.6973 meaning that around 70 % of the fraud cases were correctly identified. specificity is 0.9024 meaning that around 90% of non-fraud cases were correctly identified. precision is 0.8153 and F1 score is 0.7517,Accuracy is 0.8241 meaning that around 82 % of the labels were correctly identified by the model.

The below figure shows ROC curve for the model and AUC is 0.81.



Markov Model

The procedure was similar to above Markov model using Naive Bayes only the way of calculating the probabilities is different. The model was then tested on the test dataset.

The confusion matrix below summarized the model.

		YPRED	
		0	**1**
Y	**0**	575	43
	1	125	257

Table 3.8: Confusion Matrix for Markov Model

It can be Observed that sensitivity of Markov model with one step transition probability matrix is 0.6727 meaning that around 67% of the fraud cases were correctly identified. specificity is 0.9304 meaning that around 93% of non-fraud cases were correctly identified. precision is 0.8566 and F1 score is 0.7536, Accuracy is 0.8320 meaning that around 83% of the labels were correctly identified by the Markov model with one step TPM.

Markov Model with Gradient Boosting Method

The model was further improved by including more features into the data and retaining the some of the features as it is. i.e. no categorization was done.

One of the limitation while working with Markov Model was that the features had to be categorized into buckets. This was required otherwise the total number of states would be too high for the Markov model. However with the help of machine learning model, we can work with a very large number of states and the associated probabilities can be learnt for all the states.

In Markov model with GBM, we added two more variables in our previous five important variables namely PerProviderClmDiagnosisCode_3_count_Patient_Age, PerProviderClmProcedureCode_5_count_InscClaimAmtReimbursed

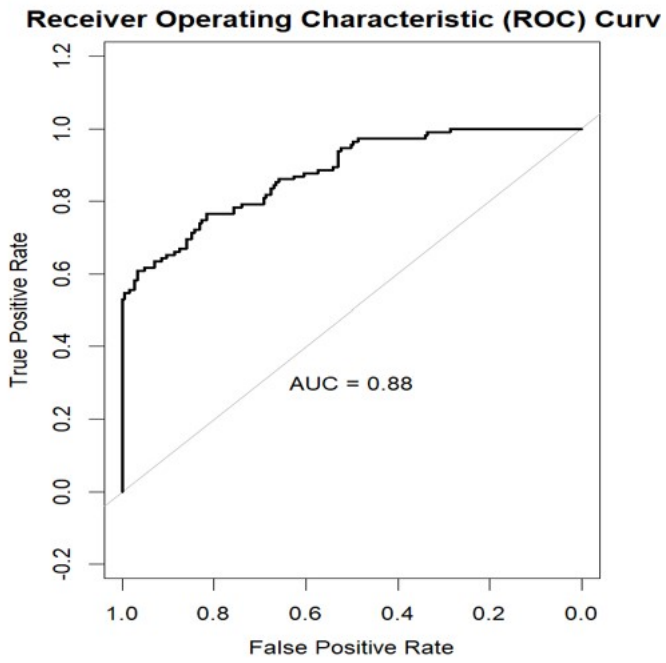
for GBM Modelling a total of 300 trees were used and interaction depth (maximum depth of each tree) is kept at 5. learning rate is kept at 0.1, In addition to usual fit 10 fold cross validation is performed. The model was tested on the test data . The confusion matrix and the statistics below summarizes the performance of the model.

		gbmClass	
		0	**1**
Y	**0**	172	13
	1	42	73

Table 3.9: Confusion Matrix for Markov Model with GBM

It can be Observed that sensitivity of Markov model with GBM is 0.6033 meaning that around 60 % of the fraud cases were correctly identified. specificity is 0.9832 meaning that around 98 % of non-fraud cases were correctly identified. precision is 0.9605 and F1 score is 0.7411, Accuracy is 0.83 meaning that around 83 % of the labels were correctly identified by the model.we can see the significant improvement in Accuracy and Specificity of the model.

The below figure shows ROC curve and AUC for the model



The below table compares the model performance of all the three methods discussed in Approach 4.

Model	**Accuracy**	**Specificity**	**Sensitivity**	**F1 Score**	**Precision**
Markov Model with Naive Bayes	0.8241	0.9024	0.6973	0.7517	0.8153
Markov Model	0.8320	0.9304	0.6727	0.7536	0.8566
Markov Model with GBM	0.8300	0.9832	0.6033	0.7411	0.9605

Table 3.10: Performance Metrics for Markov Based Models

■

Chapter 4

Conclusions

- We saw that **Markov model with GBM** and **Random forest** worked the best model for this healthcare provider fraud detection project.

- We can conclude that when machine learning model is incorporated into some of statistical models (Markov model) we can except a significant improvement in the performance.

- Markov Model shows a significant improvement when a boosting technique is used and hence,gave us Specificity of 98.32% which means that it is correctly predicting 98.32% of non-fraudulent claims as non-fraudulent which,thus takes care of minimising **Type 2 error** also. Moreover,Accuracy is 83% which is also considerably good.

- Also, random forest gave us Area under Curve of 0.96 which implies that the model achieves a high TPR while maintaining a low FPR. In other words, it has a high probability of correctly identifying fraud cases (true positives) while minimizing false alarms (false positives).With an AUC of 0.96, the model demonstrates a high level of discrimination power. It indicates that the model has a strong ability to correctly rank and differentiate between fraudulent and non-fraudulent cases across a range

of classification thresholds.

- But Markov model with GBM is more interpretable than random forest as random forest is a Black-Box model.

- We have considered to use time homogeneous Markov Model from our observation of EDA where we found out that Claim variable is Uniformly distributed.

4.1 Conclusion in Users Term:

- This project provides a comprehensive study of statistical healthcare fraud assessment. After providing a description of the health insurance data, statistical methods are discussed with focus on sampling, stochastic models and data mining.

- We present illustrations of recently proposed supervised machine learning models,deep learning models and Markov model using real world insurance health claims data.

- By comparing different statistics derived from the application of these models, Markov model with GBM and Random forest were found to be slightly better than others for the purpose of making predictions.

■

Chapter 5

Scope and Limitations

5.1 Scope

- In this work, Markov model and supervised learning methods have been used to build the model. In future, unsupervised methods or the hybrid of supervised and unsupervised methods can be used by extracting the best features from both the approaches and building a hybrid model for fraud detection.
- Also, a model can be built which processes dynamic data and detects frauds and flags them for further investigation dynamically.
- In addition, our study does not cover fraud detection in other fields such as credit card fraud, money laundering, telecommunication fraud, computer intrusion and scientific fraud.Our suggested models for fraud detection can also work in these industries for reducing fraudulent transactions.
- One can go for more sophisticated Markov models of higher orders.
- Hidden Markov Model (HMM) can be used on the same dataset and see whether their is significant improvement in the performance.

5.2 Limitations

- Our health insurance fraud data is an imbalanced classification problem, with the number of fraudulent cases being significantly lower than legitimate cases. Imbalanced data can affect the performance of fraud detection models, as they tend to be biased towards the majority class and may struggle to detect rare instances of fraud.

- Fraudsters continually adapt their tactics to evade detection, making it challenging to keep up with emerging fraud patterns. Fraud detection systems may lag behind in identifying new and sophisticated fraud schemes until they are recognized and incorporated into the system.

- Due to time limitation we were not able to perform the same analysis on more sophisticated higher order Markov Models.

- We are assuming that available data is reliable.(Secondary data)

- Some of the classification models that we have chosen for our analysis, such as Random Forest, Neural Networks, and GBM, are considered black-box models. These models are known to be less interpretable compared to simpler models like linear regression or decision trees.

■

Bibliography

[1] Khaled Gubran Al-Hashedi and Pritheega Magalingam. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, 40:100402, 2021.

[2] Mr G. Fraud detection in medinsu using machine learning algorithms - a web application. *International Journal for Research in Applied Science and Engineering Technology*, 7:689–693, 10 2019.

[3] Rohan Yashraj Gupta, Satya Sai Mudigonda, Pallav Kumar Baruah, and Phani Krishna Kandala. Markov model with machine learning integration for fraud detection in health insurance. *arXiv preprint arXiv:2102.10978*, 2021.

[4] ROHIT ANAND GUPTA. Healthcare provider fraud detection analysis @ONLINE, 2017.

[5] Healthcare Fraud Prevention Partnership. Healthcare Fraud Prevention Partnership, 2021. Accessed on May 18, 2023.

[6] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, New York, NY, 2013.

[7] A Jenita Mary and SP Angelin Claret. Analytical study on fraud detection in healthcare insurance claim data using machine learning classifiers. In

AIP Conference Proceedings, volume 2516, page 240006. AIP Publishing LLC, 2022.

[8] OpenAI. OpenAI ChatGPT, 2021. Accessed on May 18, 2023.

[9] Vipula Rawte and G Anuradha. Fraud detection in health insurance using data mining techniques. In *2015 International Conference on Communication, Information & Computing Technology (ICCICT)*, pages 1–5. IEEE, 2015.