Analysis of Fraud Detection in Health Insurance Sector

Monika (2134)
Swapnil Morkhade (2135)
Gayatri Wadghule (2153)

Guide: Dr.Mohan Kale

Department of Statistics
Savitribai Phule Pune University
Pune - 411007

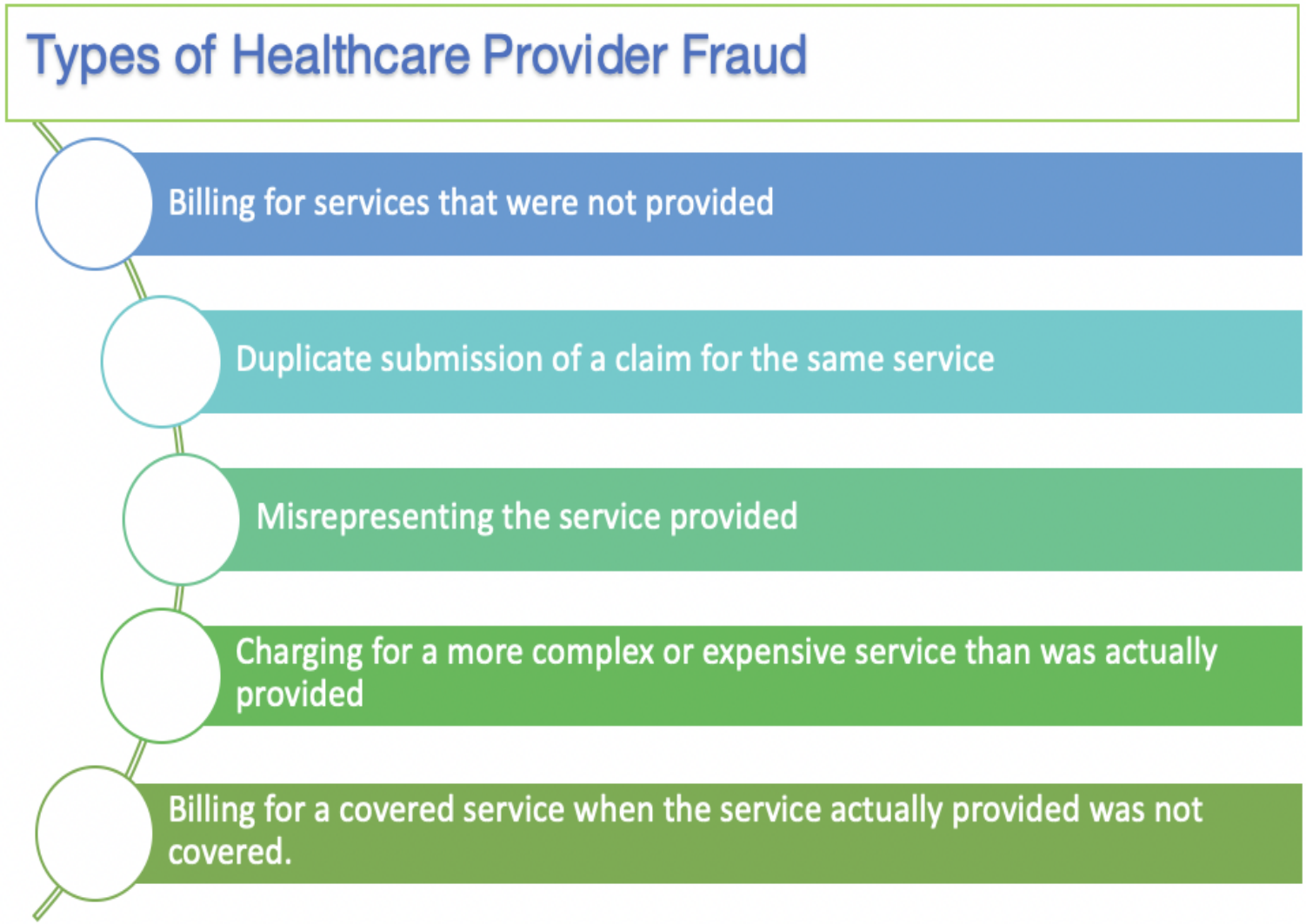
ST 402 Project
External Examination
23^{th} May 2023

Table of content

Introduction

- 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.

Types Of Healthcare Provider Fraud



Objectives

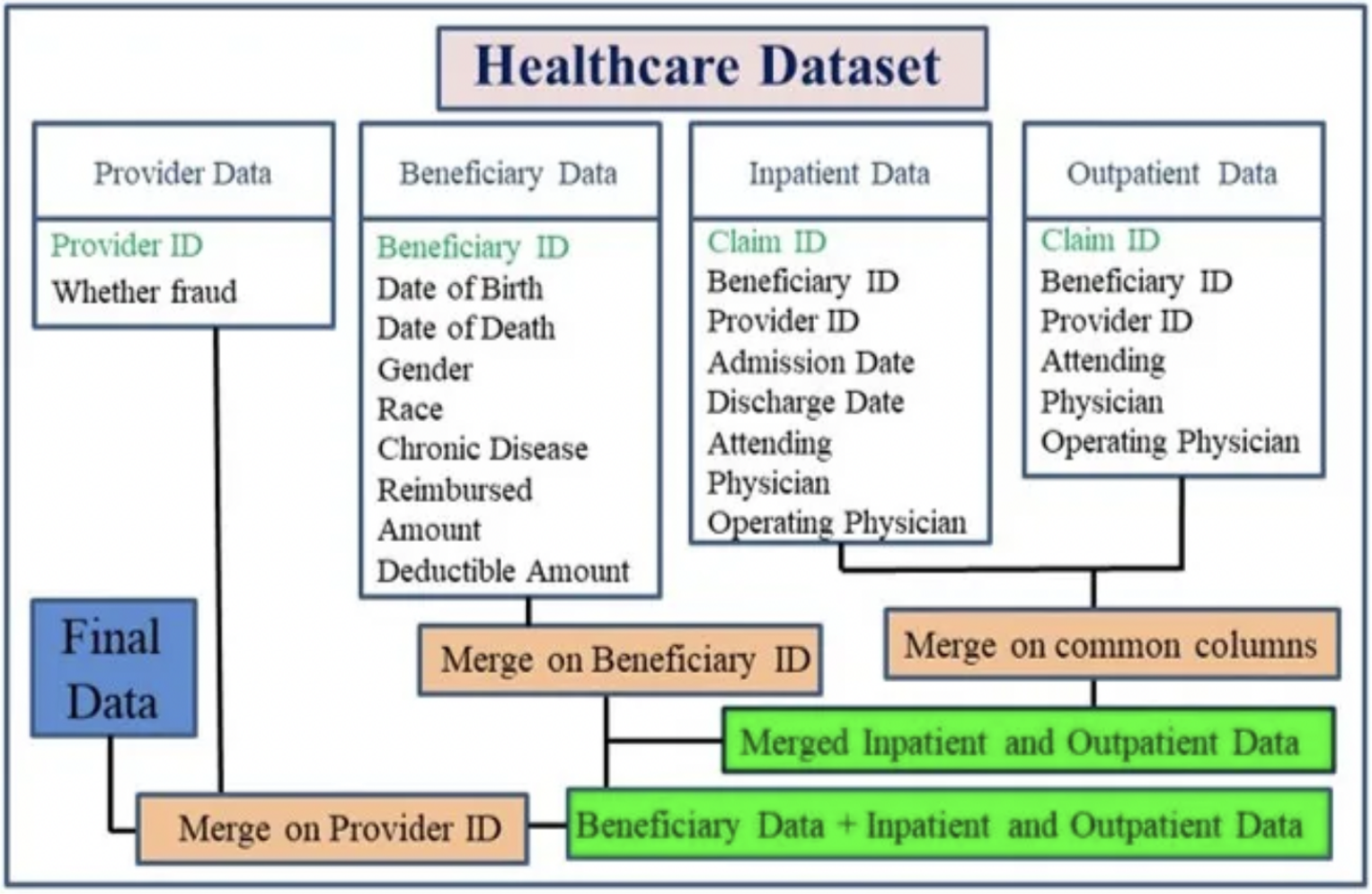
Healthcare fraud can occur in many forms, here ,we have focused on Provider's Fraud.(In the context of healthcare, providers are entities such as hospitals, clinics, or individual healthcare professionals who offer medical services to patients.)

- 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 fraudulent providers.Also,study of fraudulent patterns in the provider's claims to understand the future behaviour of providers;
- Protecting the healthcare system so that they can provide quality and safe services to legitimate patients.

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
- **Description of the variables:** There are total 55 variables with more than 7,00,000 observations. Variables ,Unique value count , Missing observations and Type of variable are shown in table below.

Data Description Cont...



Overall Representation of The Dataset

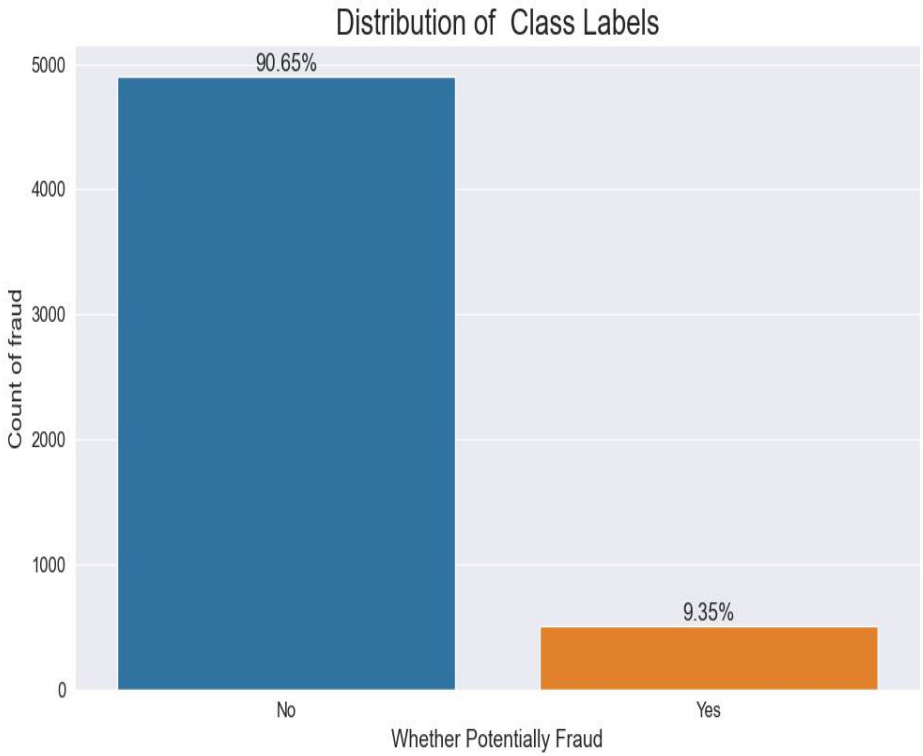
Variable	Unique	Missing Obs	Type
State	52	0	Nominal
RenalDiseaseIndicator	2	0	Nominal
Race	4	0	Nominal
Provider	6763	0	Nominal
PotentialFraud	3	135392	Nominal
OtherPhysician	57493	445235	Nominal
OperatingPhysician	43654	551963	Nominal
OPAnnualReimbursementAmt	2084	0	Count
OPAnnualDeductibleAmt	792	0	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

Literature Review

- 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. Vol. 2516. 1.AIP Publishing LLC.2022, p. 240006.]*
- Vipula Rawte and G Anuradha. "Fraud detection in health insurance using data mining techniques" .*[In: 2015 International Conference on Communication,Information Computing Technology (ICCICT). IEEE. 2015, pp. 1–5]*
- Rohan Yashraj Gupta et al. "Markov model with machine learning integration for fraud detection in health insurance" .*[In: arXiv preprint arXiv:2102.10978 (2021)]*

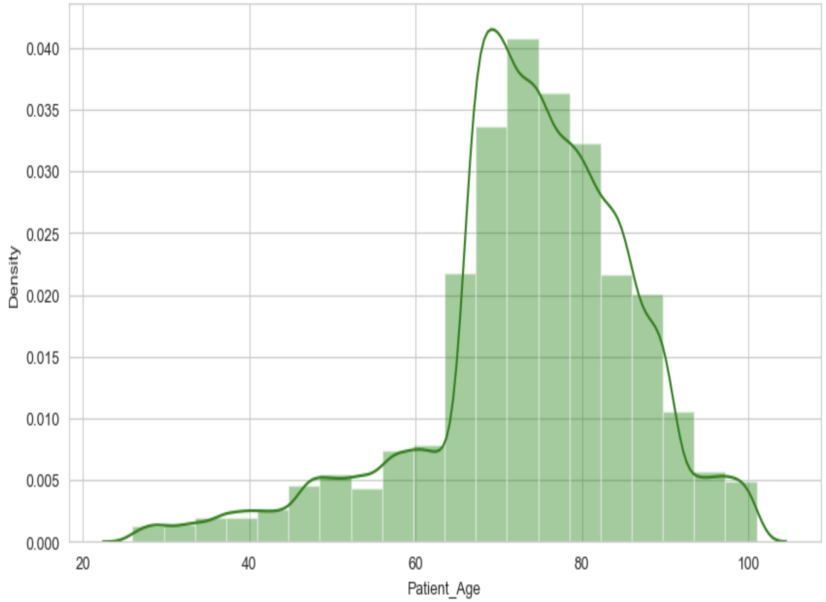
Results of Explorartory Data Analysis

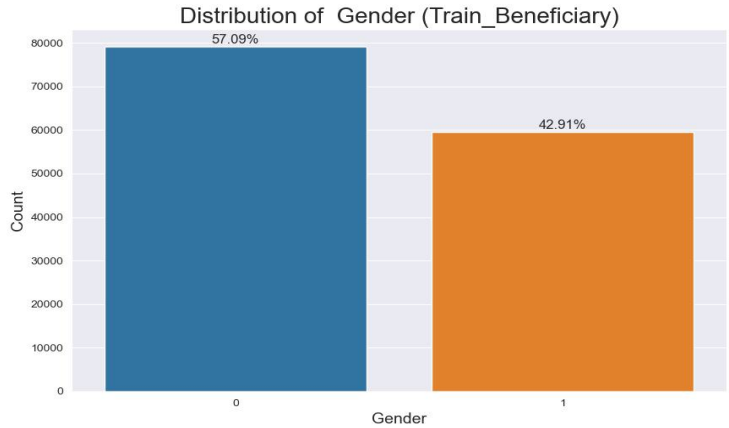
We are plotting the below plot to check the distribution of the class label.

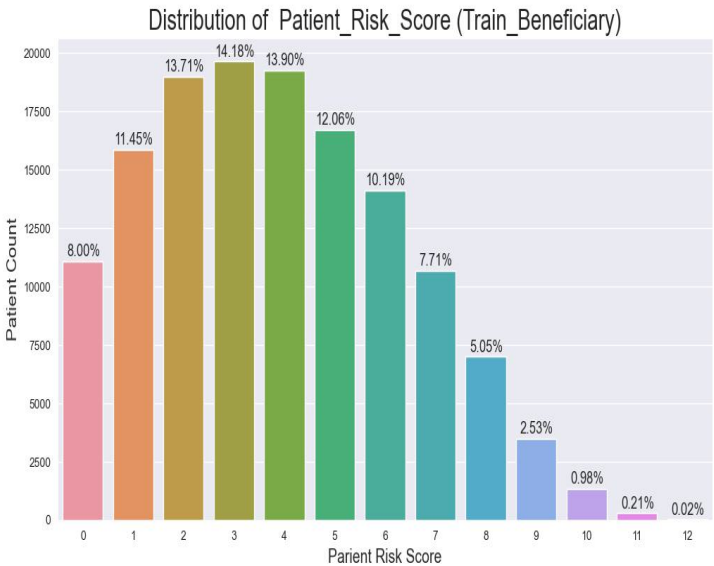


This is an highly imbalanced dataset. There are 10% fraudulent providers and 90% non-fraudulent providers in the whole dataset.

Results of Explorartory Data Analysis for Beneficiary Data







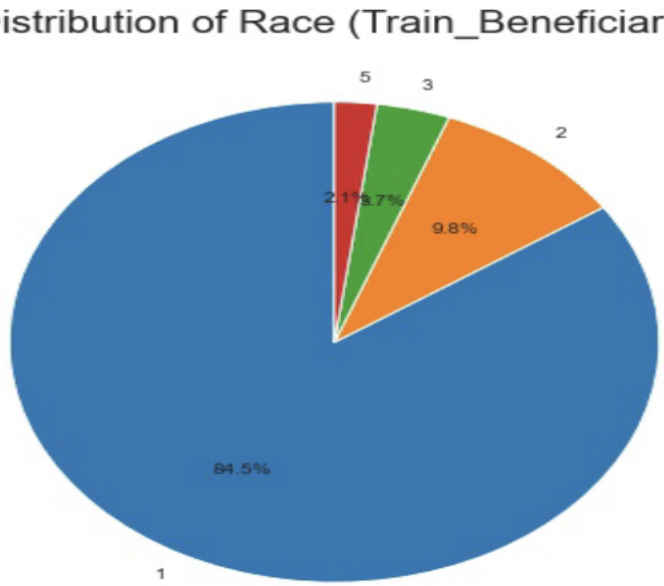
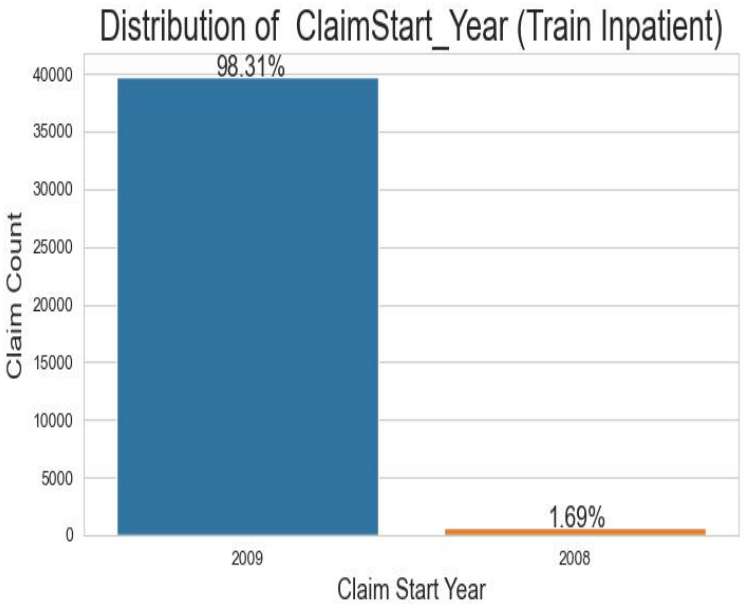
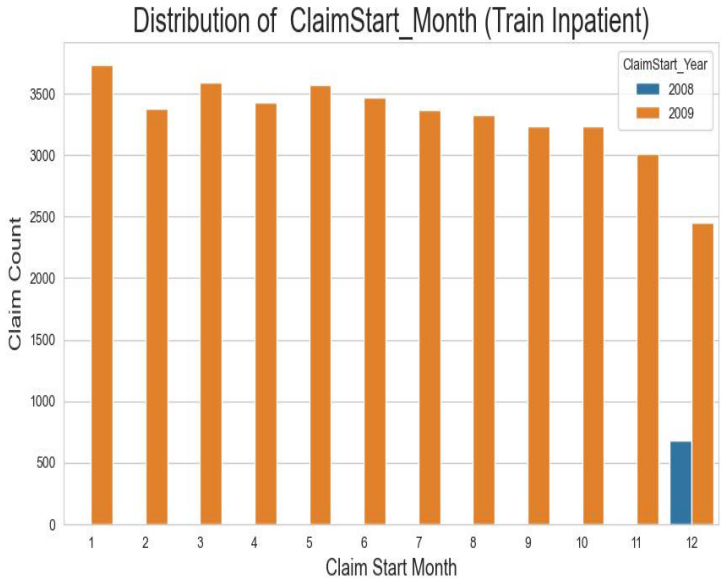
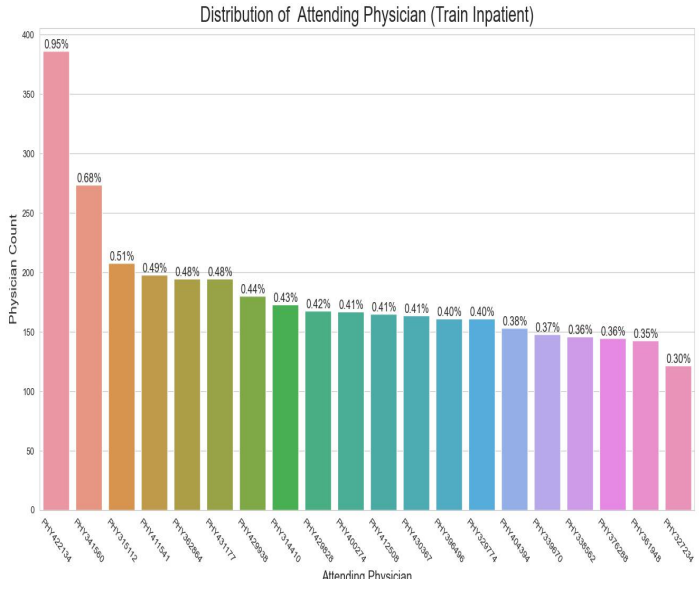


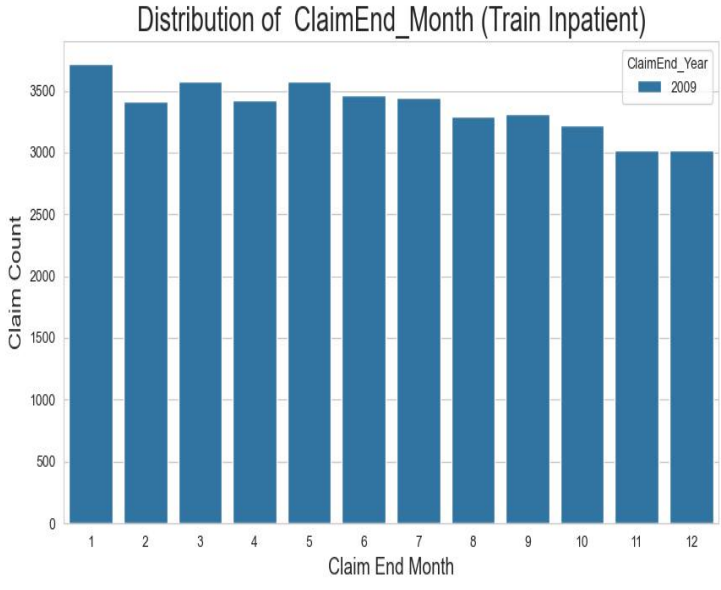
Figure 2.5: **Countplot of Race**

EDA results based on Inpatient Data

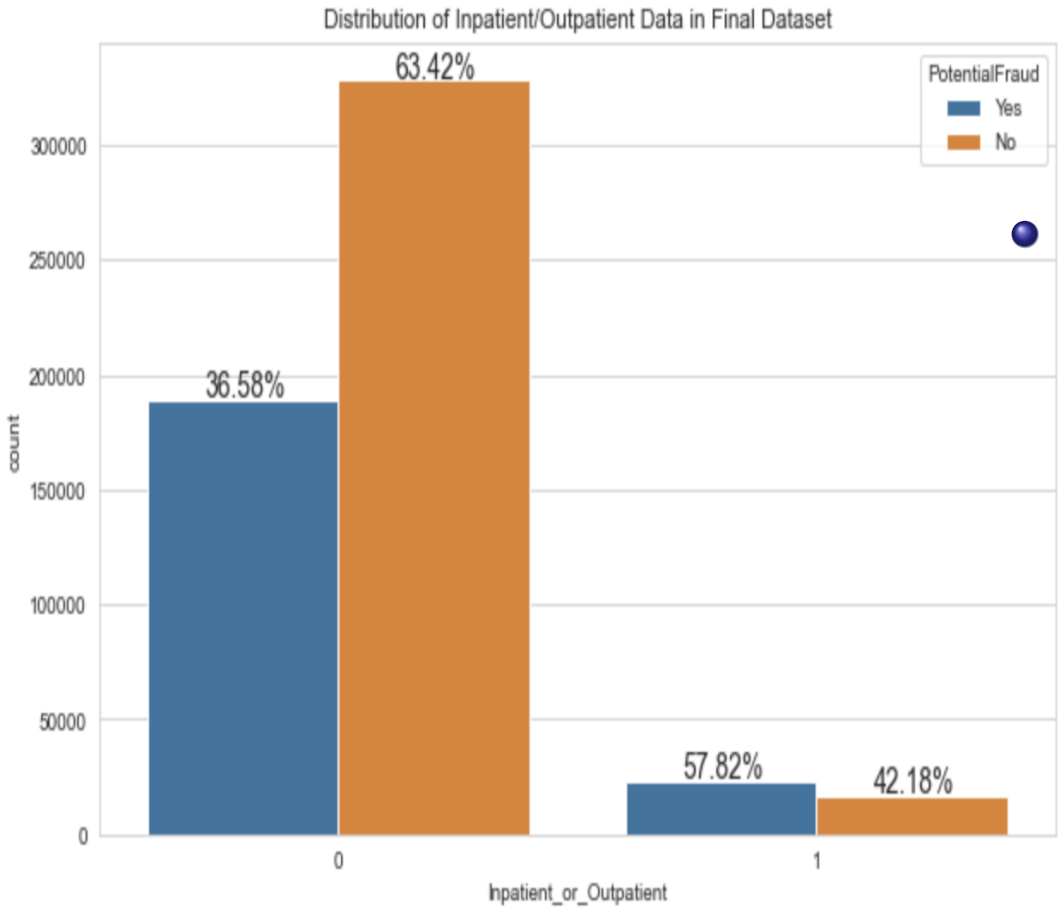






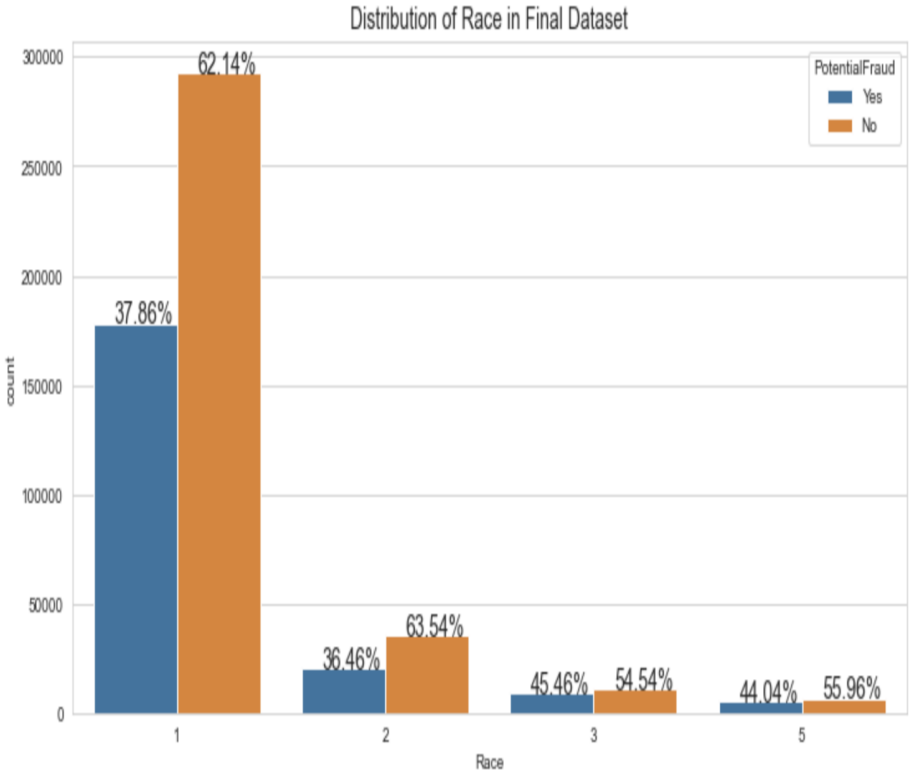


EDA Results for final merged Dataset



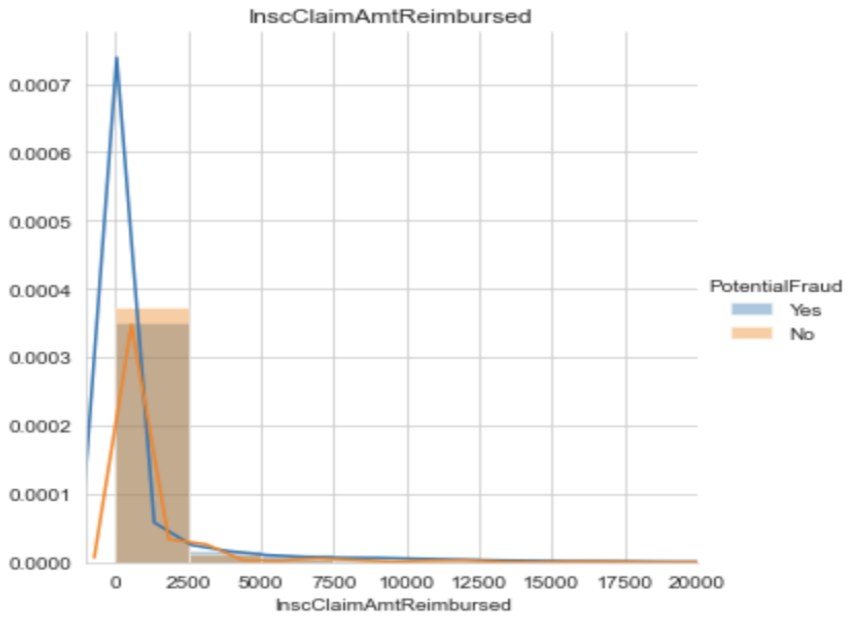
- The number of claims are less for inpatient data compared to outpatient data.
- Even though the claims are less in inpatient data, percentage of fraudulent activity is more in inpatient data(57.8%) whereas it is 36.5% in outpatient data. This is beacuse per claim reimbursement amount for IPD much higher(35 times calculated earlier) than the per claim reimbursement amount of OPD.

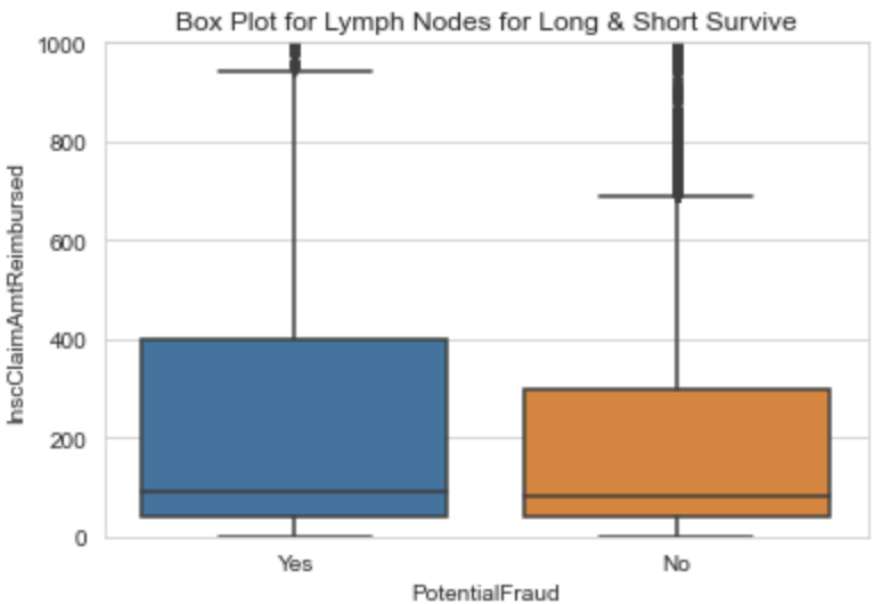
EDA Results for final merged Dataset



- Total number of transactions are more for Race 1 and 37.8% are fraudulent out of them.
- The ratio of fraudulent transaction is most for Race 3 (45.5%)
- So, race is an important feature in fraud detection.

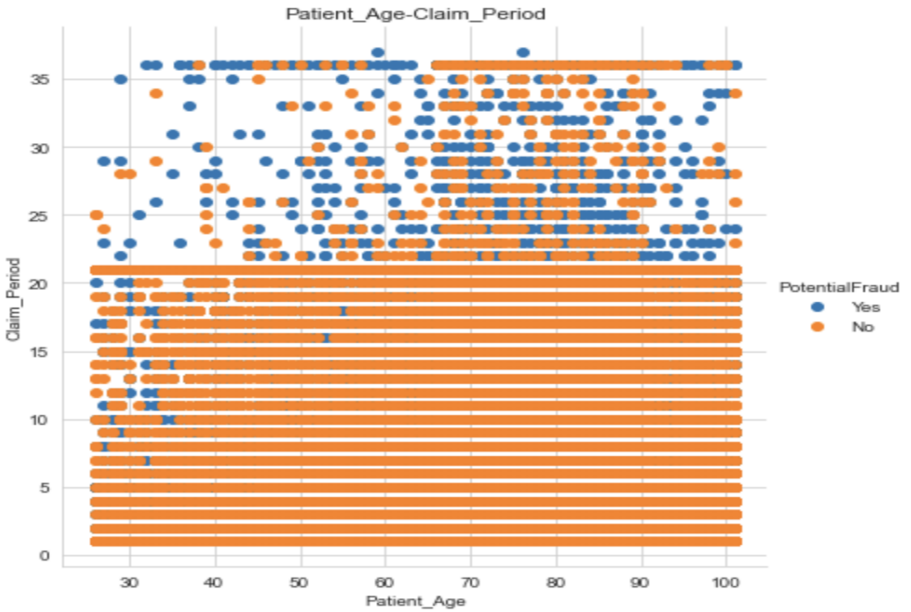
EDA Results for final merged Dataset

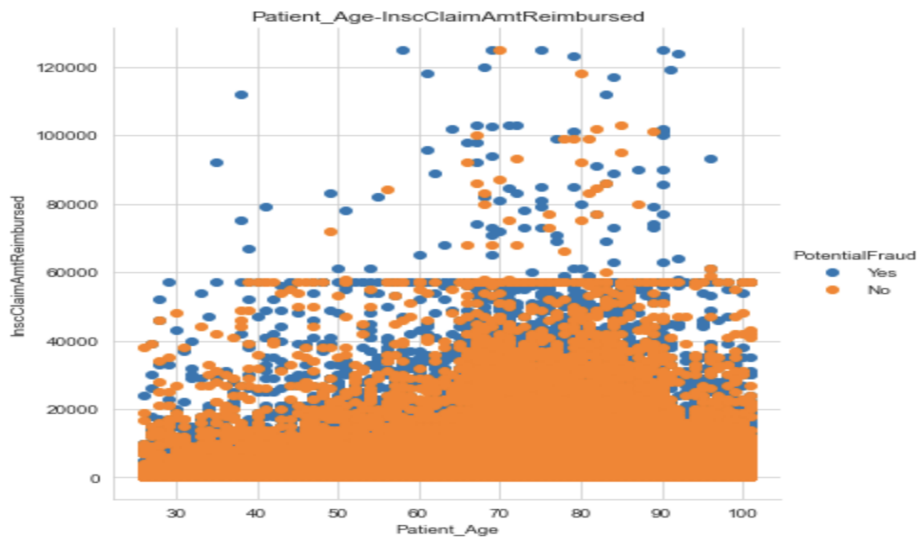




- From the histogram we can observe that when the claim amount is less, the number of fraud claims are much higher compared to legitimate claims.
- 25th,50th percentiles are very less for claim amount reimbursed. 75th percentile for fraud claims is higher than non-fraud claims.

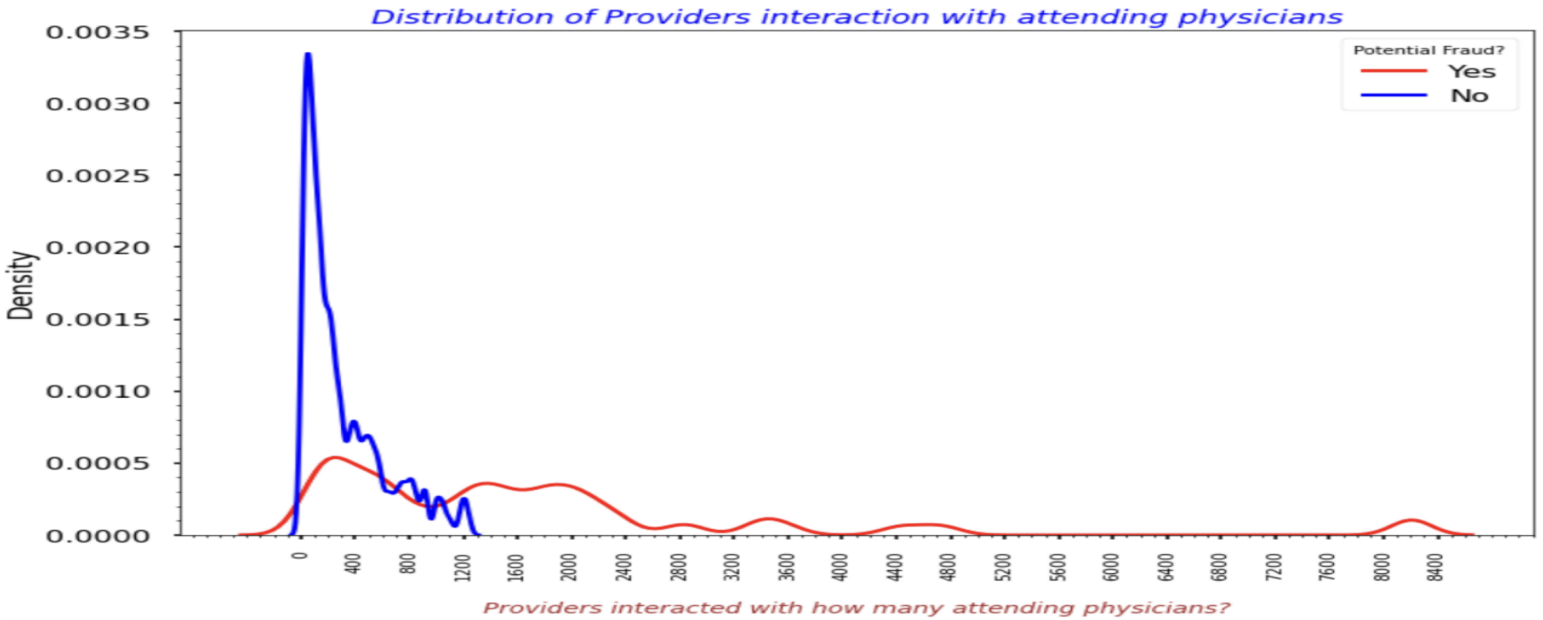
EDA Results for final merged Dataset





- From the scatter plot we can see that when patient's age <70 yrs claim period > 20 days, the probability of the transaction is fraudulent is high.
- We can see that if patient's age<60 yrs claim amount<60000 it tends to be a fraudulent transaction. If the patient's age>80 yrs claim amount>60000 the probability of claim to be fraudulent is high.

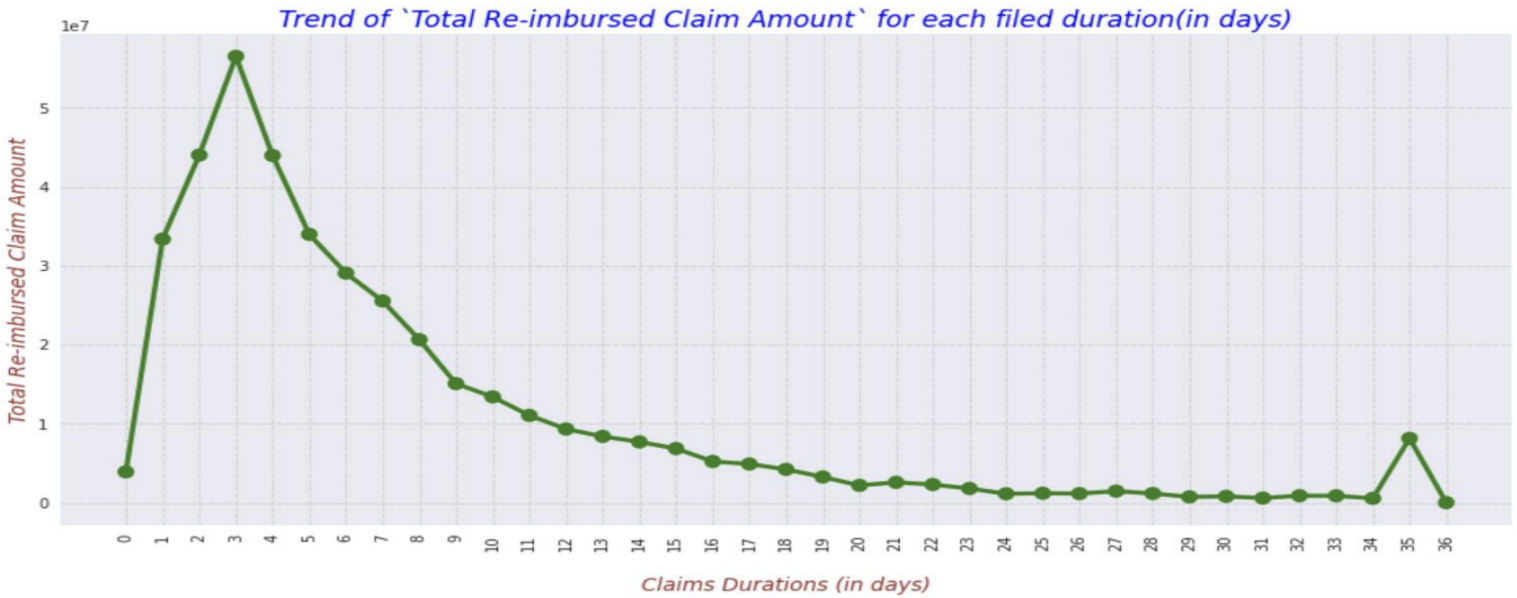
EDA Results for final merged Dataset



Plot of Providers interaction with attending physicians

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

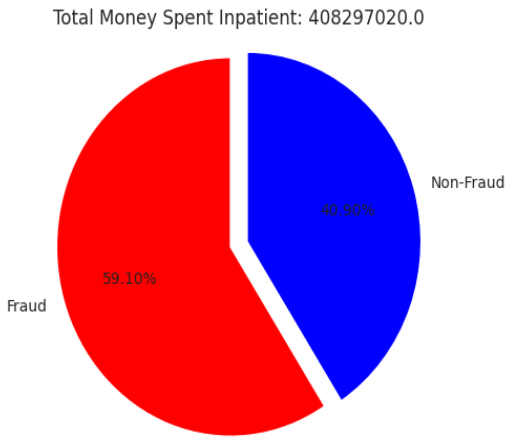
EDA Results for final merged Dataset

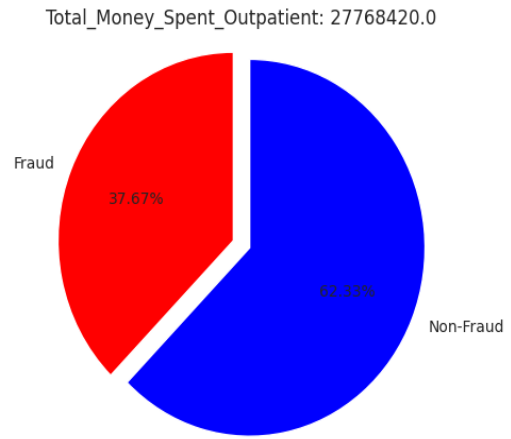


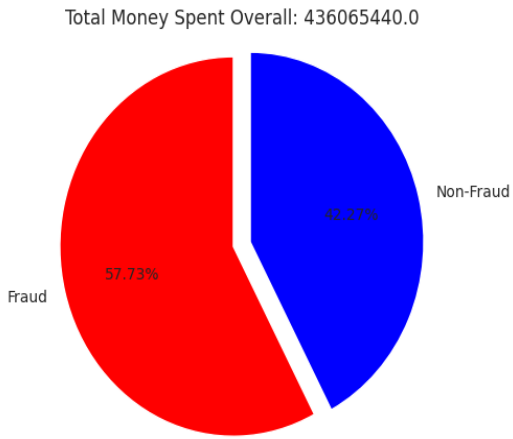
1)The above graph tells us that the Total Re-imbursed Amount is the highest for 3 days claims.

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 activity.

EDA Based on Insurance Amount Reimbursed

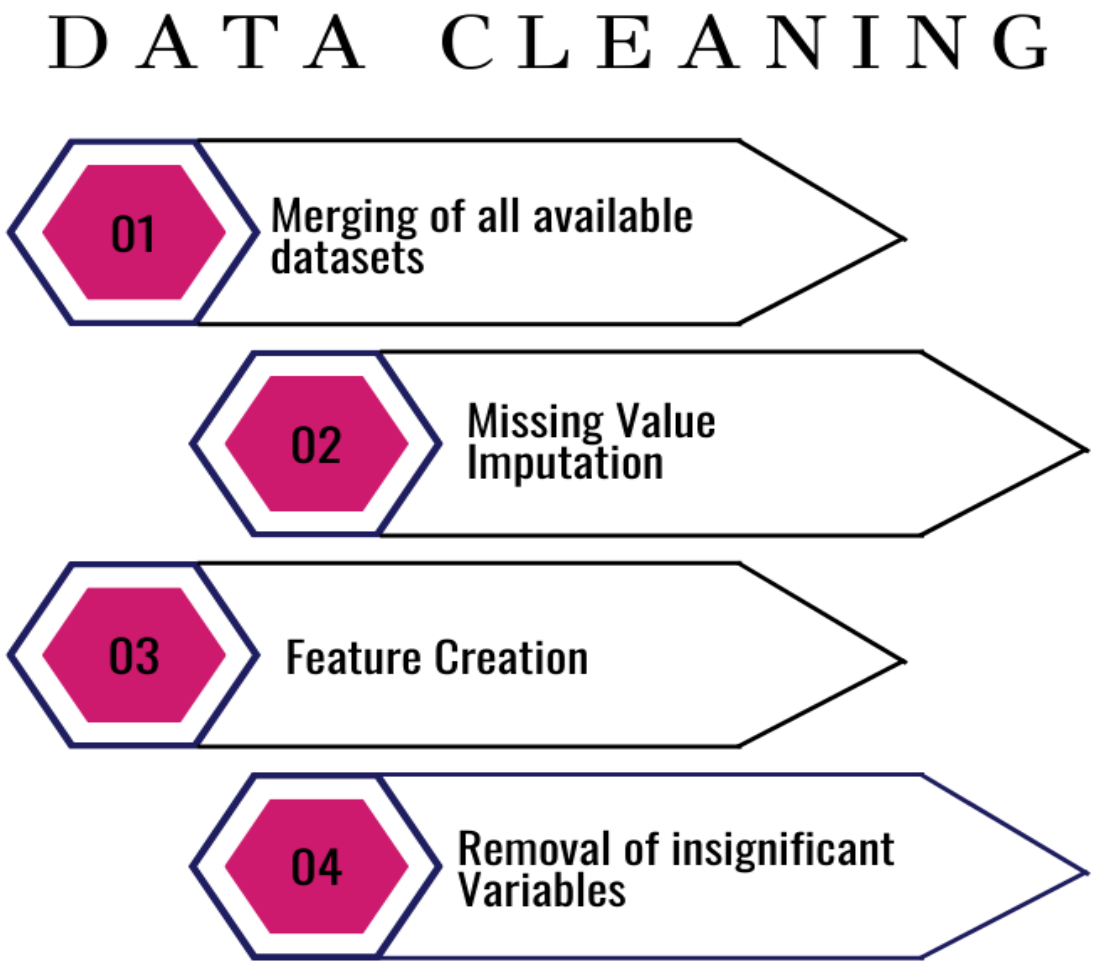




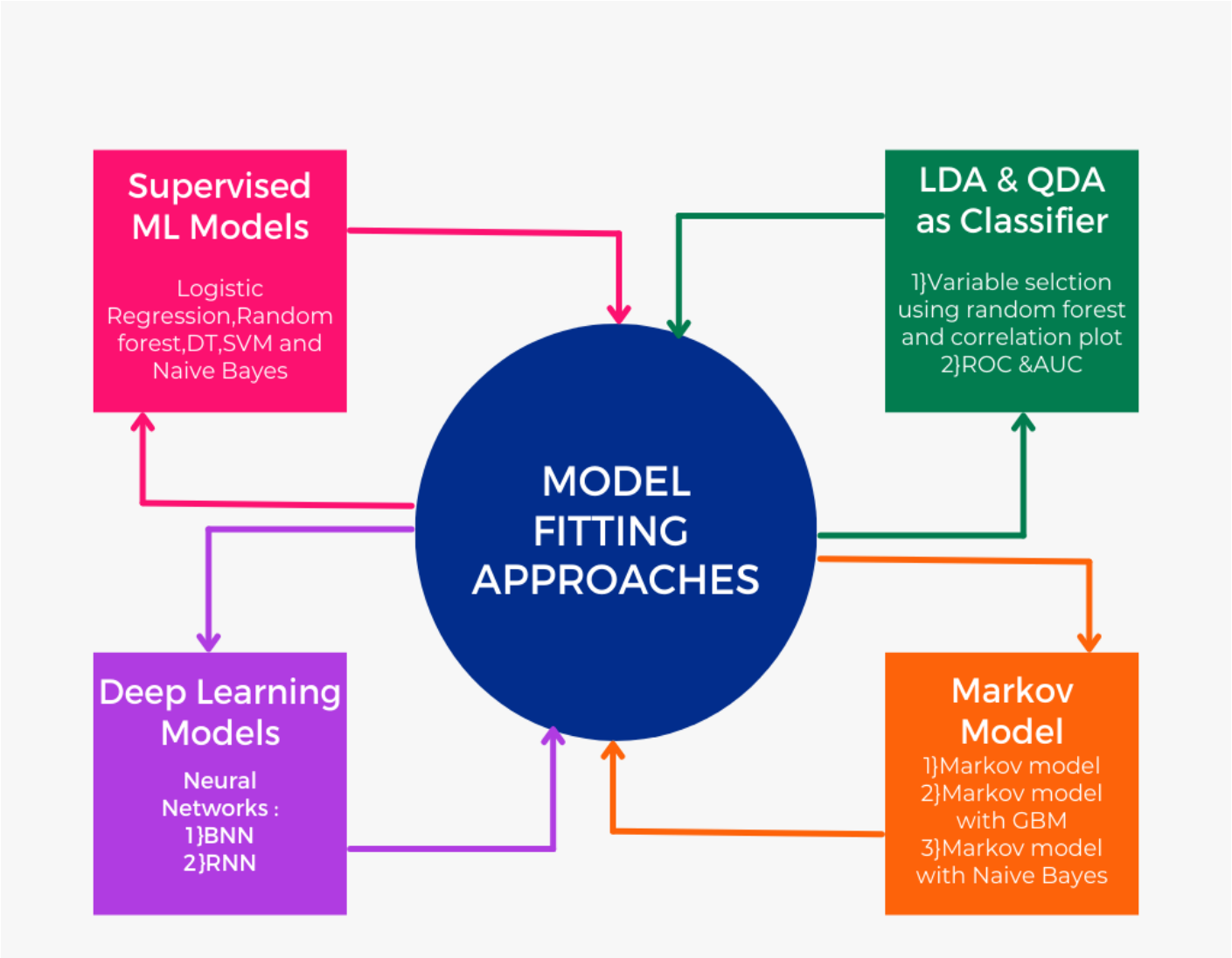


	Inpatient	Outpatient	Total
Total Money Spent	408297020.0	27768420.0	436065440.0
Average Reimbursed Amount per Patient	10088	282	-
Total Money Spent in Fraud	241288510.0	10460410.0	251748920.0
Percentage of Money Spent in Fraud	59.10%	37.67%	57.73%

Data Cleaning



Model Fitting



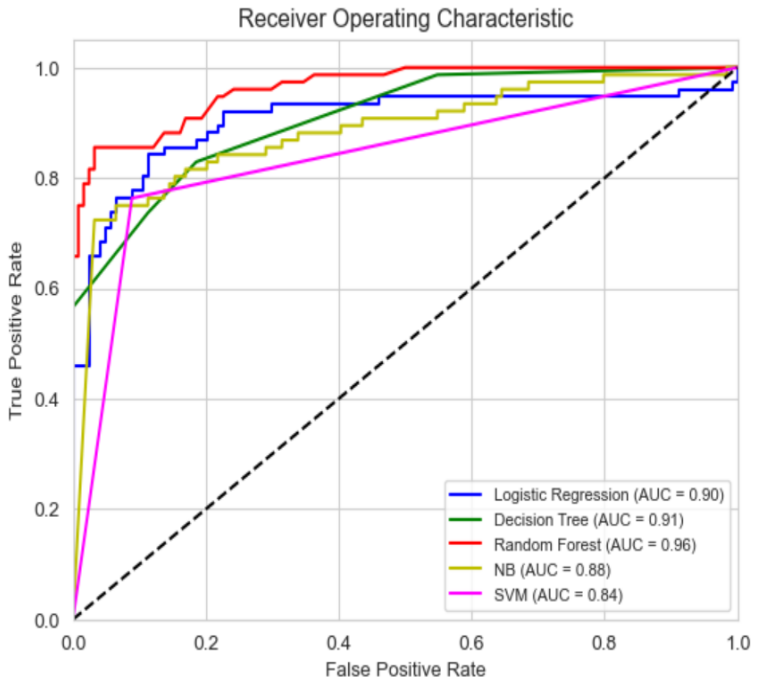
Approach 1 : Supervised Machine Learning Models

- 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 along with **10-fold CV** to avoid overfitting or underfitting. Pick the best model based on the performance score.

Supervised Machine Learning Models

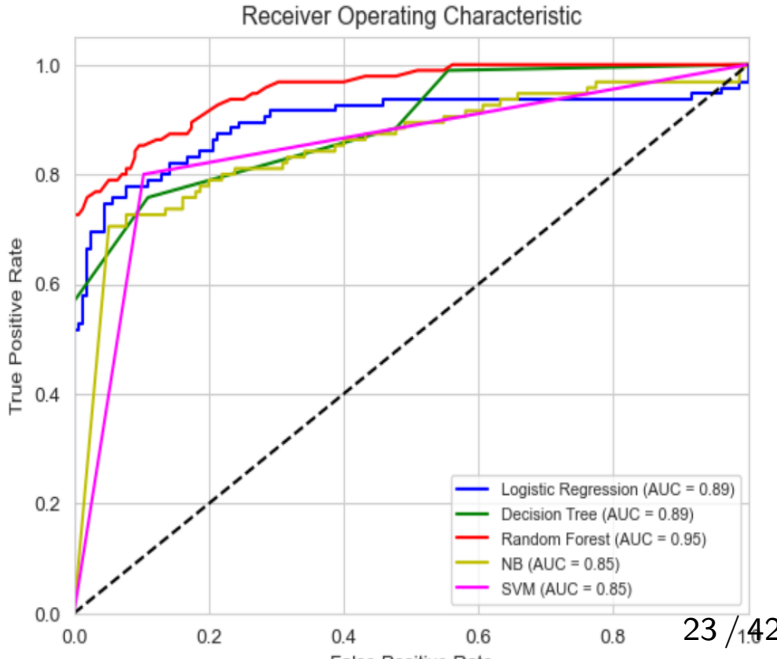
Model	Accuracy	Sensitivity	Specificity	F1 score	Rank
Logistic Regression	0.855	0.7763	0.9032	0.8027	4
Random Forest	0.895	0.75	0.9838	0.8444	1
Decision Trees	0.83	0.7368	0.8870	0.7671	5
SVM	0.855	0.7631	0.9112	0.8	3
Naive Bayes	0.87	0.7236	0.9596	0.8088	2

Table 1: 80:20 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

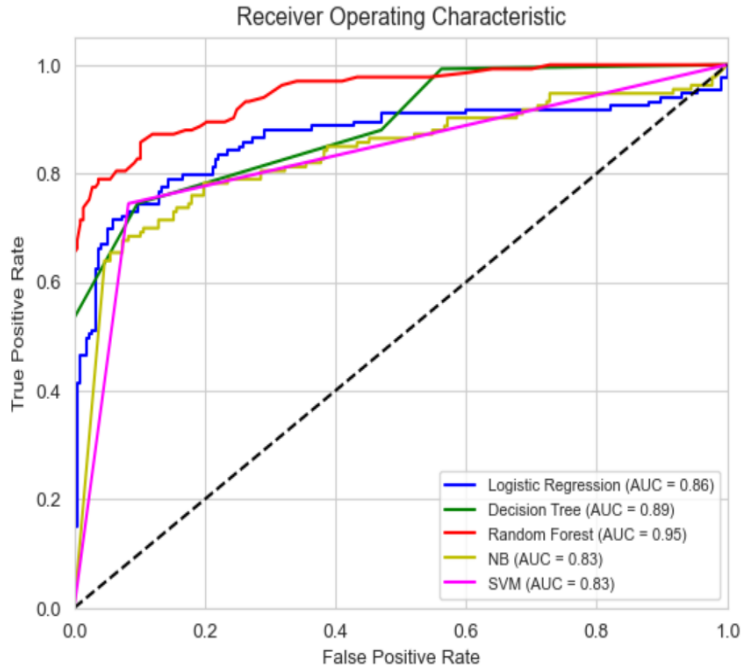
Table 2: 75:25 results



Supervised Machine Learning Models

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: 65:35 results



- These metrics suggest that the random forest model performs well in healthcare fraud detection which means that the model is very good at distinguishing non-fraud cases from the total set of negative cases.
- From all ROC curves we can see that Random forest has the highest AUC indicating that the model is highly effective in identifying and flagging potential fraudulent activities.

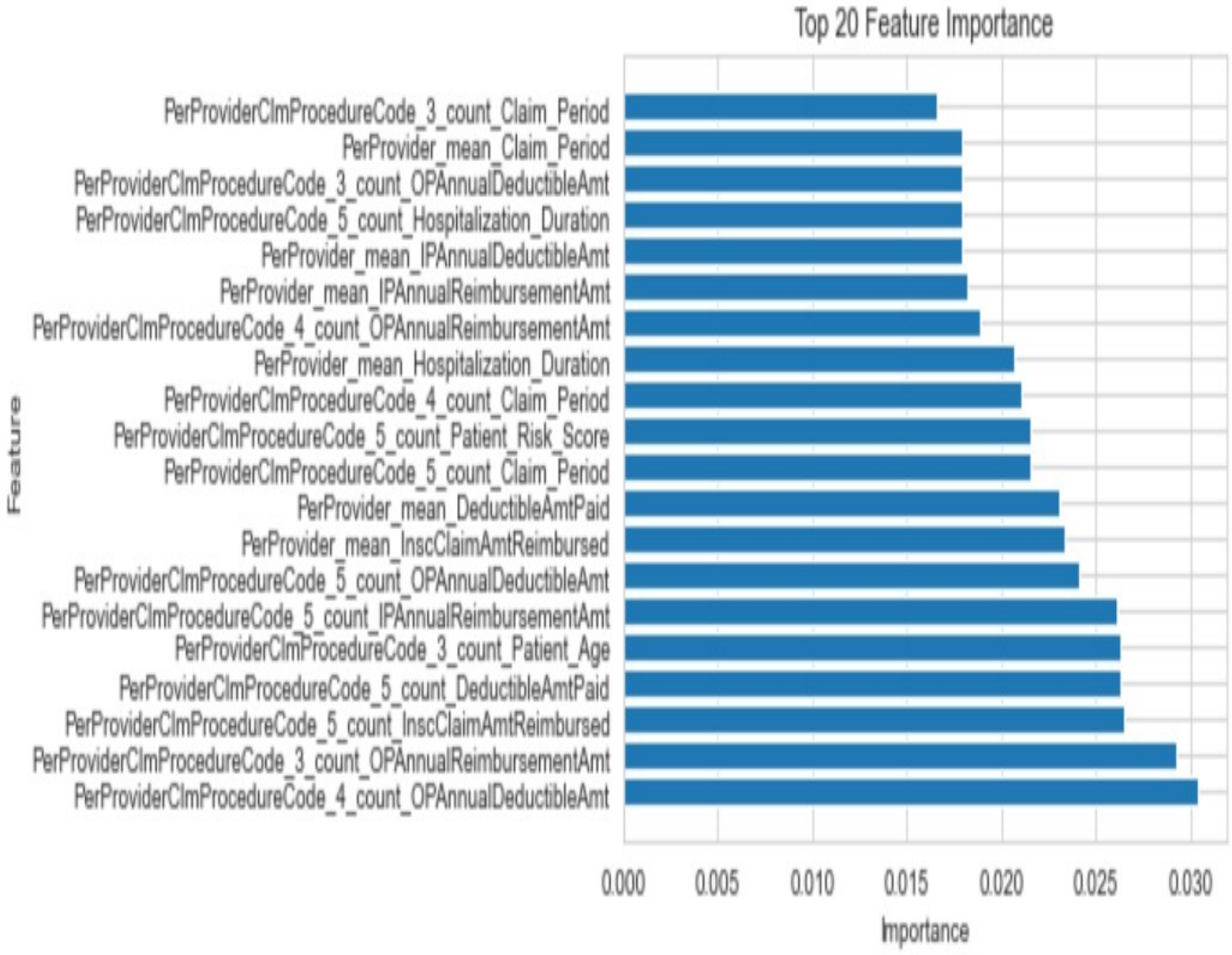
Supervised Machine Learning Models

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

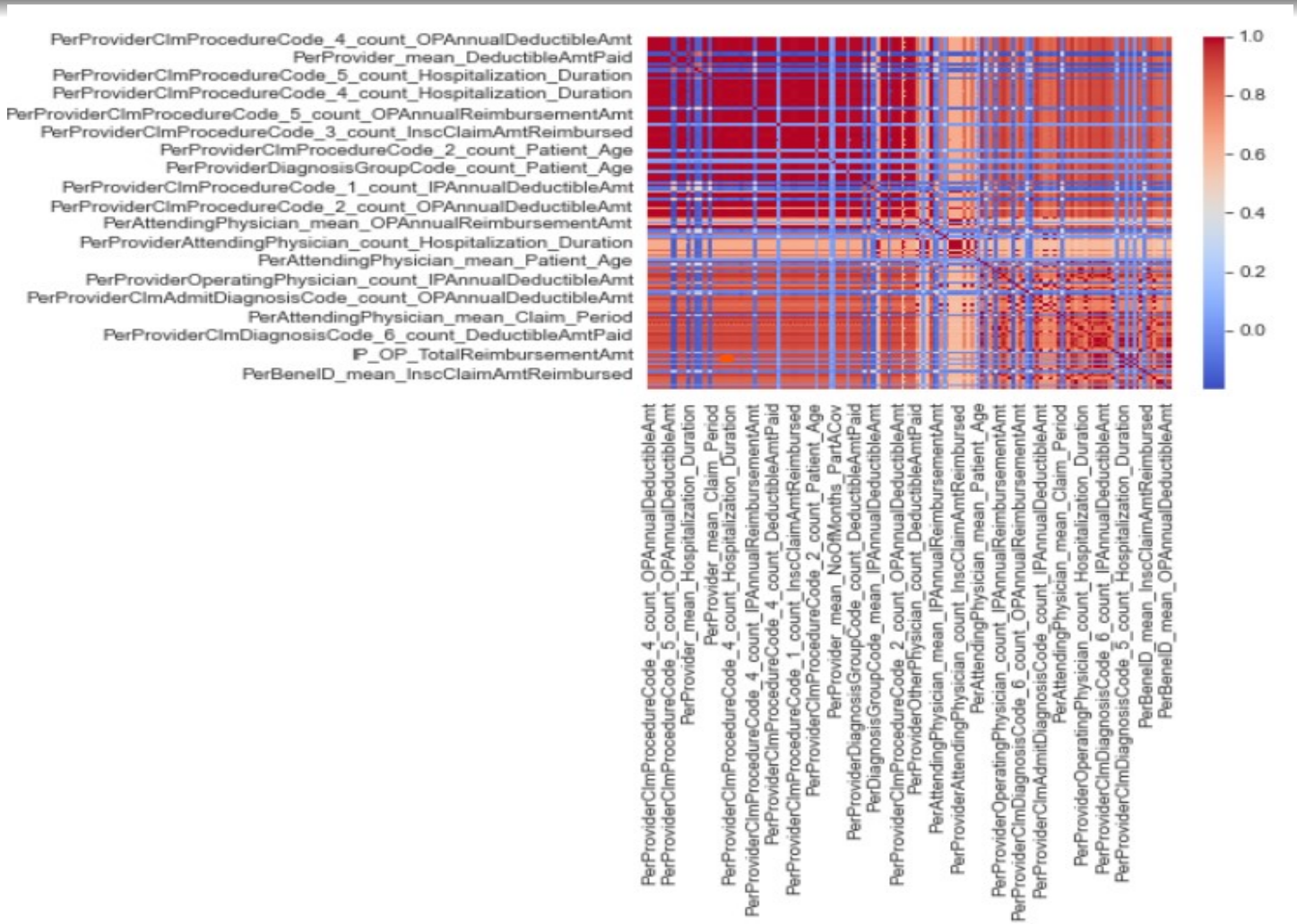
Table 4: Selection of Best Model

- Our objective is correctly identifying fraud and non-fraud claims and minimizing type 2 errors because **type 2 errors** are considered to be more severe than type 1 errors.
- Therefore from the above table random forest is working as the best model with an accuracy of 89.5% and specificity of 98.38% for 80:20 split criterion.
- Specificity of 98.38% means that it correctly identifies 98.38% of the non-fraud cases out of all the actual non-fraud cases in the dataset.

Approach 2 : Variable selection by Random Forest



Correlation Plot

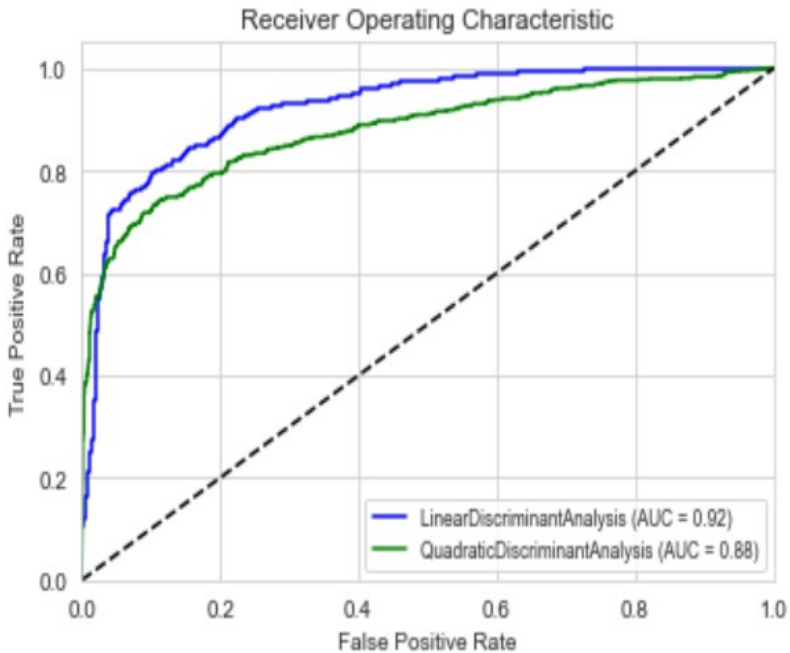


- Dark red colour between the variables shows they are highly correlated.
- We have simply removed these highly correlated variables.

Approach 2 : Results of LDA and QDA

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

Table 5: Performance Metrics for LDA and QDA



Conclusions:

- Accuracy Precision and Specificity of LDA is higher than QDA
- Sensitivity and F1 score of QDA is higher than LDA
- AUC of QDA is 0.88 and AUC of LDA is 0.92. Overall LDA perform better than QDA

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 on the final dataset:

  1) Binary Classifiction Neural Networks (BNN)

  2) Recurrent Neural Networks (RNNs)

Deep Learning Models Results

Results:

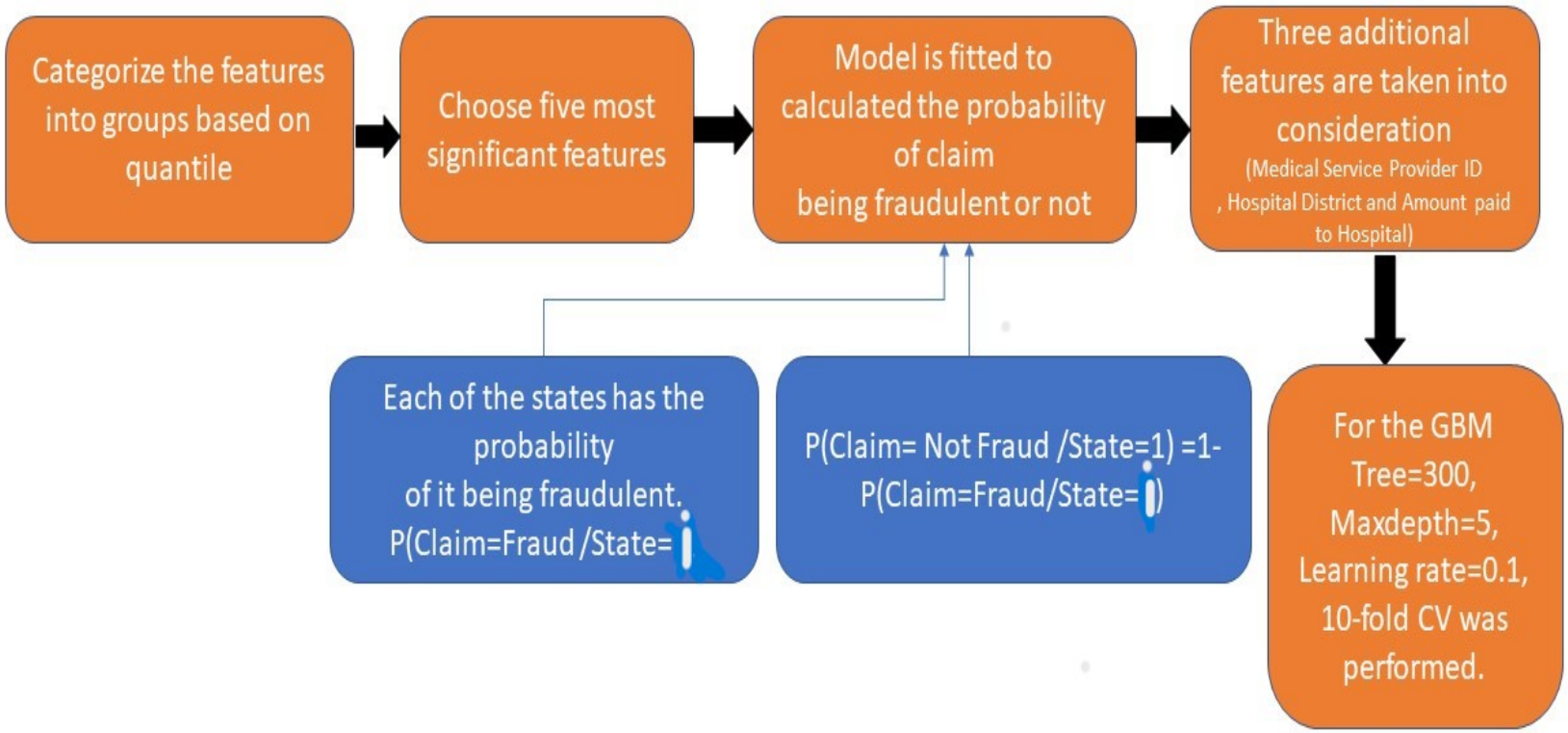
Model	Accuracy	Sensitivity	Specificity	F1 score
BNN	0.8625	0.7898	0.9076	0.8148
RNN	0.617	0.75	0.8523	0.78

Table 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.

Approach 4 :Methodology

Methodology for Markov Model With Gradient Boosting Method(GBM)



Appraoch 4: Markov Model

- Significant Variables

Variable
Patient_Age
InscClaimAmtReimbursed
Hospitalization_Duration
PerProviderClmProcedureCode_4_count_OPAnnualDeductibleAmt
Patient_Risk_Score

- Each of the above feature is categorized into groups based on quantile. Such that the groups had equal number of claims in it.
- The sequence of values taken in the feature was labelled into states
- Each of the claim has a class label as fraud or not-fraud.
- P(PotentialFraud=1 /State=1) is calculated.
- P(PotentialFraud=0/State=1) =1-P(PotentialFraud=1 /State=1) Similarly we can find the respective probabilities for all the remaining states.

Appraoch 4: Markov Model

We can calculate these probabilities by two methods.

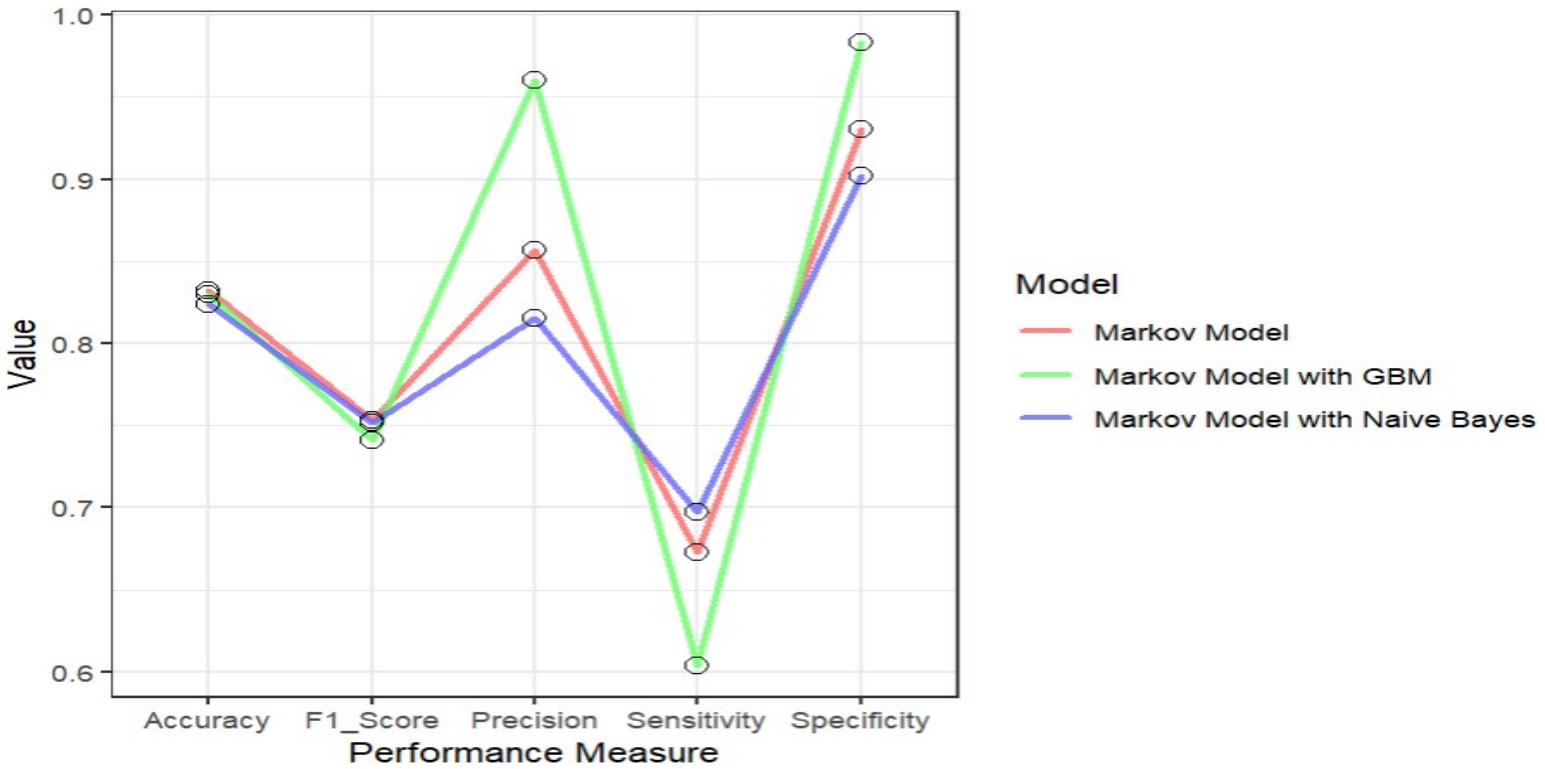
1. **By Naive Bayes Algorithm** (By Bayes Theorem)
2. **By One-Step Transition Probability Matrix.**

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

Markov Model with GBM

- Limitation of Markov Model is that the features had to be categorized into buckets.
- Here, No need of feature categorization
- Previous five important variables + PerProviderClmDiagnosisCode_3_count_Patient_Age + PerProviderClmProcedure-Code_5_count_InscClaimAmtReimbursed.
- For GBM Modelling a total of 300 trees were used
- Interaction depth (max. 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.

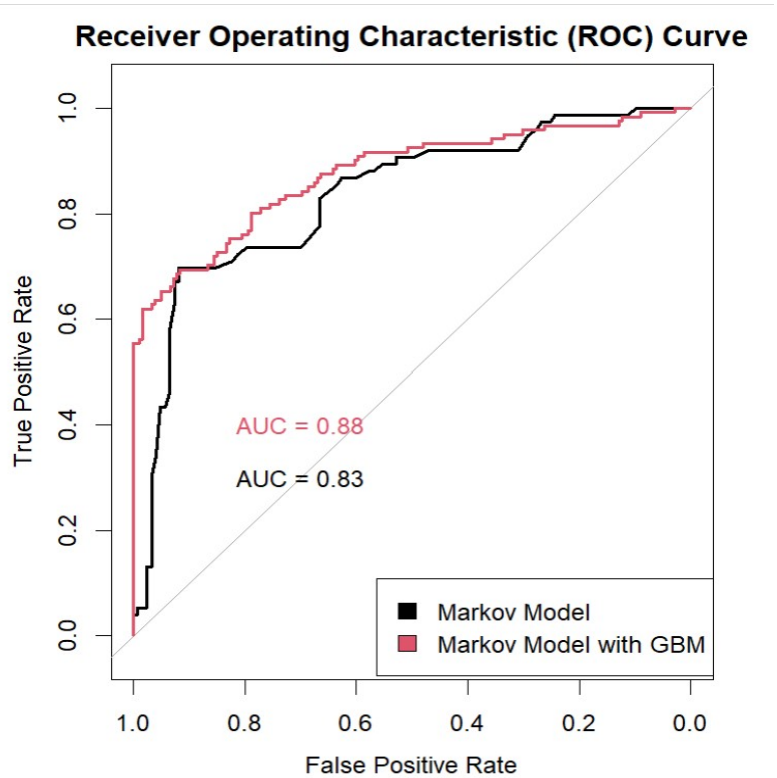
Results of Markov Model



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 7: Performance Metrics for Markov Based Models

ROC and AUC for Markov and Markov with GBM



Conclusions:

- AUC is improved from 0.83 to 0.88 after implimenting GBM with Markov model

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 with 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.

Conclusions

- Also, Random forest gave us AUC of 0.96 which implies that,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**.

Scope

- 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.
- In addition, our suggested models for fraud detection can also work in credit card fraud ,telecommunication fraud, computer intrusion and scientific fraud industries for reducing fraudulent transactions.
- One can go for more sophisticated Markov models of higher orders.Hidden Markov Model (HMM) can also be used on the same dataset and see whether there is significant improvement in the performance.

Limitations

- 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.
- Dataset use in our analysis belongs to US(Results for indian health insurance market might be different)
- 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.

References

[1] Mr G. "Fraud Detection in MeDInsu using Machine Learning Algorithms - A Web Application". In: *International Journal for Research in Applied Science and Engineering Technology* 7 (Oct. 2019), pp. 689–693. DOI: 10.22214/ijraset.2019.10105.

[2] Rohan Yashraj Gupta et al. "Markov model with machine learning integration for fraud detection in health insurance". In: *arXiv preprint arXiv:2102.10978* (2021).

[3] ROHIT ANAND GUPTA. *HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS @ONLINE*. 2017. URL: https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis.

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

[5] Healthcare Fraud Prevention Partnership. *Healthcare Fraud Prevention Partnership*. Accessed on May 18, 2023. 2021. URL: https://hfpp.org/.

[6] Gareth James et al. *An Introduction to Statistical Learning with Applications in R*. New York, NY: Springer, 2013. ISBN: 978-1461471370.

[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*. Vol. 2516. 1. AIP Publishing LLC. 2022, p. 240006.

[8] OpenAI. *OpenAI ChatGPT*. Accessed on May 18, 2023. 2021. URL: https://openai.com/research/chatgpt.

[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)*. IEEE. 2015, pp. 1–5.

Thank You !