Project Summary

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

Acknowledgement

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

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

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

TABLE OF CONTENTS

TABLE OF CONTENTS

1. INDUSTRY REVIEW

1.1 BACKGROUND RESEARCH:

Health Care Industry:

Healthcare industry primarily focuses on the maintenance and progression of health which involves the diagnosis, treatment and prevention of diseases in humans. Every country has its unique healthcare system.

Health Care Industry and Machine Learning:

Machine Learning (ML) is in boom when it comes to the Healthcare industry. The prolific increase in the application of machine learning in the healthcare industry directly reflects a glimpse of the future. Some of the applications aid in identifying the Diseases and Diagnosis, Drug Discovery and Manufacturing, Medical Imaging Diagnosis, Smart Health Records and Outbreak Prediction.

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

Health Insurance Schema in US:

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

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

Medicare program providing healthcare insurance to individuals 65 and older or those under 65 who meet eligibility requirements. In 2016, the US spent $3.4 trillion on health care expenditures [1].

Fraud in the Healthcare Industry:

Fraud is an intentional deception or misrepresentation of fact that can result in unauthorized benefit or payment.

The National Health Care Anti-Fraud Association estimates conservatively that health care fraud ranges from 3 to 10 percent of the nation's total health care expenditures. According to the above percentages, fraud, waste and abuse would range from $102 billion to $340 billion [1].

Experts project that health care expenditures will soar as high as $5.5 trillion by 2025 [2]. Tracing back to 2009, it is reported that Medicare fraud alone is $100 billion annually [3].

Types of Fraud in HealthCare [4]:

The fraudulent are happening at wide verse and stay as a critical problem to articulate.

1. **Fraud by the Service Provider:**
    - Billing for the medical services that are not actually performed
    - Billing for each stage of a medical procedure as a separate treatment
    - Billing for expensive medical services than the one actually performed
    - Performing unnecessary medical services

2. **Fraud by Insurance:**
    - Filing claim for medical services which was not actually received
    - Using another person's insurance coverage
    - Non-disclosure of Pre-Existing Diseases and manufacturing diagnosis reports to justify tests, examinations and surgeries to prove claim worthiness
    - Misrepresenting treatments that are not covered as medically necessary
    - Duplicate submission of a claim for the same service

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

1.2 CURRENT PRACTICES:

Problem Understanding:

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

Current Solution to the Problem: [5] [6]

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

Proposed Solution to the Problem: [5] [6]

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

2. LITERATURE SURVEY

A survey on statistical methods for health care fraud detection
Authors : Jing Li , Kuei-Ying Huang , Jionghua Jin , Jianjun Shi [7]

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

Three kinds of Frauds are mentioned in this paper :

1. Service Providers Fraud
2. Insurance Subscribers Fraud
3. Insurance Carriers Fraud.

The fraud committed by the Service providers is more comparatively.

The paper then tells how one should approach the fraud detection dataset. The following steps such as :

1. Goal Setting
2. Data Cleaning
3. Missing Value Treatment
4. Data Transformation
5. Feature Selection
6. Data Auditing

Once all these steps are done then statistical models for detections are built. There are three types for model building:

1. Supervised Learning
2. Unsupervised Learning
3. Semi-Supervised Learning

Post model building the performance evaluation has to be done..

3. METHODOLOGY

The Methodology followed by the cross industry standard process for Data Mining (CRISP-DM). CRISP-DM provides the complete blueprint for conducting a Data Mining project. It provides a uniform framework for experience documentation.

Fig Source : Wikipedia

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

Phase 1 : Business Understanding

Understanding the business requirement is of paramount importance. The primary activity is to find out exactly what business is trying to accomplish. One needs to go deeper into fact finding to outline the issues in the business goals task.

After understanding the requirements the next steps is to lay down every step that the data scientist intends to take until the project is accomplished and results are presented and reviewed.

Phase 2 : Data Understanding

The data understanding phase goes hand in hand with the business understanding phase. In this phase, the data will be collected, described, explored and verified.

Data Understanding phase is where Exploratory Data Analysis is performed.

Phase 3 : Data Preparation

The third phase is Data Preparation which is also called as Data Wrangling or Data Munging Phase.

It involves developing the final dataset for modelling. It covers all activities to construct the final dataset from the initial raw data. In this phase Data Cleaning, Imputation and Feature Engineering are performed

Phase 4 : Modelling

The fourth phase is Modelling, which involves selecting the actual modelling technique that needs to be used.

In this phase the Modelling technique will be selected, test design will be generated, model will be built. Several models will be built based on tuning iterations until the best model is found

Phase 5 : Evaluation

The fifth phase is Evaluation, where the model will be evaluated .

The steps executed to construct the model to ascertain it properly achieves the business objectives.

Phase 6 : Deployment

The last and final phase is Deployment. However for this project, this step will not be executed.

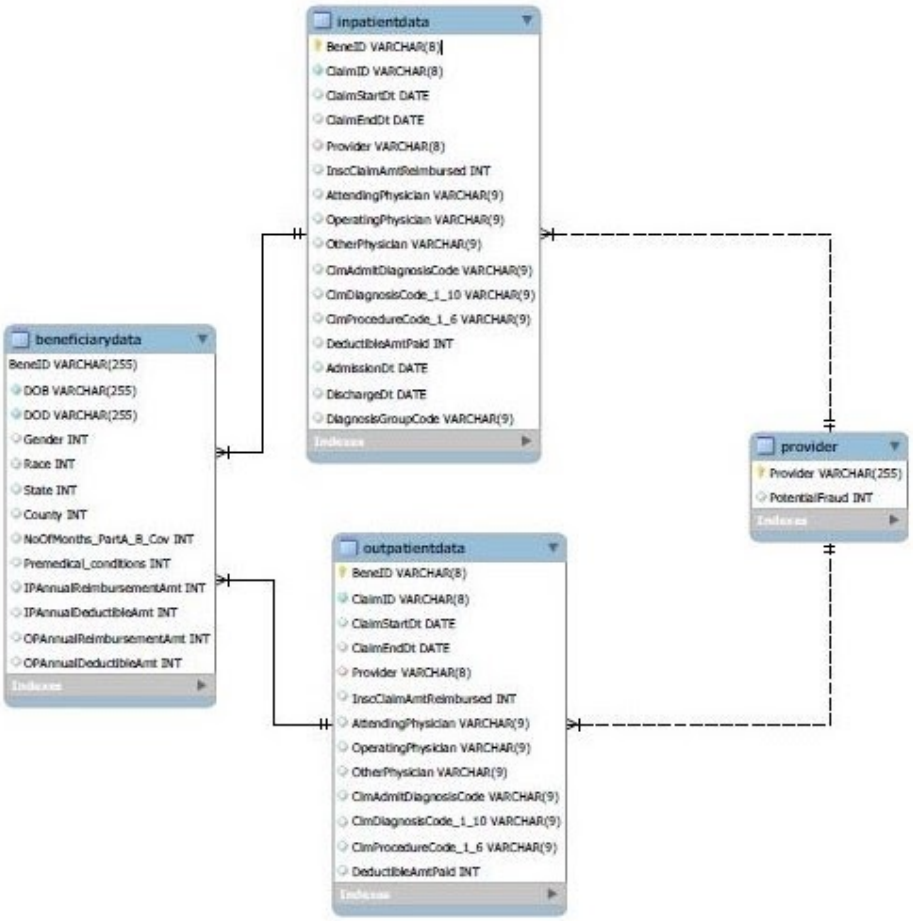
4. DATASET AND DOMAIN

4.1 DATASET SOURCE:

The data used in this project was retrieved from Kaggle which is used to analyse the Insurance Claim Fraud for Medicare[8].

4.2 DATASET OVERVIEW:

The data in its original form consisted of eight different csv files. Four of these files belong to the dataset labeled with potentially fraudulent providers: train beneficiary, train outpatient, train inpatient, and train providers flagged. The other four files belong to the unlabeled dataset (providers not tagged as potentially fraudulent) : test beneficiary, test outpatient, test inpatient, and test providers. The labeled data contained a total of 5,58,211 claims and 5,410 providers.



4.3 DATA DICTIONARY:

1. Beneficiary Details Data: It provides the KYC details of both the Inpatients and Outpatients. The corresponding file contains an exact of 1,38,556 rows and 25 columns.

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

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

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

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

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

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

Inference from the Dataset:

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

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

4.4 VARIABLE CATEGORIZATION:

Numerical Features	6
Categorical Features	8
Object Features	42

4.5 PROBLEM STATEMENT:

The dataset summarizes a set of heterogeneous features about the Insurance Claim Fraud in Medicare in 2009.

The goal of the project is to predict the potential fraudulent providers based on the claims filed by them, thereby saving enormous amounts of money from frauds which would aid those who are in real need.

4.6 TOOLS USED:

- Programming Language: Python
- Visualizations: Python

5. EXPLORATORY DATA ANALYSIS

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

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

5.1 FEATURE ENGINEERING:

S.No	Column Name	Column Description
1	IsDead	Based on the Date of Death, if a patient is Dead it is marked as 1 and if a patient is not Dead it is marked as 0.
2	Claim_Duration	Difference between the Claim End Date and the Claim Start Date
3	AdmitDuration	Difference between the Discharge Date and Admission Date
4	Claimed_Age	Difference between the Claim Start Date and the Date of Birth [10]
5	Insurance_Amount	Difference between the sum of the Patients Annual Reimbursement Amount and Patients Annual Deducted Amount
6	Avg_Ins_Amt_per_bene	Average Insurance Amount Per Beneficiary
7	Claim_Amount	Sum of the Reimbursement Amount and Deductible Amount
8	Avg_Claim_Amt_per_bene	Average Claim Amount Per Beneficiary
9	Total_Claims_Beneficiary	Total Number of Claims for Each Beneficiary

S.No	Column Name	Column Description
10	Total_Provider_Claims_Covered	Total Number of Claims Covered by Each Provider
11	Total_Unique_Physicians	Total Unique Physicians for Each Beneficiary
12	Total_Physicians	Total Number of Physicians for each beneficiary
13	DeadActions	Check if there are any action post the Beneficiary's Death
14	Total_Dis_per_bene	Total Number of Diseases for each Beneficiary
15	Total_Diag_per_bene	Total Number of Diagnosis Code for each Beneficiary
16	Total_procedure_per_bene	Total Number of Procedure Code for each Beneficiary
17	Fraudulent Providers for each Beneficiary	Total Number of Fraudulent Providers for each Beneficiary
18	Beneficiary for each Fraudulent Providers	Total Number of Beneficiary for each Fraudulent Providers

After performing Feature Engineering, we shall drop some of the Redundant columns: ClaimStartDt, ClaimEndDt, AdmissionDt, DischargeDt, DOB, DOD, IPAnnualReimbursementAmt, OPAnnualReimbursementAmt, IPAnnualDeductibleAmt and OPAnnualDeductibleAmt.

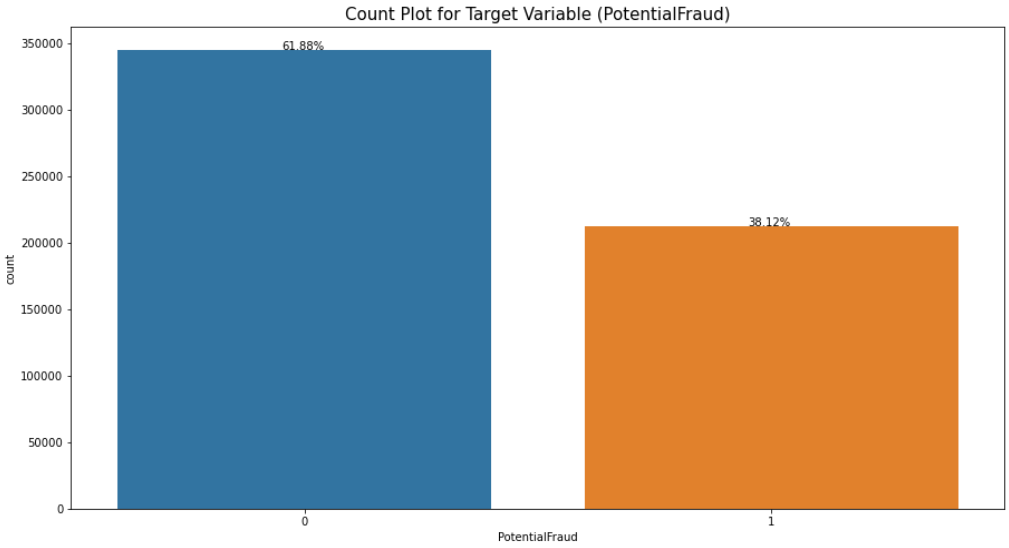
After Feature Engineering, there are **3,46,09,082** Data Points.

5.2 UNIVARIATE ANALYSIS:

Univariate analysis is the simplest form of analyzing data. It doesn't deal with causes or relationships. It's major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.

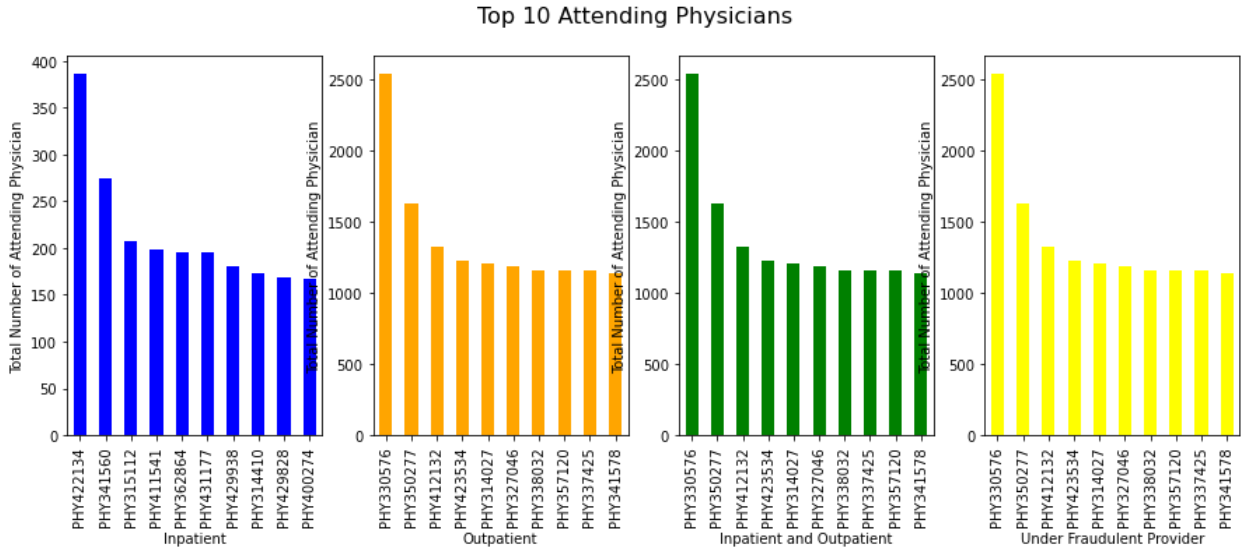
5.2.1 **Target Variable:**

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



From the above distribution, it is noticed that it is imbalanced data between the target variables 0 and 1. SMOTE analysis is not required as the imbalance difference is only 25%.

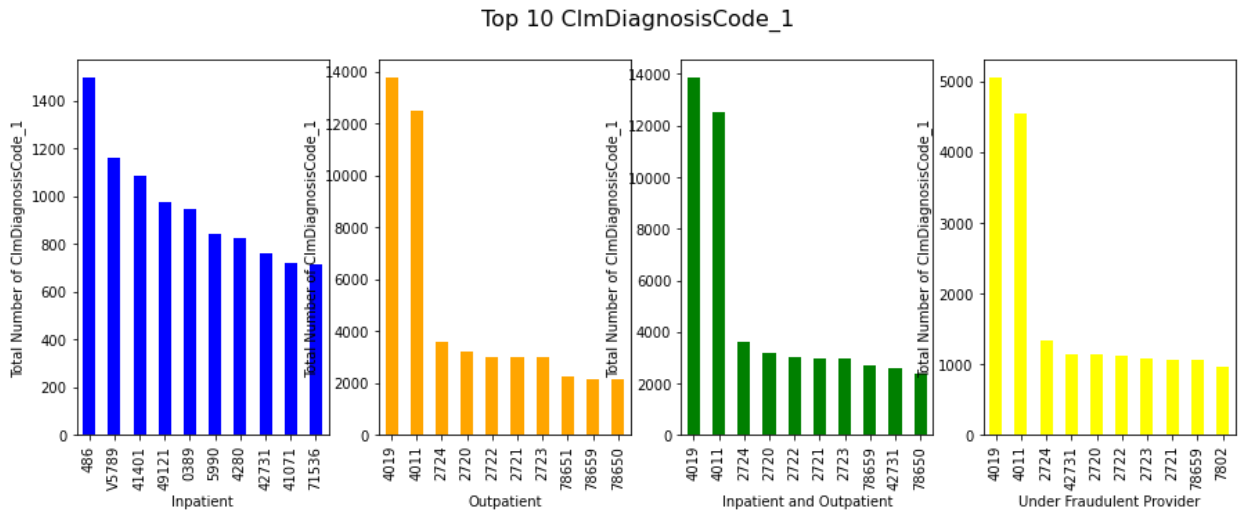
5.2.2 **Attending Physicians:**



From the above plots it is inferred that,

- The top Attending Physician for Inpatient is PHY422134
- The top Attending Physician for Outpatient is PHY330576
- The top Attending Physician for Inpatient and Outpatient is PHY30576
- The top Attending Physician under Fraudulent Provider is PHY30576.

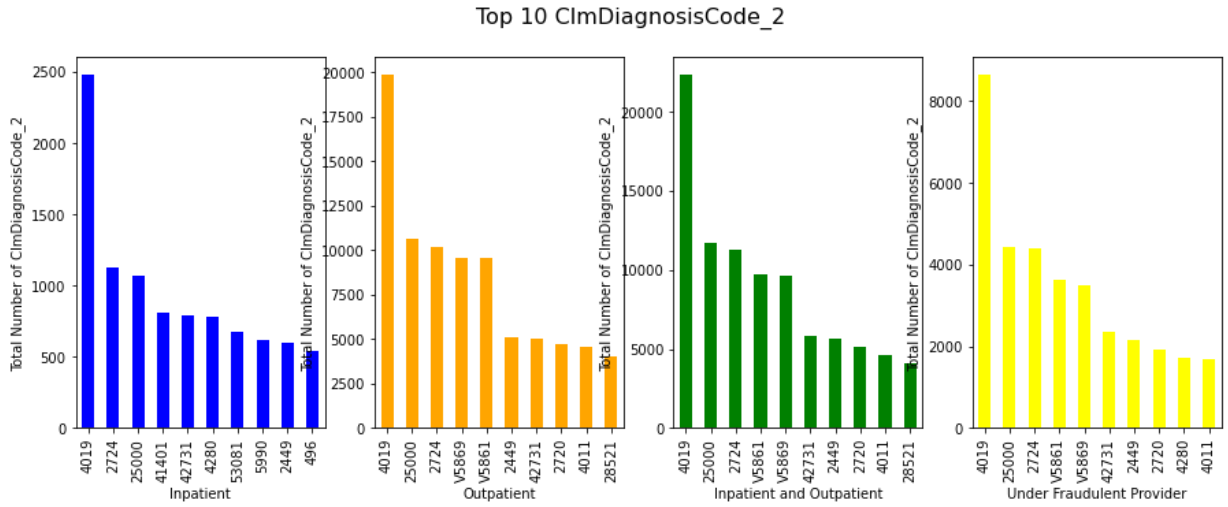
5.2.3 **ClmDiagnosisCode_1 :**



From the above plots it is inferred that,

- The top ClmDiagnosisCode_1 for Inpatient is 486
- The top ClmDiagnosisCode_1 for Outpatient is 4016
- The top ClmDiagnosisCode_1 for Inpatient and Outpatient is 4019
- The top ClmDiagnosisCode_1 under Fraudulent Provider is 4019.

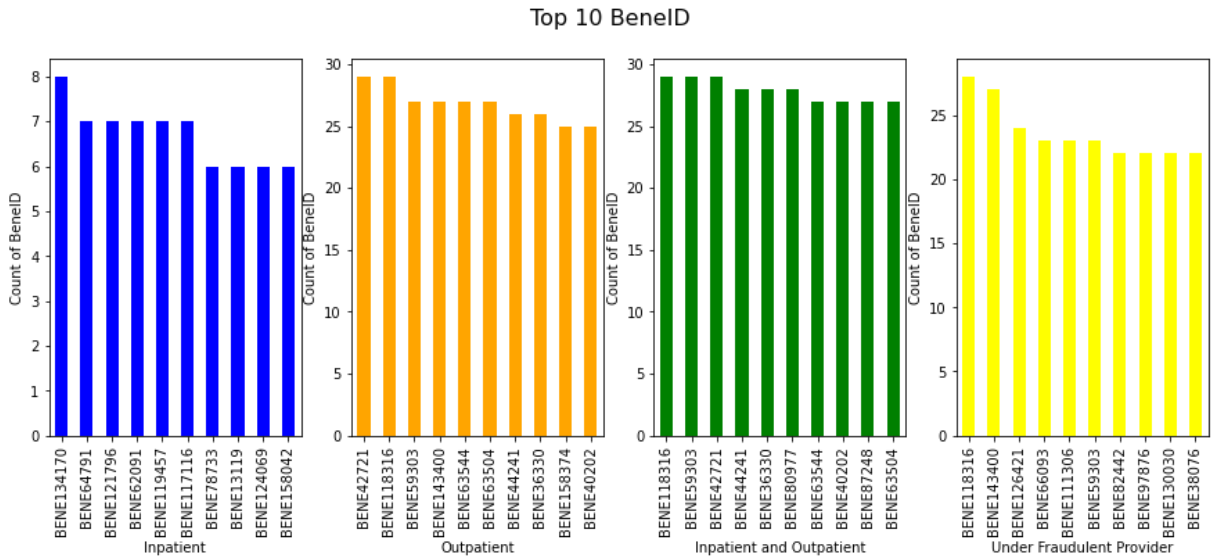
5.2.4 **ClmDiagnosisCode_2 :**



From the above plots it is inferred that,

- The top ClmDiagnosisCode_2 for Inpatient is 4019
- The top ClmDiagnosisCode_2 for Outpatient is 4019
- The top ClmDiagnosisCode_2 for Inpatient and Outpatient is 4019
- The top ClmDiagnosisCode_2 under Fraudulent Provider is 4019.

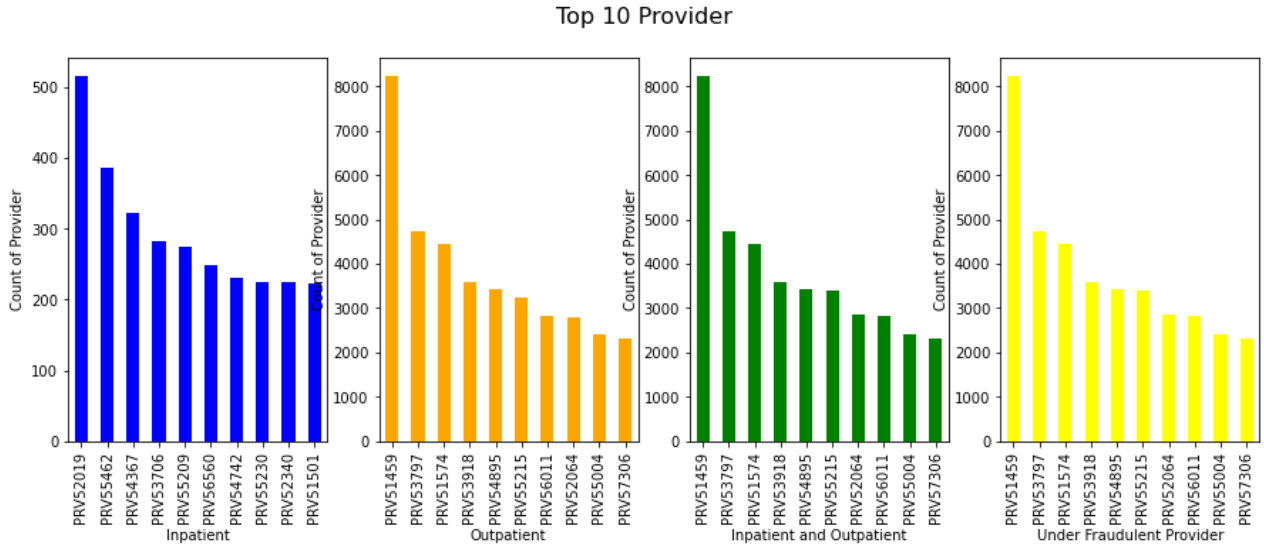
5.2.5 **BeneID:**



From the above plots it is inferred that,

- The BeneID under Inpatient BENE134170 uses this Service to Claim Insurance
- The BeneID under Outpatient BENE42721 uses this Service to Claim Insurance
- The BeneID under Inpatient and Outpatient BENE59303 uses this Service to Claim Insurance
- The BeneID under Fraudulent Provider BENE118316 uses this Service to Claim Insurance.

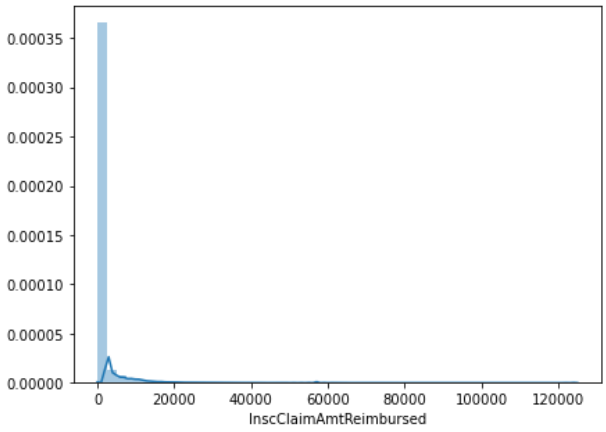
5.2.6 **Provider:**

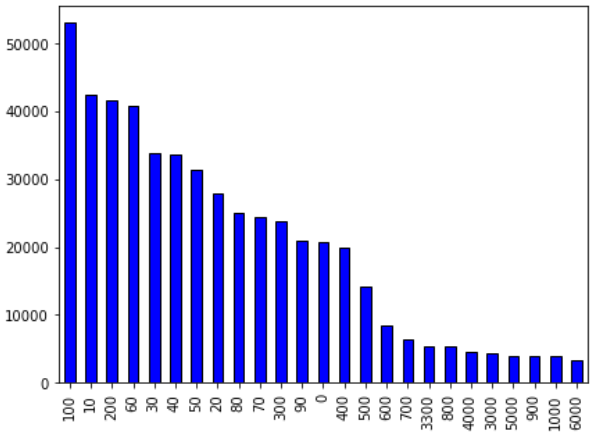


From the above plots it is inferred that,

- The Provider PRV52019 under Inpatient uses this Service to Cover the Claim by Beneficiary
- The Provider PRV51459 under Outpatient uses this Service to Cover the Claim by Beneficiary
- The Provider PRV51459 under Inpatient and Outpatient uses this Service to Cover the Claim by Beneficiary
- The Provider PRV52019 under Fraudulent Provider uses this Service to Cover the Claim by Beneficiary.

5.2.7 **InscClaimAmtReimbursed:**





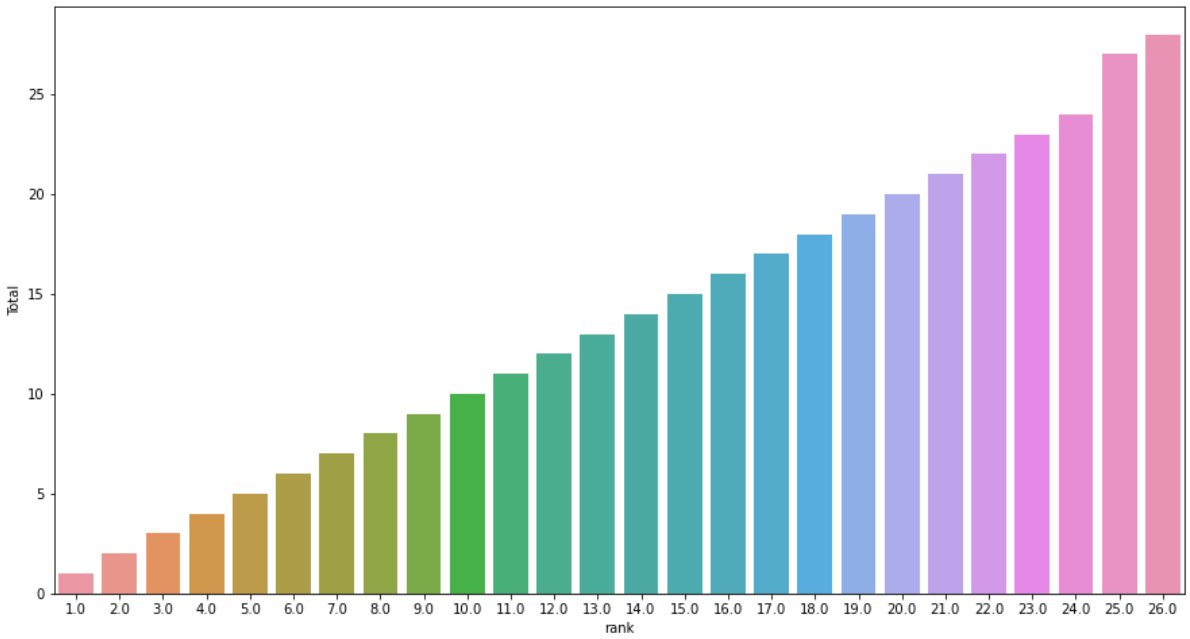
From the above plots it is inferred that,

- Distribution of Insurance Claim Amount Reimbursed seems like log normal distribution
- All most all Reimbursed amount is between 0 and 20000
- In very few cases amount more than 20000 is paid for claim reimbursement.

5.3 BIVARIATE ANALYSIS:

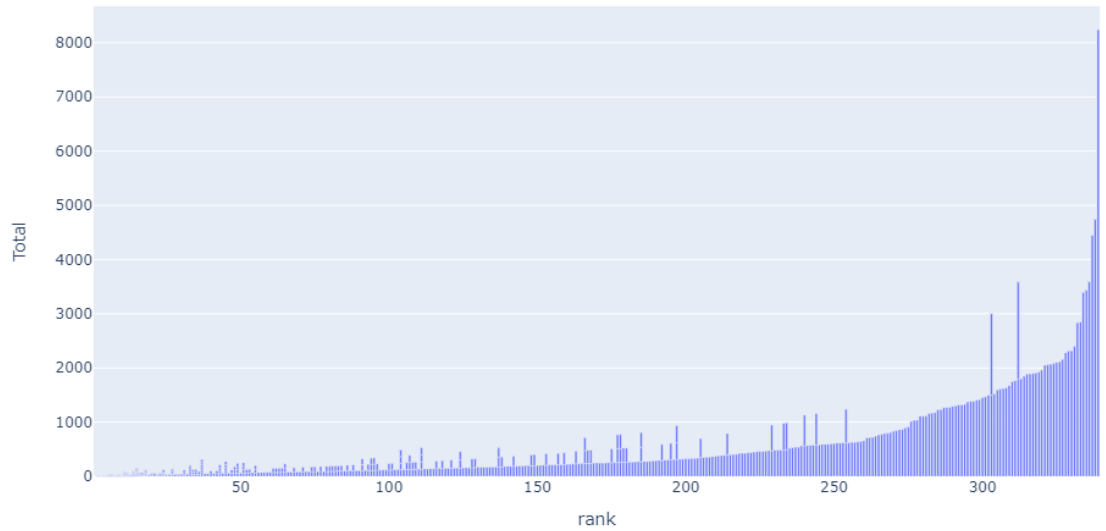
Bivariate analysis means the analysis of bivariate data. It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values. It usually involves the variables X and Y.

5.3.1 **Top Fraudulent Providers for each Beneficiary:**



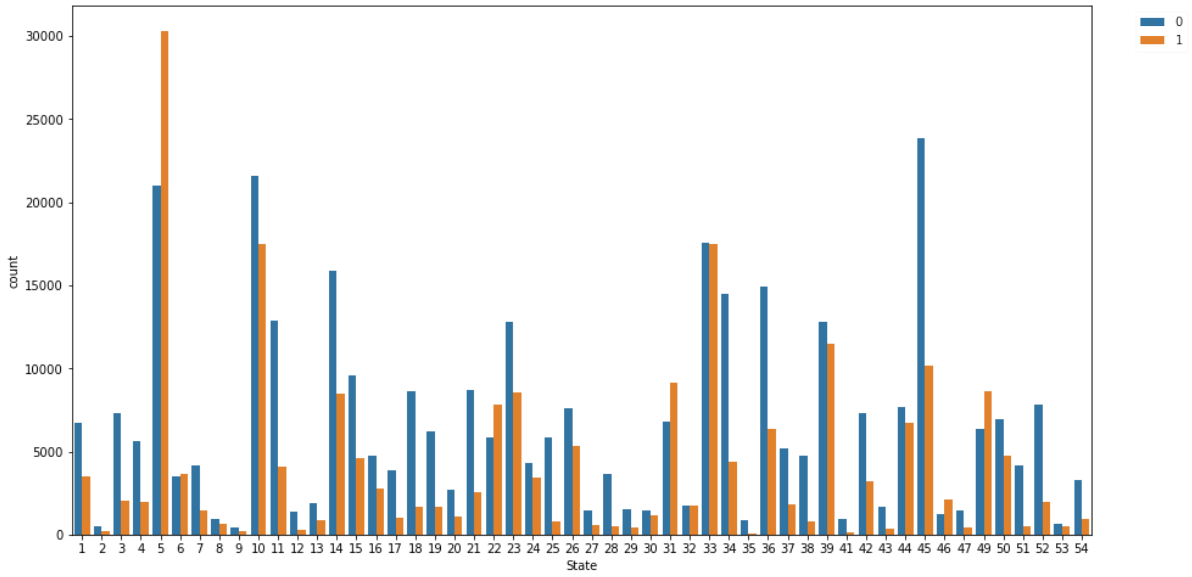
From the above plot it is inferred that, the Provider with Rank 26 has the highest Number of frauds for the Beneficiaries.

5.3.2 **Top Beneficiary for each Fraudulent Providers:**



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

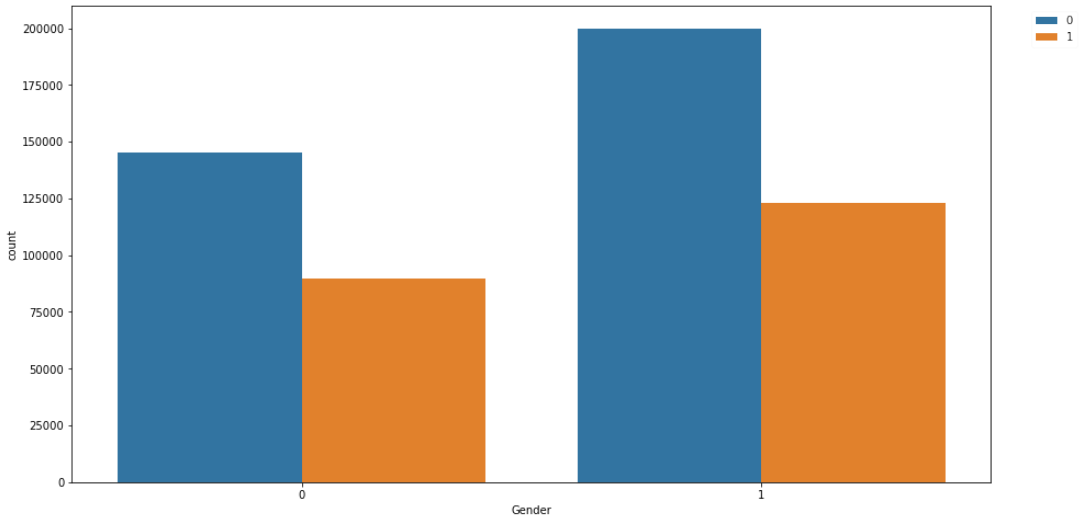
5.3.3 **State and Potential Fraud:**



From the above plot it is inferred that, State with Code 5 has the highest number of fraud claims when the potential fraud has been encountered.

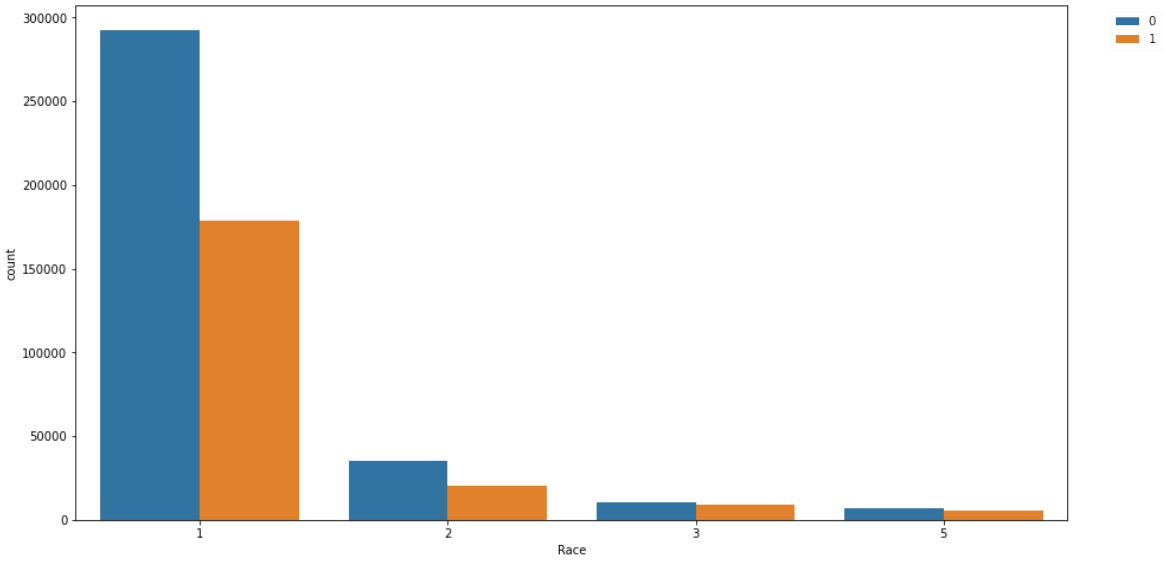
5.3.4 **Gender and Potential Fraud:**

Before plotting the distribution, we need to encode it such that the Gender variable has 0 and 1.



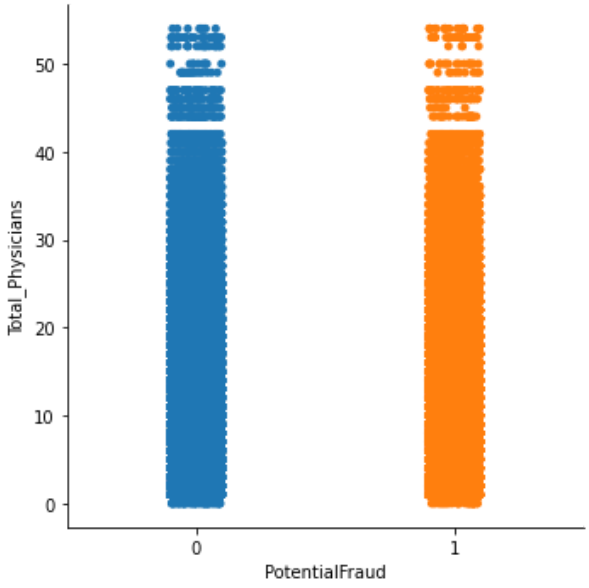
From the above plot it is inferred, that Gender Coded as 1 has the highest number of fraud claims when the potential fraud has been encountered.

5.3.5 **Race and Potential Fraud:**



From the above plot it is inferred, that Race Coded as 1 has the highest number of fraud claims when the potential fraud has been encountered.

5.3.6 **Total Physician and Potential Fraud:**



From the above plot it is inferred that Physicians have an impact on potential fraud.

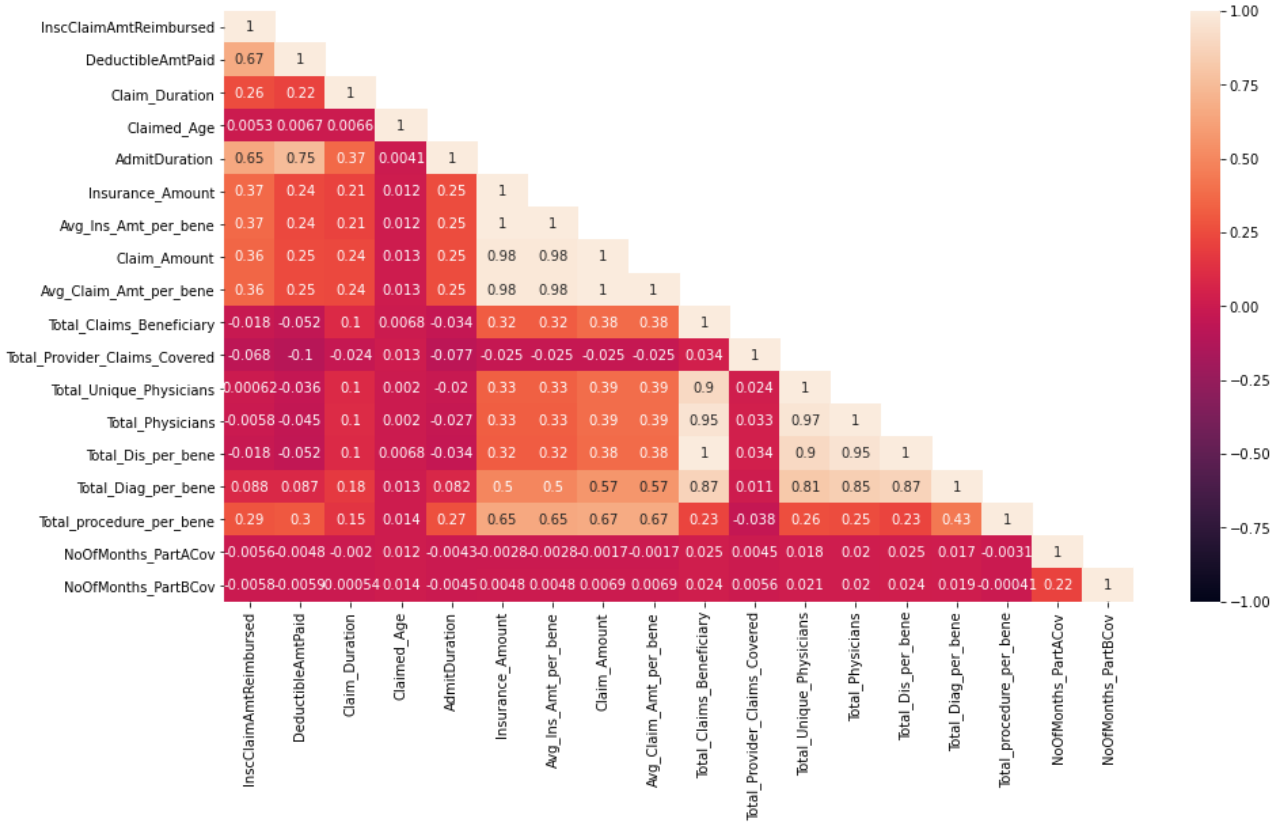
5.4 MULTIVARIATE ANALYSIS:

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

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

The heatmap in the picture below shows how independent features are correlated with each other. The scale on the right side indicates the level for the correlation. Anything close to 1 is highly correlated.

In other words, those two features represent the same insights. By design, a feature when checked for a correlation with itself will always be 1.



The corresponding features have strong correlation as their correlation value is greater than 0.8 Claim Amount, Insurance Amount, Avg Claim Amt Per Bene, Avg Ins Amt Per Bene Total Physicians, Total Claims Beneficiary, Total Dis Per Bene, Total Diag Per Bene.

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

This gives an indication for the presence of multicollinearity.

5.5 MISSING VALUE ANALYSIS:

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

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

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

Missing Values have been treated in three ways.

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

Treating the Missing Value Columns:

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

2. **AttendingPhysician**: It would be unjust to fill this column by considering the entire dataset as one single group. But in reality this dataset can be segregated into two groups. 'yes_fraud' group that will have all the data pertaining to those providers who are fraud. And a 'no_fraud' group that will have data pertaining to those providers who are not fraud. AttendingPhysican is filled in such a fashion that by taking the mode of each group separately.

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

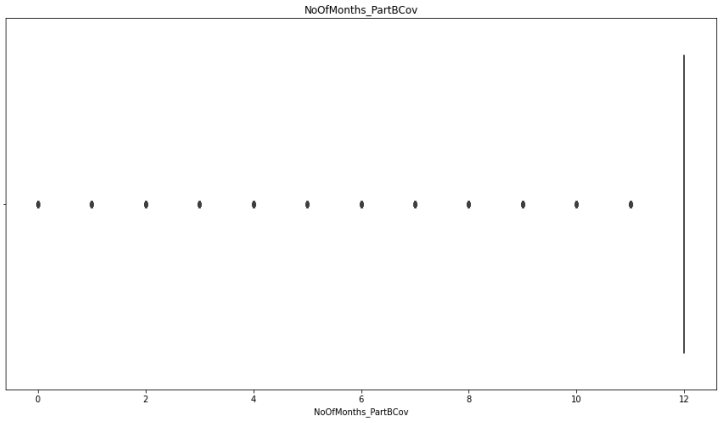
4. **ClmDiagnosisCode_2** : This feature gets filled by taking the mode of the 'yes_fraud' group separately and 'no_fraud' group separately just similar ways of filling done in AttendingPhysician.

5.6 OUTLIER TREATMENT:

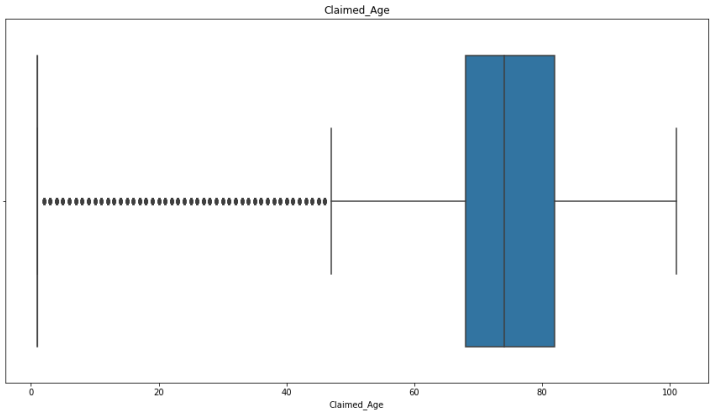
There are 18 numerical columns in this dataset. Boxplot is used to visualize these numerica columns to check for the presence of outliers.

From these 18 numerical columns , it can be seen that 11 are continuous columns and the rest are discrete numerical columns.

Example for a Discrete Column



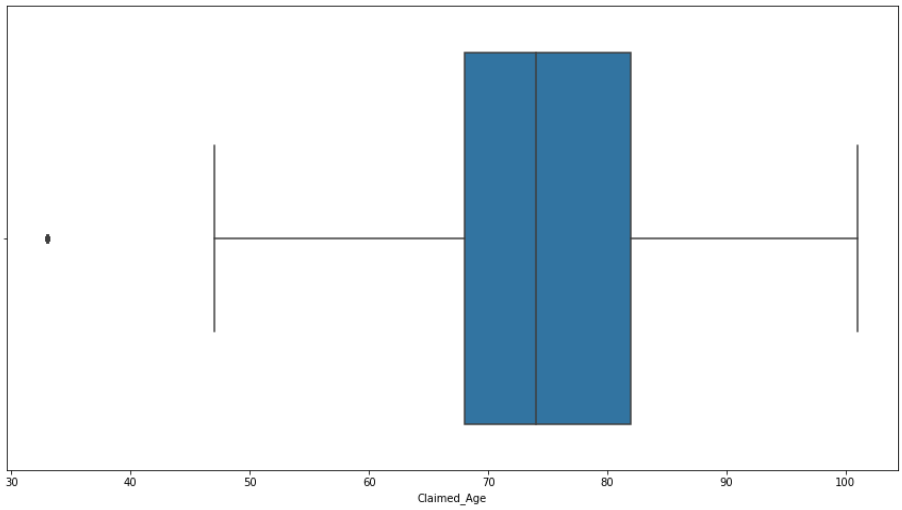
Example for a Continuous Column



11 columns requires outlier treatments, which are

1	Total_Physicians	7	Total_Claims_Beneficiary
2	Claimed_Age	8	Total_Provider_Claims_Covered
3	Insurance_Amount	9	Total_Unique_Physicians
4	Avg_Ins_Amt_per_bene	10	Total_Dis_per_bene
5	Claim_Amount	11	Total_Diag_per_bene
6	Avg_Claim_Amt_per_bene		

Winsorization technique has been used for treating the outliers.



5.7 ENCODING CATEGORICAL DATA:

This dataset has a combination of Numerical and Categorical Variables.

1. The **premedical conditions** fall under the classification of either Yes and No.

   Different methods of encoding such as Label Encoding or One Hot Encoding were tried, but integrity of the data was lost. So the ideal way to encode the variables was replacing 1 for Yes and 0 for 2.

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

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

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

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

The columns AttendingPhysician, ClmDiagnosisCode_1 and ClmDiagnosisCode_2 are dropped after the encoding is completed.

6. STATISTICAL SIGNIFICANCE OF VARIABLES

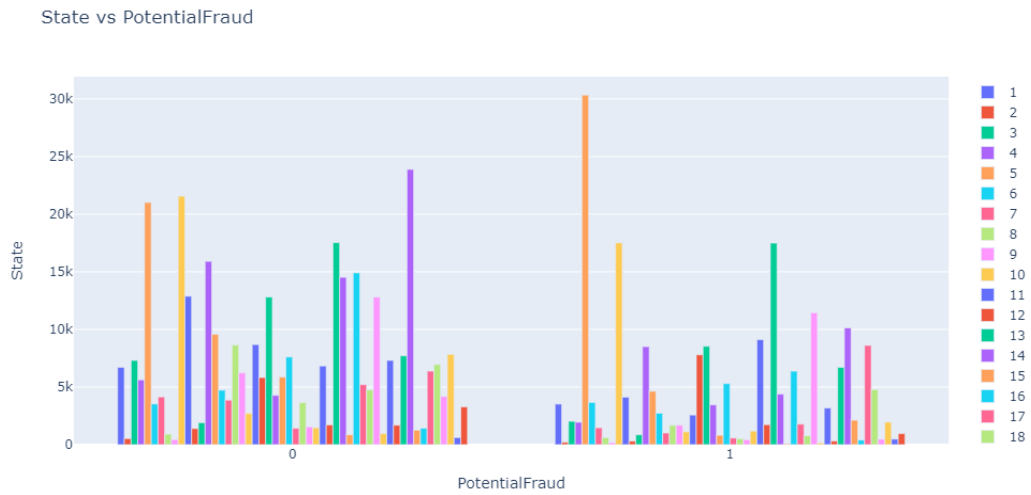
Statistical significance tests are designed to address the problem and quantify the likelihood of the samples of skill scores being observed in given the assumption that they were drawn from the same distribution. Statistical learning theory deals with the problem of finding a predictive function based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

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

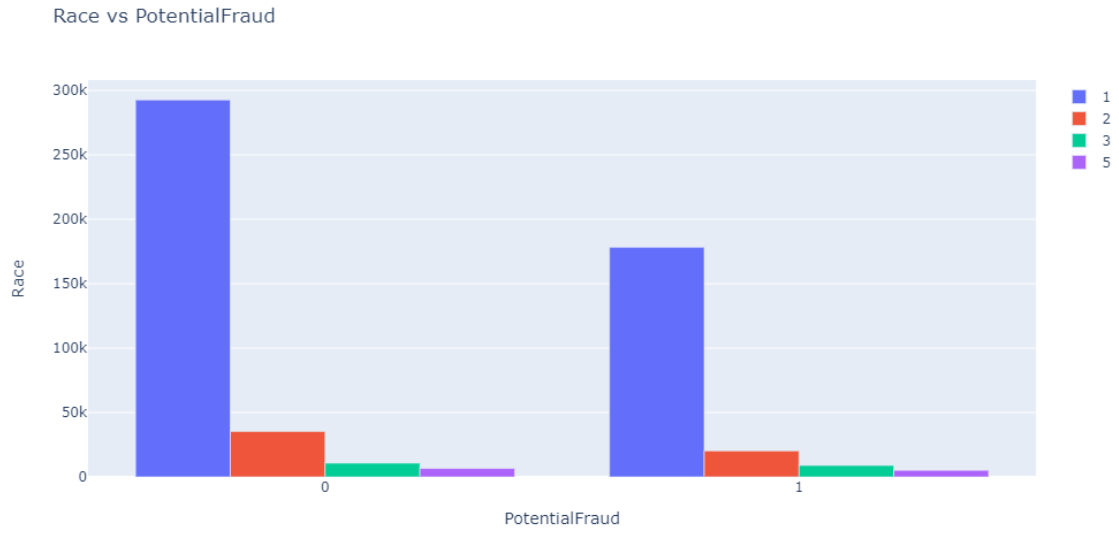
If Null Hypothesis is selected, the feature that we used for the test is not significant to build the model, because there is no use to predict the potential fraud. If it satisfies Alternate Hypothesis it is useful in predicting the potential fraud.

Statistical Representation of the Columns with Target Variable:

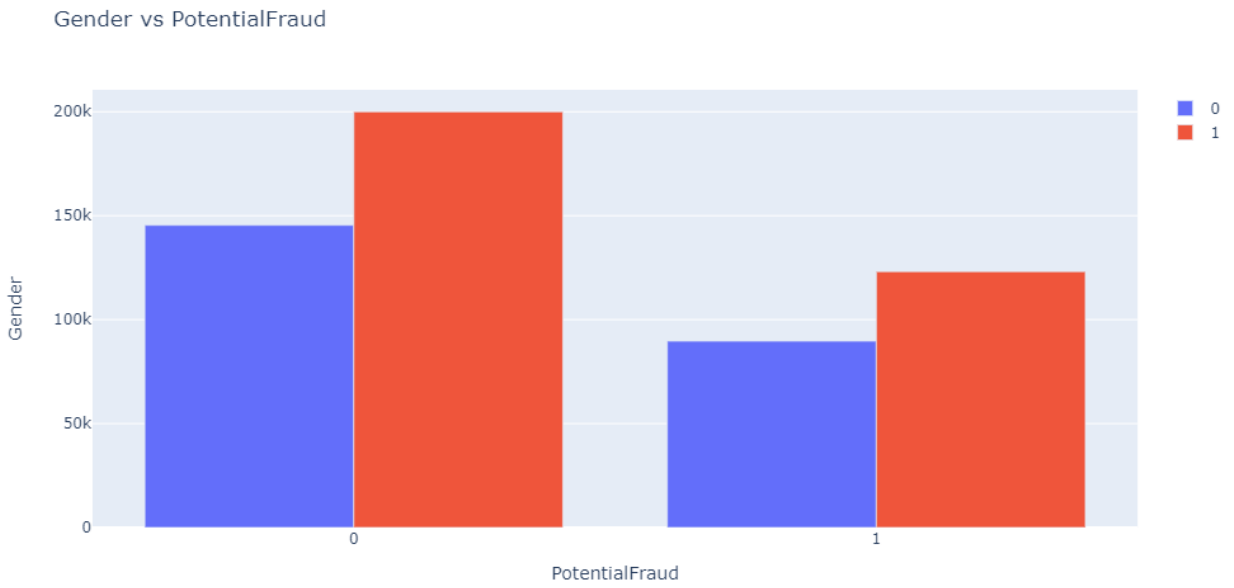
6.1 State and Potential Fraud:



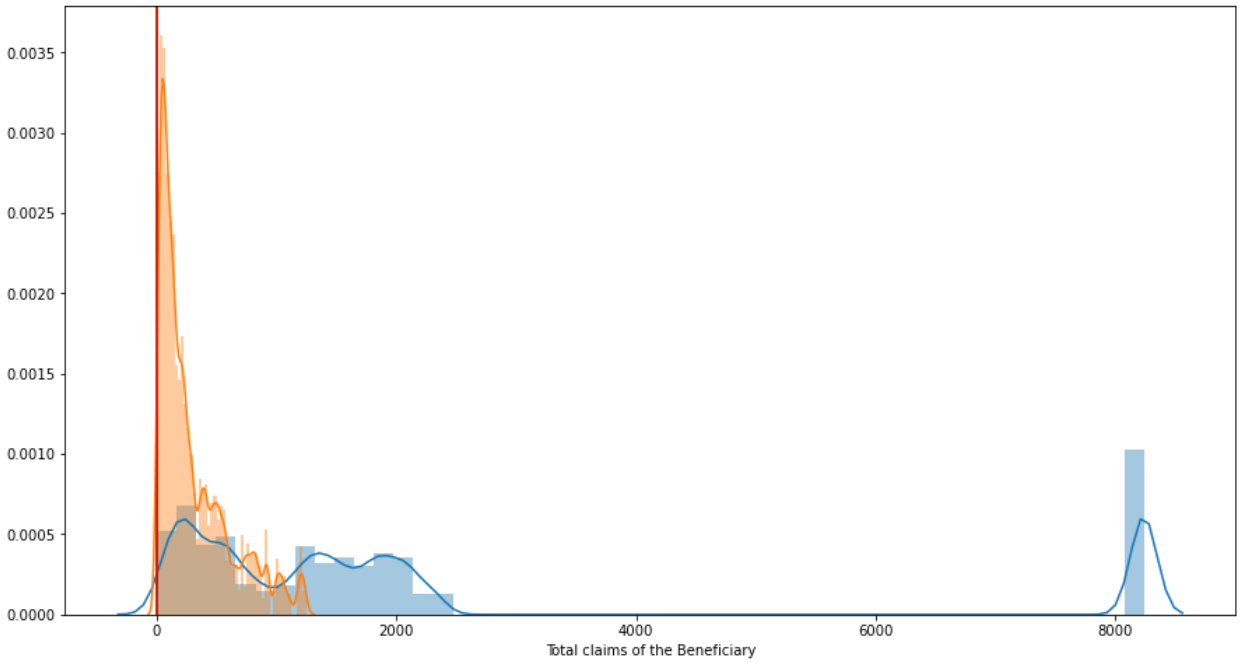
6.2 Race and Potential Fraud



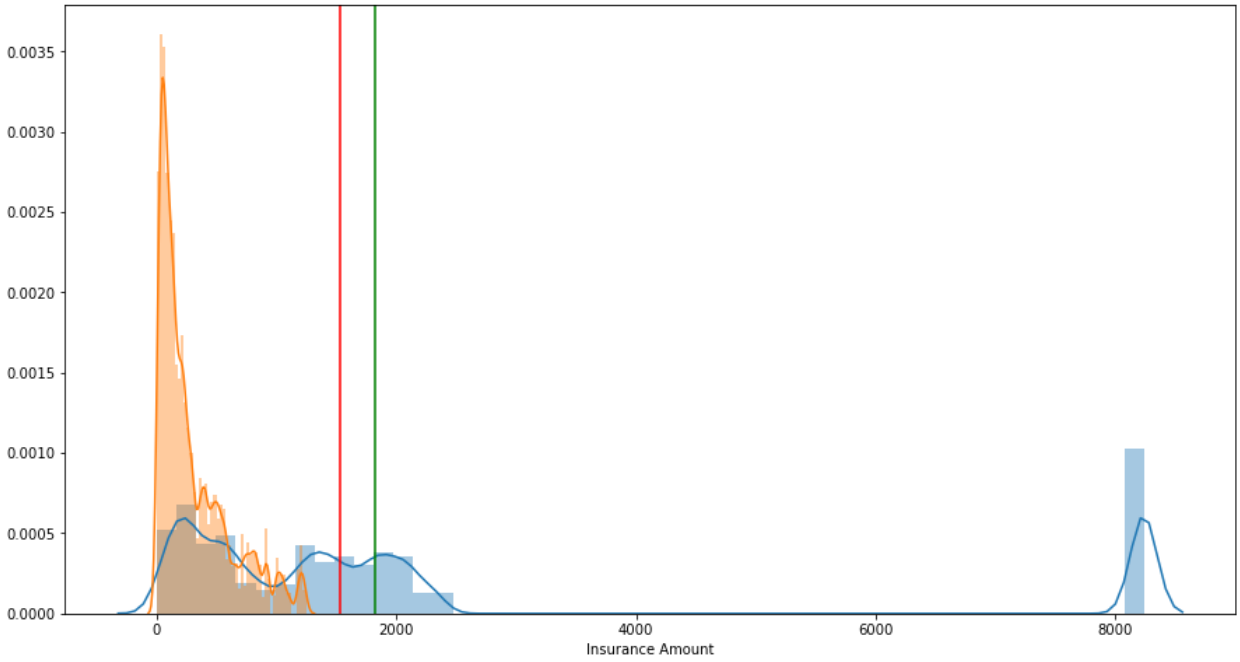
6.3 Gender and Potential Fraud



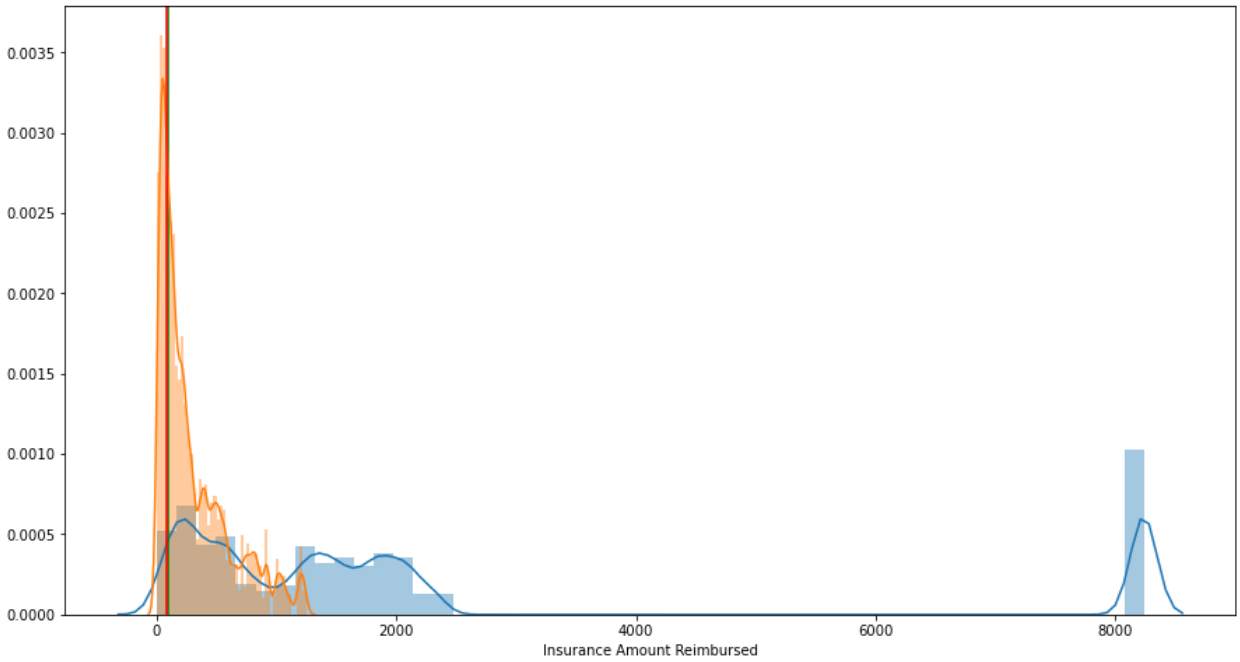
6.4 Total Claims Covered and Potential Fraud



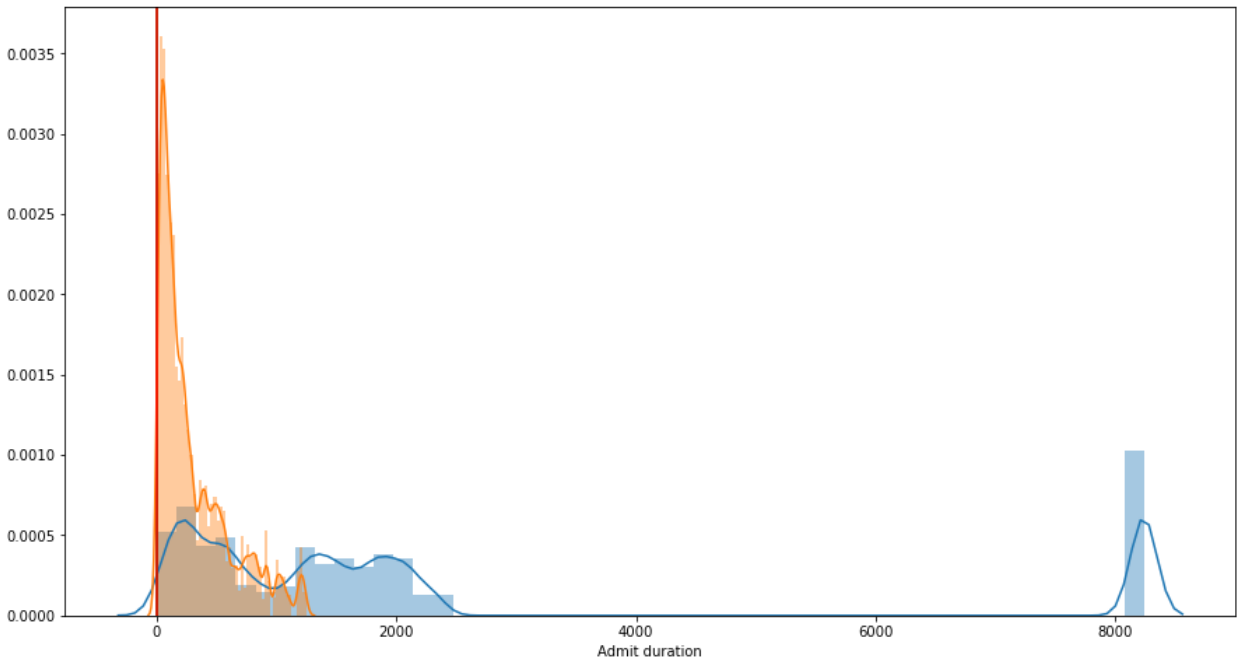
6.5 Insurance Amount and Potential Fraud



6.6 Insurance Amount Reimbursed and Potential Fraud



6.7 Admit Duration and Potential Fraud



6.7 Is Admitted and Potential Fraud

Since the Data is not normal for the Numerical Columns, MannWhitneyU test is performed and for the Categorical Columns Chi Square Contingency test is performed.

	Column	Stat Value	Pvalue	Significance
0	State	4.491223e+04	0.000000e+00	Significant
1	County	4.419137e+04	0.000000e+00	Significant
2	Race	7.039415e+02	2.933423e-152	Significant
3	Gender	1.159574e-01	7.334611e-01	Insignificant
4	Total_Provider_Claims_Covered	1.084865e+10	0.000000e+00	Significant
5	Total_Claims_Beneficiary	3.574426e+10	3.106159e-67	Significant
6	Insurance_Amount	3.520054e+10	1.557709e-155	Significant
7	InscClaimAmtReimbursed	3.462854e+10	7.575701e-290	Significant
8	AdmitDuration	3.452667e+10	0.000000e+00	Significant
9	Is_Admitted	7.177629e+03	0.000000e+00	Significant

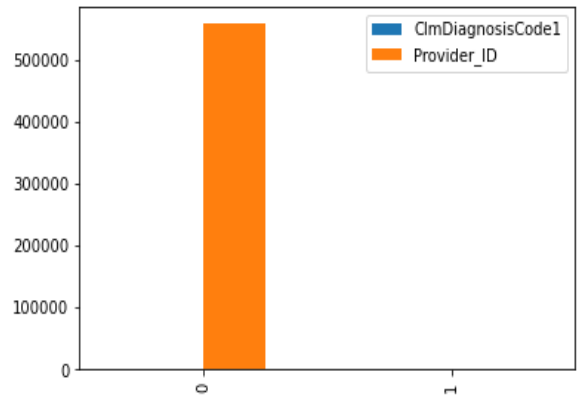
7. INTERESTING FINDINGS

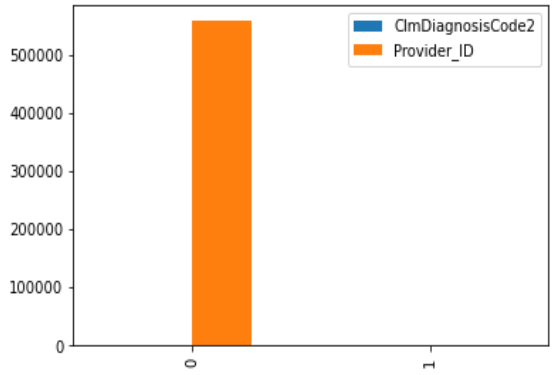
Finding 1 : Correlation with Beneficiary ID and Provider With Respect To the Claim Diagnostic Code

Correlation of the Bene_ID, Provider_ID, ClmDiagnosisCode1 and ClmDiagnosisCode2 variables has been performed inorder to identify the pattern which influences each other.

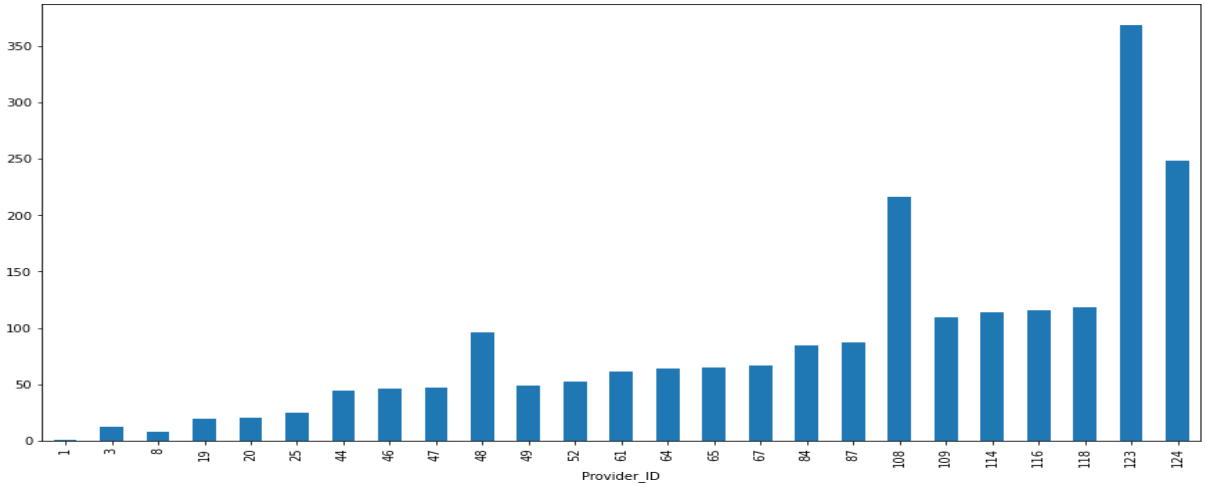
Inorder to find the correlation between the variables Label Encoding has been done for BeneID, ProviderID and Claim ID.

Bene_ID and Provider_ID has been focussed on determining the plots which affects the claim diagnosis ID on different views.

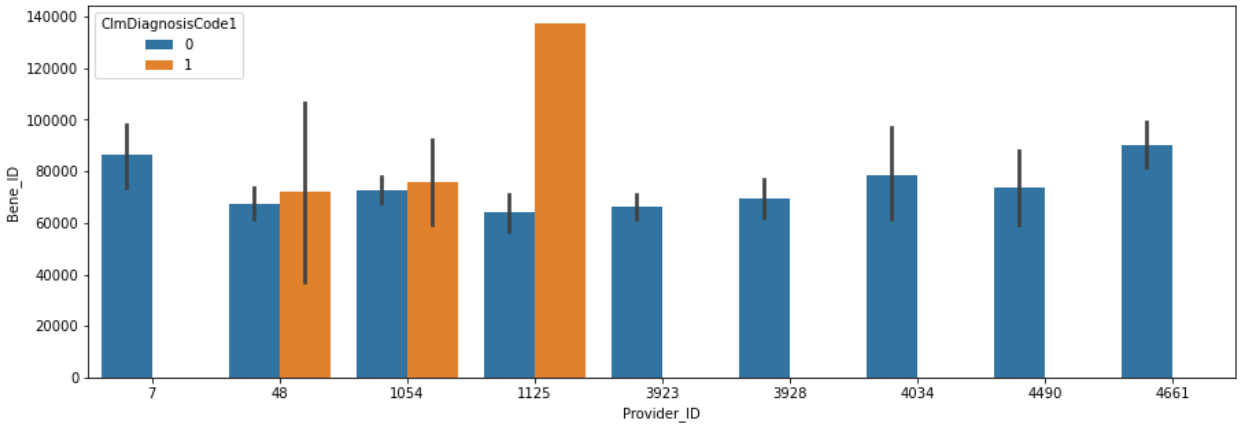




On comparing the Provider_ID with ClmDiagnosisCode1 & ClmDiagnosisCode2 , we can infer that >500000 providers have not registered for the potential fraud incident. However, < 1 % of the providers have registered for the potential fraud incidents.

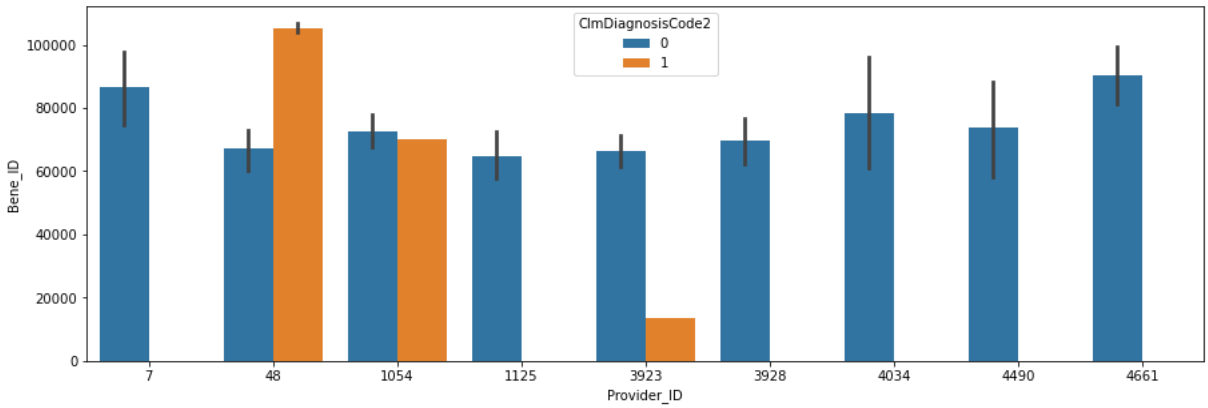


Provider_ID frequency has been plotted to infer that the provider with ID 67 has marked for highest number of fraud claims from the top providers.



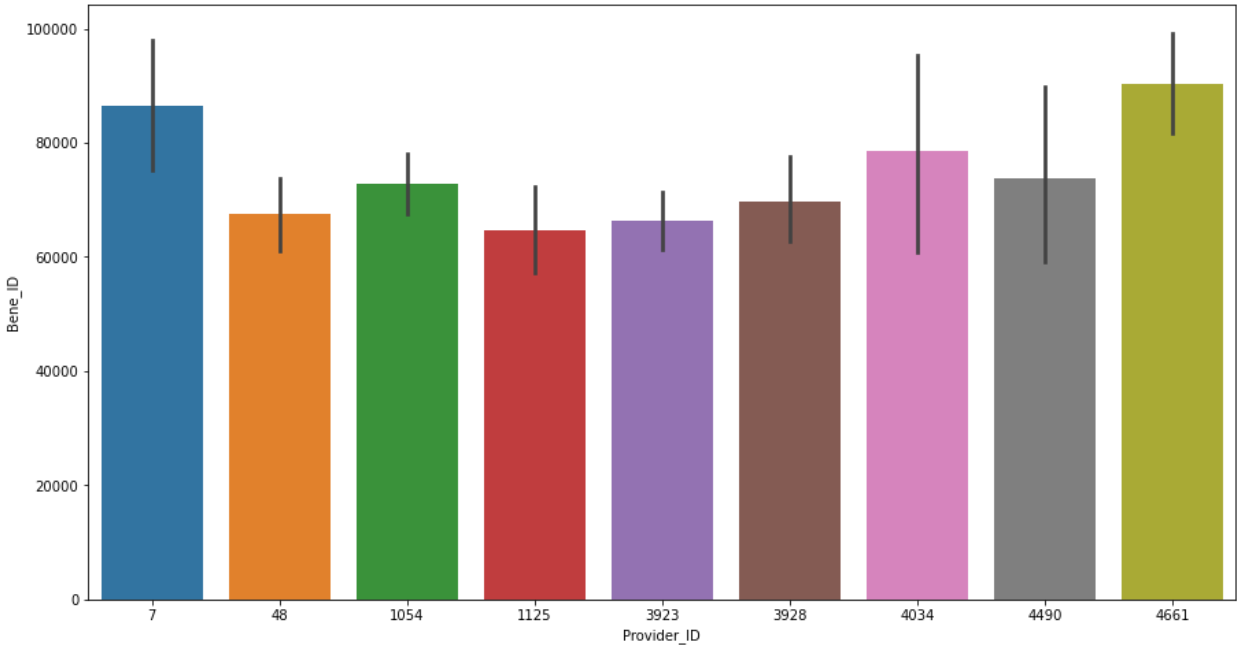
The above plot determines that segmentation of the fraud claim Clam Diagnosis Code 1 (ie... 0 or 1) on comparing with beneficiaries and the providers.

We could infer that the provider 1125 has the highest value count with respect to the beneficiary of determining the fraudulent claims.



The above plot determines the segmentation of the fraud claim (ie... 0 or 1) on comparing the beneficiary ID and the provider ID.

We could infer that the provider 48 has the highest value count with respect to the beneficiary of determining the fraudulent claims.



The above plot gives the statistics of the highest number of providers incorporating the beneficiary on claiming the fraudulent.

From that we could infer that the provider ID 4661 has quoted the highest number of fraud claims with the Beneficiary Id.

Finding 2 : Identifying the Common Beneficiary for Inpatient and Outpatient

Once the files were merged we noticed that around 5% (26713) of the total Beneficiaries were common for both the Inpatient and Outpatient Data.

Finding 3 : Calculating the Total Money Lost

Based on the information provided in the dataset, we were able to calculate the Total Money Lost is $ 29, 56, 81, 120.

The columns Bene_ID, Claim_ID are dropped after the findings. The columns BeneID, ClaimID and Provider are dropped before the model building.

Now there are 42 columns and 5,58,211 rows accounting for 2,34,44,862 Data Points.

8. LOGISTIC REGRESSION MODEL

8.1 SCALING THE DATA:

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

8.2 ASSUMPTIONS OF LOGISTIC REGRESSION MODEL:

Multicollinearity is a situation where two or more predictors are highly linearly related. In general, an absolute correlation coefficient of >0.7 among two or more predictors indicates the presence of multicollinearity.

After treating Mulit-collinearity:

	VIF	Features
10	3.593459	Total_Diag_per_bene
9	2.985237	Total_Unique_Physicians
6	2.709901	AdmitDuration
1	2.607826	DeductibleAmtPaid
0	2.089099	InscClaimAmtReimbursed
7	1.910804	Avg_Ins_Amt_per_bene
11	1.738308	Total_procedure_per_bene
4	1.212408	Claim_Duration
3	1.049248	NoOfMonths_PartBCov
2	1.049192	NoOfMonths_PartACov
8	1.010040	Total_Provider_Claims_Covered
5	1.000692	Claimed_Age

8.3 BASE MODEL BUILDING AND MODEL EVALUATION

Train Test Split:

The data is split into dependent feature Y (target) and independent features X. The data is then split into training and testing sets in order to avoid data leakage. The default 70:30 split is done.

The basic Logistic Regression Model has been built as the Target Variable falls under the Binary Classification.

Model Evaluation:

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

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

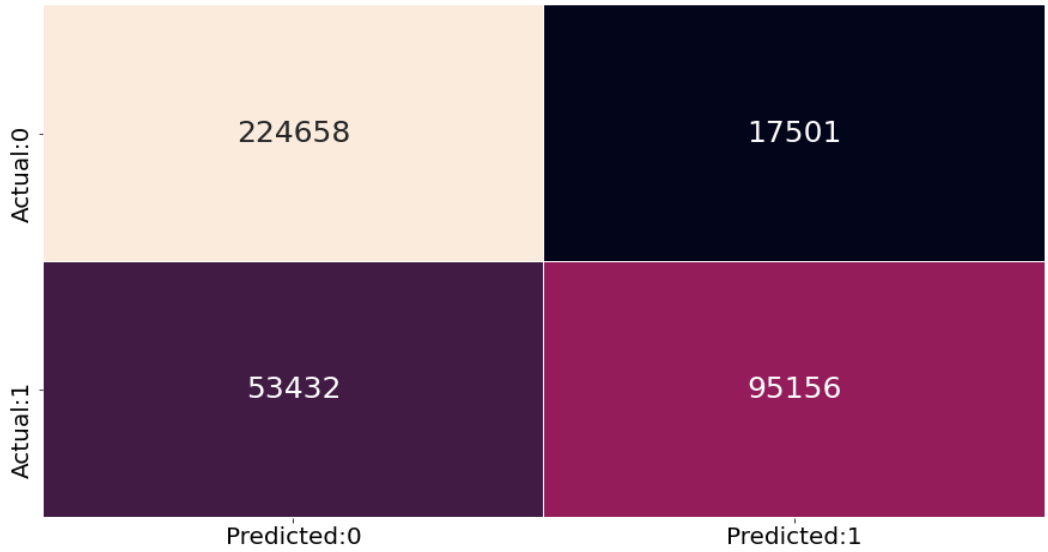
Train Dataset Evaluation:

1. Accuracy Score:

The accuracy score is calculated with a default threshold of 0.5 .

```
Train Accuracy Score : 0.8184682160067768
```

2. Confusion Matrix:



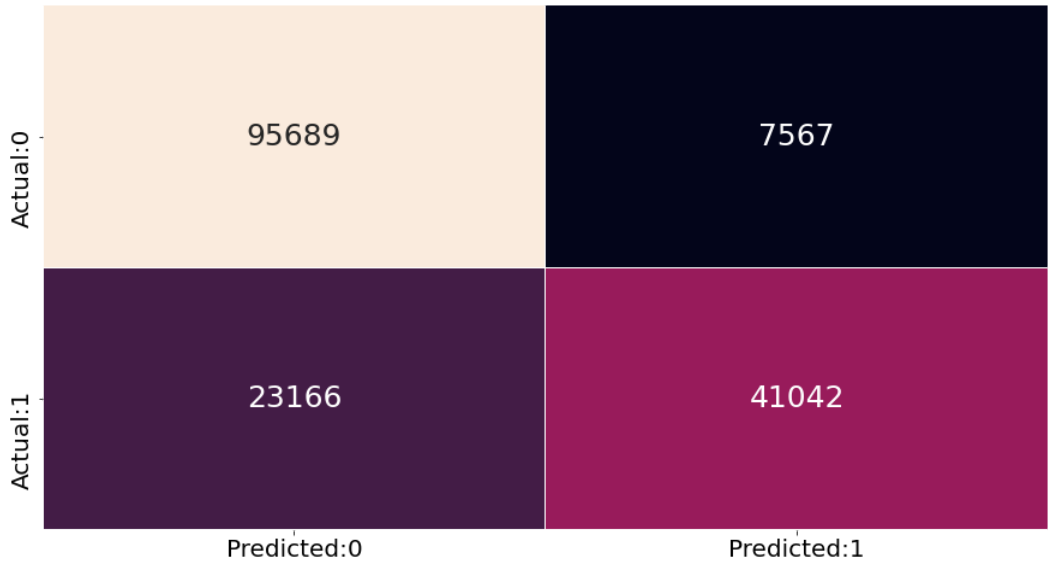
Test Dataset Evaluation:

1. Accuracy Score:

The accuracy score is calculated with a default threshold of 0.5 .

Test Accuracy Score : 0.8164799598719725

2. Confusion Matrix:



8.4 TUNING THE THRESHOLD

Using the argmax difference between the fpr and tpr the optimal threshold value has been identified as 1.

8.5 BUILDING MODEL BASED ON TUNED PARAMETERS AND EVALUATION

The Logistic Regression Model is built and evaluated.

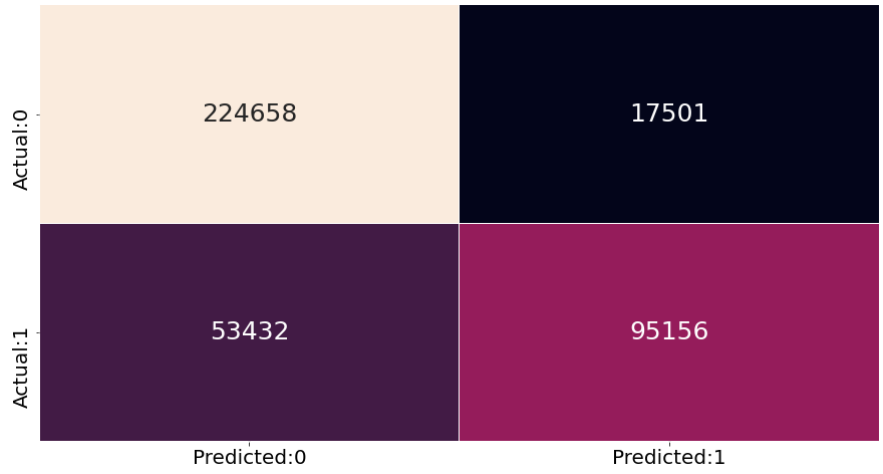
Train Dataset Evaluation:

1. Accuracy Score:

The accuracy score is calculated with a calculated optimal threshold of 1.

Train Accuracy Score : 0.8184682160067768

2. Confusion Matrix:



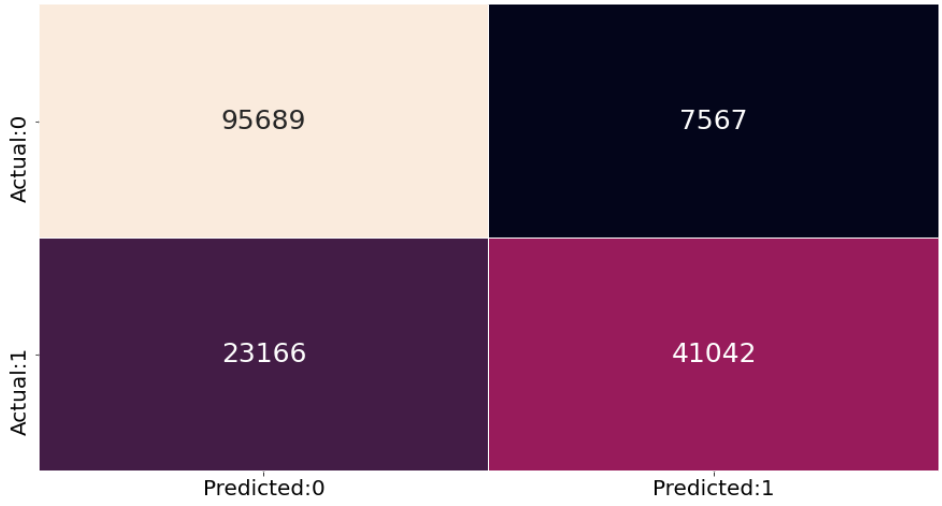
Test Dataset Evaluation:

1. Accuracy Score:

The accuracy score is calculated with a calculated optimal threshold of 1.

Test Accuracy Score : 0.8164799598719725

2. Confusion Matrix:



9. FUTURE WORK

Building other Models with optimal model parameters that are calculated using Grid Search.

The models that will be built are:

- Random forest
- Decision tree
- Naive Bayes
- Gradient Boost
- Adaboost
- XGboost

10. REFERENCES

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

[2] https://stonebridgebp.com/library/uncategorized/how-healthcare-fraud-affects-us-all/

[3] https://www.fraud-magazine.com/article.aspx?id=4294976280

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

[5] https://fin.plaid.com/articles/algorithmic-and-rules-based-fraud-models/

[6] https://www.opusconsulting.com/rule-based-vs-machine-learning-effective-fraud-prevention/

[7]https://www.researchgate.net/publication/23290716_A_survey_on_statistical_methods_for_health_care_fraud_detection

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

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

[10] https://pumas.nasa.gov/files/04_21_97_1.pdf