Healthcare provider fraud classification

Team Name: Detective Elites

Team Members

Mounika Kandru
Sai Charit Kondepati
Vaishnavi Babli

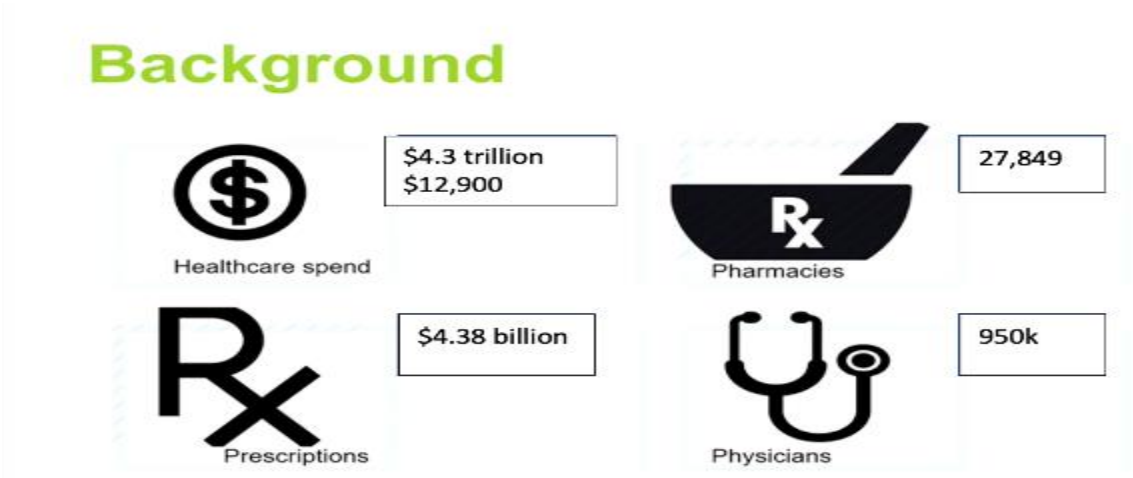
Table of Contents

Introduction:

Healthcare provider fraud is a significant concern for both government and private insurance companies, and it is estimated to cost billions of dollars each year. Fraudulent activities by healthcare providers include submitting false claims, upcoding, unbundling, and billing for services that were not provided. These activities can result in increased healthcare costs, reduced quality of care, and decreased patient trust in the healthcare system.

As a result, healthcare provider fraud classification has become an important area of research in healthcare management. Machine learning algorithms, such as logistic regression, decision trees, and neural networks, have proven to be effective in detecting fraudulent activities by healthcare providers.

The United States has one of the highest costs of healthcare in the world. We are living in a world where Medicare processes more than 4.5 million claims a day. Healthcare fraud is a type of white-collar crime wherein dishonest claims are filed to gain a profit. Fraud in U.S. Healthcare System is rampant. It is estimated that 3-10% of U.S. Healthcare annual expenditure is lost to fraud and abuse. Healthcare fraud is an organized crime that involves peers of providers (hospitals, cashiers, medical labs, nurses, lab assistants, and others), physicians, and beneficiaries acting together to make fraud claims. By dollar value, healthcare fraud is the largest category of criminal behavior in the United States today which is approximately more than $300 billions per year.



The above statistics shows that in the United States every year $4.3 trillion is spent on the healthcare and on an average every individual spends $12,900. There are 27,849 pharmacies in the United States and sound $4.38 billion prescriptions are processed every year. There are around 950k active Physicians in the United states as of 2023. CMS, the Centers for Medicare & Medicaid Services, is a federal agency within the United States Department of Health and Human Services that administers the fraudulent activities and deactivates the fraudulent Physicians.

Types of Healthcare frauds:

Provider Fraud:

- Billing for services that were not provided.
- Duplicate submission of a claim for the same service. Misrepresenting the service provided.
- Billing for a covered service when the service actually provided was not covered.

Member Fraud:

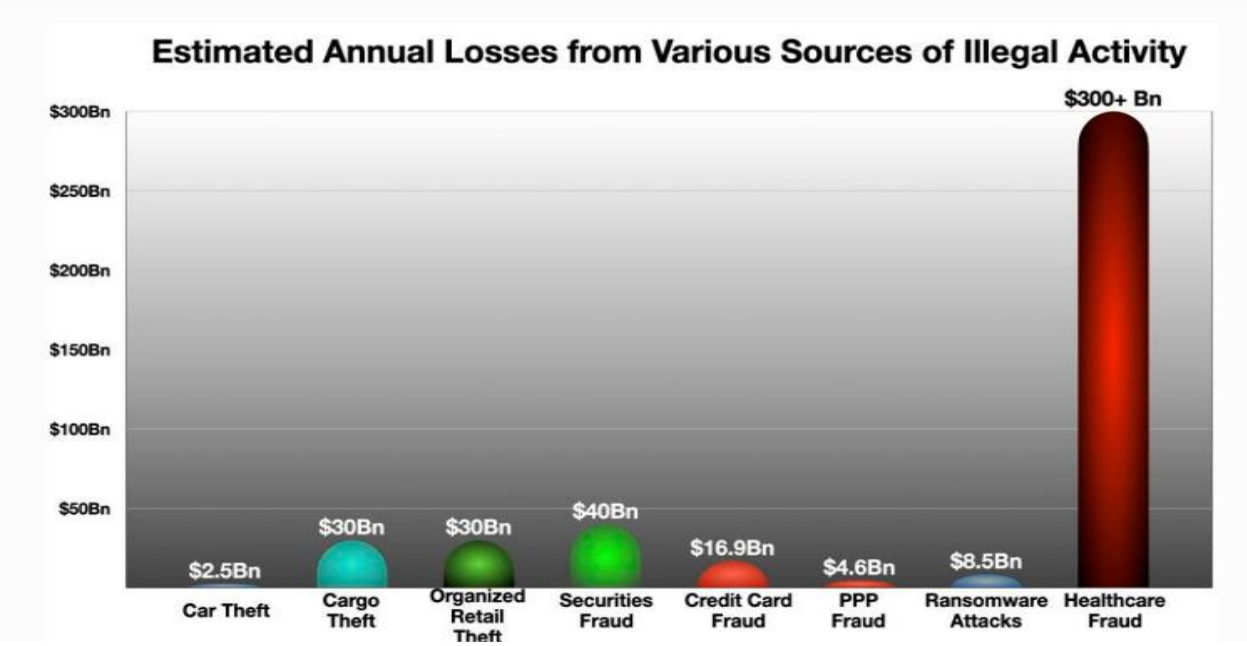
- Falsification of information such as forgery.
- Adding an ineligible dependent to the plan.
- Loaning or using another person's insurance card.

Broker and Agent Fraud:

- Alteration of documents
- Bribery and kickbacks
- Falsification of member or group information to obtain reasonable rates.
- Sale of non-existent policies

Impact on Society:

Waste of funds that would have been otherwise used for providing better medical treatments or services to actual patients. It is a criminal act as patients were either given false medicines or procedures which were not required. It increases the overall expenditure on healthcare, and it returns as a burden on the insured user because private payers increase their premiums. The false money earned by doing such frauds has also been used for carrying out various illegal activities that can be potentially harmful either to the nation or the entire world.



From the above graph it is very evident that Heathlcare fraud outnumbers many other types of frauds like credit card fraud, securities fraud and it's the most serious challenge that needs to be addressed. More than $300 billion is lost every year due to healthcare fraud.

Business Problem:

The National Healthcare Anti-Fraud Association estimated that approximately tens of billions of dollars are lost due to healthcare fraud each year. This immense financial loss places the responsibility of recovery on insurance companies, but more importantly, on patients. Patients are cheated into compensating for the cost in primarily two ways: payment of fraudulent copays and higher insurance premiums. Thus, it is pertinent to determine the patterns in healthcare fraud and take preventative measures against such crimes. Insurance companies are the most vulnerable institutions impacted due to these fraudulent claim practices. The insurance premium is also increasing day by day due to fraudulent activities and as result healthcare is becoming costly matter day by day.

Business Solution:

The goal of this project is to 'predict the potentially fraudulent providers' based on the claims filed by them that helps the Insurance company whether to accept or deny the claim or set up an investigation on that provider. Along with this, we will also discover important variables helpful in detecting the potentially fraud providers. In addition to it, we will study fraudulent patterns in the provider's claims to understand the future behavior of providers. Medicare loses billions of dollars each year due to fraud, errors, and abuse. Estimates place these losses at approximately $60 billion annually, though the exact figure is impossible to measure. Thus, building a binary classification model based on the claims filed by the provider along with In-patient data, Out-patient data, and Beneficiary details to predict whether the provider is potentially fraudulent or not which will help the insurance companies in easy detection of fraud and also would minimize patients' financial losses which ultimately better serves the society.

Data Source:

The data for this project was obtained from Kaggle and included four separate CSV files: beneficiary, outpatient, inpatient, and flagged fraudulent providers.

The beneficiary, outpatient, and inpatient datasets contained information at the individual patient and claims level, while the flagged fraudulent providers data was at the provider level. In order to effectively analyze the data and identify potential fraudulent providers, we needed to aggregate and transform the patient-level data to create a new dataset at the provider level.

This involved grouping and summarizing the patient-level data by provider, and calculating relevant metrics such as the total number of claims, total reimbursement amounts, and the prevalence of certain medical conditions or procedures among their patients. This process allowed us to gain a more comprehensive understanding of each provider's billing patterns and identify any outliers or suspicious activity.

We have reviewed the columns and their corresponding data types in the dataset and they are as below.

Inpatient Dataset:

```
inpatient_df.head()
```

	BeneID	ClaimID	ClaimStartDt	ClaimEndDt	Provider	InscClaimAmtReimbursed	AttendingPhysician	OperatingPhysician	OtherPhysician	AdmissionDt
0	BENE11001	CLM46614	2009-04-12	2009-04-18	PRV55912	26000	PHY390922	NaN	NaN	2009-04-12
1	BENE11001	CLM66048	2009-08-31	2009-09-02	PRV55907	5000	PHY318495	PHY318495	NaN	2009-08-31
2	BENE11001	CLM68358	2009-09-17	2009-09-20	PRV56046	5000	PHY372395	NaN	PHY324689	2009-09-17
3	BENE11011	CLM38412	2009-02-14	2009-02-22	PRV52405	5000	PHY369659	PHY392961	PHY349768	2009-02-14
4	BENE11014	CLM63689	2009-08-13	2009-08-30	PRV56614	10000	PHY379376	PHY398258	NaN	2009-08-13

5 rows × 30 columns

Providers Dataset:

```
fraud_or_not_df.head()
```

	Provider	PotentialFraud
0	PRV51001	No
1	PRV51003	Yes
2	PRV51004	No
3	PRV51005	Yes
4	PRV51007	No

Beneficiary Data:

```
patient_beneficiery_data.head()
```

	BeneID	DOB	DOD	Gender	Race	RenalDiseaseIndicator	State	County	NoOfMonths_PartACov	NoOfMonths_PartBCov	...	ChronicCond_Depression
0	BENE11001	1943-01-01	NaN	1	1	0	39	230	12	12	...	1
1	BENE11002	1936-09-01	NaN	2	1	0	39	280	12	12	...	2
2	BENE11003	1936-08-01	NaN	1	1	0	52	590	12	12	...	2
3	BENE11004	1922-07-01	NaN	1	1	0	39	270	12	12	...	2
4	BENE11005	1935-09-01	NaN	1	1	0	24	680	12	12	...	2

5 rows × 25 columns

We merged the Inpatient and outpatient with Providers data based on the Provider's ID, which is common across all data sets. We then combined it with Beneficiary Data based on Beneficiary ID, which is common across both data sets.

The inpatient data provides crucial details about patients who have been admitted to a hospital. The claim start date and end date help determine the duration of the hospital stay, while the amount reimbursed by the insurance company and the deductible amount paid by the patient provide insights into the financial aspect of

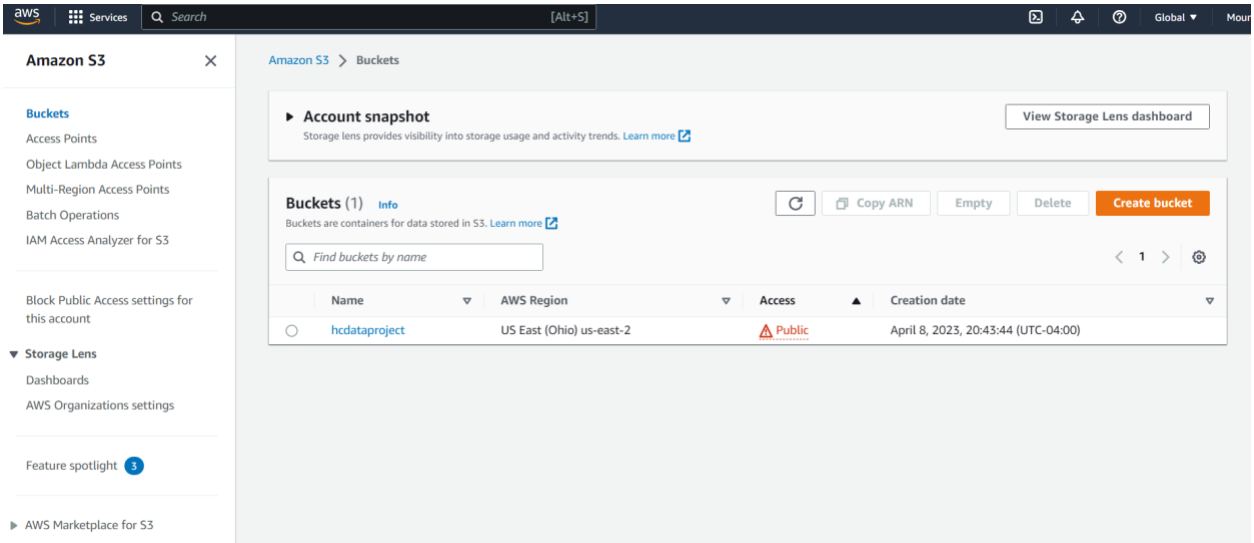
the claim. The physicians who attended to the patient are also recorded, providing information about who performed the surgery.

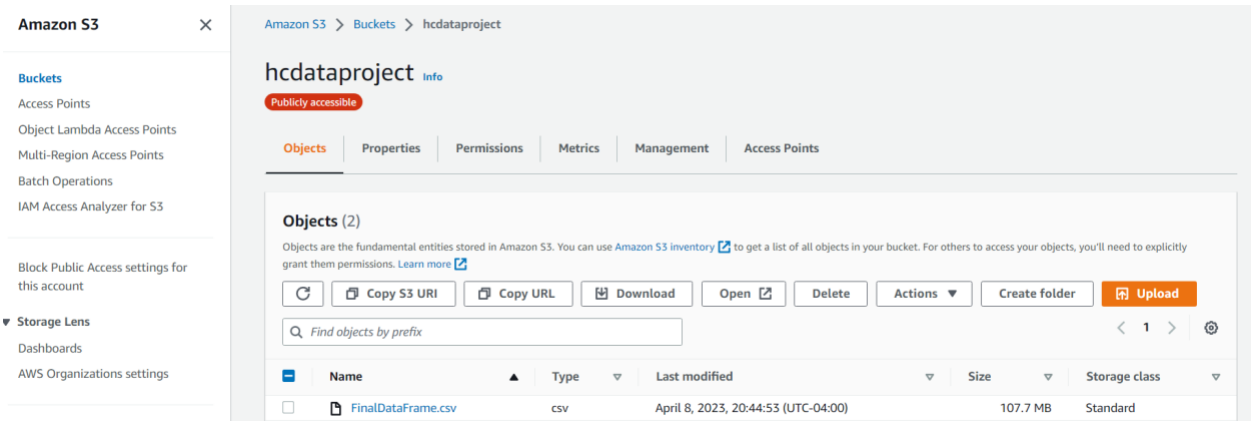
Additionally, the diagnosis and procedure codes provide information about the medical condition and treatment the patient received. The outpatient data is also important, providing information about patients who visit hospitals but are not admitted.

The number of months a patient has opted for coverage, as represented in the Coverage A and Coverage B columns, helps identify the duration a patient has opted a coverage for. The chronic conditions feature, represented in 10 different columns, indicates the different types of diseases a patient has. If the patient has a certain disease, the column's value is 1, otherwise it is 0. Finally, the state and county columns provide information about the patient's location.

Data Storage:

We used AWS S3, for storing the raw data. AWS S3 is flexible, cost efficient, Ease of migration, all-time available and simple to manage.

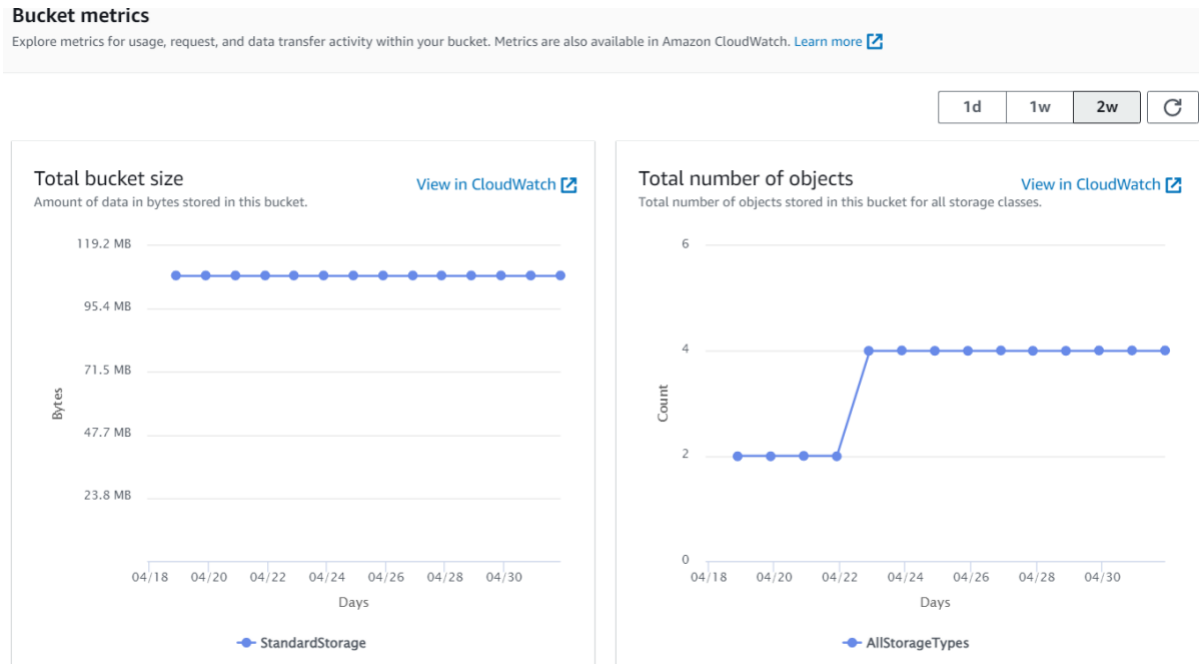




Steps followed to store the dataset in S3:

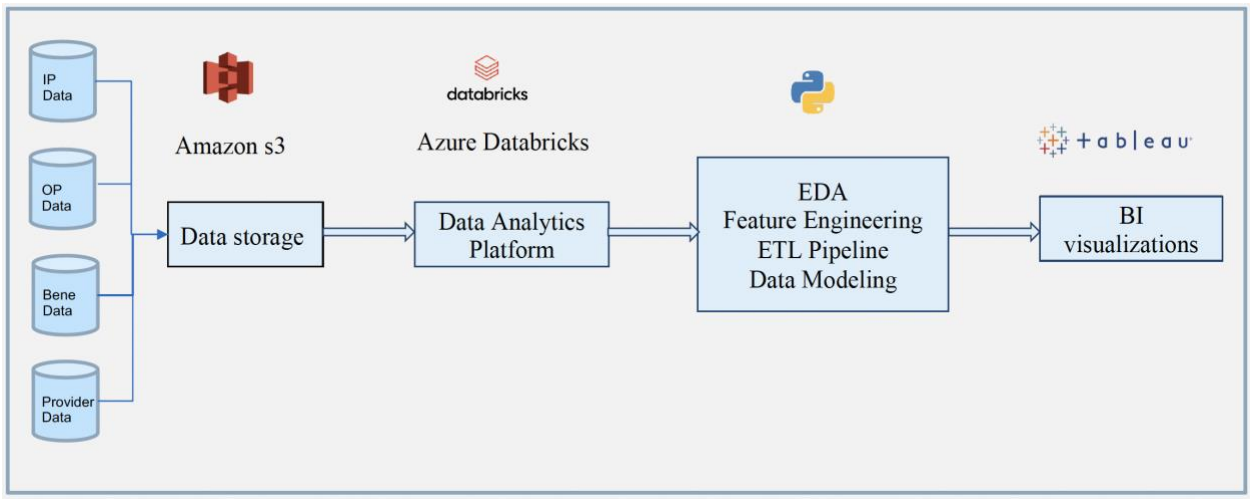
a) Created an Amazon AWS account
b) Created an S3 bucket
c) Uploaded the dataset
d) Set permissions for the file
e) Generate a URL for the file

Bucket Metrics:



Solution Architecture:

Extract Transform and Load:



Initially we joined the individual Datasets based on IDs and stored the resultant dataset in Amazon S3, which is a cloud-based storage service provided by Amazon Web Services (AWS). It is designed to provide scalable, secure, and durable storage for data objects such as files, images, videos, and application data. We later fetched the Dataset in Azure Data Bricks, which can scale up or down automatically to handle large volumes of data, making it well-suited for big data processing and analytics and it provides a collaborative workspace for teams to work together, share data, and build data pipelines. We performed Exploratory Data Analysis, Feature Engineering, Data Modeling.

a) Data Source:

The data for this project was obtained from Kaggle and included four separate CSV files: beneficiary, outpatient, inpatient, and flagged fraudulent providers.

The beneficiary, outpatient, and inpatient datasets contained information at the individual patient and claims level, while the flagged fraudulent providers data was at the provider level. In order to effectively analyze the data and identify potential fraudulent providers, we needed to aggregate and transform the patient-level data to create a new dataset at the provider level.

b) Data Storage:

We used AWS S3, for storing the raw data. AWS S3 is flexible, cost efficient, Ease of migration, all-time available and simple to manage.

c) Data analytics platform:

We will be leveraging Azure Databricks as a data analytics platform. We will be exploring Azure Databricks Notebook for visualization and constructing pipelines. As Azure is an available tool in the Microsoft suite it is cost effective and easily available.

d) Data Modeling:

We will be using Python to set up pipelines and use fit and transform functions to transform the raw data to make it ready for pipelining. We will be utilizing several libraries such as pandas, numpy, scikit etc. to build the building blocks of the pipeline. We will further analyze the data and provide insights through calculated columns.

e) Visualization tool:

We will be utilizing python libraries such as MatplotLib and Scikit to provide visualization on data analysis. Furthermore, we will be utilizing Tableau to generate dashboards to give us insight on the model performance and data exploration aspect.

Data Cleaning:

Standardizing data is an essential step in ensuring that the data is consistent across all variables, sources, and systems. In this project, we standardized the Gender field by converting the values for male and female from 1 and 2 to 1 and 0, respectively. It is crucial to handle missing data appropriately to avoid biased results. There are different techniques that can be used to handle missing data, such as imputing missing data, deleting missing data, or using machine learning models to predict the missing data.

In our project, we replaced the null values with "NaN" to handle the missing data in the dataset. This approach allowed us to retain the data without losing the entire record, and also enabled us to identify the missing values while analyzing the dataset. Overall, standardizing and handling missing data are crucial steps in ensuring the accuracy and reliability of healthcare data analysis.

We checked and treated the Null Values and replaced them with NaN.

```
pd.isnull(hc_info).sum()
```

```
new_hc_info = hc_info.fillna('NaN')
pd.isnull(new_hc_info).sum()
```

We have performed Feature Engineering for data cleaning and data preparation involving the below:

a) **Feature Encoding:**

We have used Label Encoder to converts categorical variables to numerical variables to make the data compatible with Machine Learning models. Ex: ClmDiagnosisCodes, AttendingPhysiscians etc

```
#feature engineering
from sklearn.preprocessing import OneHotEncoder
# Select the columns to be one-hot encoded
columns_to_encode = ['ClmAdmitDiagnosisCode', 'Provider', 'AttendingPhysician', 'OperatingPhysician', 'OtherPhysician',
                     'DiagnosisGroupCode', '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', 'InPatient_or_OutPatient', 'PotentialFraud']
from sklearn.preprocessing import LabelEncoder
hc_info_df = hc_info.copy()
le = LabelEncoder()
# Apply LabelEncoder on columns 'A' and 'C'
hc_info_df[columns_to_encode] = hc_info_df[columns_to_encode].apply(lambda col: le.fit_transform(col) if col.name in columns_to_encode else col)
hc_info_df.head()
```

b) **Feature Scaling:**

It is a method used to normalize the range of independent variables or features of data.

Ex: DeductibleAmtPaid and InscClaimAmtReimbursed

```
from sklearn.feature_selection import SelectKBest, chi2

numeric_cols=['InscClaimAmtReimbursed','DeductibleAmtPaid','NoOfMonths_PartACov', 'NoOfMonths_PartBCov', 'number_of_days_admitted']
minmax_scaler = MinMaxScaler()
minmax_scaler.fit(hc_model[numeric_cols])
hc_model[numeric_cols] = minmax_scaler.transform(hc_model[numeric_cols])
hc_model.head()
```

c) **Feature Selection:**

It is a process in machine learning to identify important features and reducing the input variables to the model by using only the relevant data to improve the performance of the model.

d) **Chi Square Method:**

It is a statistical method commonly used in data analysis to determine if there is a significant association between two categorical variables.

```
selected_features_indices = selector.get_support(indices=True)

selected_features_names = hc_model.columns[selected_features_indices]
print(selected_features_names)
```

```
Index(['Provider', 'AttendingPhysician', 'OperatingPhysician',
       'ClmDiagnosisCode_2', 'ClmDiagnosisCode_3', 'ClmDiagnosisCode_4',
       'ClmDiagnosisCode_5'],
      dtype='object')
```

e) Feature Importance:

Based on the Feature selection, we have the following features and we would be determining how much important each feature would be in predicting whether the given Provider is fraudulent or not?

- Provider
- Insurance Claim Amount Reimbursed
- Physician

f) Data Imbalance:

Our analysis revealed that 90% of the providers in the dataset are non-fraudulent, while only 10% are fraudulent. This data imbalance can lead to biased model predictions, where the minority class is poorly represented.

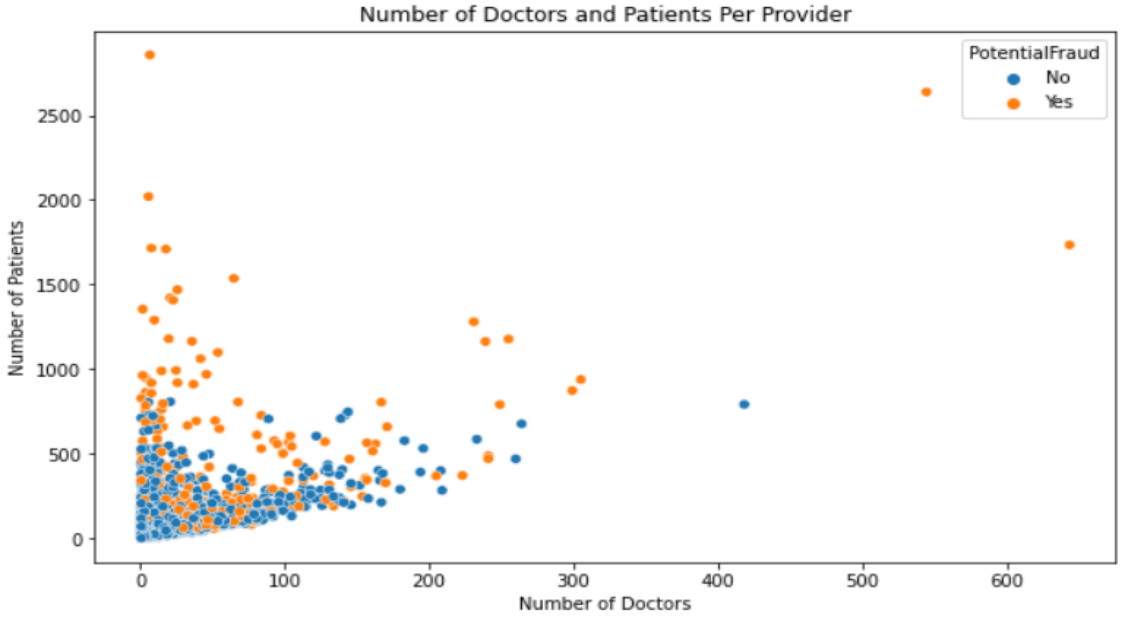
To address this issue, we plan to use the SMOTE sampling technique. It involves generating synthetic samples for the minority class by interpolating between existing minority class samples.

Exploratory Data Analysis:

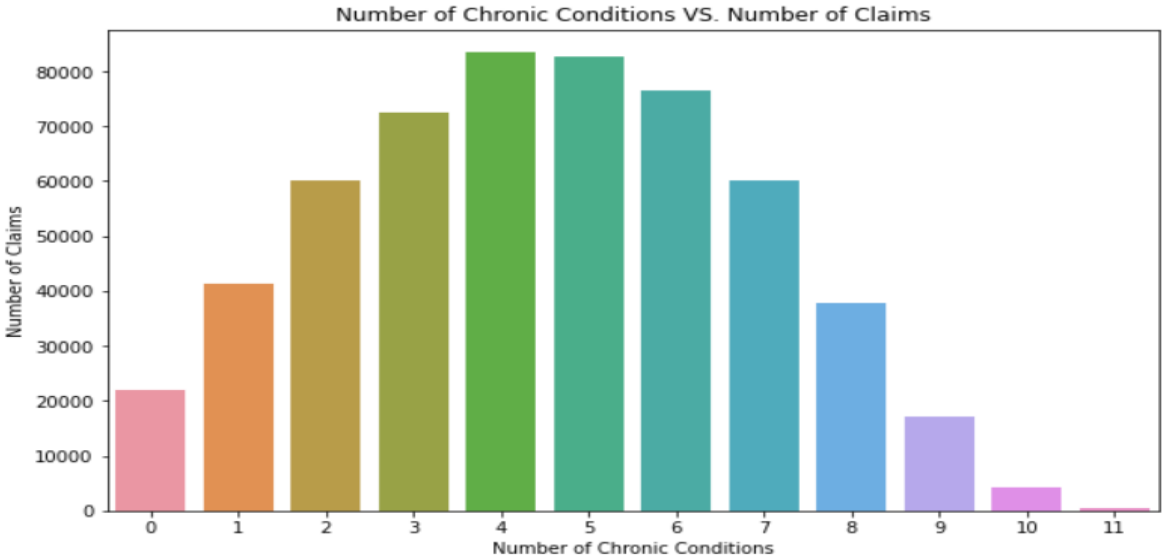
Exploratory Data Analysis is a process of analyzing and summarizing a dataset to understand its main characteristics and relationships between variables. EDA is often the first step in any data analysis project, and it helps to identify patterns, trends, and anomalies in the data. We have performed Exploratory data analysis for data preparation involving the below.

a) **Data Visualization**: To understand the main characteristics, patterns and trends of our dataset.

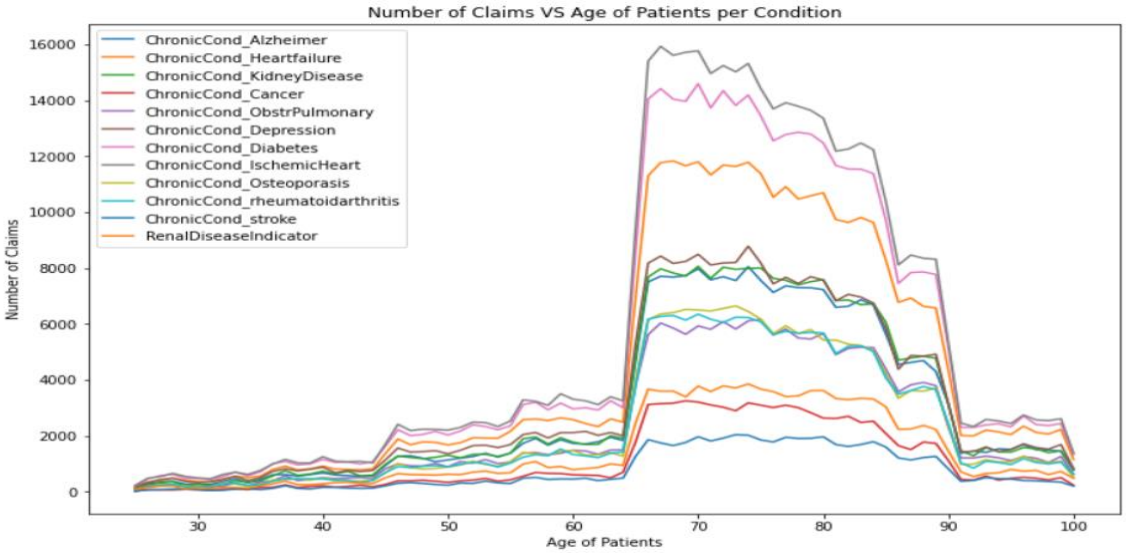
Below are some of the findings of the features of our dataset:



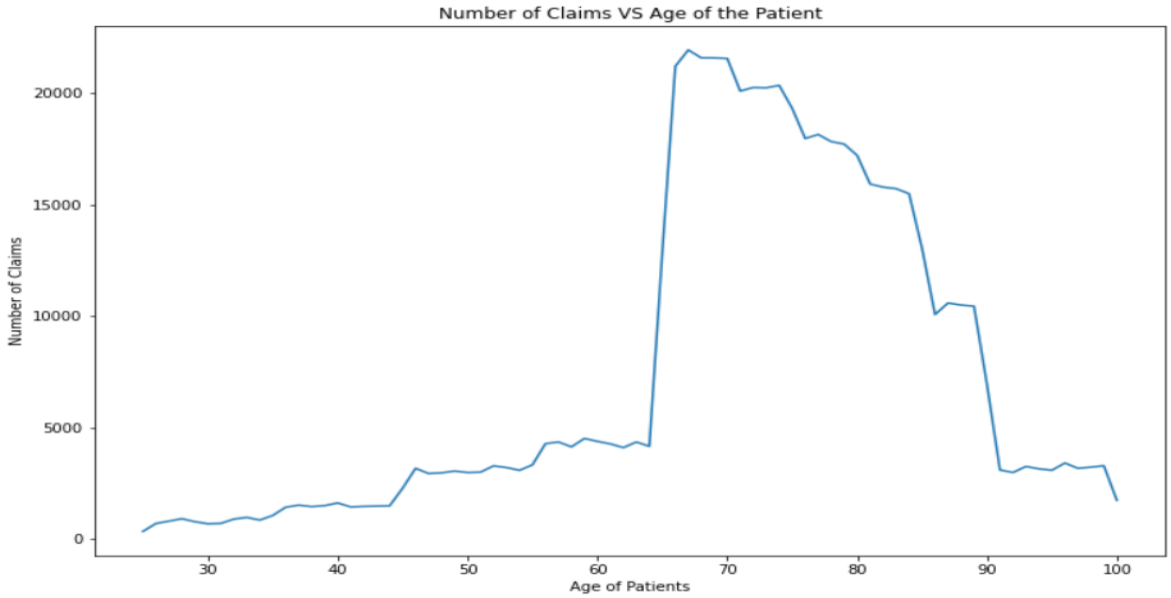
From the above graph it is clear that as the Provider's number of doctors and patients increases, so is the increase in the potential fraud.



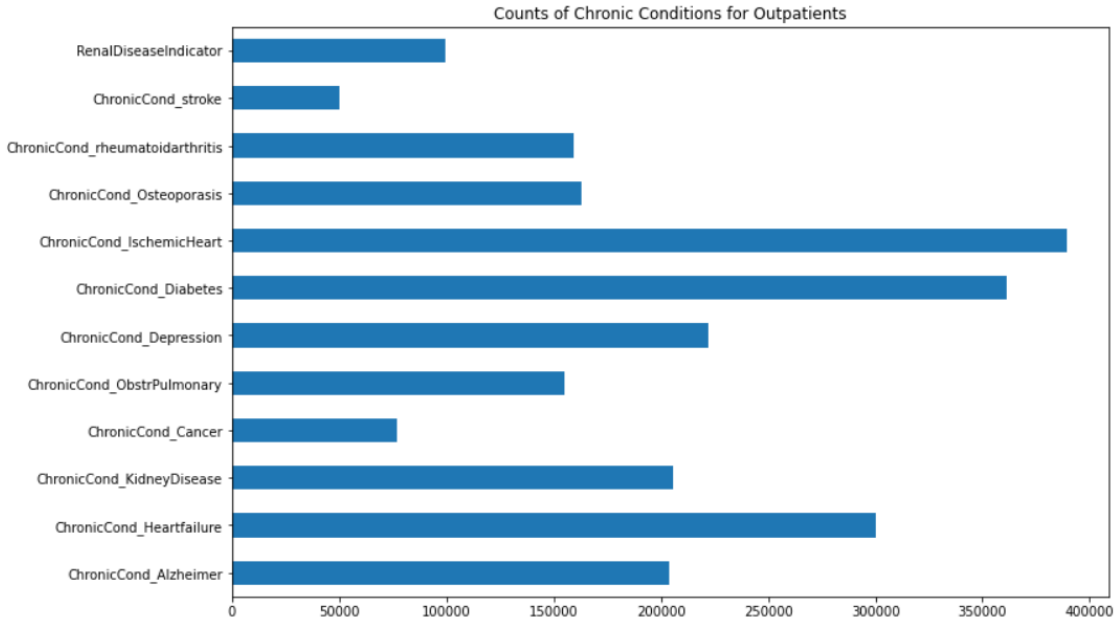
From the above graph it is evident that patients with 4 to 6 chronic conditions are the ones having a higher number of claims.
4- ChronicCond_Cancer
5-ChronicCond_ObstrPulmonary
6-ChronicCond_Depression



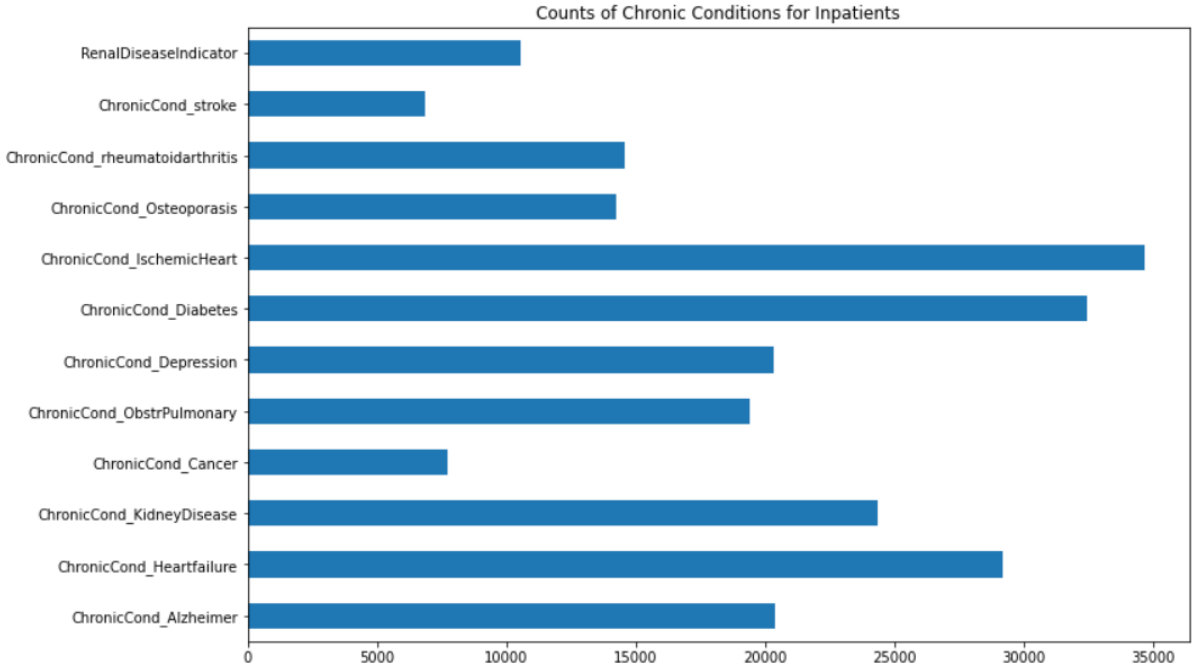
The above graph shows a relationship between patient age and the chronic conditions. It shows a definite trend with increasing ages. Top 3 Chronic Cond are Depression, Diabetes and Heart failure



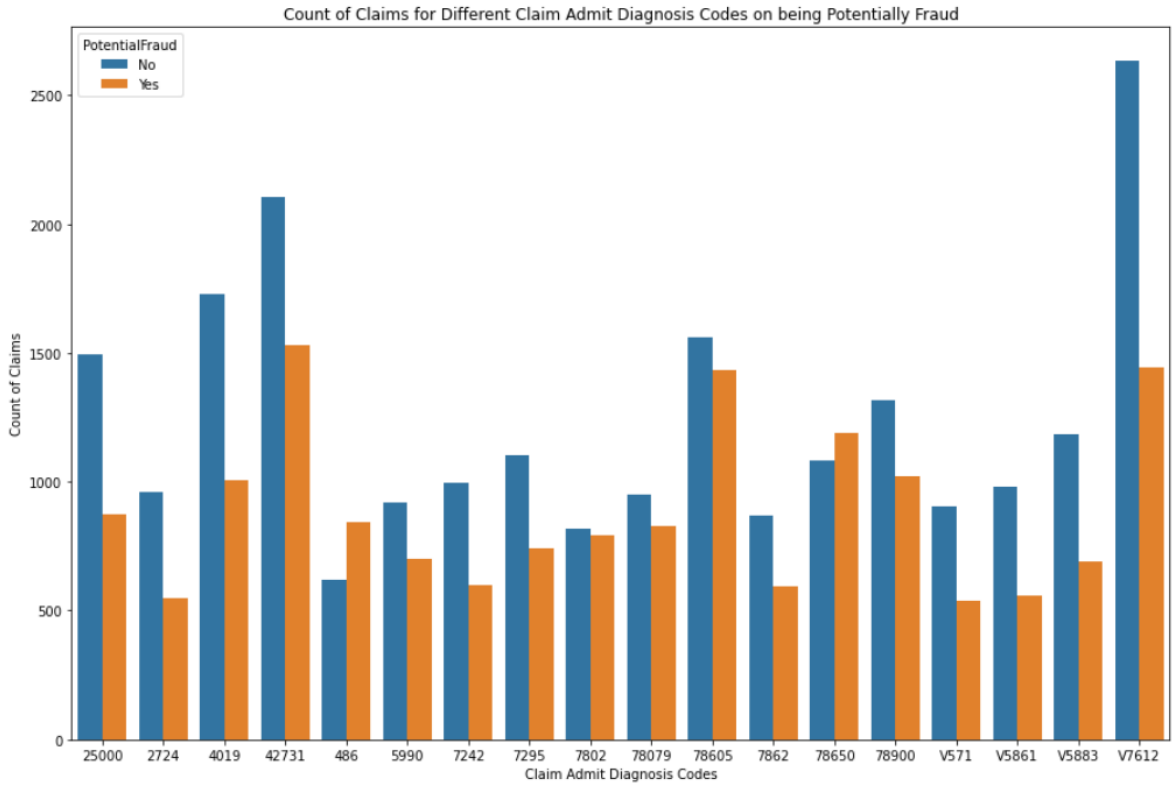
The above graph shows that patients aged 60-70 file more claims than the rest and we also see that very few claims filed by patients aged below 30.



The above graph shows the count of Chronic conditions for Outpatients. Also, it indicates what type of patients would make more outpatient visits.

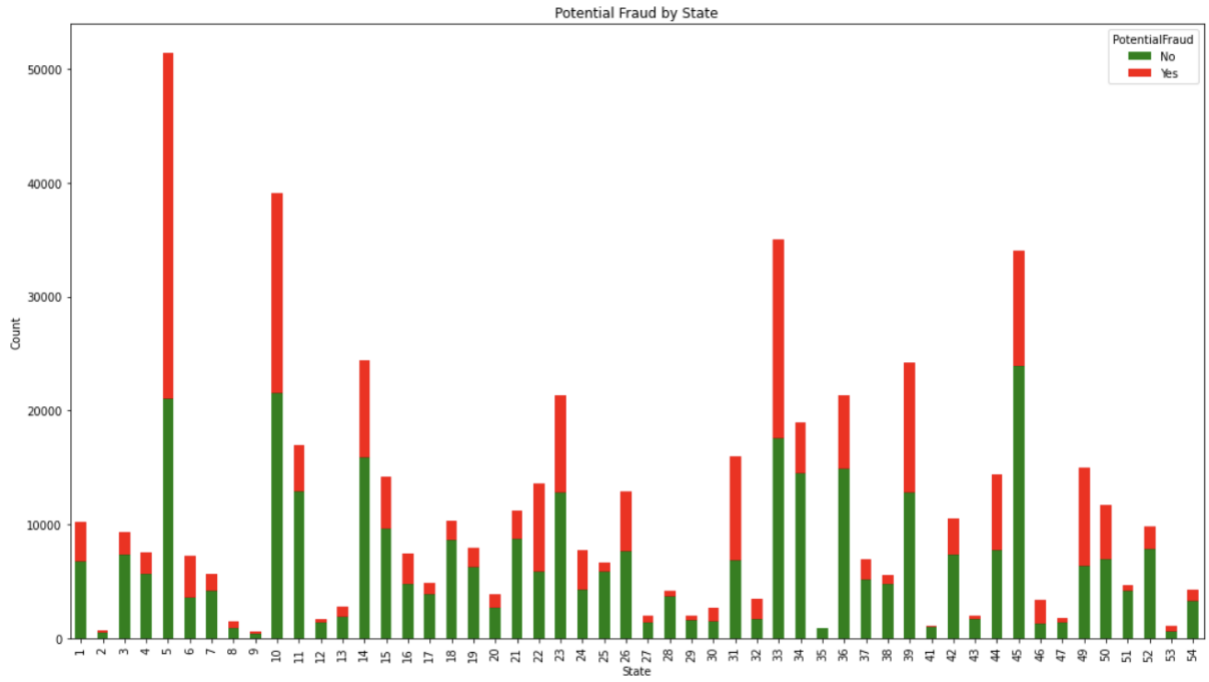


The above graph shows the count of Chronic conditions for Inpatients. Also, it indicates what type of patients would make more Inpatient visits.



From above graph we can infer the below:

- The codes with most fraudulent accounts are 42731, 78605 and v7612.
- There is one code that has more yes than no (78650).
- Code V7612 has the highest number of claims from week 1 up to week 15 and also had the highest drops of claims.
- Most of the codes shows a decline of claims as the week progresses.
- Patient with diagnosis code V5789 had the highest median duration which is around 15 days.



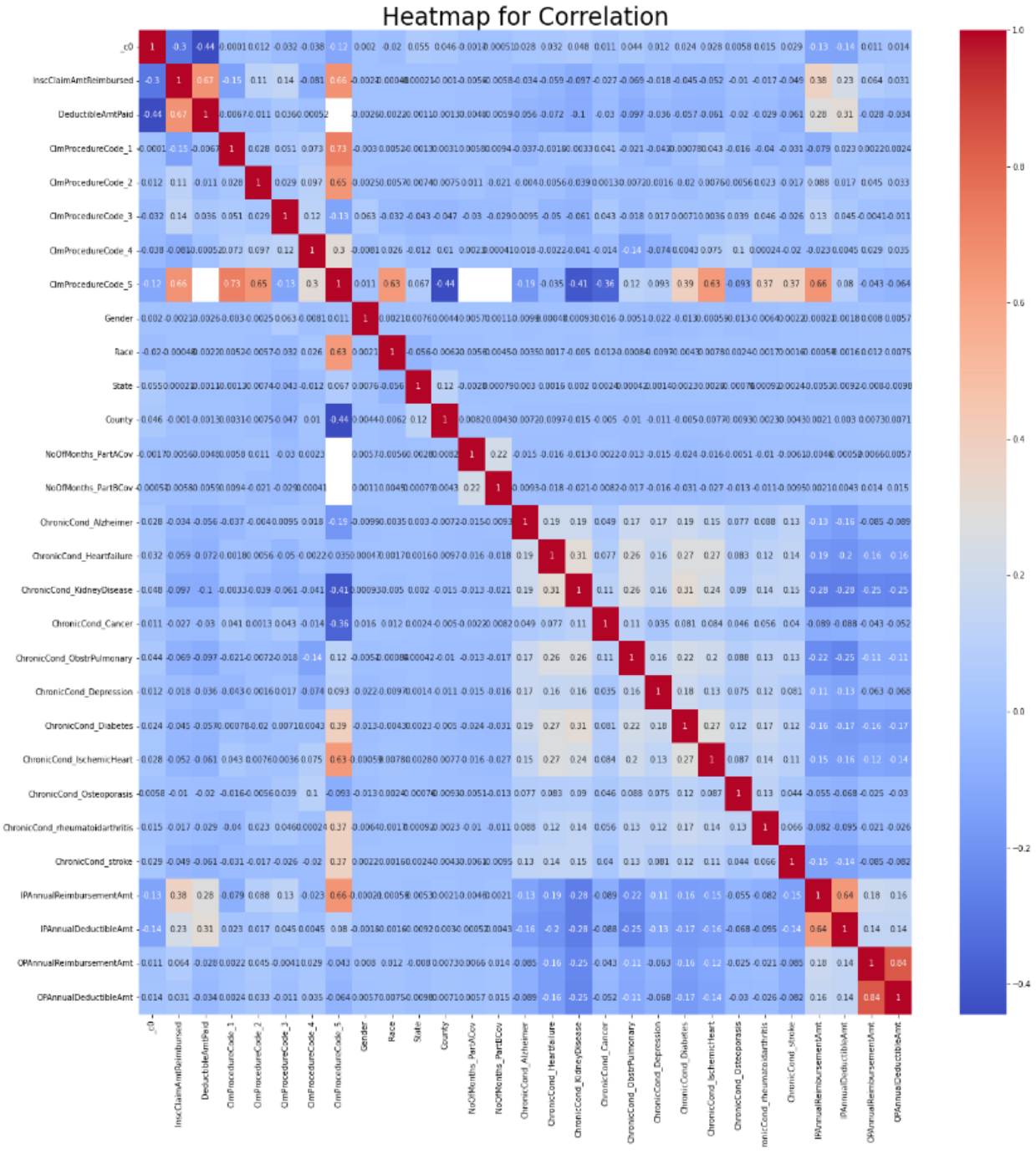
From the above graph we are trying to identify the states where most Fraud occurs. For potentially fraud cases, whether it is duplicated or not, states #5, #10, #33 and county #200 seem to have the highest count.

b) Heatmap to identify the correlation between the features:

Correlation heatmaps is used to find potential relationships between variables and to understand the strength of these relationships

We were able to determine which features are positively and negatively associated as well as which features have no correlation at all thanks to the heatmap shown below. To plot the heatmap, we utilized the matplot and seaborn packages.

To determine which features are positively and negatively associated as well as which features have no correlation at all. The InsuranceClaimAmtReimbursed and Deductible amount paid are positively corelated. County and ClmProcedureCode_5 is negatively correlating. We would be excluding one of the features which is positively corelating as it is not too significant for modeling.


Heatmap for Correlation
_c0
InscClaimAmtReimbursed
DeductibleAmtPaid
ClmProcedureCode_1
ClmProcedureCode_2
ClmProcedureCode_3
ClmProcedureCode_4
ClmProcedureCode_5
Gender
Race
State
County
NoOfMonths_PartACov
NoOfMonths_PartBCov
ChronicCond_Alzheimer
ChronicCond_Heartfailure
ChronicCond_KidneyDisease
ChronicCond_Cancer
ChronicCond_ObstrPulmonary
ChronicCond_Depression
ChronicCond_Diabetes
ChronicCond_IschemicHeart
ChronicCond_Osteoporasis
ChronicCond_rheumatoidarthritis
ChronicCond_stroke
IPAnnualReimbursementAmt
IPAnnualDeductibleAmt
OPAnnualReimbursementAmt
OPAnnualDeductibleAmt
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Data Modeling:

Data modeling involves identifying the data elements that are important to the organization or system, defining relationships between them, and creating a structure that allows for efficient data retrieval and manipulation. It is a critical aspect of database design and is used in a wide range of applications, from business and finance to healthcare and scientific research.

• SMOTE technique to address the problem of imbalanced data distribution.
• Logistic Regression, Random Forest Classifier, KNN and XGB Classifier.
• Used performance metrics such as Confusion Matrix, Precision, Recall and F-1 score.

Model	Accuracy	F1	Precision	Recall
LogisticReg	54%	0.41	0.40	0.42
RandomForest	88%	0.84	0.82	0.85
KNN	77%	0.71	0.68	0.75
XGBoost	94%	0.92	0.95	0.89

Logistic Regression:

A statistical technique called logistic regression is used to predict one of two probable outcomes in binary classification. It is a well-known and frequently applied algorithm in machine learning, particularly when the dependent variable is dichotomous, or only accepts one of two possible values.

```
from imblearn.over_sampling import SMOTE
#SMOTE logistic
pipe_SMOTE_LOG = make_imb_pipeline(SMOTE(), LogisticRegression())

pipe_SMOTE_LOG.fit(standard_scaled_train_inputs, standard_scaled_train_targets)



standard_scaled_test_targets_pred = pipe_SMOTE_LOG.predict(standard_scaled_test_inputs)

scores = cross_validate(pipe_SMOTE_LOG,
                        standard_scaled_train_inputs, standard_scaled_train_targets, cv=10,
                        scoring=('roc_auc', 'average_precision'))
```

Random Forest Classifier:

An ensemble learning technique for classification known as a "Random Forest Classifier" constructs a large number of decision trees and then combines their outputs to produce a final classification outcome.

```
#SMOTE randomforest
pipe_SMOTE_RFC = make_imb_pipeline(SMOTE(), RandomForestClassifier())

pipe_SMOTE_RFC.fit(standard_scaled_train_inputs, standard_scaled_train_targets)

print(pipe_SMOTE_RFC.score(standard_scaled_test_inputs, standard_scaled_test_targets))
standard_scaled_test_targets_pred = pipe_SMOTE_RFC.predict(standard_scaled_test_inputs)


scores = cross_validate(pipe_SMOTE_RFC,
                        standard_scaled_train_inputs, standard_scaled_train_targets, cv=10,
                        scoring=('roc_auc', 'average_precision'))
```

K-Nearest Neighbors:

When used for classification, KNN locates a new data point's k-nearest neighbors from the training set and then assigns the majority of those neighbors' class labels to the new data point.

```
#SMOTE randomforest
pipe_SMOTE_RFC = make_imb_pipeline(SMOTE(), RandomForestClassifier())

pipe_SMOTE_RFC.fit(standard_scaled_train_inputs, standard_scaled_train_targets)

print(pipe_SMOTE_RFC.score(standard_scaled_test_inputs, standard_scaled_test_targets))
standard_scaled_test_targets_pred = pipe_SMOTE_RFC.predict(standard_scaled_test_inputs)


scores = cross_validate(pipe_SMOTE_RFC,
                        standard_scaled_train_inputs, standard_scaled_train_targets, cv=10,
                        scoring=('roc_auc', 'average_precision'))
```

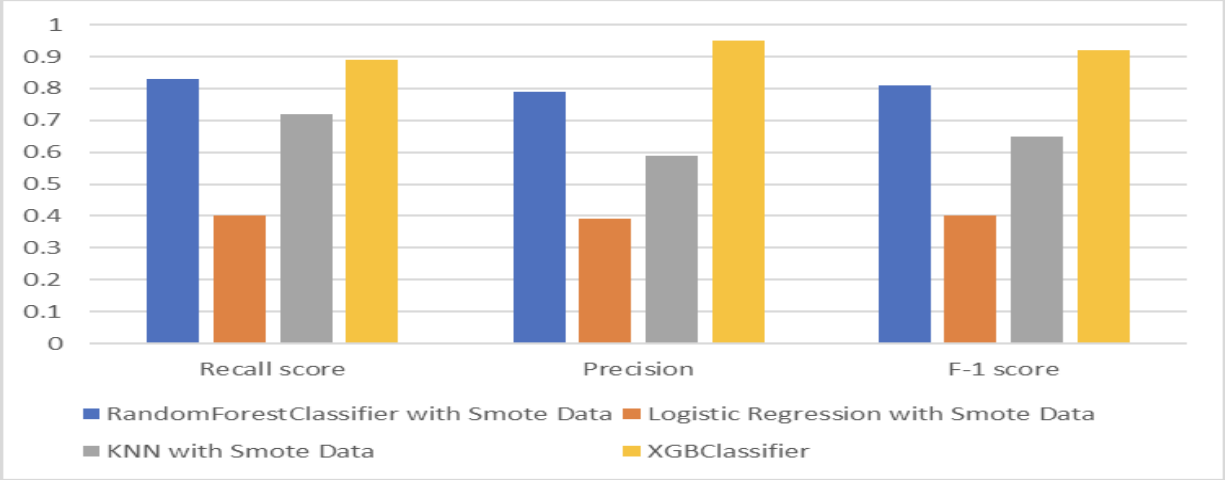
X Gradient Boost:

A series of decision trees are trained iteratively by XGBoost, with each tree being trained to rectify the flaws of the one before it. The approach computes the gradient and second-order derivative of the loss function after each iteration, and then modifies the tree weights to minimize the loss function.

```
import xgboost as xgb
from sklearn.model_selection import train_test_split, cross_validate
from sklearn.metrics import f1_score, precision_score, recall_score, classification_report, confusion_matrix
from collections import Counter

# Define XGBoost classifier model
model = xgb.XGBClassifier(objective='binary:logistic', n_estimators=1000, learning_rate=0.1, max_depth=3)

# Split data into training and validation sets
raw_inputs_train_USD_XG, raw_inputs_valid_USD, raw_targets_train_USD_XG, raw_targets_valid_USD = train_test_split(raw_inputs_select, raw_targets, test_size=0.2,
random_state=42)

# Train the model
model.fit(raw_inputs_train_USD_XG, raw_targets_train_USD_XG, early_stopping_rounds=10, eval_set=[(raw_inputs_valid_USD, raw_targets_valid_USD)])

# Make predictions on the test set
XG_test_targets_pred = model.predict(standard_scaled_test_inputs)

# Compute performance metrics using cross-validation
scores = cross_validate(model, standard_scaled_train_inputs, standard_scaled_train_targets, cv=10, scoring=('roc_auc', 'average_precision'))
```

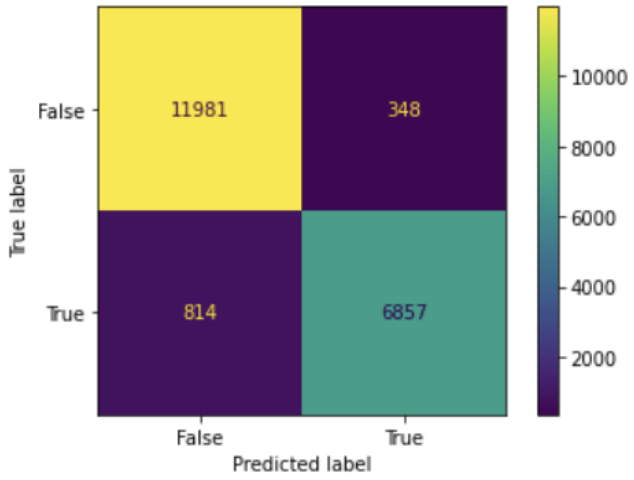
Performance of Models:

XGBoostClassifier Model is the best performing model with an accuracy of 94%



Evaluation Metrics for XGB Classifier:

```
test_roc_auc 0.9812967194202618
test_average_precision 0.9749210913781969
F1 score =  0.9153182308522114
Precision score =  0.9496362618914381
Recall score =  0.8833940655908381
Test Performance:
              precision    recall  f1-score   support

           0       0.93      0.97      0.95     12316
           1       0.95      0.88      0.92      7684

    accuracy                           0.94     20000
   macro avg       0.94      0.93      0.93     20000
weighted avg       0.94      0.94      0.94     20000

Actual =  Counter({0: 12316, 1: 7684})
Predicted =  Counter({0: 12852, 1: 7148})
Out[80]: array([[11956,   360],
       [  896,  6788]])
```



Conclusion:

• Healthcare provider fraud classification is a complex problem that requires careful consideration of various factors, such as data quality, feature selection, and model selection.
• A successful fraud classification system should be able to accurately identify fraudulent healthcare providers while minimizing false positives and false negatives.
• In conclusion, healthcare provider fraud classification is an important area of research and has significant implications for healthcare fraud detection and prevention. It is important to continue to explore and develop new approaches to healthcare provider fraud classification in order to stay ahead of evolving fraud schemes and protect the integrity of the healthcare system.

Learnings and challenges:

• **Learned new terminology of healthcare domain**:

Healthcare data has its own unique characteristics and terminology, so gaining a deeper understanding of the industry was helpful. This may involve learning about healthcare regulations, reimbursement systems, medical codes, and clinical workflows.

• **Data Integrity and quality:**

We understood the importance of having quality data to find more insights from the data available. Healthcare data is frequently inconsistent, inaccurate, and incomplete, which might compromise the precision of fraud detection models. To address these problems, data cleaning and normalization techniques are needed to make sure the data is correct and trustworthy.

• **Continuous Monitoring and Adaption:**

Healthcare fraud is a constantly evolving problem, with fraudsters constantly developing new schemes and tactics. To stay ahead of these threats, healthcare fraud detection systems must be continuously monitored and adapted to respond to new and emerging risks.

References:

1. https://www.beckersasc.com/asc-news/number-of-active-physicians-by-specialty.html#:~:text=The%20report%20is%20based%20on,2021%2C%20according%20to%20a%20Jan.
2. statista.com/statistics/261303/total-number-of-retail-prescriptions-filled-annually-in-the-us/
3. https://www.pgpf.org/blog/2023/01/why-are-americans-paying-more-for-healthcare#:~:text=How%20Much%20Does%20the%20United,to%20about%20%2412%2C900%20per%20person.
4. https://nycdatascience.com/blog/student-works/data-analysis-on-healthcare-fraud/
5. https://www.smpresource.org/Content/Medicare-Fraud/Dollars-Lost-to-Fraud.aspx#:~:text=Medicare%20fraud%20is%20big%20business,figure%20is%20impossible%20to%20measure.
6. https://medium.datadriveninvestor.com/medicare-provider-fraud-detection-f551dd941947
7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7579458/

Personal Viewpoint on the Project
-Vaishnavi Babli

The expense of healthcare in the US is among the highest in the world. With an annual cost of more than $300 billion, healthcare fraud is currently the greatest category of criminal activity in the United States.

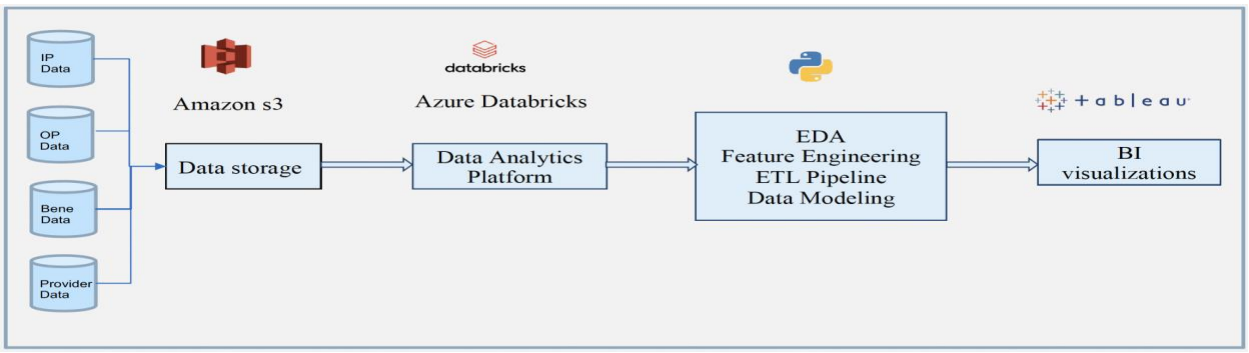
Business Problem:

The National Healthcare Anti-Fraud Association estimated that approximately tens of billions of dollars are lost due to healthcare fraud each year. The insurance companies, and patients are the most vulnerable who get impacted because of this immense financial loss.

Business Solution:

Based on the claims they submit, the project's objective is to "predict the potentially fraudulent providers," which aids the insurance company in deciding whether to accept the claim, reject it, or launch an investigation into that provider.

Solution Architecture:



There are 4 different source datasets inpatient, outpatient, beneficiary, provider datasets which are combined to form a combined dataset which is stored in the Amazon s3 bucket because it is highly durable, secured and scalable.

The dataset is then imported into the Azure Databricks platform as its highly powerful and flexible for the data analytics, data engineering and data modeling purposes. Further EDA, feature engineering and pipelines are built for the data pre-processing and data preparation.

Exploratory Data Analysis: EDA is the process of examining and condensing a dataset to comprehend its key features and correlations among variables. Every data analysis project begins with EDA, which aids in finding patterns, trends, and anomalies in the data involving data cleaning, data visualization.

Data Imbalance:

Our analysis revealed that 90% of the providers in the dataset are non-fraudulent, while only 10% are fraudulent. This data imbalance can lead to biased model predictions, where the minority class is poorly represented. To address this issue, we planned to use the SMOTE sampling technique. It involves generating synthetic samples for the minority class by interpolating between existing minority class samples.

This was the most important step in the project as it will help in accurately identifying the fraudulent claims by solving the data imbalance issue.

Feature Engineering: We have implemented the following techniques.

Feature Encoding**:** We have used Label Encoder to converts categorical variables to numerical variables to make the data compatible with Machine Learning models.
Feature Scaling: It is a method used to normalize the range of independent variables or features of data.
Feature Selection**:** It is a process in machine learning to identify important features and reducing the
input variables to the model by using only the relevant data to improve the performance of the model.

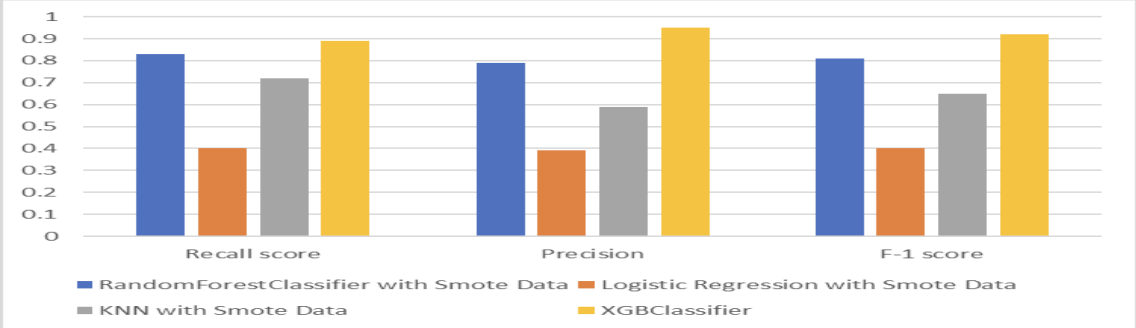
Feature Importance**:** Based on the Feature selection, we have the following features, and we would be determining how much important each feature would be in predicting whether the given Provider is fraudulent or not?

Data Modeling: We have used the following Models:

Logistic Regression
Random Forest Classifier
KNN
XGBoost

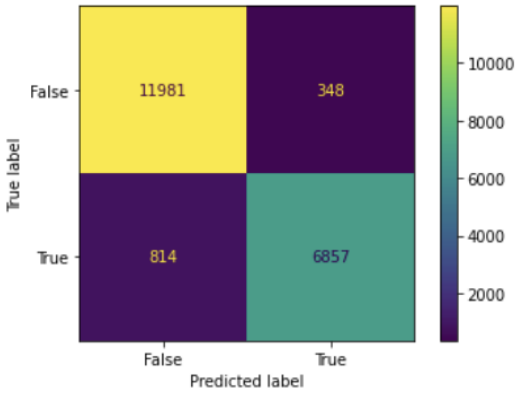
Performance of Models:

XGBoostClassifier Model is the best performing model with an accuracy of 94%



Performance Metrices for XGBoostClassifier Model:

```
test_roc_auc 0.9812967194202618
test_average_precision 0.9749210913781969
F1 score =  0.9153182308522114
Precision score =  0.9496362618914381
Recall score =  0.8833940655908381
Test Performance:
              precision    recall  f1-score   support

           0       0.93      0.97      0.95     12316
           1       0.95      0.88      0.92      7684

    accuracy                           0.94     20000
   macro avg       0.94      0.93      0.93     20000
weighted avg       0.94      0.94      0.94     20000

Actual =  Counter({0: 12316, 1: 7684})
Predicted =  Counter({0: 12852, 1: 7148})
Out[80]: array([[11956,   360],
       [  896,  6788]])
```



Conclusion:

With different samples of data, comparisons between the models are made. The X Gradient model, which has an accuracy of 94%, is the best one, according to our research. The F1 score, Precision score, and recall score were also taken into account, and we discovered that the X-Gradient Boosting Classifier has the highest levels of all of these performance indicators. We only took into account these measures since they show how well the model anticipated the minority class, or the quantity of fraudulent claims, and so show how the model has performed.

The classification of healthcare provider fraud is a challenging issue that calls for careful evaluation of a number of variables, including feature choice, model choice, and data quality. A good fraud classification system should have a low rate of false positives and false negatives and be able to accurately identify dishonest healthcare practitioners. In conclusion, classification of healthcare provider fraud is an important topic of research with substantial ramifications for preventing and detecting healthcare fraud. To keep up with emerging fraud schemes and safeguard the credibility of the healthcare system, it is crucial to keep researching and creating new methods for identifying healthcare provider fraud.