April 22, 2023

Healthcare Provider Fraud Detection

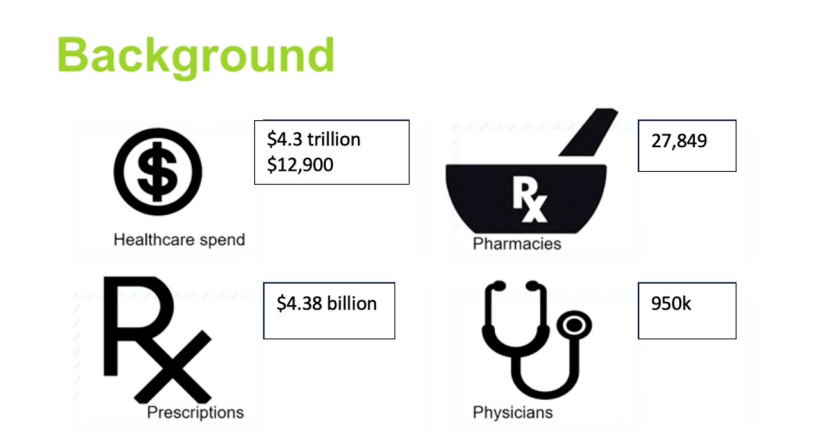
Team Name: Detective Elites

Team Members: Mounika Kandru
Sai Charit Kondepati
Vaishnavi Babli

Table of contents

Introduction

- 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, medical labs, nurses, lab assistants, and others), physicians, and beneficiaries acting together to make fraudulent 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.



We're doing it #thestateway

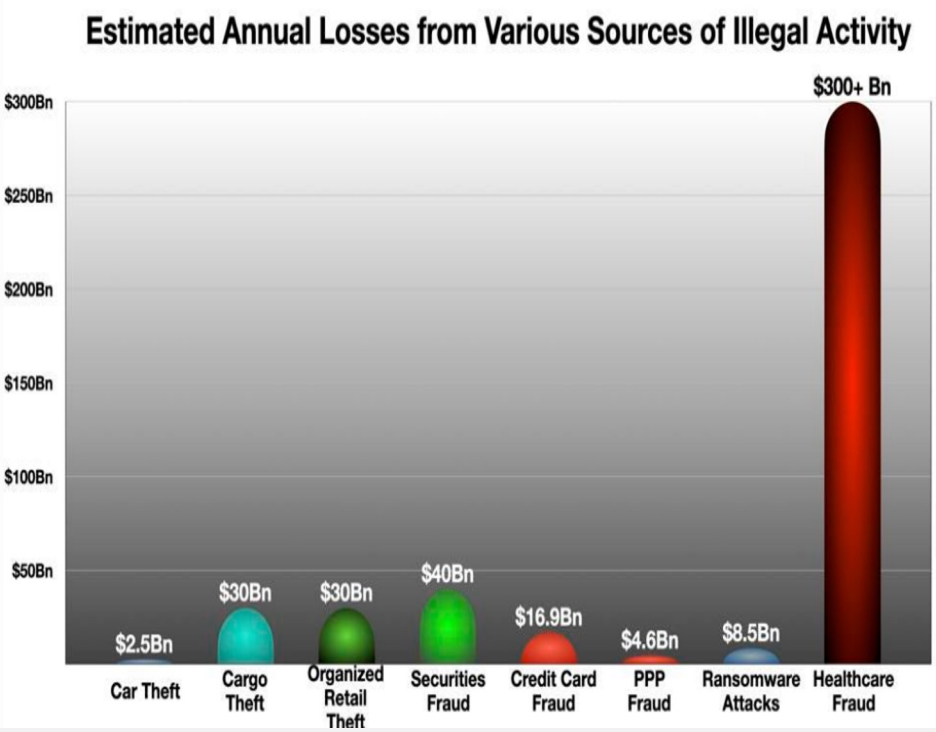
Introduction

Impact on Society:

• Waste of funds that would have been otherwise used for providing better medical treatments to actual patients.

• It is a criminal act as patients were either given false medicines or procedures which were not required.

• 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.



Business Problem

- The National Healthcare Anti-Fraud Association estimated that approximately tens of billions of dollars are lost due to healthcare fraud each year and as result healthcare is becoming costly matter day by day.
- This immense financial loss places the responsibility of recovery on insurance companies which are the most vulnerable institutions impacted due to these frauds, and also more importantly the patients. Patients are cheated into compensating for the cost in primarily two ways: payment of fraudulent copays and higher insurance premiums.

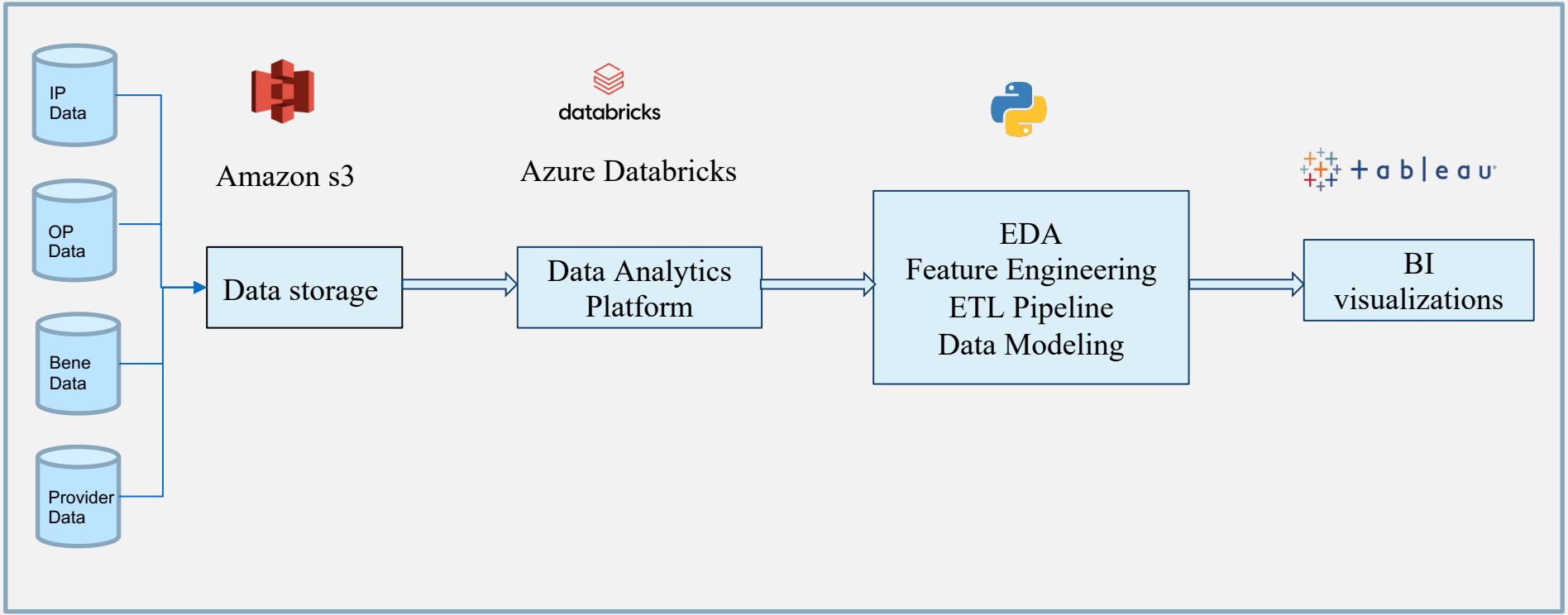
Business Solution

- The goal of this project is to ‘predict the potentially fraudulent providers’ based on the claims filed by them that helps the Insurance company decide whether to accept or deny the claim or set up an investigation on that provider.
- 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 will help the insurance companies in easy detection of fraud and also would minimize patients’ financial losses which ultimately better serves the society.

Recapitulation

We're doing it #thestateway

Solution Architecture



We're doing it #thestateway

Data Source and Data Storage

Data Source:

• The data used in our project is retrieved from:
https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis

Introduction to the Dataset:

For the purpose of this project, we are considering Inpatient claims, Outpatient claims and Beneficiary details of each provider.

• **Inpatient Data**: This data provides insights about the claims filed for those patients who are admitted in the hospitals. It also provides additional details like their admission and discharge dates and admit diagnosis code etc.

• **Outpatient Data**: This data provides details about the claims filed for those patients who visit hospitals and not admitted in it.

• **Beneficiary Details Data**: This data contains beneficiary KYC details like health conditions, region they belong to etc.

• **Provider Data:** This data has the Provider and corresponding details whether it is fraudulent or not.

Data Storage:

•We will merge all the datasets and store it in Amazon S3 and would be accessing it in azure Databricks.

Exploratory Data Analysis

We have performed Exploratory data analysis for data preparation involving the below:

- **Data Cleaning:** We checked and treated the Null Values and replaced them with NaN.

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

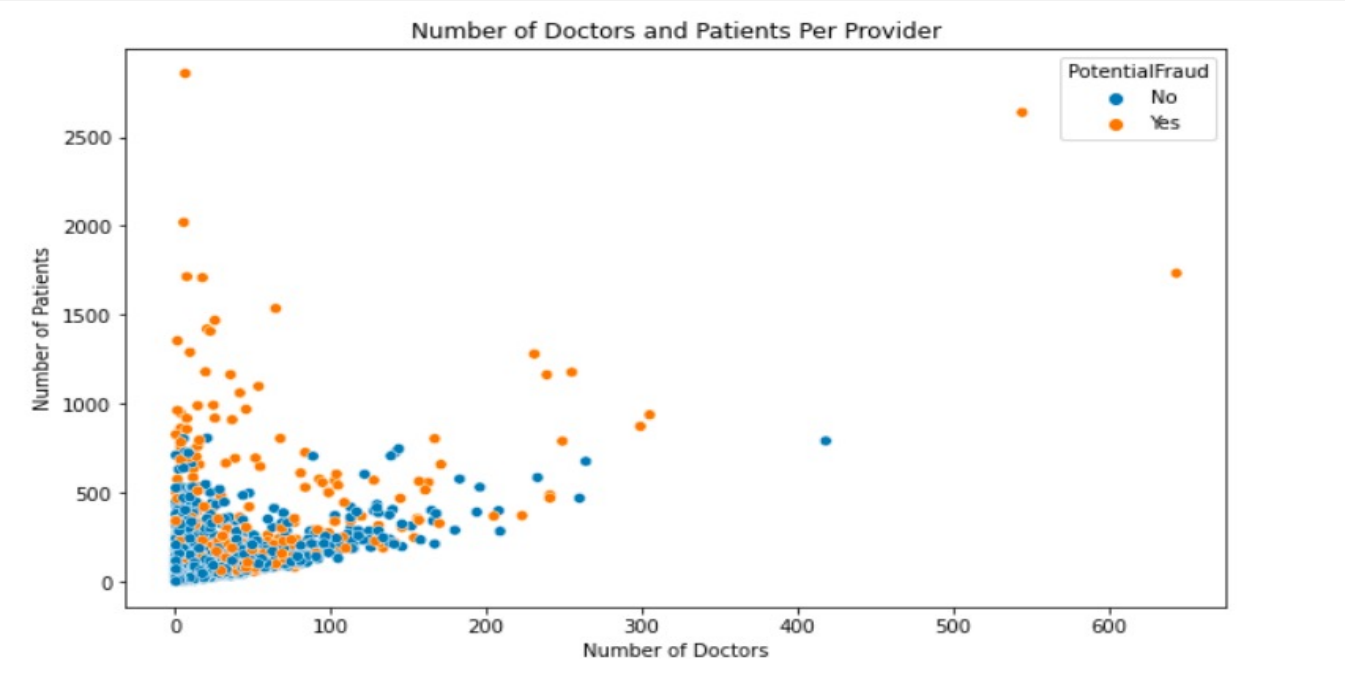
```
new_hc_info = hc_info.fillna('NaN')
pd.isnull(new_hc_info).sum()
```

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

- **Heatmap for correlation:** To determine which features are positively and negatively associated as well as which features have no correlation at all.

Data Visualization

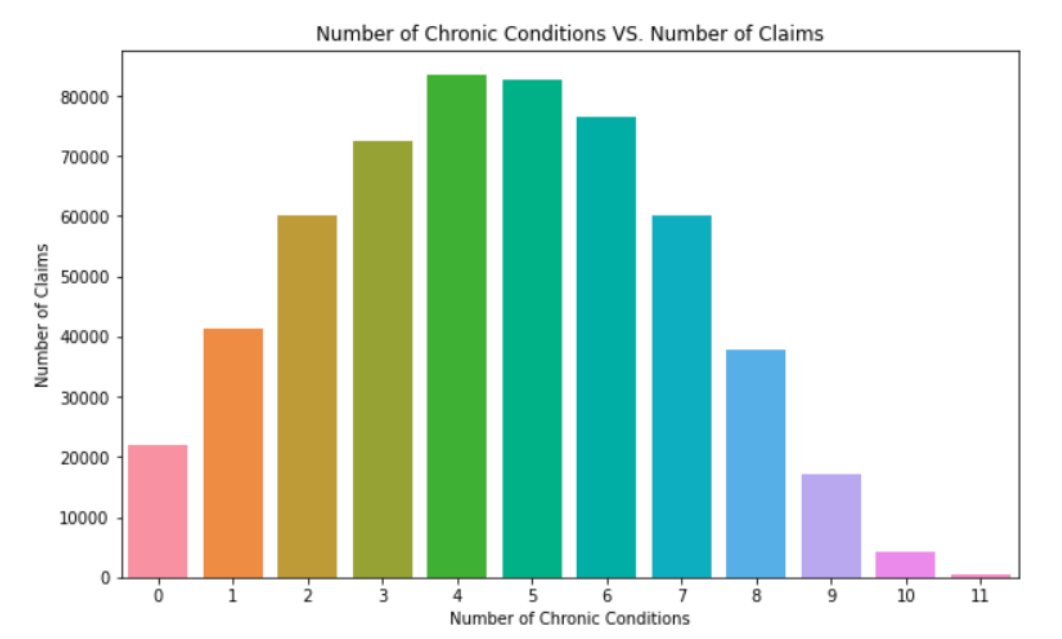
How does providers' number of doctors and patients help detect fraud?



We're doing it #thestateway

Data Visualization

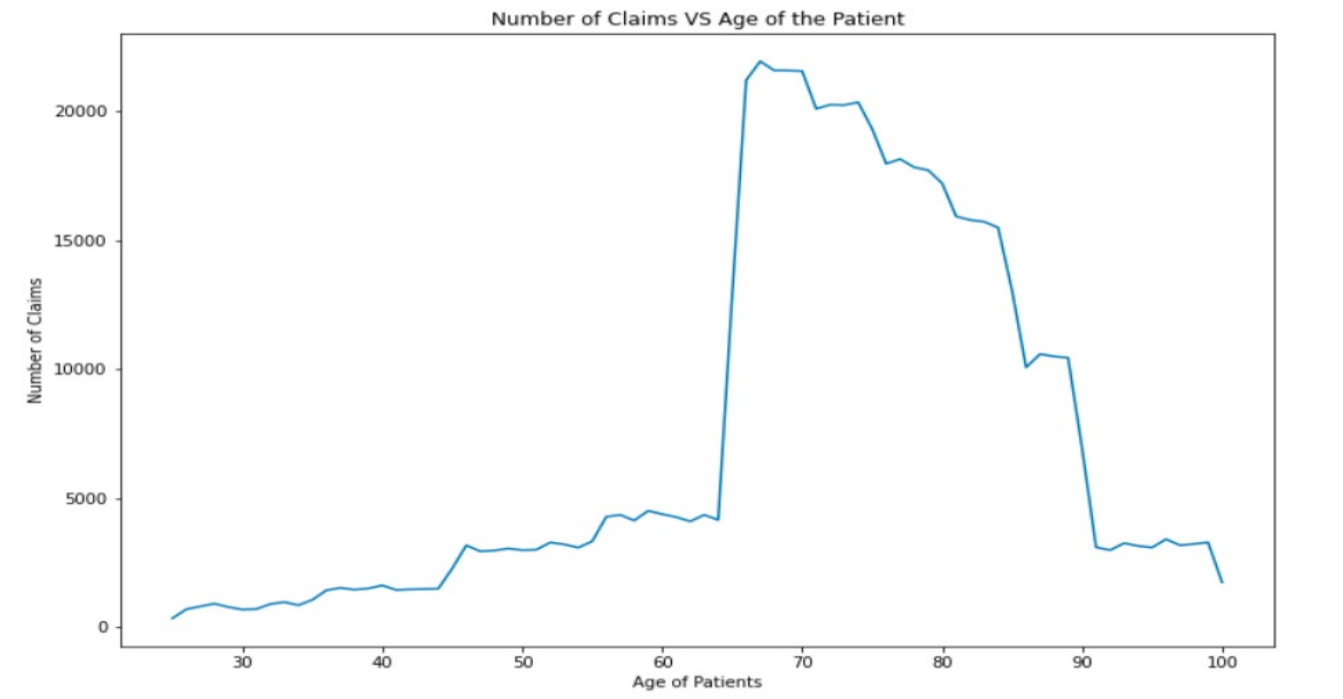
What is the relationship between the Chronic conditions and the Claims?



4. ChronicCond_Cancer
5. ChronicCond_ObstrPulmonary
6. ChronicCond_Depression

Data Visualization

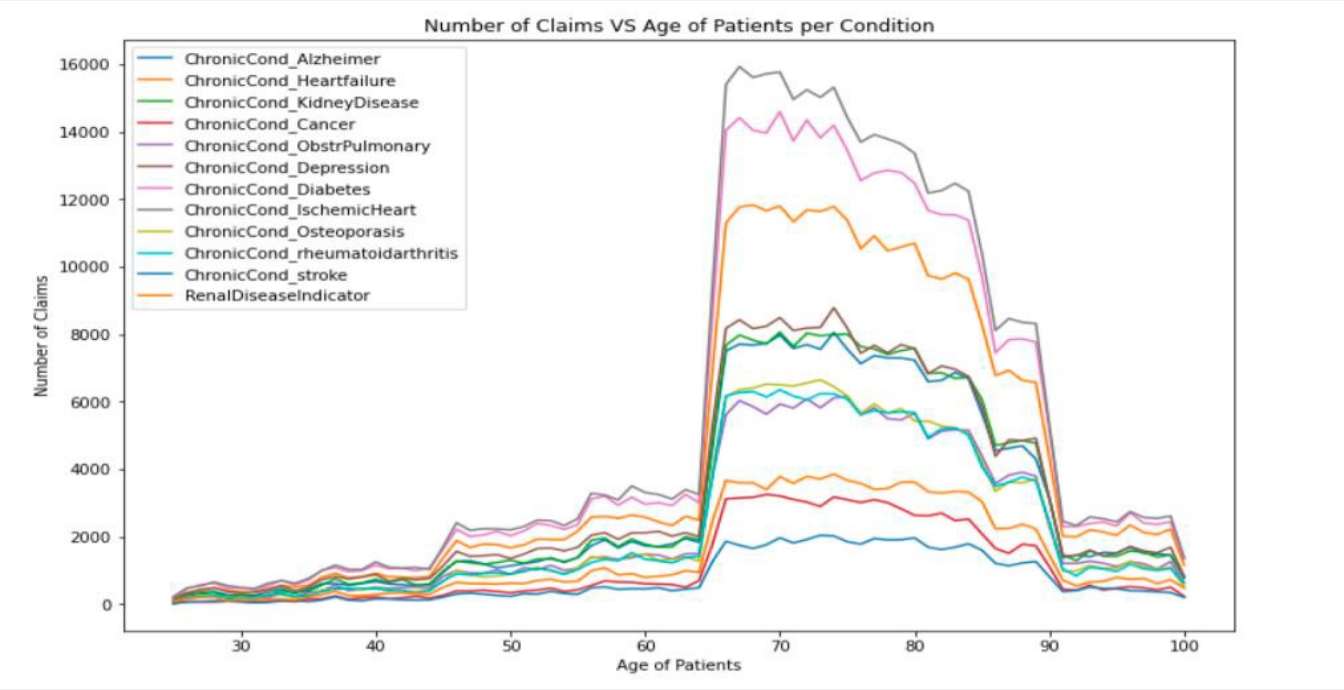
Relationship between Age of Patients and Claims



We're doing it #thestateway

Data Visualization

What is the trend between Age of Patients their Chronic conditions and Claims?

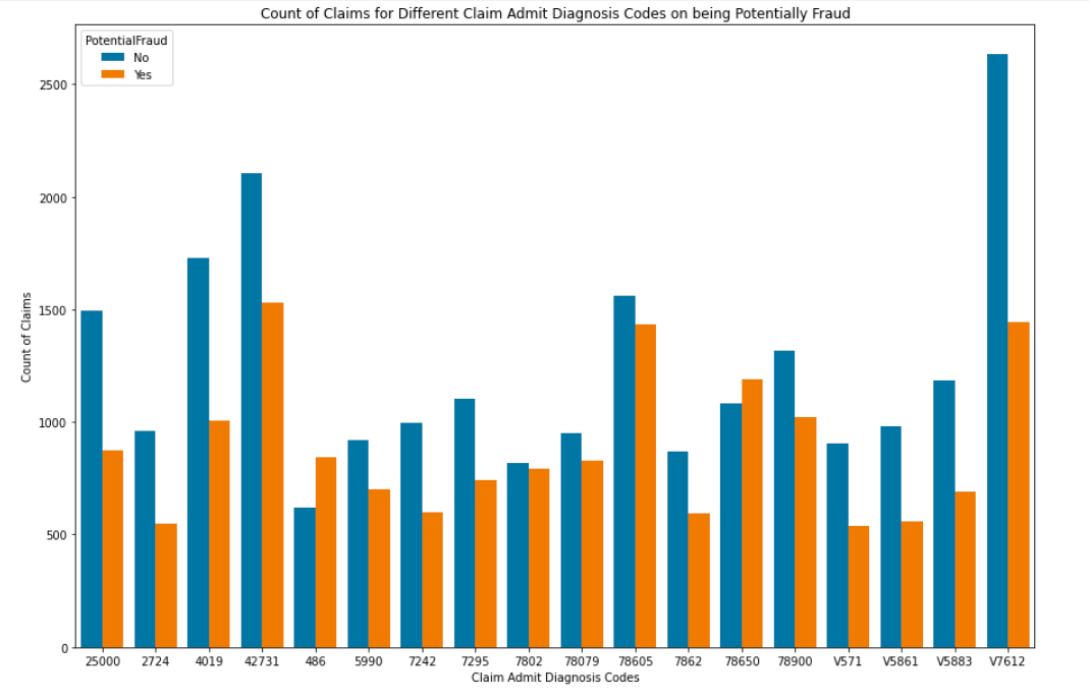


Top 3 Chronic Cond
Depression
Diabetes
Heart failure

We're doing it #thestateway

Data Visualization

Identifying Top 3 Admit Diagnosis Codes

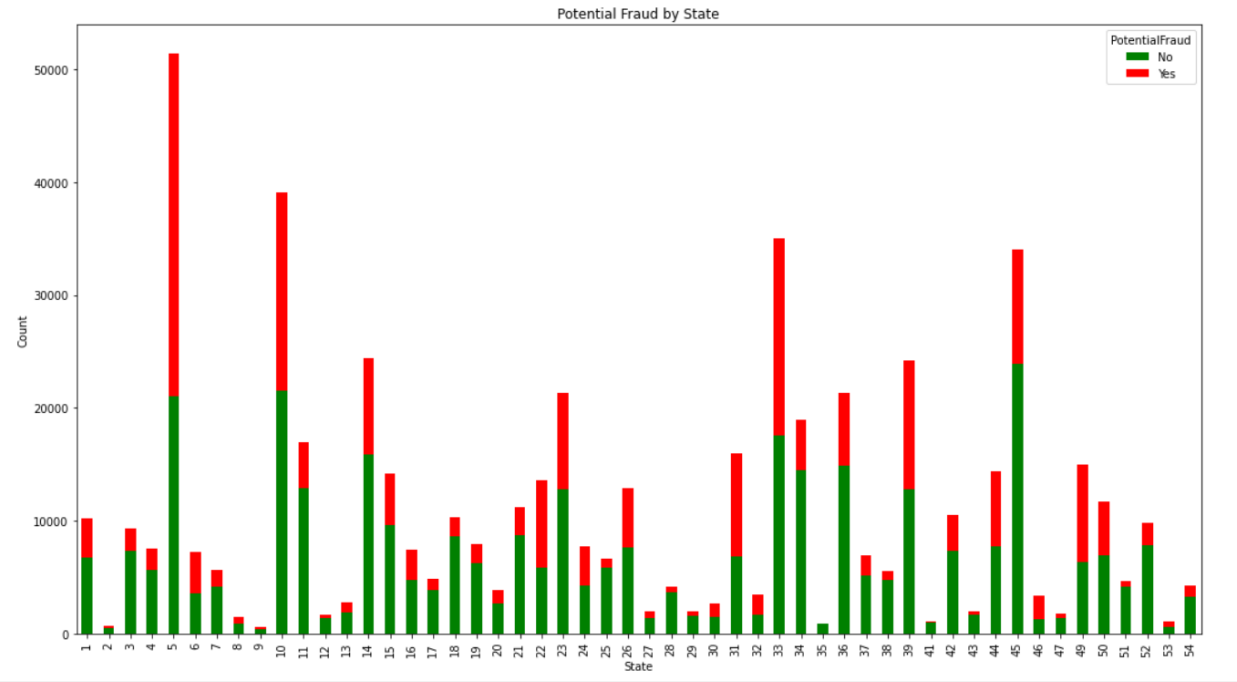


Top 3 Admit Diagnosis Codes
42731
78605
V7612

We're doing it #thestateway

Data Visualization

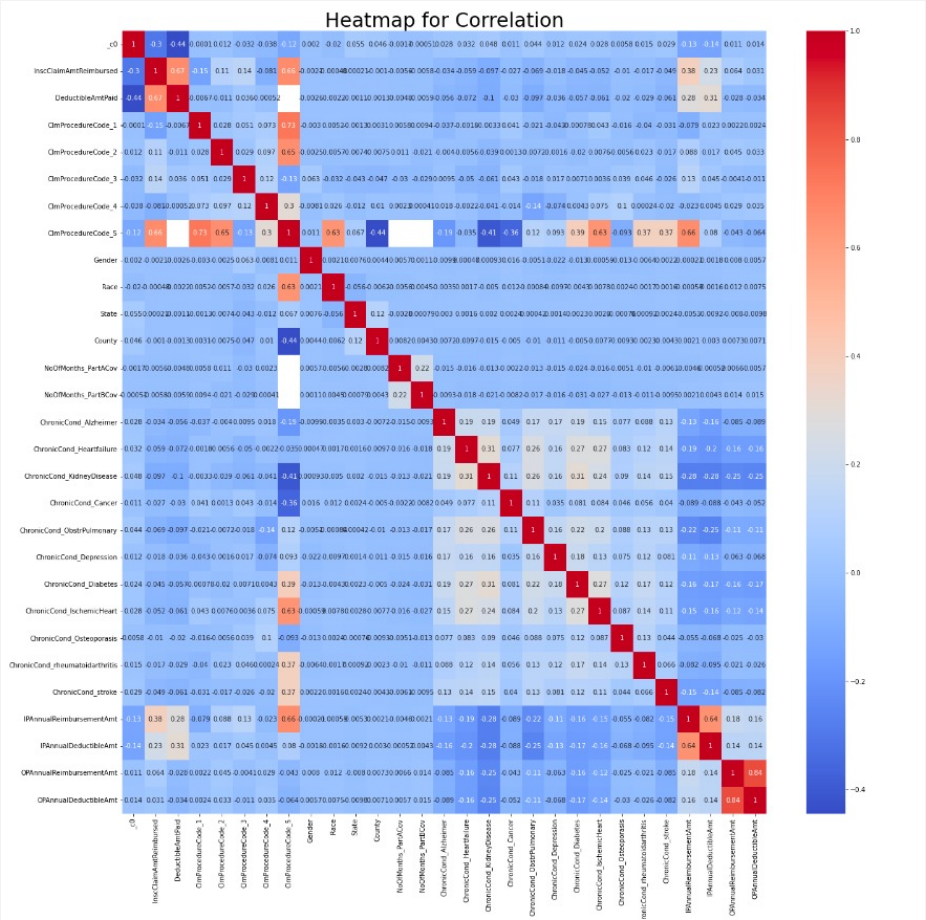
In which states most fraud occurs?



Top 3 state codes:
#5
#10
#33

We're doing it #thestateway

Data Visualization



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

We're doing it #thestateway

Feature Engineering

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

- **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()
```

Feature Engineering

- **Feature Scaling:** 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()
```

Feature Engineering

• **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.

• **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')
```

Feature Engineering

• **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

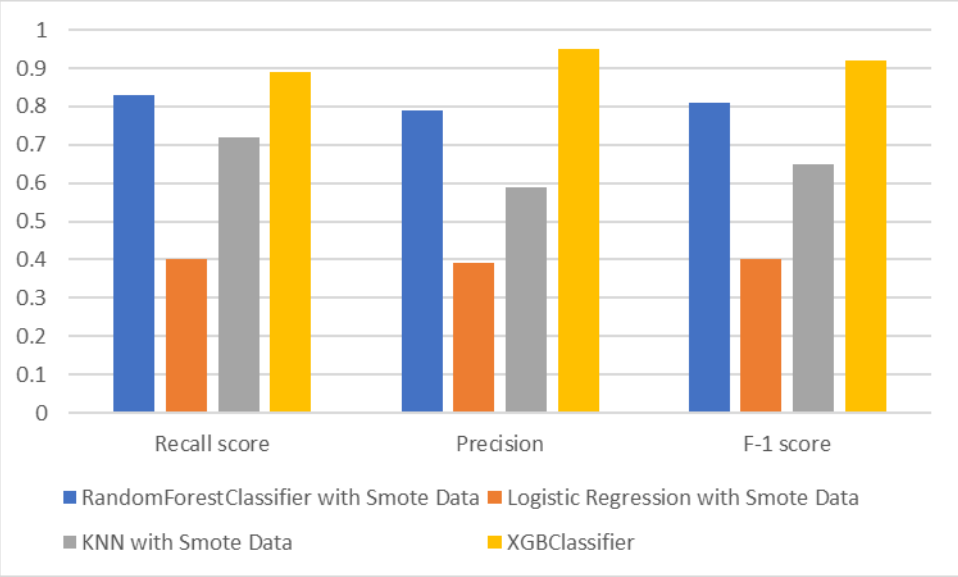
Data Modeling

- 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

Performance of Models

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



Data Modeling

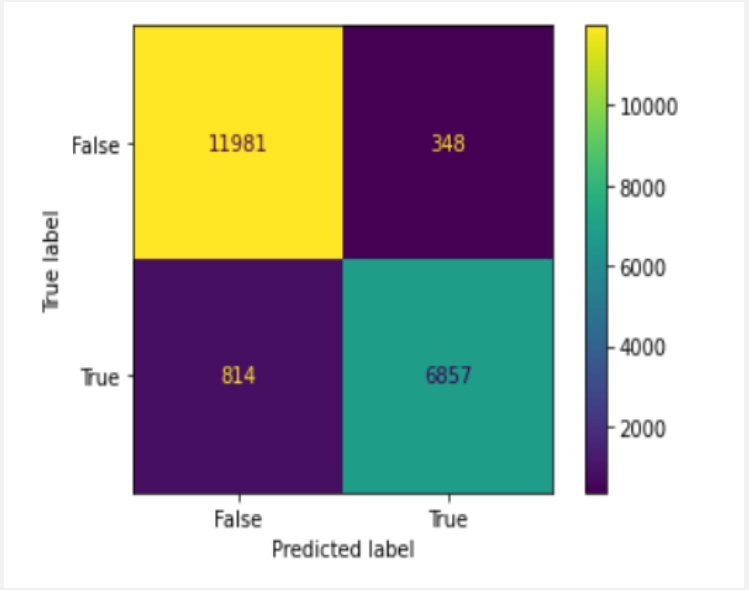
Evaluation Metrics for XGB Classifier:

```
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.

We're doing it #thestateway