WCD Machine Learning Project

•••

# Healthcare Provider Fraud Detection Analysis

Cheryl Chien

05.13.2023

- Intro
- ML in Healthcare & Types of Health Insurance Fraud
- ML workflow
- Target Variable
- Dataset EDA
- Features Used for Modeling
- Data Preprocessing
- Model Selection
- Summary

# Intro

## Why this topic?

- Want to work on healthcare IT and analysis
- Financial factors play an important role in healthcare decisions.
- Preventing health insurance fraud can ensure that resources are used fairly
- Prevent insurance premium increase

# ML in Healthcare & Types of Health Insurance Fraud

## ML in Healthcare

1. Fraud prediction
2. Risk assessment
3. Dx and Rx evaluation
4. Healthcare workforce automation

## Health Insurance Fraud

1. By provider, beneficiary, prescription
2. Billing for services that were not provided.
3. Duplicate submission of a claim for the same service.
4. Misrepresenting the service provided.
5. Charging for a more complex or expensive service than was actually provided.

# ML Project Workflow


**1. Planning**

Decide project topic and choose dataset

**2. Prepare Data**

Data review and cleaning

**3. EDA**

Use SQL and Python to explore the dataset and find trends

**4. Preprocessing & Modeling**

Tried two different processing and models

**5. Evaluation**

Test the accuracy

# Target Variable

1. Predict the potentially fraudulent providers-- YES/NO
2. Discover patterns of potentially fraudulent providers
3. Health insurance beneficiaries who are at greater risk of abuse, ex. Age, Health condition...

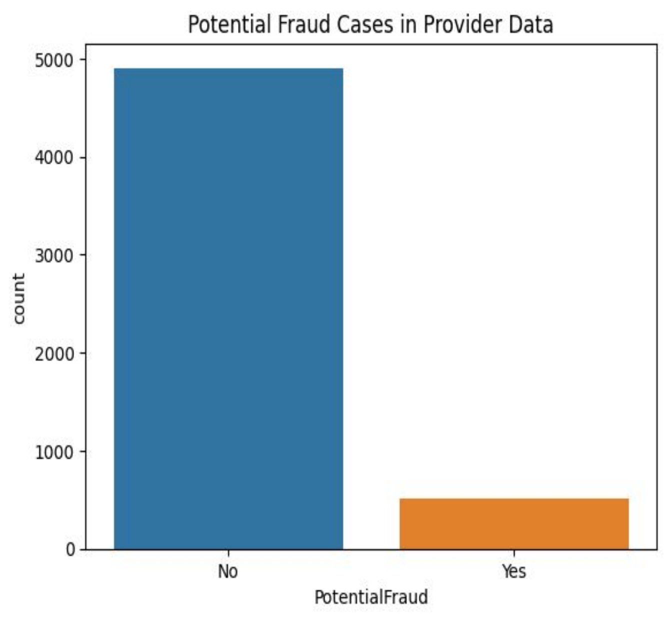
# About Dataset

- **Data source:** Kaggle [HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS]
- 4 Tables: Provider data, Beneficiary Details Data, Inpatient Data, Outpatient Data

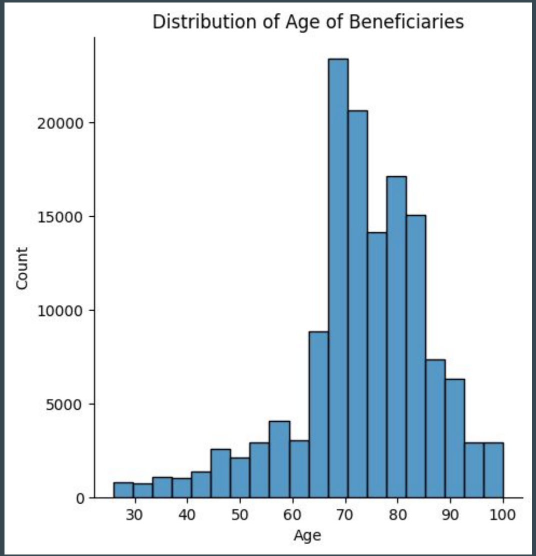
| | | |
|---|---|---|
| 01 | **Provider Data** | • Shape: 5410, 2<br>• 5410 unique providers<br>• Potential Fraud Yes/No |
| 02 | **Beneficiary Details Data** | • Shape: 138556, 25<br>• 138556 unique beneficiaries<br>• KYC like demographic, health conditions |
| 03 | **Inpatient Data** | • Shape: 40474, 30<br>• 40474 unique claimID<br>• Dates, provider, diagnosis, $ Reimburse |
| 04 | **Outpatient Data** | • Shape: 517737, 27<br>• 517737 unique claimID<br>• Dates, provider, diagnosis, $ Reimburse |

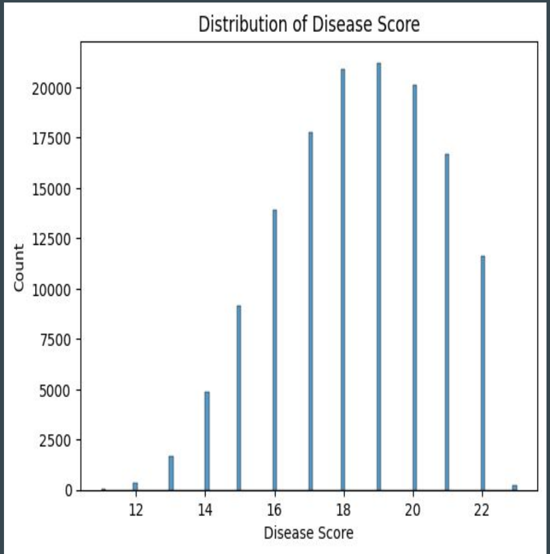
# Exploratory Data Analysis--1

## Provider Data



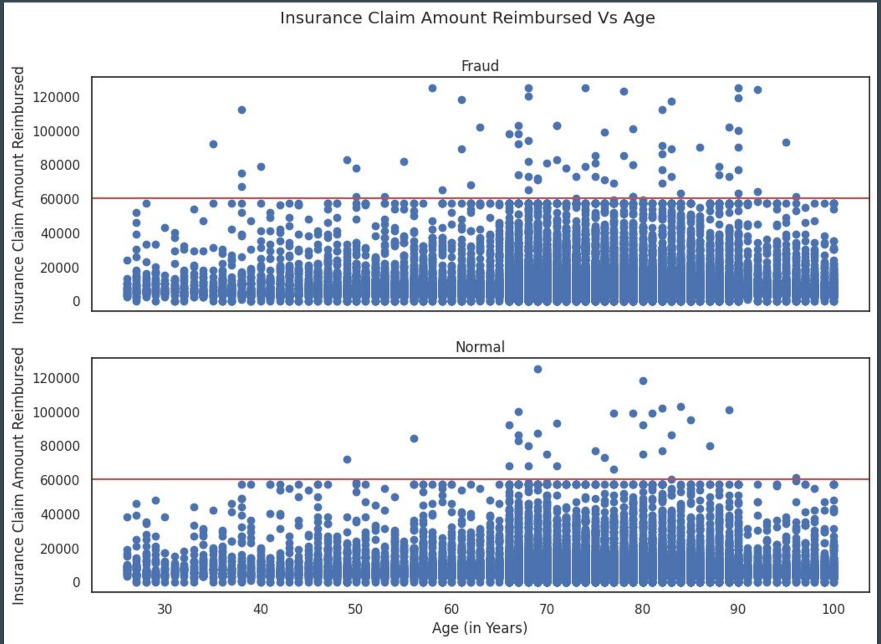

## Beneficiary Data

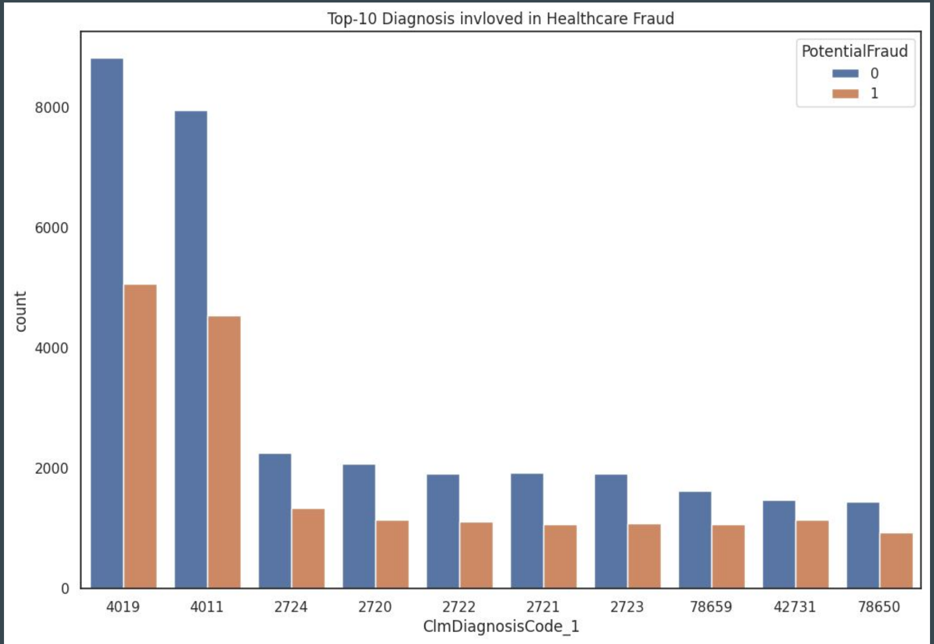


# Exploratory Data Analysis--2

## Joining all table:  Provider, Beneficiary, IP, OP Data



# Features Used for Modeling

Original Merged dataset shape: 558211 rows, 61 features
dtypes: bool(1); datetime64[ns](4), float64(8), int64(27), object(21)

>> Drop features

- **10 Unrelated features:** 'BeneID', 'ClaimID', 'ClaimStartDt', 'ClaimEndDt', 'has_claimed_op', 'DOB', 'DOD', 'NoOfMonths_PartACov', 'NoOfMonths_PartBCov', 'has_claimed_ip'

- **20 More than 50% null:** 'AttendingPhysician', 'OperatingPhysician', 'OtherPhysician', 'ClmDiagnosisCode_4', 'ClmDiagnosisCode_5', 'ClmDiagnosisCode_6', 'ClmDiagnosisCode_7', 'ClmDiagnosisCode_8', 'ClmDiagnosisCode_9', 'ClmDiagnosisCode_10', 'ClmProcedureCode_1', 'ClmProcedureCode_2', 'ClmProcedureCode_3', 'ClmProcedureCode_4', 'ClmProcedureCode_5', 'ClmProcedureCode_6', 'ClmAdmitDiagnosisCode', 'AdmissionDt', 'DischargeDt', 'DiagnosisGroupCode'

# Dataset used for Training

After dropping the features, the dataset will be used for training has 558211 rows and 31 features

>> **Separate X and y data**

X = Drop Provider and PotentialFraud columns

y = PotentialFraud

>>> **Split the data into train and test sets**

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Data Preprocessing 1

1. Replace PotentialFraud value from YES/NO to 1 and 0
2. Replace RenalDiseaseIndicator value from YES/ 0 to 1 and 0
3. **TargetEncoder:**

   Transform Object ClmDiagnosisCode_1, ClmDiagnosisCode_2, and ClmDiagnosisCode_3
4. **Filling missing value:**

   ClmDiagnosisCode_1, ClmDiagnosisCode_2, and ClmDiagnosisCode_3, and DeductibleAmtPaid

# Model Selection

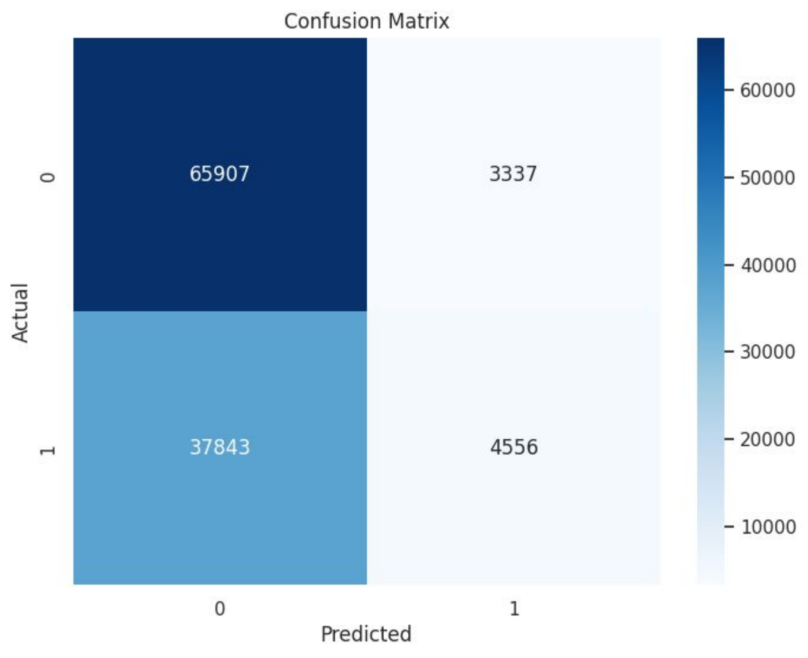
**Target variable:** Predict the potentially fraudulent providers-- **YES/NO**

**Features are labeled>> Supervised machine learning**

Tried **Logistic Regression and Random Forest Classifier**

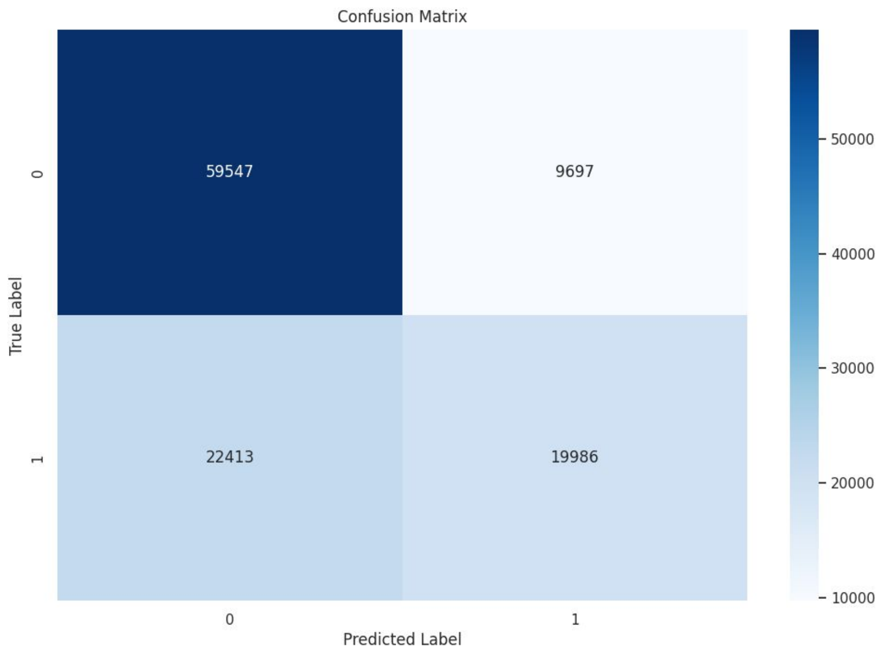
# Logistic Regression

Accuracy: 0.6311457055077345

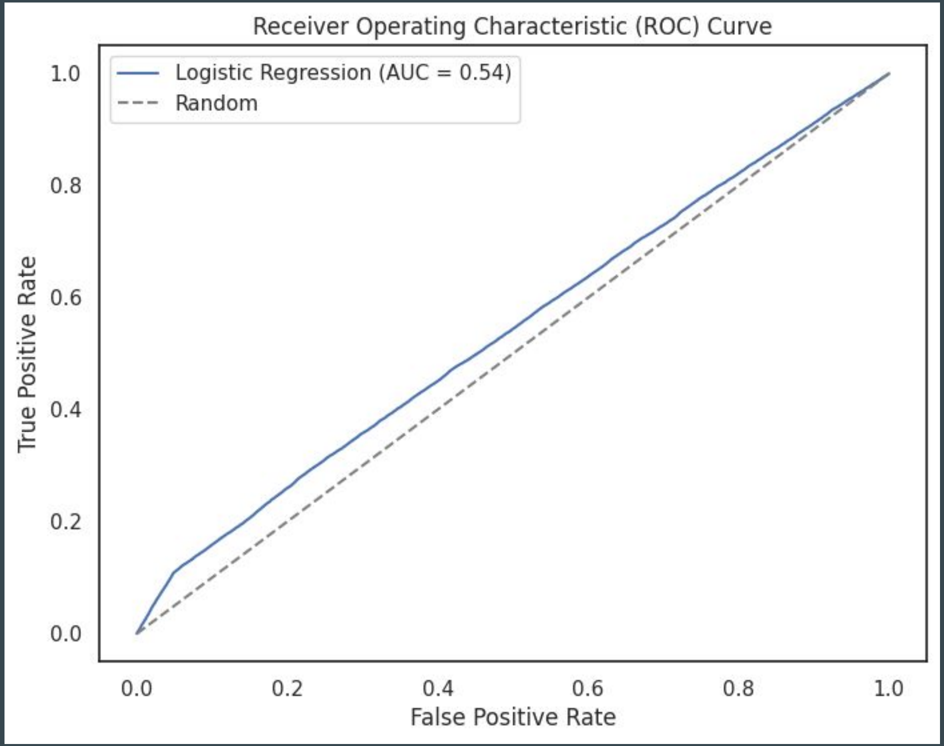



# Random Forest Classifier ✔

Accuracy: 0.7123868043674928



# Evaluation



Receiver Operating Characteristic (ROC) Curve
Logistic Regression (AUC = 0.54)
Random
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

# Data Preprocessing 2

1. StandardScaler
2. Imblearn.over_sampling

   > RandomOverSampler

   >> SMOTE

| Logistic Regression | Random Forest Classifier |
|---|---|
| Accuracy: 0.5517049882213843 | Accuracy: 0.6877457610418924 |

# Summary

1. From EDA:
   a. Overall fraud rate in inpatient and outpatient data is about 38.12%
   b. Older beneficiaries have a slight higher risk to be involved in fraud
   c. Most common codes those also involved in fraud are 4019, 4011
2. Data processing & Model selection

| | Logistic Regression | Random Forest Classifier |
|---|---|---|
| Data Processing 1 | Accuracy: 0.6311457055077345 | Accuracy: 0.7123868043674928 |
| Data Processing 2 | Accuracy: 0.5517049882213843 | Accuracy: 0.6877457610418924 |

# Limitation

1. Not sure the source and date of the dataset
2. Missing definition of some features, ex. Race, State, Chronic diseases
3. Could drop more features provider and Has_ChronicCondition
4. Could test more models like xgboostclassifier

# Thank You

# Reference


1. Health Care Fraud — FBI
2. Medicare Fraud & Abuse: Prevent, Detect, Report
3. HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS | Kaggle
4. GitHub - Atharvak19/Big-Data-Medicare-Fraud-Detection