Fraud Detection in Healthcare

Data Science in Healthcare - Final Project

Sakshi Shende, Jose Gerala

Problem Statement

01 Provider Fraud in Medicare is causing financial **losses exceeding $100 billion annually** and a surge in healthcare costs.

02 Health care fraud is a crime that involves misrepresenting information, concealing information, or deceiving a person or entity, and leads to reduced benefits, coverage, and increased insurance expenses.

Our goal is to **predict fraudulent providers**, enhancing detection through claims analysis and key variable identification.

Survey of Existing Solutions

- **Rule-Based Systems** use predefined criteria to detect anomalies
- **Data Analytics and Machine Learning** aims to identify patterns indicative of fraud for more effective detection
- **Predictive Modeling** utilizes historical data to foresee future instances of fraud
- **Anomaly detection** detects abnormal billing patterns or unexpected patient behaviors.
- **Social Network Analysis** examines relationships between entities like providers and patients to uncover suspicious connections

Data Description

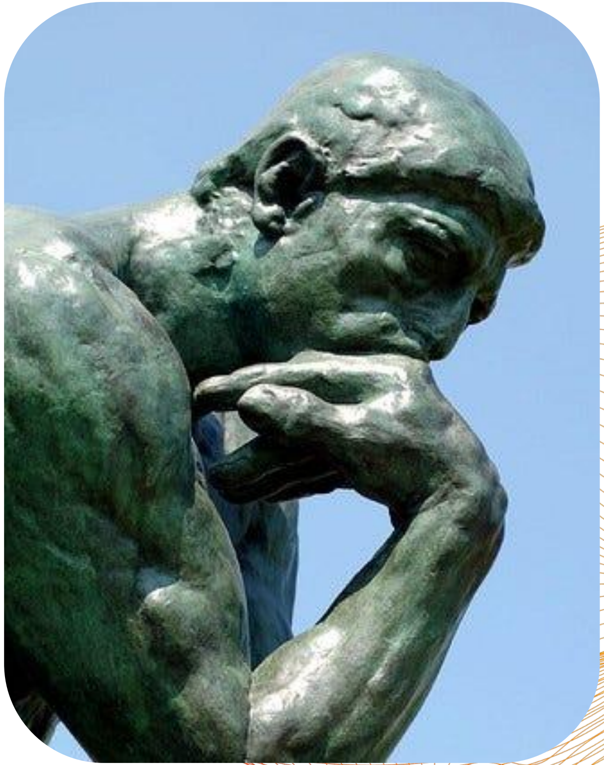
- 4 Datasets
  - **Beneficiary Data**: Beneficiary KYC details like DOB, DOD, Gender, Race, health conditions, etc.
  - **Inpatient Data**: Claim details of the patients who were admitted into the hospitals.
  - **Outpatient Data**: Claim details of the patients visited the hospital but were not admitted.
  - **Provider Information**: Healthcare provider information and its corresponding fraud status.

Source: https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis

Assumptions and Hypotheses

01 Assumptions:

- The **data** from inpatient claims, outpatient claims, and beneficiary details **adequately represent diverse patterns associated with healthcare fraud.**

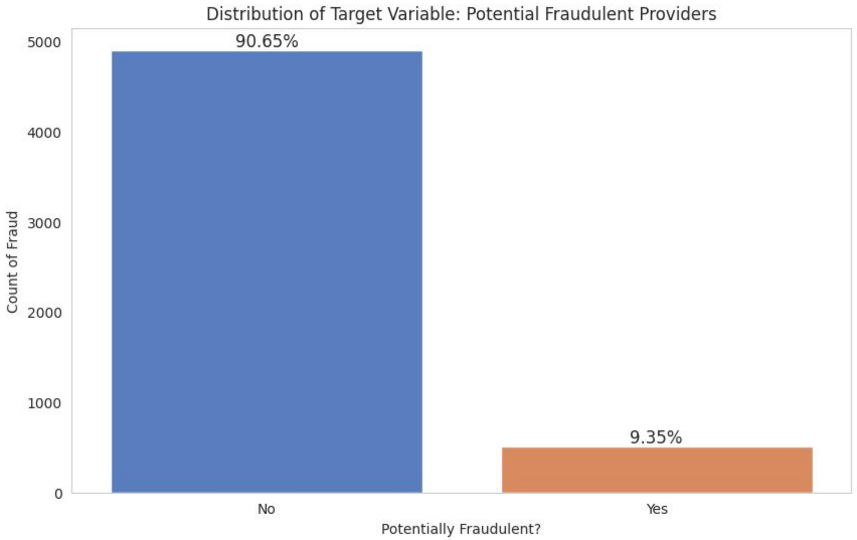


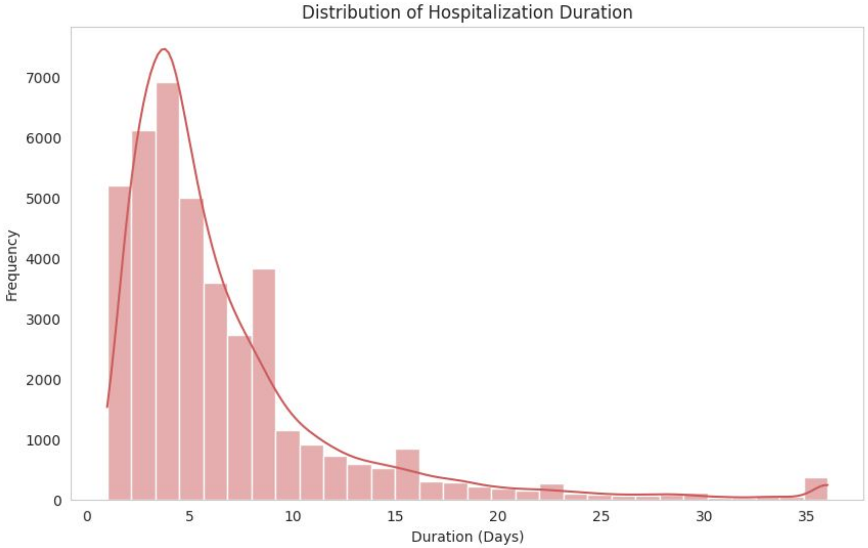
02 Hypotheses:

- Fraudulent behavior among healthcare providers exhibits **temporal consistency**, enabling the model to leverage historical data for future predictions.

- Specific provider features are crucial in identifying potentially fraudulent providers.

Exploratory Data Analysis

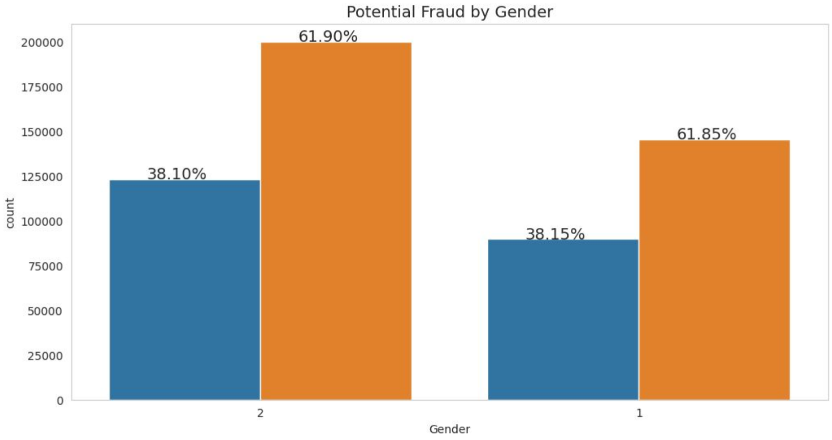


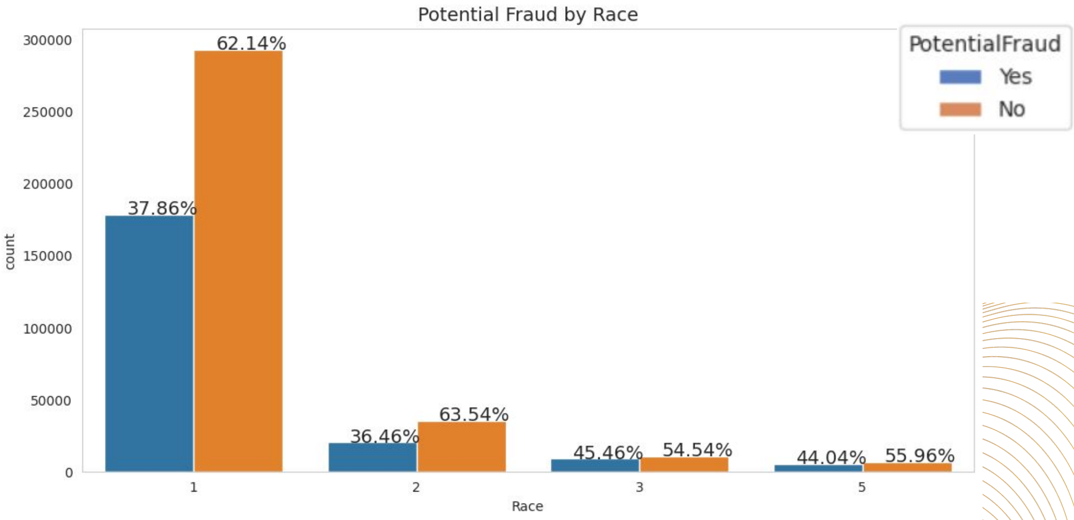


- The target variable is highly imbalanced.
- There are 9.3% fraudulent providers and 90% non-fraudulent providers.

- Distribution of hospitalization duration is left skewed
- May indicate fictitious admission, where providers create false records to generate fraudulent claims.

Exploratory Data Analysis





- Gender 2 showcased higher number of fraudulent activities than Gender 1.
- This can be viewed as individuals of gender 2 are more likely to be targeted by fraudsters.

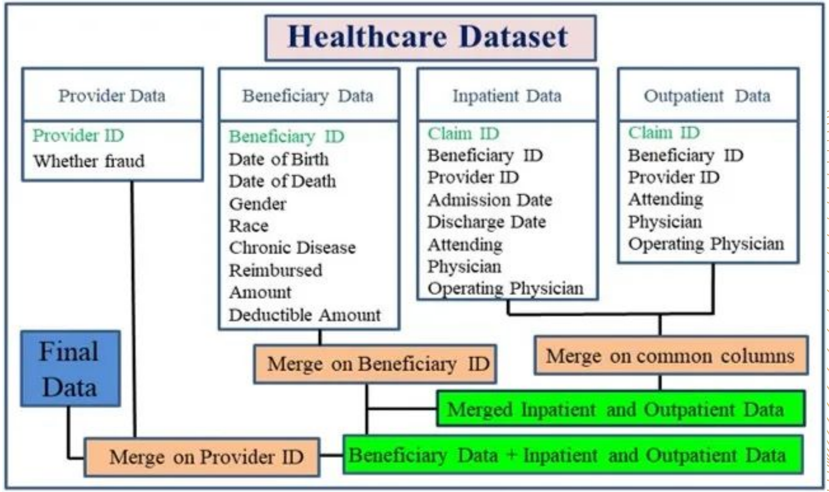
- Individuals from Race 1 are associated with higher likelihood of engaging in fraudulent activities.
- Ratio of fraudulent transaction is highest for Race 3.

Feature Engineering

01 We initially had 4 datasets. Inpatient, Outpatient, Beneficiary and Provider data. Fraudulent providers were targeted.

02 What we did in a nutshell:

- Imputed zeros for missing data
- One-hot encoding
- Merged all datasets together
- Grouped by provider applying mean and count operations to all numeric variables.



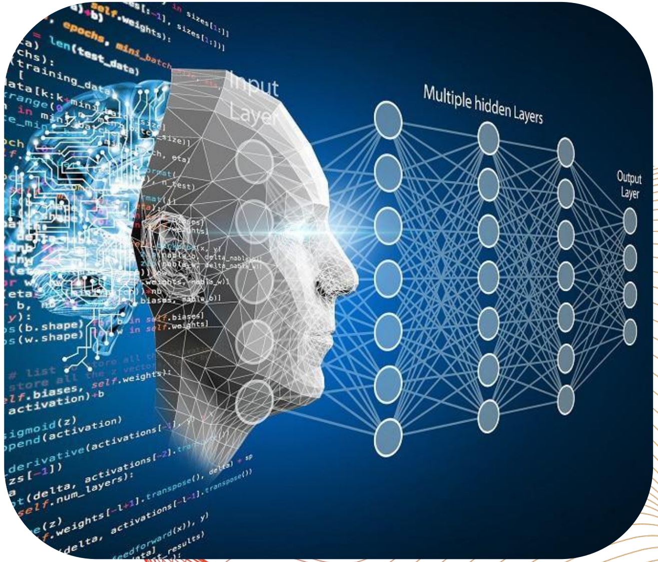
Custom Models

01 Custom Deep NN 1

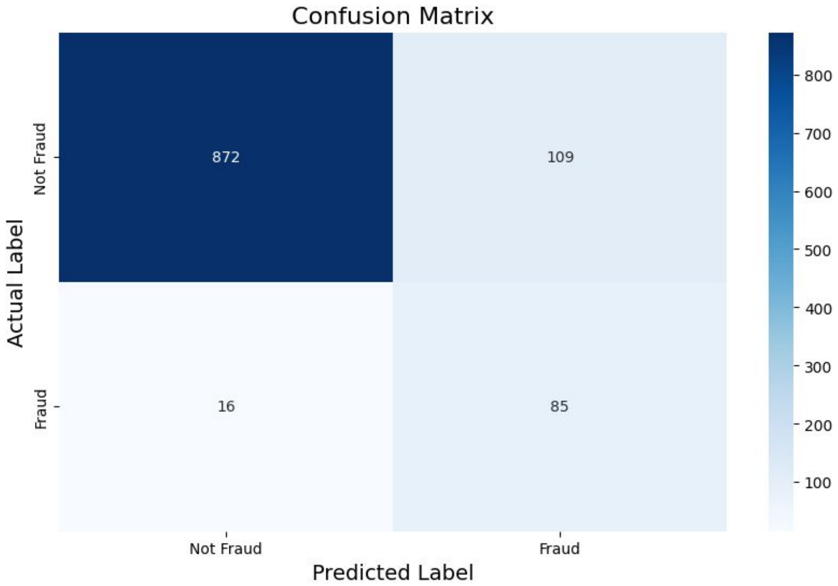
- **Strengths**: Batch normalization improved precision for both fraudulent and non-fraudulent transactions.
- **Weaknesses**: Limited interpretability and limitation in adapting to new fraud patterns.

02 Custom Deep NN 2

- **Strengths**: High precision for non-fraudulent transactions and precision improvement for fraudulent transactions by almost double (64%).
- **Weaknesses**: Interpretability challenges typical of deep learning models.



Baseline Model: Random Forest

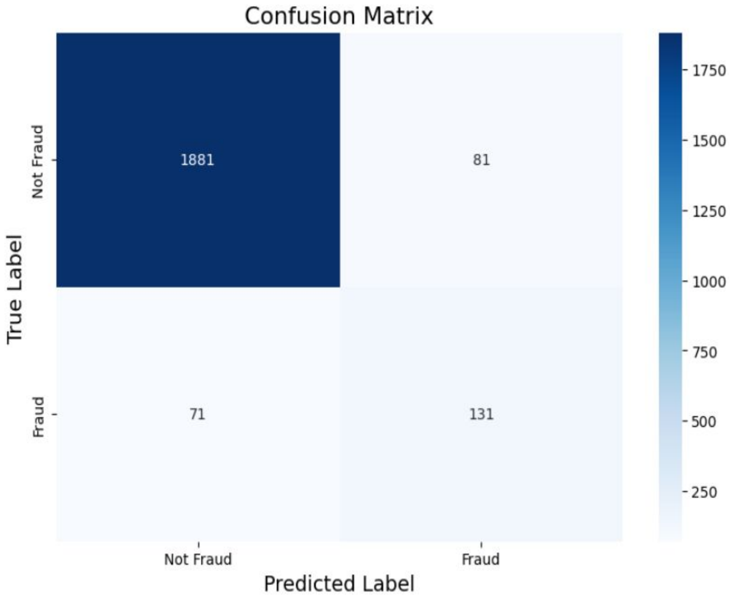


Classification Report	Precision	Recall	F1-Score	Support
0	0.98	0.89	0.93	981
1	0.44	0.84	0.58	101

Best Hyperparameter

Hyperparameter	Value
max_depth	5
n_estimators	200
max_features	auto
random_state	42

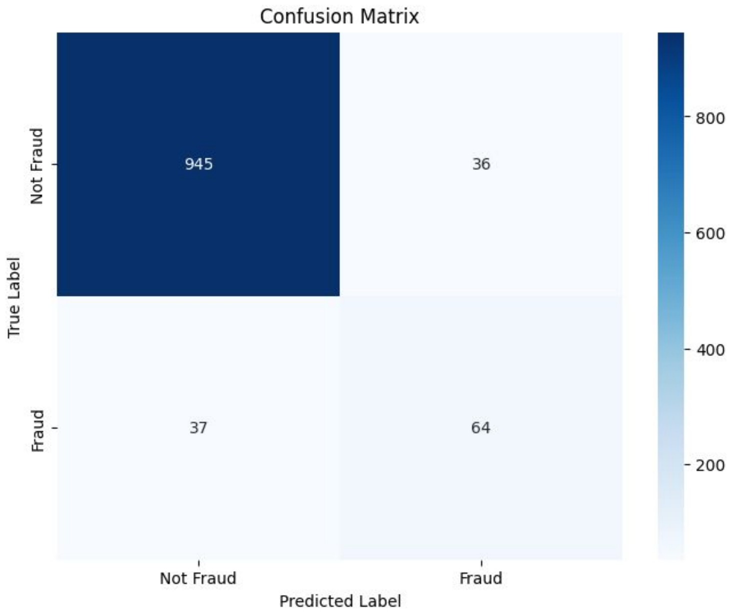
Custom NN Model 1



Classification Report	Precision	Recall	F1-Score	Support
0	0.96	0.96	0.96	1962
1	0.62	0.65	0.63	202

```
Model: "custom_nn_model_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_1 (Dense)             (None, 512)               120832

 batch_normalization_1 (Bat  (None, 512)               2048
 chNormalization)

 dense_2 (Dense)             (None, 128)               65664

 batch_normalization_2 (Bat  (None, 128)               512
 chNormalization)

 dense_3 (Dense)             (None, 32)                4128

 batch_normalization_3 (Bat  (None, 32)                128
 chNormalization)

 dense_4 (Dense)             (None, 1)                 33

=================================================================
Total params: 193345 (755.25 KB)
Trainable params: 192001 (750.00 KB)
Non-trainable params: 1344 (5.25 KB)
```

Custom NN Model 2

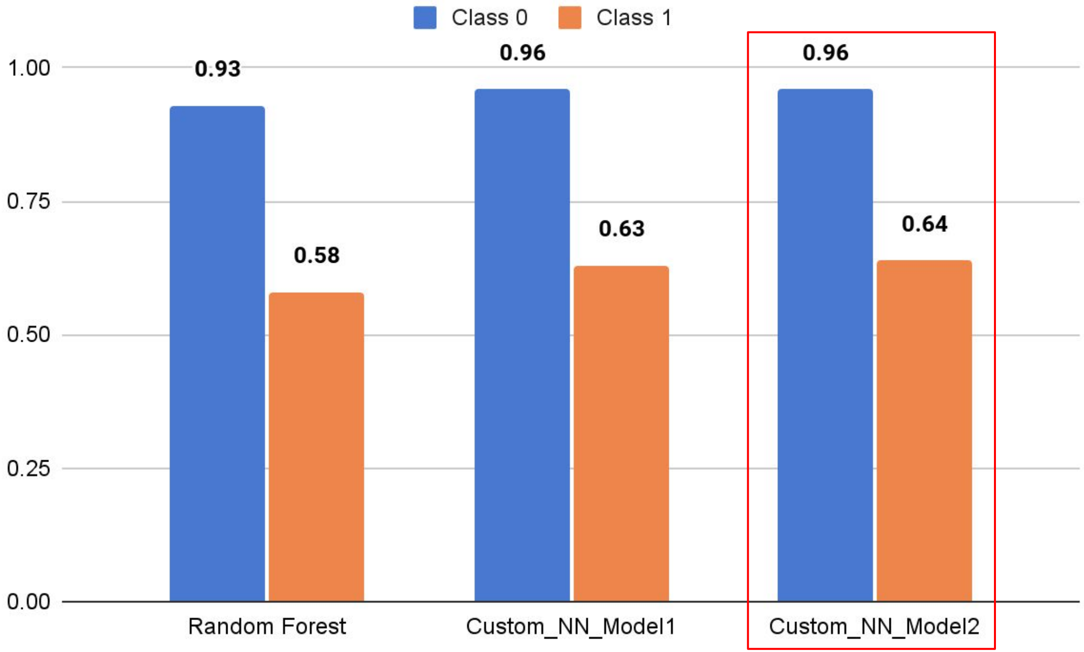


Classification Report	Precision	Recall	F1-Score	Support
0	0.96	0.96	0.96	981
1	0.64	0.63	0.64	101

```
Model:  "custom_nn_model_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_5 (Dense)             (None, 236)               55932

 dense_6 (Dense)             (None, 128)               30336

 dense_7 (Dense)             (None, 64)                8256

 dense_8 (Dense)             (None, 32)                2080

 dense_9 (Dense)             (None, 1)                 33

=================================================================
Total params: 96637 (377.49 KB)
Trainable params: 96637 (377.49 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
Batch size: 128
Epochs: 300
```

Model Evaluation

For imbalanced dataset, F1 Score, Classification Report and Confusion Matrix are vital to evaluate and understand model performance.



Proposed Solution and Results

Custom_NN_Model2 showcased superior performance in comparison to other models.

Model Justification

- **Optimal Dimensions:** Initial layer with 236 neurons aligns with dataset columns for efficient information processing.
- **Architecture Empirical Superiority:** Experimentation confirmed superior performance compared to alternative architectures.
- **Normalization Not Vital:** Standardizing data offered minimal benefit, indicating the model's adaptability to feature scales.
- **Dropout Unnecessary:** Best F1 score achieved without dropout, balancing complexity and overfitting.
- **Effective Class Weights:** Class weights address imbalance, enhancing predictions for the minority class without compromising on the majority.

Trade-offs

- **Reduced Interpretability:** Increased depth sacrifices some interpretability, challenging intuitive understanding of layer significance.
- **Hyperparameter Sensitivity:** Performance may be sensitive to hyperparameter values, requiring careful tuning.
- **Computational Demands:** Deeper architecture may demand more computational resources during training and inference.

Healthcare Impact

01 Benefits:

- Curbing financial losses.
- Safeguarding financial integrity.
- Preserving principles of affordable and effective healthcare.

02 The solution integrates into healthcare decision systems, offering **real-time monitoring** and **automated alerts** through advanced claims analysis. It serves as a decision support tool for administrators and financial analysts.

Limitation and Future Work

01 Incorporate Additional Data

- A **broader dataset** could provide a more holistic view, potentially leading to improved detection accuracy and a deeper understanding of fraudulent patterns.

02 Feature Importance Challenges

- **Understanding feature importance** in neural networks challenging due to complex, non-linear relationships. Consider Model-Agnostic Techniques like Permutation Importance or SHAP (SHapley Additive exPlanations) for identifying key fraud-related features.

Thank you!

Please feel free to ask any questions. 😄

APPENDIX

Data Description

Beneficiary Data (Train and Test)

Column Name	Description
BeneID	Unique identifier of the beneficiary (patient)
DOB	Date of Birth of the beneficiary
DOD	Date of Death of the beneficiary
Gender, Race, State, Country	Gender, Race, State and Country of the beneficiary
RenalDiseaseIndicator	Indicates whether the patient has existing kidney disease
ChronicCond_*	Indicates if the patient has that particular disease existing. Also indicates the risk score
IPAnnualReimbursementAmt	Maximum reimbursement amount for hospitalization annually
IPAnnualDeductibleAmt	Premium paid by the patient for hospitalization annually
OPAnnualReimbursementAmt	Maximum reimbursement amount for outpatient visits annually
OPAnnualDeductibleAmt	Premium paid by the patient of outpatient visits annually

Outpatient Data (Train and Test)

Column Name	Description
BeneID	Unique identifier for each beneficiary (patient)
ClaimID	Unique identifier of the claim submitted by the provider
ClaimStartDt	Date when the claim started (yyyy-mm-dd)
ClaimEndDt	Date when the claim ended (yyyy-mm-dd)
Provider	Unique identifier of the provider
InscClaimAmtReimbursed	Amount reimbursed for that particular claim
AttendingPhysician	Unique identifier of the physician who attended the patient
OperatingPhysician	Unique identifier of the physician who operated on the patient
OtherPhysician	Unique identifier of the other physician
ClmDiagnoisiCode	Code of diagnosis performed by the provider on the patient
ClmProcedureCode	Code of procedure of the patient for treatment
DeductibleAmtPaid	Amount paid by the patient.

Inpatient Data (Train and Test)

Column Name	Description
AdmissionDt	Date on which the patient was admitted (yyyy-mm-dd)
DischargeDt	Date on which the patient was discharged from the hospital (yyyy-mm-dd)
DiagnosisGroupCode	Group code for the diagnosis done on the patient

Provider (Train and Test)

Column Name	Description
ProviderID	Unique identifier for the provider
PotentialFraud	Fraud status (Yes/No)