## Insurance Frauds:

**About Healthcare Frauds:**

There are few organized health care groups who commit frauds. They take advantage of the loop holes present in the industry and do fraudulent medical claims which results in exponential increase of Medicare spending. These frauds may/ may not involve(s) a service providers, physicians, beneficiaries. Medical codes represent the treatment provided. These codes are not understandable by layman. Fraudsters take advantage of this and manipulate the medical codes and submit the claim

Most common frauds committed in health care are as follows:

Set up claims for procedures and services that did not take place. Billing for services using patient information that are obtained through identity theft. Submitting the same claim multiple times for the same service. Treatments which are not covered by insurance are falsely represented as procedures which are covered by insurance. This type of crime happens in cosmetic procedures. Do irrelevant medical procedures on patients in order to get payment from the insurance companies. Charge extra from the insurance providers and defer the patient's payment

**Potential Clients:**

Insurance providers/ industry.

**Dataset:**

Open source public data from Kaggle that has over a million data, each record containing hospital visits in the US region. The data set can be found [here](https://kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis?utm_medium=social&utm_campaign=kaggle-dataset-share&utm_source=linkedin)

The dataset is of .CSV format and has been divided into 3 csv files. They contain Inpatient claims details, Outpatient claims details and Beneficiary details. Another csv classifies whether the service providers (hospitals) are fraudulent or not. Using this we can train the model.

**Dataset Information:**

1) Beneficiary Details Data

**Dataset Name: Train\_Beneficiarydata-1542865627584.csv**

**(Rows, Columns)**: (138556, 25)

**Header**: BeneID; DOB; DOD; Gender; Race; RenalDiseaseIndicator; 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

**Description**: This data set contains the details of each patient

2) Inpatient Data

**Dataset Name: Train\_Inpatientdata-1542865627584.csv**

**(Rows, Columns**): (40474, 30)

**Header**: BeneID;ClaimID;ClaimStartDt;ClaimEndDt;Provider;InscClaimAmtReimbursed;AttendingPhysician;OperatingPhysician;OtherPhysician;AdmissionDt;ClmAdmitDiagnosisCode;DeductibleAmtPaid;DischargeDt;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\_

**Description**: This dataset contains the details of the claims filed. The patients are admitted in the hospital

3) Outpatient Data

**Dataset Name: Train\_Outpatientdata-1542865627584.csv**

**(Rows, Columns)**: (517737, 27)

**Header**: BeneID;ClaimID;ClaimStartDt;ClaimEndDt;Provider;

InscClaimAmtReimbursed;AttendingPhysician;OperatingPhysician;

OtherPhysician;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;DeductibleAmtPaid;

ClmAdmitDiagnosisCode

**Description:** This dataset contains the details of the claims filed. The patients are not admitted in the hospital

**To be Predicted:** Provider

**Approach:**

* Variable Identification
* Merge 3 datasets together - consolidation (either by joins or unions)
* Each record in the dataset resembles a visit made by the patient to the service provider (hospital/ physician). But we have to predict if the service provider is fraudulent or not. So we need to remove the hospital visit grain and group by the service providers
* Categorical columns (flags) are either 1 or 2. Need to replace 2 by 0
* Clean the dataset - check for missing values and outliers
* Univariate analysis
* Bivariate analysis
* Missing value treatment
* Outlier treatment
* Feature engineering
* Build ML models to predict whether the service provider is fraud

**Input Columns:**

|  |  |  |
| --- | --- | --- |
| **Input** | | |
| **Beneficiary DataFrame** | **InPatients DataFrame** | **OutPatients DataFrame** |
| BeneID | AdmissionDt | AttendingPhysician |
| ChronicCond\_Alzheimer | AttendingPhysician | BeneID |
| ChronicCond\_Cancer | BeneID | ClaimEndDt |
| ChronicCond\_Depression | ClaimEndDt | ClaimID |
| ChronicCond\_Diabetes | ClaimID | ClaimStartDt |
| ChronicCond\_Heartfailure | ClaimStartDt | ClmAdmitDiagnosisCode |
| ChronicCond\_IschemicHeart | ClmAdmitDiagnosisCode | ClmDiagnosisCode\_1 |
| ChronicCond\_KidneyDisease | ClmDiagnosisCode\_1 | ClmDiagnosisCode\_10 |
| ChronicCond\_ObstrPulmonary | ClmDiagnosisCode\_10 | ClmDiagnosisCode\_2 |
| ChronicCond\_Osteoporasis | ClmDiagnosisCode\_2 | ClmDiagnosisCode\_3 |
| ChronicCond\_rheumatoidarthritis | ClmDiagnosisCode\_3 | ClmDiagnosisCode\_4 |
| ChronicCond\_stroke | ClmDiagnosisCode\_4 | ClmDiagnosisCode\_5 |
| County | ClmDiagnosisCode\_5 | ClmDiagnosisCode\_6 |
| DOB | ClmDiagnosisCode\_6 | ClmDiagnosisCode\_7 |
| DOD | ClmDiagnosisCode\_7 | ClmDiagnosisCode\_8 |
| Gender | ClmDiagnosisCode\_8 | ClmDiagnosisCode\_9 |
| IPAnnualDeductibleAmt | ClmDiagnosisCode\_9 | ClmProcedureCode\_1 |
| IPAnnualReimbursementAmt | ClmProcedureCode\_1 | ClmProcedureCode\_2 |
| NoOfMonths\_PartACov | ClmProcedureCode\_2 | ClmProcedureCode\_3 |
| NoOfMonths\_PartBCov | ClmProcedureCode\_3 | ClmProcedureCode\_4 |
| OPAnnualDeductibleAmt | ClmProcedureCode\_4 | ClmProcedureCode\_5 |
| OPAnnualReimbursementAmt | ClmProcedureCode\_5 | ClmProcedureCode\_6 |
| Race | ClmProcedureCode\_6 | DeductibleAmtPaid |
| RenalDiseaseIndicator | DeductibleAmtPaid | InscClaimAmtReimbursed |
| State | DiagnosisGroupCode | OperatingPhysician |
|  | DischargeDt | OtherPhysician |
|  | InscClaimAmtReimbursed | Provider |
|  | OperatingPhysician |  |
|  | OtherPhysician |  |
|  | Provider |  |

**Pre-Processed Columns**:

|  |
| --- |
| **Output** |
| **Final DataFrame** |
| Provider |
| Age\_median |
| Age\_mode |
| Deceased\_sum |
| Gender\_mode |
| Race\_mode |
| State\_code\_mode |
| County\_code\_mode |
| No\_days\_admitted\_median |
| No\_days\_admitted\_mode |
| No\_days\_ClaimProcessing\_median |
| No\_days\_ClaimProcessing\_mean |
| Attending\_Physician\_mode |
| OperatingPhysician\_mode |
| Other\_Physician\_mode |
| RenalDiseaseIndicator\_mode |
| ChronicCond\_Alzheimer\_median |
| ChronicCond\_Alzheimer\_mode |
| ChronicCond\_Cancer\_median |
| ChronicCond\_Cancer\_mode |
| ChronicCond\_Depression\_median |
| ChronicCond\_Depression\_mode |
| ChronicCond\_Diabetes\_median |
| ChronicCond\_Diabetes\_mode |
| ChronicCond\_Heartfailure\_median |
| ChronicCond\_Heartfailure\_mode |
| ChronicCond\_IschemicHeart\_median |
| ChronicCond\_IschemicHeart\_mode |
| ChronicCond\_KidneyDisease\_median |
| ChronicCond\_KidneyDisease\_mode |
| ChronicCond\_ObstrPulmonary\_median |
| ChronicCond\_ObstrPulmonary\_mode |
| ChronicCond\_Osteoporasis\_median |
| ChronicCond\_Osteoporasis\_mode |
| ChronicCond\_rheumatoidarthritis\_median |
| ChronicCond\_rheumatoidarthritis\_mode |
| ChronicCond\_stroke\_median |
| ChronicCond\_stroke\_mode |
| NoOfMonths\_PartACov\_median |
| NoOfMonths\_PartACov\_mode |
| NoOfMonths\_PartBCov\_median |
| NoOfMonths\_PartBCov\_mode |
| no\_of\_service\_provided |
| InscClaimAmtReimbursed\_sum |
| InscClaimAmtReimbursed\_median |
| OPAnnualDeductibleAmt\_sum |
| OPAnnualDeductibleAmt\_median |
| OPAnnualReimbursementAmt\_sum |
| OPAnnualReimbursementAmt\_median |
| IPAnnualDeductibleAmt\_sum |
| IPAnnualDeductibleAmt\_median |
| IPAnnualReimbursementAmt\_sum |
| IPAnnualReimbursementAmt\_median |
| DeductibleAmtPaid\_sum |
| DeductibleAmtPaid\_median |
| PotentialFraud(Target) |