**HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS**

**Project By: Siddharth Satyakam (50293092,** [**s59@buffalo.edu**](mailto:s59@buffalo.edu)**)**

**Milestone 1:**

**Problem Statement**

The goal of this project is to " predict the potentially fraudulent providers " based on the claims filed by them along with this, we will also discover important variables helpful in detecting the behavior of potentially fraud providers. and we will study fraudulent patterns in the provider's claims to understand the future behavior of providers.

By using the dataset where we are having the sets of Beneficiary Data, Inpatient data and Outpatient Data and finally the tags relating to each beneficiary Id whether its potentially fraud or a valid case, we will focus on defining the model.

**DATA:**

We would be using the data from the source [https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis#Train\_Beneficiarydata-1542865627584.csv](https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis%23Train_Beneficiarydata-1542865627584.csv). This source has the data in the format:

* Train\_Beneficiarydata-1542865627584: The beneficiary data that would be used for training the model.
* Train\_Inpatientdata-1542865627584: The in Patient data those covered under the Medicare Part B cover
* Train\_Outpatientdata-1542865627584: The Patient data covered under the Medicare Part A cover.
* Train-1542865627584: The training tags of potential Frauds and cases.
* Test\_Beneficiarydata-1542969243754: These are the Test Beneficiary Data.
* Test\_Inpatientdata-1542969243754: These are the Test in-patient Data.
* Test\_Outpatientdata-1542969243754: These are the Test outpatient Data.
* Test-1542969243754: Final test tags.

The reason to choose these specific group of data is that this data has a very large group of collection of patients so more data points for better modelling. Moreover, I was not able to find any other collection having such combination of Beneficiary, Inpatient and outpatient data. **Lastly no other dataset in the public domain had the group of potential fraud cases dataset.**

**Method of Approach**

The procedure we will be following to reach to the goal is:

1. Data Cleaning and Discretization for analysis and insights.

2. Data Transformation (Removal of outliers, correlation within variables and definition of new variables etc.).

3. Data Standardization (Bringing all the data on the same scale for better model representation)

4. Data Balancing (Usually the problem that we would be facing in the fraud type of cases are there are very few actual cases for use in the data training as a result there a case of minority +ve cases, so we need to balance the data using various procedures)

5. Modelling of the Data.

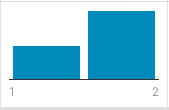
*Data Cleaning and Discretization:*

As the data cleaning was started it was seen that the data source: [https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis#Train\_Beneficiarydata-1542865627584.csv](https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis%23Train_Beneficiarydata-1542865627584.csv)

Had a lot of irregularities that caused very serious delays in all the proceedings. Firstly, the data provided doesn’t mention the country or the state the data belongs to moreover it doesn’t have any descriptions of the columns, what the values represent. As a result, there was no reference for the data cleaning and balancing, which caused very serious delays. Overall the data that we are extracting here are completely based on the assumptions that we had to infer after very long internet browsing.

The assumptions and the changes made for the data are:

1. Country: Now as the country was not mentioned in the dataset it was required to infer the country as the other values used in the dataset were also not clear and required country for research. So as the only available data was that the country whose data is provided is having 54 States and 5 races in majority so using the same the closest any nation was able to come close was Africa (54 Countries and 5 races accumulation.)
2. Date-of-Birth: Now the DOB column was not formatted correctly and was also having a lot of empty values, so we first removed the empty values then formatted it using the pandas in python 3 in total accumulating a total of 138556 data points remaining.
3. Date-of-Death: Here we saw that around 1438 data points in the whole dataset had DOD from the whole of 138556, others were still alive so as a result in case of dead persons the method used for Frauds done by the Providers must be different so we must separate the 2 datasets, for separate analysis. So we separated the dataset into 2 parts work\_dataset\_notDead and work\_dataset\_dead. And also formatted the DOD.
4. Race and Gender: In this case we didn’t need to make any changes as all the values were already there though the values were not defined.
5. RenalDiseaseIndicator: Here luckily all the +ve cases were marked as “Y” so we could easily convert the tags to respective values.
6. State and Country: Now in this case though the values are well defined and discrete but with a lagging index as which value represents which state it was impossible to deduce the respective countries and counties in Africa, hence will cause a lot of problem when we would try to deduce some insights from the given data.
7. Noofmonths\_PartACov: As this states the number of months in the cover in the Part A i.e. inpatient cover, as this is well discrete and values are understandable we will not make any changes.
8. Noofmonths\_PartBCov: As this states the number of months in the cover in the Part B i.e. outpatient cover, as this is well discrete and values are understandable we will not make any changes.
9. ChronicCond\_Alzimers: Now in case of the Alzheimer it was seen that every 1 out of 10 individual faces Alzheimer according to the CDC. So we are considering 1/10 of total data i.e. 138556 which is 13000 so close to it was 42026(label 1) rather than 92530(Label 2) of the data.



So we consider the 1= Yes and 2= NO.

1. ChronicCond\_Heartfailure: According to the CDC as the Heart Failure ios a very common disease so here we followed the same principle as the Alzheimer: 1=YES and 2=NO

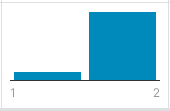


1. ChronicCond\_KidneyDisease: In this case, the share of adults 30 and older who have CKD is projected to rise from 13.2 percent today to 16.7 percent in 2030, according to a 2015 report in the American Journal of Kidney Diseases. So basically from the total population that we are having in our datasets of 138556, all are born on or before 1983, so above 30 easily hence our ailing population should be 15.6 to 17.7 percent of the total population of 138556(i.e. 21,000 to 23,000).



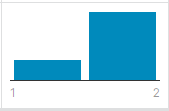
In our case the label 1 is more close to the 15-17% mark of total population than label 2 which is close to 70% of population. SO we would consider the label 1 as Yes and 2 as NO.

1. ChronicCond\_Cancer: As per the National Cancer Institute, it is seen that Approximately 38.4% of men and women will be diagnosed with cancer at some point during their lifetimes (based on 2013–2015 data). So basing on it similarly 38.4% of 138556 is 53,205.



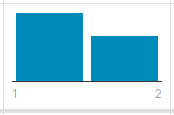
According to the figure the label 1 is more close to 38.4 percent of the data so 1 =YES and 2=NO.

1. ChronicCond\_ObstrPulmonary: We see according to survey conducted by Rupert Jones in European Respiratory Journal 2018 51:1702562 it is seen that around 17% people face the issue if lung obstruction in Africa primarily due to inhaling bio mass burnt gases,

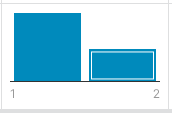


similar trend is seen that Label 1 represents Yes (i.e. 17 percent) and 2 represents NO.

1. ChronicCond\_Depression: We maintain the same trend as previous as this is a common disorder in Africa.
2. ChronicCond\_Diabates: According to WHO the rate of Diabates is 60% of total population so we would similarly consider the 1=YES and 2=NO

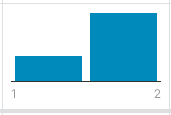


1. ChronicCond\_ischmeicHeartDisease: According to the article "Cardiovascular diseases in Tanzania: The burden of modifiable and intermediate risk factors" by Wilfrida P. Roman1, Haikael David Martin1, Elingarami Sauli2 it is mentioned that 81% of the population die due to cadio vascular diseases(30% by Ischemic Heart Disease in Africa) faced by the African subcontinent.



Hence as mentioned in the dataset label 1 represents the best case senario representing the same so 1=Yes and 2=No

1. ChronicCond\_ Osteoporasis: One of the most least cases of seath in the African Subcontinent ,hence 1= YES and 2= NO follows here.



1. ChroniCond\_ Stroke and ChroniCond\_ rheumatoidarthritis: As all the other columns are following the principle of 1=YES and 2-=NO, we will assume the same here\

All other columns from the Train\_Beneficiarydata-1542865627584 are maintaining the well discrete and distinct values in the various cover columns of monetary covers(IPAnnualReimbursementAmt, IPAnnualDeductibleAmt, OPAnnualReimbursementAmt, OPAnnualDeductibleAmt.) so we are not making any changes to the same.

Further the data cleaning of the rest of the datasets are in progress but unfortunately based on the assumptions that these are data of Africa and all the stats of various columns found follow the same values as found in the net.

And followed by the data transformation, data standardization, data balancing and modelling. And lastly formation of Dashboards.

**Future Changes:**

I would be consulting the TAs and Instructor regarding the assumptions and possibly the chances of changing the datasets for better and more fact based datasets and not assumptions.