Health Care Provider Fraud Detection Insured with Medicare

1. Introduction:

Medicare is an U.S Federal government program which provides hospitalization insurance and voluntary medical insurance for elderly person over 65 years. Healthcare provider Fraud is one of the biggest problems for Medicare. Government report claims, the total Medicare spending is increasing exponentially due to frauds in Medicare claims. In general, healthcare fraud is an organized crime in which providers, physicians and beneficiaries act together to make fraud claims.

Succinct, rigorous analysis of Medicare data has yielded many physicians who indulge in fraud. They adopt ways in which an ambiguous diagnosis code is used to adopt costliest procedures and drugs. Insurance companies are impacted badly due to these bad practices. Due to this reason, insurance companies increased their premium and as a result healthcare is becoming costlier day by day.

Healthcare fraud and abuse take place in many forms. Some of the most common types of frauds by providers are:

a) Billing for services that were not provided.

b) Duplicate submission of a claim for the same service.

c) Misrepresenting the service provided.

d) Charging for a more complex or expensive service than was actually provided.

e) Billing for a covered service when the service actually provided was not covered.

2. Problem Statement:

The goal of this project is to build a Machine Learning (ML) model to predict/classify a Health Care Provider as Potential Fraud or not based on the claims filed by the Provider. To obtain the best model, different ML classifier models will be built and evaluated using ROC AUC score (Receiver Operating Characteristic - Area Under Curve score) and the model with highest score will be chosen for deployment.

3. Data:

For this project, I collected the data available in Kaggle at the following link:

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

There are four datasets: Beneficiary, Inpatient, Outpatient and Train. Here is a short description about each dataset.

A. Beneficiary data

This dataset contains Medicare beneficiaries KYC (Know Your Customer) details; their health conditions, gender, Insurance coverage, race and region (county and state) they belong to.

B. Inpatient data

This dataset provides claims details filed by the Providers for Beneficiaries those treated in hospitals. It lists Physicians, admission and discharge dates, diagnosis and procedure codes, and Claim amount reimbursed.

C. Outpatient data

This dataset provides claims detail filed by the Providers for Beneficiaries who visited hospitals but not admitted there for treatment. It includes data on Physicians, Diagnosis and Procedure codes, claims amount reimbursed.

D: Train data

This dataset lists Providers as Potential Fraud or not.

All the four datasets are large with following rows and columns:
Beneficiary: (138556, 25)
Inpatient:(40474, 30)
Outpatient: (517737, 27)
Train: (5410, 2)

4. Data Wrangling:

The raw data obtained from Kaggle was not clean enough to carry out Exploratory Data Analysis (EDA) or further Machine Learning(ML) buildding, hence Data wrangling was carried out on them. These are the few issues observed and fixed them as follows:

Problem-1: Datetime columns were listed as object.
Solution: These are converted to Timestamp.

Problem-2: The renal disease indicator column has listed two values “Y” and 0.
Solution: The “Y” was assigned to 1 and then the converted to converted to numeric.

Problem-3: In the ‘DOD’ column, most of the values (97%) were missing.
Solution: The entire column was dropped from the dataframe.

Problem-4: In all three Physicians columns (Attending Physician, Operating Physician, Other Physician), there were many null values.
Solution: The null values were replaced with 0 (here it was assumed that missing values represent physicians were absent, not data is missing). Also, the entries with a value were replaced with 1, thus the column became categorical.

Problem-5: For diagnosis/procedure columns, some entries were missing.
Solution: The missing values were replaced with 0 (assuming diagnosis/procedures were really not carried out, because for a treatment all diagnosis/procedures need not have to be carried out). The entries values were replaced with 1.

Problem-6: The Gender column values were 1 and 2.
Solution: The entries with 2 were replaced with 0.

Problem-7: In Inpatient dataset, few of the Deductible amount paid entries were missing, these null values were constituting only 2% of the total.
Solution: The null values were replaced with median in that column because the entries in that column was only one value, hence its mean, median and mode were same.

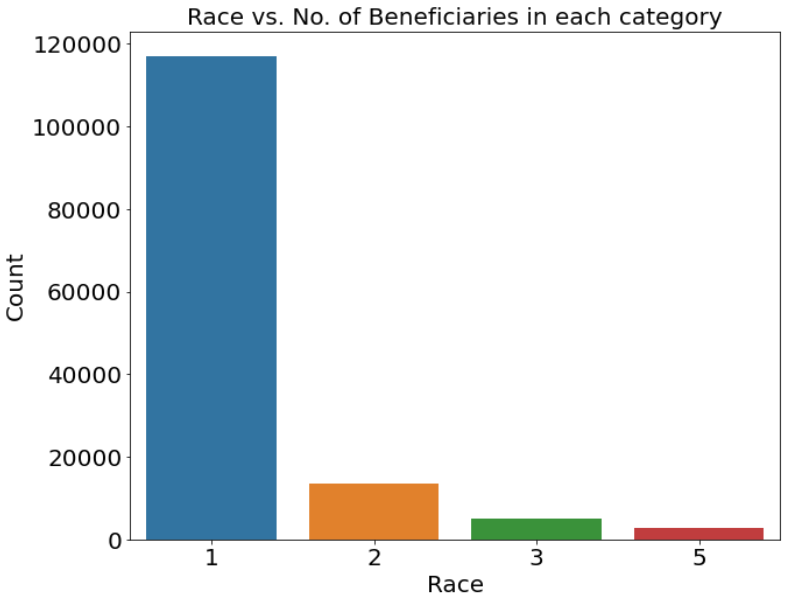
Problem-8: The Potential Fraud column had two values “Yes” and “No”.
Solution: The “Yes” and “No ” were replaced with 1 and 0 respectively.

5. Exploratory Data Analysis (EDA)

The EDA showed

- Mean IPAnnualReimbursementAmt, Mean OPAnnualReimbursementAmt: 3660, 1298
- Mean IPAnnualDeductibleAmt, Mean OPAnnualDeductibleAmt: 340, 377
- There are 9.3% Potential Fraud Providers as compared to 90.7% non-Potential Fraud Providers. Here is bar plot below showing both categories:

- Beneficiaries from state 40 and 48 are missing. Among all states, state-5 has maximum number of Beneficiaries (total beneficiaries: 12052), whereas state-2 has least number (total beneficiaries: 196).
- County-200 has highest number of beneficiaries (3943), whereas few counties has only 1 beneficiary.
- The category-4 in Race column is missing. Among all the categories, category-1 has maximum number of Beneficiaries. The bar plot below shows the different categories with their counts:



After data wrangling and Exploratory Data analysis, Beneficiary dataset was merged with Inpatient and Outpatient datasets. The two newly formed datasets (Beneficiary + Inpatient) and (Beneficiary + Outpatient), now contain four different types of data; Numeric (not categorical), binary categorical, multiple categorical, categorical with unique values in each row.

The data was preprocessed as follows:

1. The numerical columns data was grouped under each provider and their average was taken.
2. The Categorical columns were handled in three different ways, grouping under each provider:
    a. For binary category column, the fraction of one category was taken w.r.t to total was taken.
    b. For multiple categories columns, category with maximum occurrence was taken by using mode, then dummies of those categories were obtained.
    c. For columns with unique values, the total number of entries under each Provider was taken.

Finally all four groupby dataframes were merged together, followed by merging with Train dataset which contains Providers' information regarding Fraud or not.

The jupyter notebook with Data wrangling, Exploratory Data Analysis and Preprocessing has been uploaded at the link:

https://github.com/damayantinaik/Health_care_Fraud_Detection_Medicare/blob/main/Final_report/Data_Wrangling_and_EDA_final_completed.ipynb

6. Feature Engineering

The feature engineering on Preprocessed data was carried out after splitting the dataset into train and test datasets:

1. I replaced the outliers in each numeric columns with 90th percentile of that column assuming that the entries are not mistakenly entered, hence a value at the high end >= 90 percentile was chosen to do the imputation.
2. Then all the numerical columns were standardized.

The Jupyter notebook for feature engineering can be found at:

https://github.com/damayantinaik/Health_care_Fraud_Detection_Medicare/blob/main/Final_report/Prepocessing_and_Feature_Engineering_final_completed.ipynb

7. ML models building, Evaluation and Optimization

Machine Learning models were built to predict/classify the Providers as Potential Fraud or not. Different classification algorithms were built on train dataset and evaluated on test datasets.

There are different types of ML classification models available in scikit-learn; Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, Support Vector

Classifier, Naïve Bayes and few more. I chose to work with the first five classifiers for my problem and considered roc auc score to evaluate the model performance in all cases.

Below, I'll give a brief description on the Model building:

- When ML models were built on the feature engineered processed data, all models' roc auc scores were > 64%, however, among all, for Logistic regression it was highest with score ~ 70%.
- The data was highly imbalanced with

  Potential Fraud cases (No): 90%

  Potential Fraud cases (Yes): 10%

  Hence SMOTE (Synthetic Minority Oversampling Technique) was applied on the data, made the data balanced, and all the Models were rebuilt.

  With SMORT transformed data, the performance of all models improved with once again Logistic Regression with highest roc auc score 76%.

- Henceforth, further model building was carried out on SMOTE transformed data.
- As Logistic regression was with highest performance, GridSearchCV and RandomizedSearchCV were applied to obtain the best Logistic Regression performance and the related hyperparameters. GridSeachCV could calculate the highest performance Logistic Regression with roc auc score of 76%.
- To improve the model running time and performance, Best Features (i.e most important features) were sorted out and applied to Logistic Regression to obtain maximum performance and see how many features give the maximum performance. During this exploration I noted that there are many constant and quasi-constant features(columns), which needed to be dropped from the dataset. I cleaned the dataset again dropping the constant and quasi-constant features.
- All five ML models were again built on cleaned data and again Logistic regression performance was highest with roc auc score 76%.
- Finally PCA (Principal Component Analysis) was applied and Logistic Regression was built on the PCA transformed data. The model with 10 PCA component was with highest roc auc score of 76%.

The table below lists all the ML models along with their roc auc score:

Model	Roc_auc_score	Roc_auc_score (in Logistic Regression, applying PCA) (n_components = 10)
Logistic Regression	0.76	0.76
Decision Tree Classifier	0.71	
Random Forest Classifier	0.74	
Gradient Boosting Classifier	0.75	
Support Vector Classifier	0.75	

Among all the models, the Logistic Regression with PCA and without PCA have same roc auc score. However, the model with PCA was chosen as best for deployment taking into consideration the run time in future if it is applied with large dataset.

The table below summarizes the best Logistic regression model with its various features:

Model	N_components	C	solver	Penalty	Max_iter	roc_auc_score
Logistic Regression	10	0.1	liblinear	L1	100	0.758

The Jupyter notebook on Model building can be found at:

https://github.com/damayantinaik/Health_care_Fraud_Detection_Medicare/blob/main/Final_report/Model_building_final_completed.ipynb

8. Future Recommendation

In this project, ML models were built with < 1500 Providers, whereas there are thousands of providers. To build more efficient model, more amount of data is needed. So, my recommendation here is to collect more data for model building. There is no information regarding FDA approved diagnosis/procedure related to a disease, some resources related to that would be also very useful for better model building.

9. Acknowledgement

I am very much thankful to Python community for providing very rich packages for data analysis and Machine Learning models building. I am very grateful to my mentor Yadunath Gupta with whom I have in-depth discussion regarding the project building, which helped me to complete this project taking into account many aspects of the data analysis and ML model building.