Medicare Fraud Detection Analysis

Background

Health care expenditures constitute a significant portion of governmental budgets. In 2020, National health expenditure grew to $4.1 trillion and accounted for 19.7% of Gross Domestic Product (GDP)[1]. However, health care fraud costs the nation $230 billion annually, contributing 7 percent of the nation's health care spending[2]. Hence, our team's fraud analysts built prediction models to detect claims that have a high probability of being fraudulent. This would benefit the taxpayers of this country and potentially help the government save billions of dollars of losses on fraud and abuse annually in the Medicaid and Medicare system. Our goals are as follows.

- **Predict the potentially fraudulent providers based on claim information they performed and prescription drugs they administered.**
- **Figure out features with highest importance in predicting fraudulent providers.**

Data

The data is from Centers for Medicare and Medicaid Services, which includes part B, part D, DMEPOS and the label data set LEIE described as below

+36M
Raw Data Points

The raw data are from the center of Medicare and Medicaid, which includes part B, part D, DMEPOS of year 2019.

Part B: Services provided to beneficiaries by hospitals
Part D: Prescription drugs prescribe by healthcare provider
DMEPOS: Use, payments and submitted charge & place of service

+250k
Providers

Providers are uniquely identified by their the National Provider Identifier (NPI) number.

Providers covers 7 countries, 8717 cities
72 specialty types
+0.8 billion total Submitted Charge

+76k
Data Points on Fraud Exclusions

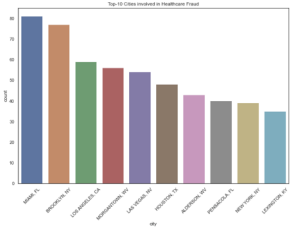
The fraud information are from the Office of Inspector General LEIE dataset

31 categories of violation
Pursuant to section 1128 of the Social Security Act (Act)

EDA

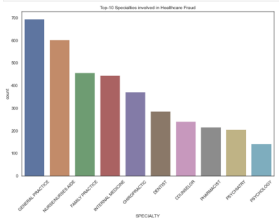
Geographic Features

- Top three states involved in healthcare frauds are **CA**, **NY**, and **FL**
- For cities, **Miami**, **FL** and **Brooklyn**, **NY** particularly stand out in healthcare frauds.

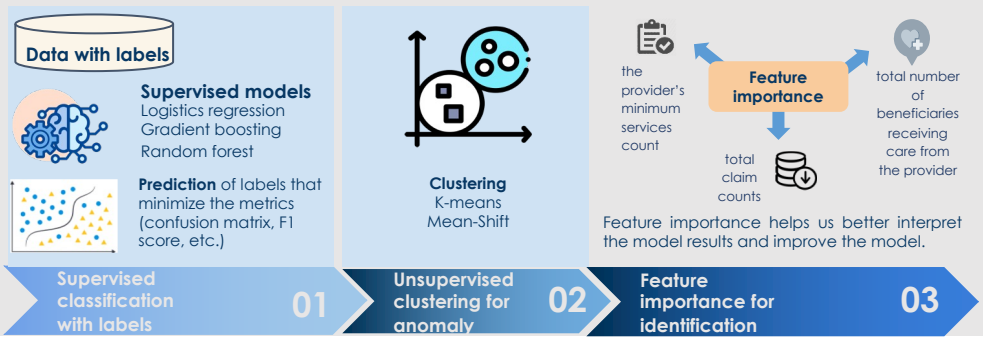


Clinical Features

- In terms of general, **physician (MD, DO)** and **IND-LIC HC SERV PRO** appear most in frauds.
- For speciality, **general practice** and **nurse** are two specialities involved most frequently in frauds.



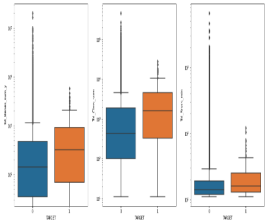
Models



Results

The problem is binary classification with highly unbalanced data, so we use under sampling to deal with skewness and logistic regression, gradient boosting as well as random forest to build our models. The random forest performs the best among these 3 models, with an F1-score at 0.75. In more details, the false negative rate (the provider who is fraud but is predicted as non-fraud) and the false positive rate (the provider who is not fraud but is predicted as fraud) are 0.188 and 0.34 respectively.

The top 3 features are **the provider's minimum services count**, **total number of beneficiaries receiving care from the provider** and **total claim counts** which indicate significant difference between fraud and non-fraud groups.



Conclusion

Best model: Random forest performs the best in identifying rare cases in our problem after undersampling the imbalanced data, with the assumption is that we want to minimize both the false positive rate and false negative rate.

Application: The analysis provides us with a pipeline for fraud detection and can be used as a template in other scenarios (e.g. credit card transactions)
The model can be customized depending on business requirement, such as bootstrapping the model so that it performs better in terms of false negative rate.

Impact: Our team's fraud analysis help the security of Medicare and Medicaid systems by predicting which provider has a high probability of being fraudulent. This would benefit the taxpayers of this country and potentially help the government save billions of dollars of losses on fraud and abuse annually in the Medicaid and Medicare system.