Healthcare Provider Fraud Detection

Capstone Project 1 Final Report

Zheng Zhang

May 28, 2020

Introduction

Healthcare provider fraud is one of the biggest problems facing Medicare, and it contributes to the total Medicare spending growth. Healthcare fraud is an organized crime which involves peers of providers, physicians, beneficiaries acting together to make fraudulent claims.

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 the most vulnerable institutions impacted. To recoup those losses, insurance companies increase the insurance premiums and as a result healthcare becomes more costly for everyone.

Healthcare fraud and abuse take many forms. Some of the most common types of fraud 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.

Problem Statement

We want to identify the potentially fraudulent providers based on the claims that they filed. In addition, we will also discover important variables helpful in detecting the behaviour of those providers. Further, we will study fraudulent patterns in the provider's claims to understand the future behaviour of providers.

Dataset

A. Dataset Description

The datasets include provider data, beneficiary data, inpatient claims data and outpatient claims data. There are separate training dataset and test dataset.It can be found here : https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis

The training provider data include 5410 providers with a binary flag for potential fraud. The training beneficiary data contains 138,556 subjects with 25 variables. The inpatient training data include 31,289 subjects and 40,474 claims with 30 variables. The outpatient training data include 133,980 subjects and 517,737 claims with 27 variables. The test provider data include 1353 providers. The test beneficiary data contains 63,968 subjects. The inpatient test data include

8,351 subjects and 9,551 claims. The outpatient test data include 59,608 subjects and 125,841 claims.

The beneficiary data variables include ID, date of birth, gender, race, state, county and chronic disease status, inpatient annual reimbursement and deductible amount, and outpatient annual reimbursement and deductible amount. The inpatient dataset variables include beneficiary ID, claim id, claim start/end date, provider, reimbursement amount, admission date, admit diagnosis code, deductible amount, discharge date, diagnosis code and procedure code. The outpatient dataset variables include beneficiary ID, claim id, claim start/end date, provider, reimbursement amount, admit diagnosis code, deductible amount, diagnosis code, and procedure code.

B. Missing Data Handling

1. Missing entries in the variable (date of death) in the beneficiary data are expected.
2. Missing entries observed on physician ID and diagnosis/procedure code variables were left as they were.
3. Missing data found on the variable "deductible amount paid" (inpatient claim dataset, 899 claims) were imputed with value 0 for those claims.

C. Outliers Detection:

There are only ten variables that are continuous. Of those, 2 are related to the number of months that medicare coverage is in effect. The range of those two variables are 0-12, which are expected. The other eight continuous variables are related to deductible and insurance reimbursement amounts. Positive outliers are expected, however, there are values that are negative, which are not expected, in those cases, the negative values were changed to 0.

D. Construction of the Provider Claims Dataset

Although the claim data is provided either at the claim level or at the beneficiary level, in order for us to build the model to predict which provider is fraudulent, we need a claim dataset built on the provider level. We hence constructed a provider dataset that includes the number of either inpatient or outpatient claims each provider submitted and the total claims amounts per provider.

Exploratory Data Analysis of the Training Datasets

A. Frequency counts:

Providers: There are 5410 providers, of which 506 (9.35%) are labelled as potential fraud.

Beneficiaries: There are 138,556 beneficiaries, of which 59,450 (42.9%) are men. Race compositions are 117,057 (84.5%) white, 13,538 (9.8%) black, 5,059 (3.7%) asian and 2902 (2.1%) others. They are from 52 states and 314 counties. There are 11 pre-existing chronic conditions listed, including Alzheimer (33%), heart failure(49%), kidney disease(31%), cancer(12%), obstructive pulmonary disease(24%), depression(36%), diabetes(40%), ischemic heart disease(32%), osteoporosis (27%), rheumatoid arthritis(26%), and stroke(8%).

Inpatient claim data: 31,289 (22.6%) beneficiaries generated 40,474 inpatient claims, most have one claim, but one person has 8 visits. There are 2092 providers, the median number of claims each provider submitted is 8, but some had submitted as many as 516 claims.

Outpatient claim data: 133,980 (96.7%) beneficiaries generated 517,737 outpatient claims, the median number of claims per beneficiary is 3, but some had as many as 29 claims. There are 5012 providers, the median number of claims each provider submitted is 31, but the most claims a single provider submitted is 8240.

B. Medicare coverage and annual reimbursements

137,389/136,902 beneficiaries had full 12 month coverage for part A/B, 1000/675 had no coverage for part A/B, and the rest had less than full year coverage.
36,030 had inpatient reimbursement, the max reimbursement is $161,470.
134,339 had outpatient reimbursement, the max reimbursement is $102,960, median value is $570.

C. Distribution of inpatient and outpatient claim amounts

As expected, the distribution of inpatient claims is heavily skewed on both ends, with 77.4% of all beneficiaries have no claims, and the mean claim amount is $3660, but the maximum claim amount is $161,470.
The distribution of the outpatient claims is also skewed, with mean claims $1298, median claim $570 and the maximum claim amount is $102,960.

D. Relationship between the claim amounts and patient demographic information

1. Gender: No significant differences observed by boxplots.
2. Race: only two of the four racial groups have inpatient claims, and the outpatient claims are similar across all racial groups.
3. Co-mobidity conditions: People with comobidity conditions incur almost all inpatient claims, they also tend to have higher outpatient claim amounts.

E. Relationship between the claim amounts and whether the provider is flagged as fraudulent

The potentially fraudulent providers had substantially more claims submitted, with the median claims per provider as 24 inpatient and 99 outpatient compared with 0 for the non-fraudulent providers. In addition, the fraudulent providers had substantially more annual claim amounts.

Statistical Analysis of the Training Datasets

In order to identify the factors that could distinguish the potentially fraudulent providers from the others, we looked at both provider level variables (such as the indicator variable that flags the providers as potentially fraudulent) as well as patient level variables such as gender and baseline chronic conditions. After the extensive exploratory data analysis, we have identified several factors that may impact the claim amount, specifically:

I. Patient level factors

We used two-sample t-test to compare inpatient or outpatient mean claim amounts between men and women. There is no statistical significant difference for mean inpatient claim amount ($3641 vs. $3675, P=0.51). However, there is a statistical significant difference for mean outpatient claim amount ($1278 vs. $1313, P=0.009), albeit the difference is small ($35).

We also used two-sample t-test to compare inpatient or outpatient mean claim amounts between patients with the condition vs. those without. Our analysis showed that across the board, the patients with the condition have significantly higher claim amounts compared with the patients without on average, as shown in the following tables.

Inpatient claims:

Condition	Yes	No	P-value
Alzheimer	5371	2809	<0.0001
Heart Failure	5422	1943	<0.0001
Kidney Disease	7501	1916	<0.0001
Cancer	6068	3332	<0.0001
Obstructive Pulmonary	7362	2510	<0.0001

Depression	5022	2909	<0.0001
Diabetes	4871	1831	<0.0001
Ischemic Heart	4698	1497	<0.0001
Osteoporosis	4608	3301	<0.0001
Rheumatoid Arthritis	5102	3162	<0.0001
Stroke	8111	3278	<0.0001

Outpatient claims:

Condition	**Yes**	**No**	**P-value**
Alzheimer	1623	1136	<0.0001
Heart Failure	1671	935	<0.0001
Kidney Disease	2062	951	<0.0001
Cancer	1788	1231	<0.0001
Obstructive Pulmonary	1811	1139	<0.0001
Depression	1604	1130	<0.0001
Diabetes	1607	831	<0.0001
Ischemic Heart	1521	834	<0.0001
Osteoporosis	1495	1224	<0.0001
Rheumatoid Arthritis	1548	1212	<0.0001
Stroke	1925	1244	<0.0001

II. Provider level analysis: Potentially Fraud vs. Not

Question: Is there difference between the mean number of claims and the total claim amount between the providers who are flagged as “potentially fraud” vs. those that are not?

Approach: We used two-sample t-test to compare inpatient or outpatient mean claim number and the total amount between providers who are flagged as “potentially fraud” vs. those who are not.

Result: Across the board, the providers that are potentially fraudulent have filed significantly more claims and significant higher total claim amounts compared to those who are not flagged on average, as shown in the following tables.

	Fraud	Not-Fraud	P-value
Number of claims - inpatient	46.2	3.5	<0.0001
Number of claims - outpatient	374	67	<0.0001
Claim amount - inpatient($)	476,855	34,056	<0.0001
Claim amount - outpatient($)	107,495	19,138	<0.0001

In-depth Analysis of the Training Datasets - Machine Learning Technique to Predict Provider Fraud

In order to identify the factors that could distinguish the potentially fraudulent providers from the others, we looked at both provider level variables (such as the indicator variable that flags the providers as potentially fraudulent) as well as patient level variables such as gender and baseline chronic conditions. After the extensive exploratory and statistical data analysis, we have identified several factors that may differentiate the fraudulent providers: number of inpatient claims submitted, number of outpatient claims submitted, total amount of inpatient claims, and total amount of outpatient claims. We assembled a database with those four features and employed several machine learning algorithms to build the prediction model. The model building process using different machine learning algorithms share some of the following common steps:

1. Split the dataset into a 70% training set and 30% test set, stratified by the fraud status.
2. Some algorithms have hyperparameters, where GridSearch has been used to identify the best choice for such hyperparameters (hyperparameter tuning) by using 5-fold cross-validation.
3. The percentage of potential fraud providers is less than 10%, which means our data is imbalanced. Machine learning algorithms by default aim to minimize the overall mis-classification rate, which would mean a dis-proportional emphasis on the classification accuracy of the majority class, which is the class of honest providers. However, our goal is to identify as many potentially fraudulent providers as possible, so it is important to maximize sensitivity while maintaining a reasonable specificity. To this end we chose to oversample the minority class from the training set so the percentage of fraud providers is at 50% of the resampled data.

Results from the machine learning algorithms:

	Hyperparameter Tuning	Correction for Imbalanced Data	Sensitivity	Specificity	Misclassification
K Nearest Neighbor	K=8	No	47%	98%	112(6.9%)
Logistic Regression		No	46%	99%	104(6.4%)
Random Forest	N=50	No	47%	97%	125(7.7%)
Logistic Regression		Yes	81%	90%	162(10.0%)
Random Forest	N=67	Yes	88%	84%	255(15.7%)
Balanced Bagging		Yes	77%	88%	212(13.1%)
Rusboosting	N=73	Yes	86%	83%	273(16.8%)
Adaboosting		Yes	86%	85%	241(14.8%)

As expected, without correcting for imbalance, we achieve a low overall misclassification rate (6-8%), but the sensitivities are all below 50%, which are not desirable.

We compared the performance of five different algorithms, all of which have sampling mechanisms that use balanced bootstrap samples: Logistic regression, random forest, bagging and boosting (Rusboosting and Adaboosting). Of those five, random forest achieves highest sensitivity at 88%, but with a lower specificity at 84%. The overall misclassification rate is also high at 15.7%. The boosting methods offer similar performance as random forest. Logistic regression has the highest specificity at 90% and lowest misclassification rate of 10%, but the sensitivity is only 81%. The bagging method is similar to logistic regression with better specificity than sensitivity, however, its performance is not as good as logistic regression.

Overall, we will recommend logistic regression with over-sampling or balanced random forest for our datasets. Those two methods have different strengths and could be applied to different evaluation criteria. If the overall goal is to maximize the probability in detecting the fraud, and the cost of false positives is not a huge concern, then balanced random forest is a good choice. However, if we would like to keep the overall misclassification rate low, then logistic regression with balanced sampling is a good option.

Discussion

Although not reported above, we have tried using support vector machine, smote sampling method, and beneficiary features (both raw data as well as components from principal components analysis), however, the performances are poor. In conclusion, a naive oversampling of the minority class to achieve balanced sampling in conjunction with logistic regression works well. The other method that stands out is balanced random forest.