HealthCare Provider Potential Fraud Detection

MACHINE LEARNING AND DATA ANALYSIS: CSCI 6671-01

Ashesh Byanju

University of New Haven

Date: December 14, 2020

Abstract

Provider Fraud is one of the biggest problems facing Healthcare. According to the government, the total Medicare spending increased exponentially due to frauds in Medicare claims. The largest healthcare provider fraud takedown in US history was announced just recently, resulting in charges against 400 defendants in 41 federal districts for schemes totaling $1.3 billion, according to HHS and the Office of the Inspector General (OIG). Healthcare fraud is an organized crime which involves peers of providers, physicians, beneficiaries acting together to make fraud claims.

Table of Contents:

Abstract 2

Introduction 4

Methodology 5

Data Preprocessing and Feature Engineering 5

Data Modeling 6

Results 7

Conclusion 8

References 9

Introduction

Provider Fraud is the biggest problem faced in Healthcare in the US specially in Medicare programs. Providers involved in fraud relied on creative schemes such as forgery, bribes, fake patients, and falsified billings to financially benefit from administrative vulnerabilities in public programs such as Medicare and Medicaid. Due to this reason, insurance companies increased their insurance premiums and as a result healthcare is becoming costly day by day.

The objective of this project is to use machine learning models to predict potential fraudulent providers based on claims filed, and understand the future behaviour of providers by studying fraudulent patterns in the provider's claims beforehand and prevent increasing healthcare costs and insurance premiums.

Data Source

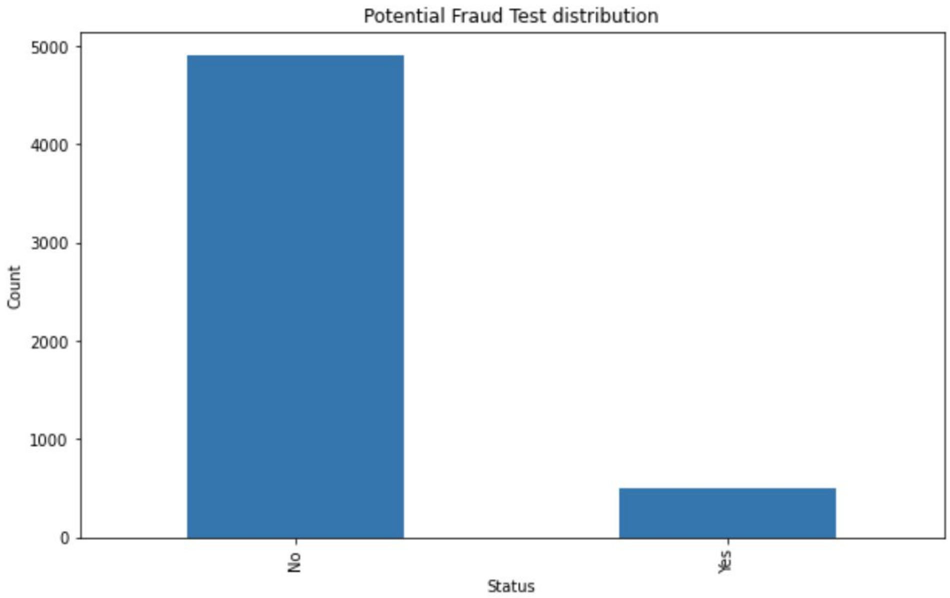
After exploring different sources, I found a dataset in kaggle that is most relevant to what I initially intended to do. We have three datasets consisting of Inpatient claims, Outpatient claims and Beneficiary details of each provider. Inpatient data provides insights about the claims filed for those patients who are admitted in the hospitals. It also provides additional details like their admission and discharge dates and admit d diagnosis code. Outpatient dataset provides details about the claims filed for those patients who visit hospitals and are not admitted in it. Beneficiary data contains details of patient personal information like health conditions, region they belong to etc.

Methodology

Data Preprocessing and Feature Engineering

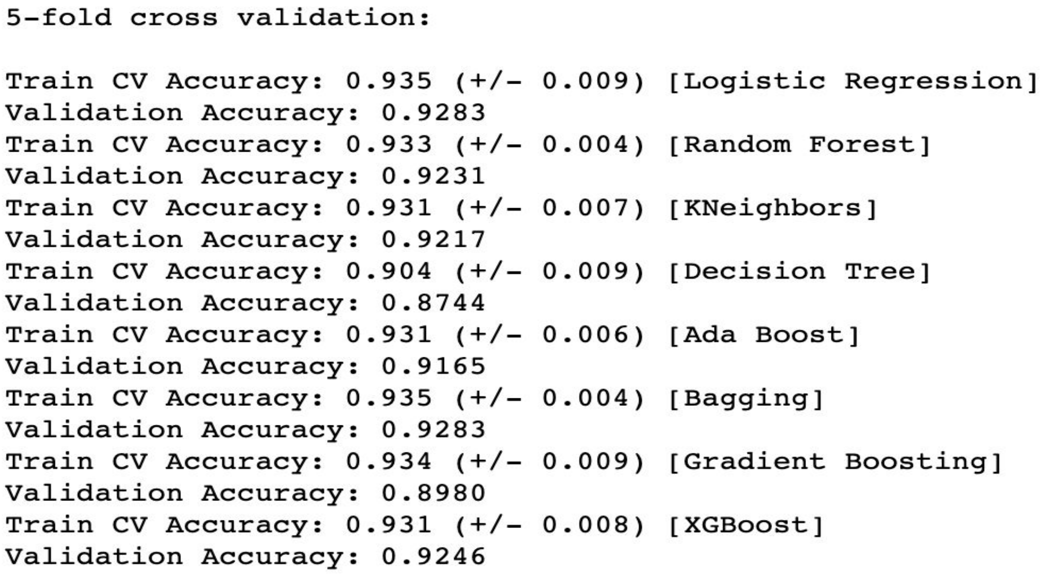
Firstly, I went through the data where I found that many of the features needed to be pre processed as they didn't have uniform values. Although features had binary values most of them have data like 0 or yes which is irrelevant for the machine to understand, so to make it uniform I change them to 0 and 1. Similar was the case for others where values were 1 or 2. For the next step, I looked into null values in the data and found out that most of the columns could be merged. So, as the next step I created a new column and merged data from different features and deleted all others features. For example, we had 2 features: date of birth and date of death, so I created a new feature age from these features and deleted both. After that I merged all three datasets. Since our target values were based on providers data so it was necessary for me to group the data as per providers using mean, sum or unique whichever was suitable. Finally I checked for null values in the data and replaced it with zero.

Here is the snapshot of the bar graph grouped by providers based on potential fraud on the train data. From the bar graph we found out that most of the providers were not fraud. Since data consist of a lesser number of fraud providers this could affect our model.



Data Modeling

Dataset had seperate train and test dataset, with test dataset without outcomes where we need to do our prediction, so firstly I split train data into train and validation so we can view how well our data performed in the validation dataset. I used Logistic Regression and Random Forest as the main model to compare the results and also tried to find how well other models perform in this dataset. In order to get the optimal hyperparameter of the model I use grid search and cross validation, to get the best performing parameter within the model for better accuracy and performance. Below is the figure showing performance of each models on the train dataset.



```
5-fold cross validation:

Train CV Accuracy: 0.935 (+/- 0.009) [Logistic Regression]
Validation Accuracy: 0.9283
Train CV Accuracy: 0.933 (+/- 0.004) [Random Forest]
Validation Accuracy: 0.9231
Train CV Accuracy: 0.931 (+/- 0.007) [KNeighbors]
Validation Accuracy: 0.9217
Train CV Accuracy: 0.904 (+/- 0.009) [Decision Tree]
Validation Accuracy: 0.8744
Train CV Accuracy: 0.931 (+/- 0.006) [Ada Boost]
Validation Accuracy: 0.9165
Train CV Accuracy: 0.935 (+/- 0.004) [Bagging]
Validation Accuracy: 0.9283
Train CV Accuracy: 0.934 (+/- 0.009) [Gradient Boosting]
Validation Accuracy: 0.8980
Train CV Accuracy: 0.931 (+/- 0.008) [XGBoost]
Validation Accuracy: 0.9246
```

Results

Among all Logistic regression performed very well with 92.83% accuracy. Performance of random forest was also not that bad with 92.38% accuracy, slightly lower than that of logistic regression but not that too much. There were altogether 1353 providers in test data and with logistic regress we predict that 1282 were not fraud and 71 were predicted as fraud providers. Similarly as per random forest 1291 were not fraud and 62 were fraud providers. Below is the side by side fraud prediction bar graph of logistic regression vs random forest.

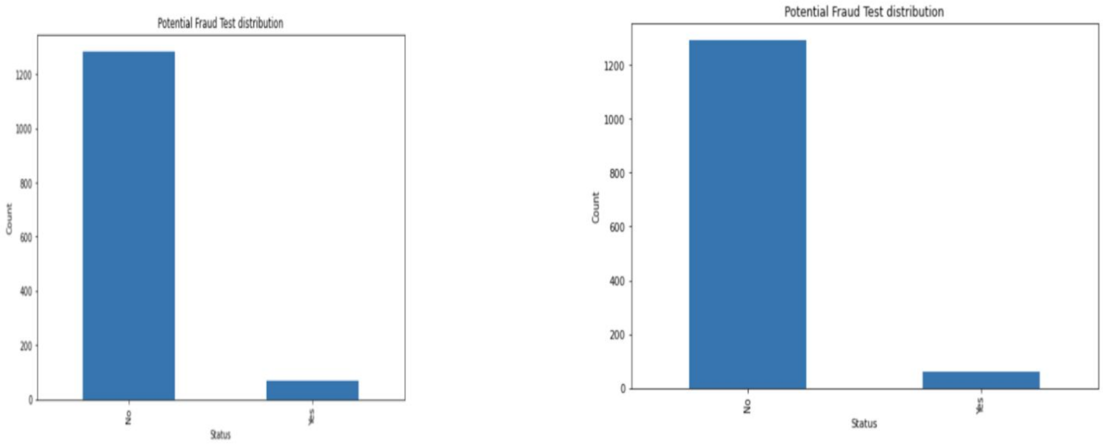


Figure: Logistic regression vs Random forest potential fraud distribution on test data

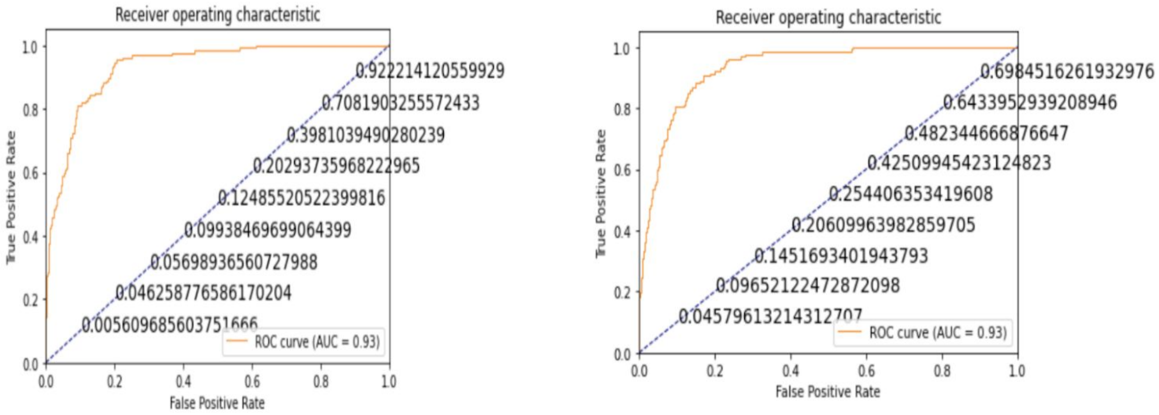


Figure: Logistic regression vs Random forest ROC curve

Conclusion

This project is the sample concept/example of how machine learning models can be used in the healthcare sector especially in the insurance claims to minimize the potential providers fraud by detecting it beforehand as a motive to save million dollars and prevent increasing insurance premium costs. This project is applicable in real world to data analytics organizations, health insurance companies and also the US healthcare department. For further improvement, we can use this process or model on insurance claim dataset to detect potential fraudulent claims in healthcare.

References

[1] Data set. https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis