Semi-Supervised Medical Fraud Detection Using Autoencoders

Rayhaan Rasheed[1]
[1]Department of Data Science, The George Washington University,
Washington, D.C., USA

Abstract

Every day, patients seek medical counsel and assistance from physicians, but access to affordable healthcare is still a pressing issue. For those who qualify, the United States government provides financial support in the form of its Medicare and Medicaid programs. These programs allow citizens to receive quality care, and physicians are paid from the government's pocket. Since physicians are not billing the patient directly, some have resorted to committing fraud to get more money. Large abnormalities can easily be detected, but many fraudulent physicians are difficult to detect since they could identically resemble normal physicians. Leveraging publically available data, anomaly detection systems can be developed to point out fraudulent physicians while working under a highly imbalanced class setting. With only 0.017% as the fraudulent class size, a deep anomaly detection system is developed to understand the minute differences between the two classes and gain a good understanding of the overall feature space. The model of choice is an Autoencoder since it is commonly used for anomaly detection problems due to its ability to understand the fundamental components of the input data and be able to reconstruct the data using these components. Three versions that differed in the number of hidden layers were compared to find the best network architecture for detecting fraudulent physicians. The shallow network, which only contained one hidden layer between the input and output layers, was the best model because it detected the most fraudulent cases maintaining the least number of false negatives.

Keywords: Autoencoders, Semi-supervised Learning, Medicare, CMS, LEIE, Fraud Detection

1. Introduction

Millions of Americans rely on federally subsidized healthcare to afford medical procedures, medication, and assistive devices. According to the Center for Medicare and Medicaid Services (CMS), the United States government spent more than a trillion dollars on its healthcare system in 2018, and they predict more each year after. [1] Instead of billing the patient directly, physicians and medical institutions get paid directly from insurance companies and government funds. With a tremendous amount of money in the system, it is natural for some to take unlawful actions. [2] Fraudulent schemes in the healthcare system range from billing for services that were not provided to organized crime infiltrating the Medicare program. The Federal Bureau of Investigation estimates more than ten percent of total health spending consists of fraudulent billing. [3]

1.1 Data

To address the issue of detecting fraudulent entities, CMS released large, multidimensional datasets for the different parts of their Medicare/Medicaid program. For this experiment, only the Provider Utilization and Payment Data (Part B) is used to understand and detect healthcare fraud. Part B helps cover doctors' services and outpatient care. This is where a high percentage of fraud exists since overbilling is one of the most common forms of fraudulent activities. [3] Each observation in the dataset corresponds to a procedure performed by a specific physician, where each physician is given a unique NPI code.

To know which physicians and organizations are fraudulent, the Department of Health and Human Services' Office of the Inspector General (OIG) established the List of Excluded Individuals and Entities (LEIE), which is maintained monthly. The OIG is required by law to exclude anyone from federally

funded healthcare programs if they are convicted of Medicare/Medicaid fraud. [4] The LEIE is a dataset where each person has a set of features including their unique NPI code.

1.2 Autoencoders

Hinton and Salakhutdinov first introduced the Autoencoder in 2006 as a way to describe a nonlinear generalization of PCA. The method uses a multilayer network to reduce high-dimension data to a lower level then attempt to reconstruct the input data from the low-level code (See Figure 1). [5] There are four parts to a typical Autoencoder: encoder, bottleneck, decoder, and reconstruction loss. The encoder reduces the high-dimensional input to the low-level code, which is called the bottleneck. From the bottleneck layer, the decoder attempts to reconstruct the data from the learned encoding. Finally, the reconstruction loss measures how well the decoder performed in creating an output similar to the input.

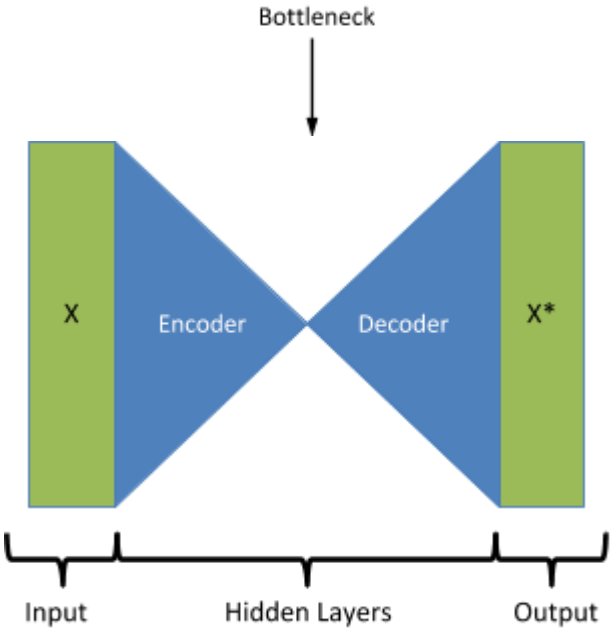

Figure 1: Autoencoder Architecture

Autoencoders are very common in unsupervised learning since they do not backpropagate based on the target labels but on the reconstruction loss. [6] Today, Autoencoders are in a variety of situations ranging from image reconstruction to denoising. The most common use case is anomaly detection. The model can understand the fundamental components of a non-fraudulent physician, and detect fraudulent ones if the reconstruction loss is high.

2. Related Works

There have been other studies that focused on identifying healthcare fraud. Some developed unique methods for fraud detection or created novel ways to preprocess the data. Using the Part B dataset, Bauder et al. compared a variety of anomaly detection models including isolation forest (IF), local outlier factor (LOF), k-nearest neighbor (KNN), and Autoencoders (AE), with LOF coming out on top with an AUC of 0.65. [7] Their version of the Autoencoder always included dropout nodes, which may have helped with generalization but gave the encoder a disadvantage in fully understanding the feature space. Likewise, Herland et al. created a useful, in-depth method for preprocessing the Part B data. By doing so, they were able to map the LEIE data and triple the feature space. In terms of model selection, they decided to make use of Logistic Regression, Gradient Boosted Trees, and Random Forest. [8] Unlike Bauder et al., they were able to achieve an AUC of 0.805 using the Logistic Regression classifier. The reason for this high AUC was the use of K-Fold Cross Validation. This resampling method lets the model train and validate on a set number of subgroups within the dataset. In the end, this results in a less biased estimate of the model. Similar to the previous groups, Johnson and Khoshgoftar tested different sampling methods on a Multiple Linear Perceptron (MLP) to find a way around the class imbalance in the Part B data. They found that random under-sampling outperforms the baseline MLP and tremendously decreases training time. [9] Finally, even if Part B data was not used, Lazaga and Santhana developed a way to detect medical treatment fraud using Restricted Boltzmann Machines. These shallow, two-layer networks are generative, stochastic neural networks that preceded similar methods like Generative Adversarial Networks (GAN) and Variational Autoencoders (VAE). Using the RBM method, Lazaga and Santhana were able to reach an AUC of 0.95 and a recall of 50%. [10]

3. Method

3.1 Data Preprocessing

The preprocessing method is similar to the one developed by Herland et al. The first step is to remove all rows that have a null value or a value of "000000". The next step is to expand the feature space, which consists of adding aggregate features and label encoding. Since physicians can have multiple observations in the Part B data, a way to connect them is by adding shared aggregate features based on all the continuous variables. This includes adding metrics such

as mean, standard deviation, minimum, and maximum for each unique NPI code. To further expand the feature space, categorical variables are separated and binarized using One Hot Encoding from the scikit-learn library in Python. As a result, the find dataset contains 8,910,479 observations and 98 features. Afterward, using the NPI codes in the LEIE dataset, the fraudulent physicians in the Part B dataset can be flagged – creating two classes. Only 0.017% of the resultant Part B dataset is fraudulent, which is a massive class imbalance. Finally, after the data has been cleaned and processed, it is split into the training and testing set. Since the end objective is to detect an extremely small group of fraudulent physicians, the Autoencoder needs to learn and understand the complexities and fundamental components that represent normal, non-fraudulent physicians. Therefore, the training set is comprised of only normal physicians. The testing set contains all of the fraudulent physicians and the remaining normal physicians.

3.2 Network Architecture

There are three Autoencoders that are tested and compared to one another, and aside from the depth of the network (i.e. number of hidden layers), the three models all have a training batch of 1,000 observations, equal input and output shapes, a hyperbolic tangent activation function in between each layer, and the same hyperparameters. By making the numbers of hidden layers vary, we can test for the optimal depth that fully understands the Part B data. The first model is a shallow Autoencoder (98-12-98) with one hidden layer in between the input and output layers. The second model (98-35-12-35-98) has three hidden layers – the bottleneck, one for the encoder, and one for the decoder. Lastly, the third model (98-48-24-12-24-48-98) contains five hidden layers.

3.3 Model Training & Reconstruction Error

For the Autoencoders to understand and represent that normal physician data to flag fraudulent ones, they need to be trained with only non-fraudulent data. This is not a supervised learning approach since the models are not backpropagating based on the target labels, but the train and test sets are created with the target labels in mind. Instead, this is a semi-supervised learning method. As a result, the models learn minute discriminative boundaries between the two classes, and test examples that do not fall within the normal bounds are flagged as fraudulent. [11]

Even though this is a semi-supervised approach, the model is trained to minimize the difference between the original input and reconstructed output. This is called the reconstruction error (RE). All three models calculate the RE using the Mean Squared Error (MSE). Ideally, after each epoch, the models decrease the MSE meaning they have learned and successfully reconstructed the training data.

3.5 Model Evaluation

After obtaining the RE for each sample in the test set, a domain-specific threshold (DST) is needed to assign class labels. A DST can be created using subject matter experts (e.g. Other physicians, OIG, CMS), and they tend to be more valuable than a simple binary classification. The distance a fraudulent point is from the mean of the normal points can observe the severity of fraud. [12] It is important to remember that because the objective of the models is to identify malicious intent, it is better to safe than sorry. Therefore, a higher recall is more important than a higher precision. A high recall means that a model was able to capture most of the fraudulent activity even if it means flagging some non-fraudulent physicians. With this notion in mind, a recall of at least 50% is a great benchmark to strive for. Multiple REs are evaluated until the model reaches 50% recall. At that point, an adequate threshold is found. Any value below the threshold is considered normal, and everything above is flagged as "fraudulent". These predicted values are used to construct a confusion matrix and plot the receiver-operating characteristic (ROC) curve. The area under the ROC curve (AUC) is a good indication for overall model performance

4. Results

In this section, the fraud detection results from the three Autoencoders are discussed. Each model is trained using a learning rate of 0.01 and an Adam optimizer. The train and test run across 10 epochs, for each model are shown in Figures 2a, 3a, and 4a, respectively. The Shallow Autoencoder (Model 1) had a fairly high training RE of 0.85, and the other two models ended with a training RE slightly above 0.1. After running the models on each observation in the test set, the average RE for fraudulent and non-fraudulent physicians is calculated to know the general range of thresholds. Next, thresholds were found where the recall score is 50%. Using these thresholds, the ROC and AUC for each model can be produced. The ROC curve can be seen in Figures 2b, 3b, and 4b, respectively. Table 1 shows the performance and other metrics from each model. Model 1 detected the most fraudulent cases, followed

by Model 2 and Model 3. It also had the lowest number of false negatives, which means it lets the least number of criminals slip away.

Table 1: Model Evaluation

	Model 1	Model 2	Model 3
AUC	0.5084	0.4847	0.5039
TP	772	706	430
TN	1,351,308	1,350,535	1,938,167
FP	1,311,836	1,322,609	734,977
FN	749	815	1,091
Average Fraud RE	0.7617	0.0444	0.1012
Average Normal RE	0.8433	0.1320	0.04025
Threshold at 50% Recall	0.23	0.03	0.03

Not only did Model 3 have the lowest number of true positives, but it also had the highest number of false negatives. This goes to show that a very low training loss does not necessarily indicate a high accuracy in anomaly detection situations. Additionally, these results show that as the complexity of the Autoencoder increases, the number of detected fraudulent physicians decreases. Based on these findings, Model 1 has the best overall performance, followed by Model 3 that contains 2 hidden layers in the encoder and decoder. However, all three Autoencoders are very close to 0.50, which is almost a random guess.

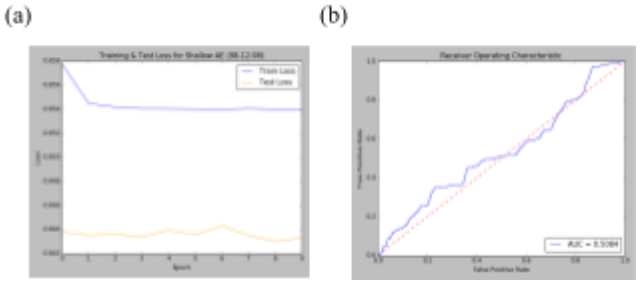

Figure 2: (a) Model 1 train and test loss per epoch (b) ROC curve

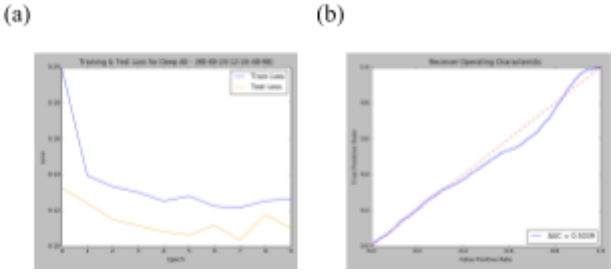

Figure 3: (a) Model 2 train and test loss per epoch (b) ROC curve

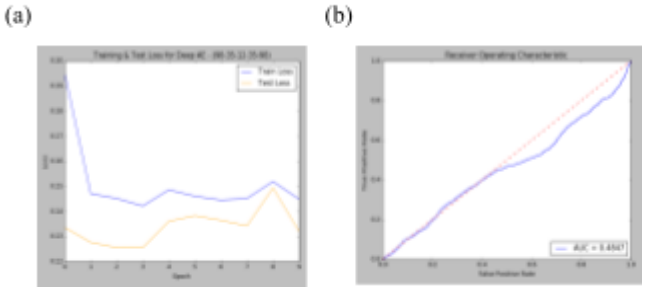

Figure 4: (a) Model 3 train and test loss per epoch (b) ROC curve

5. Conclusion

With all the money the U.S. government invests into its healthcare programs, there are physicians who will find ways to take advantage of the medical system. Patients confide in medical practitioners with their ailments and medical needs. However, some physicians abuse their patient's trust by performing unnecessary procedures under the banner of "medical necessity". Fraud detection algorithms and models flag possible fraudulent cases that can then be further investigated by authorities such as OIG or the Federal Bureau of Investigation. This explains why the

false positive is very high for all three models. It is better to err to the side of caution when billions of dollars of government money are at stake. Although the Autoencoders produced low AUC scores, there are ways to improve them by incorporating dropout layers or describing the representation using distributions for each attribute.

References

[1] U.S. Government, U.S. Centers for Medicare & Medicaid Services. The Official U.S. Government Site for Medicare. https://www.medicare.gov/. Accessed 10 Oct 2019.

[2] Morris L. Combating fraud in health care: an essential component of any cost containment strategy. Health Aff. 2009;28:1351–6. https://doi.org/10.1377/hlthaff.28.5.1351.

[3] Medicare fraud & abuse: prevention, detection, and reporting. Centers for Medicare & Medicaid Services. 2019. https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/Downloads/Fraud_and_Abuse.pdf. Accessed 10 Oct 2019 .

[4] U.S. Government, U.S. Department of Health and Human Services, Office of the Inspector General. List of Excluded Individuals/Entities. https://oig.hhs.gov/exclusions/. Accessed 10 Oct 2019.

[5] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. science. 2006;313(5786):504–7. doi: 10.1126/science.1127647.

[6] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407, 2019.

[7] Bauder RA, Khoshgoftaar TM. The detection of medicare fraud using machine learning methods with excluded provider labels. In: FLAIRS conference. 2018. p. 404–9.

[8] Herland *et al. J Big Data (2018) 5:29* https://doi.org/10.1186/s40537-018-0138-3.

[9] Johnson, J.M., Khoshgoftaar, T.M. Medicare fraud detection using neural networks. *J Big Data* **6,** 63 (2019) doi:10.1186/s40537-019-0225-0

[10] Daniel Lasaga and Prakash Santhana. Deep learning to detect medical treatment fraud. In *KDD 2017 Workshop on Anomaly Detection in Finance*, pages 114–120, 2018

[11] B Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *arXiv preprint arXiv:1801.03149*, 2018.

[12] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Mu ¨ller, and Marius Kloft. Deep one-class classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4393–4402, PMLR. http://proceedings.mlr.press/v80/ruff18a.html