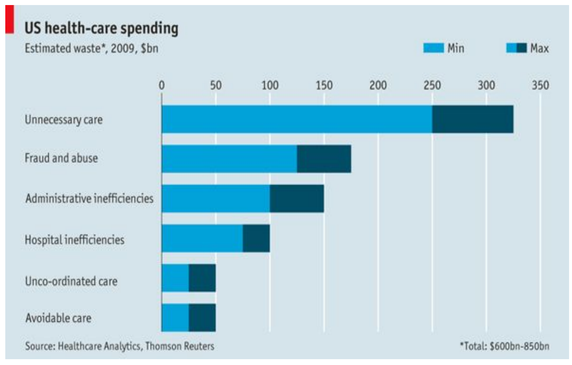
Fraud Detection in the Health Insurance Industry Using AI/ML Techniques

Keely Sweet
Kenyon College, AI for the Humanities

INTRODUCTION

Healthcare is a massive industry. In 2011, the United States spent $2.27 trillion on a little over four billion health insurance claims. According to the National Health Care Anti-Fraud Association (NHCAA), fraudulent claims cost insurers tens of billions of dollars each year.[1] Unfortunately these non-legitimate claims lead to higher insurance premiums and sometimes even reductions in coverage for responsible customers. As more insurance claims go through, the more expensive it is for employers and the government to cover all the health care needs of its clients. Most healthcare frauds take the form of billing for services that were never used, billing for more expensive services than were used, or performing altogether unnecessary services. However, there is hope with the emergence of machine learning algorithms that help to separate the fraudulent claims from the legitimate claims. Kirontech, a company based out of England, is attempting to do just that. They use everything from natural language processing to reinforcement learning to dynamic, unsupervised graph-based learning in order to detect otherwise unrecognizable patterns in the data.[3] With the sheer amount of data needed to be sifted through and analyzed, machine learning is better equipped to find and flag fraudulent healthcare claims.



2

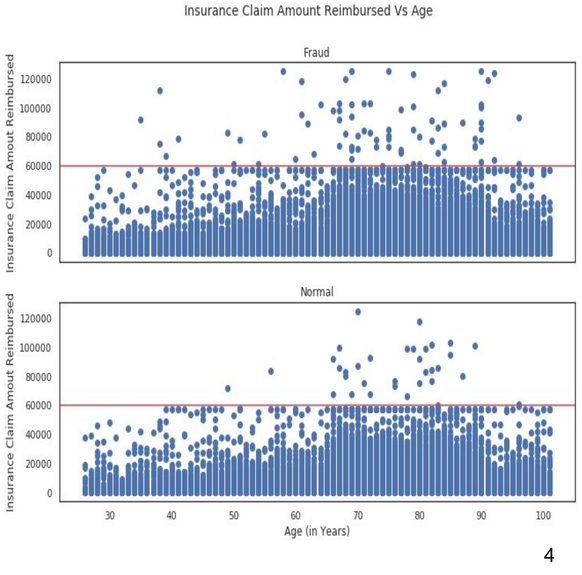
METHODOLOGY

In order to separate fraudulent claims from legitimate claims, a wide variety of variables need to be tested and assessed to determine whether they are significant. To begin my analysis, I use a Kaggle Notebook and data by Rohit Anand Gupta that compares many machine learning techniques.[4] After cleaning and organizing the data, it is important to examine and visualize it. In this dataset, we have information on race, gender, date of birth and date of death (if applicable), various disease indicators, location by state and county, diagnosis of patient, deductible amount paid, how much the patient was reimbursed, who the doctor was, and many more. Luckily, a lot of information is collected on patients and insurance claims, so we have many variables readily available.

However, we need to figure out which variables are relevant and which variables are just noise. To do this, we can look at our labeled training data and compare fraudulent claims to legitimate claims to see if there are any noticeable patterns in the data. After we decide which variables are significant, we can run our testing data through various machine learning methods to determine which one performs the highest. When deciding which technique does the best, we need to consider which type of error is the least devastating. Is it better to flag a claim as fraudulent when it is in fact legitimate, or is it better to have a few fraudulent claims pass as legitimate? In this scenario, I think it is more important to protect law abiding doctors from being wrongly accused, and as a result, some fraudulent claims will go unnoticed. In other words, we should strive for a low false positive rate.

DATA

When first looking at the data, we notice that fraudulent claims tend to come from the younger age group (30 to 70 years old). However, there is no noticeable difference between the deductible amount paid for fraudulent claims compared to legitimate claims. The same is also true for amount of the claim that was reimbursed. By just looking at the data, it is hard to find reliable patterns that determine whether a claim is fraudulent or not accurately.

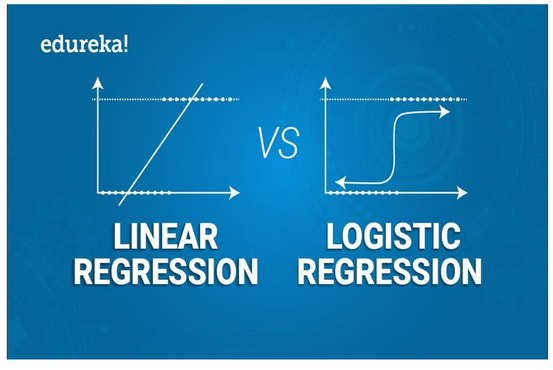


4

RESULTS

Linear Regression: A linear probability model uses an ordinary least squares regression to provide an output in the form of a probability. This model may not be very effective since the output values may go outside the bounds of 0 and 1. Another issue with an ordinary least squares regression with a limited dependent variable is heteroskedasticity. Since all the true values of the dependent variable lie at either 0 or 1, the error term is dependent on what the value of X is. This often leads to a large standard error which is symptomatic of imprecise estimations of the y-values. Errors should be randomly and equally scattered around the regression line, but this is not the case with an indicator dependent variable using ordinary least squares regression.

Logistic Regression: Considering that our desired output is binary, either a claim is fraudulent or legitimate, logistic regression is one technique that can be exploited. Logistic regression provides us with the probability that a claim is fraudulent with values ranging from 0 and 1. It also does a better job at fitting the data and has a smaller standard error. After training with a .60 probability threshold, we find that logistic regression on this testing data is 91.3 percent accurate with a specificity of 93.7 percent.



5

Random Forest: A random forest could also be an effective way to approach this issue. Each decision tree in the forest gives a classification, whether the claim is fraudulent or legitimate, and the probability is constructed from an aggregation of these individual tree classifications. However, we find that a random forest on this data is only 88.9 percent accurate with a specificity of 87.7 percent. This may be a result of overfitting, or training on variables that are not truly important. Just because we have data on a certain variable does not necessarily mean that it is important in determining whether a claim is fraudulent. To increase accuracy, we could try using a larger dataset to reduce sample bias. In addition, we could do some dimensionality reduction before feeding the data into the random forest as to prevent overfitting.

Auto-Encoder: An auto-encoder is a type of neural network that can determine which variables are important and which variables are just noise. Again, since we have so many variables, this property is extremely appealing. This neural network uses unsupervised learning and trains on non-fraudulent claims. We can then look at the threshold for the reconstruction error for the fraudulent claim data. An auto-encoder with two hidden layers and trained on 3927 epochs produce an accuracy of 91.0 percent and specificity of 94.6 percent. When it comes to sensitivity, this model did not do well with a value of 58.1 percent. This means that many legitimate claims were predicted to be fraudulent.

Overall, logistic regression performed the best on our testing data. When you have data that is relatively high in noise, logistic regression tends to outperform random forests because of the random forests' tendency to overfit. In addition, decision trees tend to be better at classifying data rather than generating probabilities. While auto-encoders are good at ignoring noise, in this instance, legitimate claims were often tagged as anomalies. This could be a result of training bias where not enough normal cases were examined and understood. Perhaps if there was more data, the auto-encoder would have a better idea of what the normal range of legitimate claims looked like.

FUTURE

Going forward, we can look to improve our methods by pinpointing which set of variables are significant when determining if a claim is fraudulent. Our random forest may have suffered from overfitting due to noisy data. In addition, we could add more data to our sample to reduce bias. When it comes to the logistic regression, we could also create an interaction term between two variables to see if a combination of the two creates a stronger impact on the probability of the claim being fraudulent or not. With more and more data being collected everyday, it may be easier to secure information on variables we have not even considered yet. Perhaps there is a variable or two, such as the number of doctors visits the patient has had in the past or the patient's distance from the doctor's office, that is indicative of fraudulent behavior.

RESOURCES

1. "The Challenge of Health Care Fraud." The National Healthcare Antifraud Association, www.nhcaa.org/resources/health-care-anti-fraud-resources/the-challenge-of-health-care-fraud.aspx.
2. "Waste Measurements." The Economist, The Economist Newspaper, 17 June 2011, www.economist.com/blogs/dailychart/2011/06/us-health-care-spending&fsrc=nwl.
3. "Kirontech." https://www.kirontech.com/
4. "Healthcare Provider Fraud Detection Analysis." Kaggle Kernel. https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis/kernels
5. "Linear Regression vs Logistic Regression." https://www.youtube.com/watch?v=OCwZyYH14uw