Fraud in Health Insurance: how do you detect it with Machine Learning?

Keywords: #ai #artificialintelligence #ml #machinelearning #python #tensorflow #fraud #frauddetection #deeplearning #dnn

1. Abstract

In this white paper we depict a problem occurring in Health Insurance, namely fraudulent claims. The challenge lies in sieving them out while retaining the legitimate ones. There are plenty of potential techniques for Fraud Detection, ranging from Supervised Learning to Unsupervised Learning. Due to the availability of abundant training data we decide to go with Supervised Learning in general and Deep Learning in particular. A number of technical considerations are discussed.

2. Problem statement

We want to define and implement an AI/ML-enriched process, which would enable us to detect fraudulent insurance claims and highlight them to relevant parties, who can take appropriate actions on them.

3. Background

Let’s suppose we run an insurance company with an automated system of receiving claims from policyholders. Let’s also suppose that we have lots of clients - the policyholders - and, therefore, lots of claims and, regrettably, much fraud.

In this Machine Learning Fraud Detection project we want to prevent fraud in Healthcare by detecting which medical claims were fraudulent. For example, a person might say they had a broken leg and claim money for it, whilst, in fact, no such damage to their body had been done. We shall discuss design for a software system, which could carry out such detection. It involved various Machine Learning techniques, such as Random Forest, Logistic Regression, and Deep Learning. Python and its associated technical stack was used for this project.

The system in question processes much data related to the policyholder making the claim, the incident causing the loss of health and the consultation with the doctor confirming the incident (the data might include, for example, the PMB - Prescribed Minimum Benefits).

4. Solution

As for most Machine Learning projects we have at our disposal some training data. This training data provides answers to the following questions:

1. Was the claim marked as fraudulent?
2. Was the claim rejected?
3. What proportion of claimed amount was paid out?

The training data also contains various information about the patient and the consultation with the doctor, such as: doctor name and id, identification and address details of the patient, history of claims of both the patient and the doctor (including fraudulent ones), timeline of injury vs. treatment vs. claim, etc.

For this challenge we shall consider one of three Machine Learning techniques: Random Forest, Logistic Regression, and Deep Learning. Please note that each of these is a Supervised Learning technique, despite the fact that Unsupervised Learning techniques are available for Fraud Detection. The reason is the abundance of training data: with ample training data Supervised Learning is bound to return more accurate results.

After some deliberation we decide to go with Deep Learning for this challenge. As you might have heard, Deep Learning has shown increasingly good performance in a number of applications, including Fraud Detection.

What is Deep Learning? It’s a type of learning utilizing a Neural Network with many layers. Neural Networks are function approximators, so let’s say that you know there is a function, which will map claims to one of two answers: fraudulent and legitimate. But given the enormous number of the claim space, you can’t implement this function exactly. Instead of implementing the function, you can, however, implement a function approximator, which will give approximate answers to the questions of whether a claim is fraudulent or not. One of such function approximators is the Neural Network, consisting of a number of layers https://en.wikipedia.org/wiki/Artificial_neural_network. And a Deep Neural Network consists of many layers https://en.wikipedia.org/wiki/Deep_learning#Deep_neural_networks.

This approach enables us to highlight fraudulent claims using all the provided details mentioned above. Thanks to dynamic design of the system, we can also retrain our model with little to no human intervention and redeploy it with little to no downtime.

5. Conclusion

Fraud Detection is a very challenging topic. Fortunately, however, there are many methods which come to our aid in handling it. The state of the art method of Deep Learning has been identified as an effective solution to the case of identifying fraudulent health insurance claims. This white paper described some of the considerations of this problem along with a candidate solution.

Fraud Detection solutions come in various guises. One of them concerns Health Insurance and works towards revealing fraudulent claims submitted by policyholders.

Fraud is a major problem for insurers. In health it frequently manifests as policyholders claiming that they underwent an illness or accident, which they didn't. With state-of-the-art Machine Learning solutions we can provide early detection of such cases and save significant costs in the process.

6. References

1. https://en.wikipedia.org/wiki/Machine_learning
2. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning
3. https://www.python.org/
4. https://arxiv.org/abs/2012.03754
5. https://en.wikipedia.org/wiki/Deep_learning
6. https://en.wikipedia.org/wiki/Fraud#Detection
7. https://www.tensorflow.org/

Would you like to hear more? Please get in touch with us via www.jagan.solutions.