Multiple inputs neural network for medicare fraud detection

Mansour Zoubeirou A Mayaki[1] and Miche Riveil[1]

[1]Université Côte d'Azur, CNRS, Inria, I3S, France

Abstract

Medicare fraud results in considerable losses for governments and insurance companies and results in higher premiums from clients. According to Insurance Europe, detected and undetected fraud costs around 13 billion euros per year to European citizens[3]. In the field of healthcare insurance, in France the compulsory scheme detected over 261.2 million euros of fraudulent services in 2018, mainly due to healthcare professionals and healthcare establishments[1]. In the United States, according to the FBI, medicare fraud costs insurance companies between 21 billion and 71 billion US dollars per year[5]. In a context where reducing management costs is a real issue for healthcare insurers, the fight against fraud is a real expectation of the customers of professionals in the sector so that everyone receives a fair return for their contributions.

This study aims to use artificial neural network based learners to detect fraudulent activities in medicare. We tested several neural network architectures and compared them to some baseline machine learning classifiers (logistic regression, random forests, Gradient boosting) based on the ROC AUC metric. We have used two types of neural network architectures: single input layer neural network models and multiple input models. We use the Medicare Provider Utilization and Payment Data from the Centers for Medicaid and Medicare Services (CMS) of the US federal government [2]. The CMS provides publicly available data that brings together all of the cost price requests sent by American hospitals to health insurance companies. The CMS data we use in this study has two parts : Part B that contains information on utilization, payment (allowed amount and Medicare payment) and Part D that provides information on prescription drugs prescribed by individual physicians and other health care providers. The main difficulty in applying machine learning methods in fraud detection is that the data sets are imbalanced and the classifiers tend to favor the majority class. The CMS data is particularly highly imbalanced with fraud rates between 0.038% and 0.074% [4]. The constructed classifiers should thus be able to take into account this issue of imbalance and give more importance to the minority class which is often the class of interest.

We first merged the two data sets and used it as input to a baseline multilayer perceptrons (MLP) neural network. In this architecture, all the features are used without making any distinction between them to predict fraudulent activities. Secondly, we separated the data set in two parts: Part 1 contains all features related to the provider and Part 2 contains the features related to the claim itself. We then used a MLP neural network with two distinct input layers, the first input layer takes Part 1 as input vector and the second input layer takes Part 2 as input. The final model is thus composed of two blocks which meet. Each block has its own input layer, hidden layers and an output layer. The outputs of the two blocks are then concatenated to form a single vector used to predict the probability of fraud. Such an architecture makes it possible to simultaneously take into account the information on the claims and that on the provider without mixing them. The second block in this architecture can be a classical MLP or

an auto encoder. When it's an auto encoder, we pretrain the auto encoder separately and use its weights in the final model in the form of transfer learning.

Our results show that although baseline MLP neural network (AUC=0.859) outperform the baseline machine learning classifiers (logistic regression = 0.849, random forests =0.769, Gradient boosting =0.692), they are outperformed by the multiple inputs neural networks with auto encoder (AUC=0.876). We have shown in this study that artificial neural networks based learners make it possible to detect fraudulent activities in medicare and in addition that multiple inputs learners give better results.

Keywords: Medicare fraud detection . Anomaly detection . Imbalanced data . Machine learning . Deep neural networks.

References

[1] *Bilan 2018 des actions de lutte contre la fraude et actions de contrôles*. Caisse nationale de l'Assurance Maladie. URL: https://www.ameli.fr/sites/default/files/2019-10-01-dp-controles-fraudes.pdf. (accessed: 10.05.2021).

[2] *Centers For Medicare Medicaid Services. Medicare fee-for-service provider utilization payment data physician and other supplier public use file: a methodological overview.* Centers For Medicare Medicaid Services. 2018. URL: https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/medicare-provider-charge-data/physician-and-other-supplier. (accessed: 01.07.2021).

[3] *Fraud prevention*. insurance europe. URL: https://www.insuranceeurope.eu/priorities/23/fraud-prevention. (accessed: 10.08.2021).

[4] Matthew Herland, Taghi M Khoshgoftaar, and Richard A Bauder. "Big data fraud detection using multiple medicare data sources". In: *Journal of Big Data* 5.1 (2018), p. 29.

[5] Justin M Johnson and Taghi M Khoshgoftaar. "Medicare fraud detection using neural networks". In: *Journal of Big Data* 6.1 (2019), pp. 1–35.