A Comparison of Machine Learning Methods Applicable to Healthcare Claims Fraud Detection

Nnaemeka Obodoekwe and Dustin Terence van der Haar(✉)

Academy of Computer Science and Software Engineering, University of Johannesburg,
Cnr Kingsway and University Road, Johannesburg 2006, Gauteng, South Africa
nnaemekaobodo@gmail.com, dvanderhaar@uj.ac.za

Abstract. The healthcare industry has become a very important pillar in the modern society but has witnessed an increase in fraudulent activities. Traditional fraud detection methods have been used to detect potential fraud, but for certain cases they have been insufficient and time consuming. Data mining which has emerged as a very important process in knowledge discovery has been successfully applied in the health insurance claims fraud detection. We implemented a prototype that comprised different methods and a comparison of each of the methods was carried out to determine which method is most suited for the Medicare dataset. We found that while ensemble methods and neural net performed, the logistic regression and the naive bayes model did not perform well as depicted in the result.

Keywords: Healthcare · Fraud detection · Machine learning

1 Introduction

The health insurance industry, a pillar in the modern-day society, that serves the purpose of providing affordable healthcare to individuals. Healthcare has become a necessity for households, hence the cost of healthcare forms a part of household expenditure. The increased cost of healthcare has made it a luxury rather than a basic need [1]. One of the reasons for the increased cost of health insurance can be attributed to the money lost through fraud in the healthcare system [2].

Fraud has impacted several aspects of life and the healthcare industry is no exception. Fraud occurs in the medical billing process leading to loss of funds by the insurance company which leads to the insurance company charging higher premiums to make up for the lost funds. The impact of fraud in healthcare has other implications aside from the monetary implications such as health risks that can arise from the altering of a patient's record by the physician [3].

Traditional fraud detection methods such as rule based statistical methods have been applied to detect possible fraud in the healthcare claims process, but

Á. Rocha et al. (Eds.): ICITS 2019, AISC 918, pp. 548–557, 2019.
https://doi.org/10.1007/978-3-030-11890-7_53

these methods no longer suffice due to the large number of claims to be processed and the variety of patterns these fraudulent activities take. Machine learning methods have been applied to other fraud detection problems such as credit card fraud detection and has also been applied to fraud detection in healthcare claims [4].

In the research study, we unpack the problem of healthcare claims fraud detection as well as the impacts it has. We then analyse how machine learning has been applied to the healthcare fraud claims detection problem by discussing the similar systems. With an understanding of the current research being done in the area, we create a data mining model that takes an exploratory approach to solving the problem of healthcare claims fraud detection by comparing and analysing different methods. We implement these methods in a prototype and then analyse the results to see which methods performed best with the Medicare dataset.

2 Problem Background

Abraham Maslow states the physiological needs of any individual are the most basic innate human needs that need to be satisfied and has the highest priority [5]. Maintaining a healthy body condition, eating, basic security is tantamount to satisfying the safety needs of an individual. To have good health, one needs access to adequate and affordable healthcare. Unfortunately, the cost of healthcare has been on the rise, making healthcare more of a luxury than a basic need [1]. One of the factors that have contributed to the increased cost of healthcare is the impact of the funds lost to fraudsters through healthcare claims fraud.

Before going deeper into the depth of health insurance, a definition of health insurance is needed. Health insurance represents a contract that a person pays an agreed premium to an insurance provider for a designated healthcare cover. The health insurance industry involves the transfer of funds and has been affected by fraudulent activities perpetrated by individuals that seek to gain illegal access to these funds.

Health insurance waste in healthcare is most times unrelated to fraud as it mainly the provision of unnecessary health services. Health insurance waste can only be seen as fraud and abuse when the act is intentional. Waste can occur when services are over utilized and then results in unnecessary expenditure [6].

Health insurance abuse is the billings of practices that either directly or indirectly is not consistent with the goals of providing patients with services that are medically necessary and these practices meet professionally recognized standards as well as being fairly priced [6].

Health insurance fraud is purposely billing for services that were never performed and or supplies not provided, medically unnecessary services and altering claims to receive higher reimbursement than the service produced [6].

To tackle the problem of fraud in the medical billing process, health insurance companies make use of traditional rule-base models, but these models do not suffice anymore due to several factors such as the large volume of claims to be processed which makes the medical billing process prone to error, slow and sometimes inefficient. Machine learning methods can be used to improve the detection of possible healthcare claims fraud. The next section discusses the related works on the applications of machine learning methods in the detection of possible fraudulent healthcare claims.

3 Review of Related Work

Machine learning methods have been effective in automatically extracting patterns from data to derive knowledge which yields meaningful results such as detecting which submitted claims are likely fraudulent. The first work we consider is the Outlier-based health insurance fraud detection for US Medicaid data presented by Thornton et al. [7]. They made use of Medicaid data for dental services which is a healthcare provider in the US that caters to low income people. Their model made use of 3 different univariate machine learning methods which are the linear regression, time series plot as well as box plot. They also used a multivariate method through clustering to detect possible health insurance fraud. The dataset used contained a case study for 500 dentists and they successfully identified 17 activities that can be deemed fraudulent among the 360 records analysed.

We also reviewed "Graph Analytics for Healthcare Fraud Risk Estimation" by Branting et al. which made use of a graph to link providers, drug prescriptions and the procedures [8]. They used two algorithms where the first algorithm was used to calculate the similarity to predetermined fraudulent and non-fraudulent providers while the second algorithm calculates the estimated fraud risk through location of practitioners. They achieved an F-score of 0.919 and an impressive AUC of 0.960.

Bauder et al. also carried out several works in the area of detecting fraud in the health insurance process. One of the systems, a multivariate outlier detection in Medicare claims payments applying probabilistic programming methods [9]. They created a base for what the expected Medicare payments should look like for each type of provider. Outliers were then identified by comparing payment amounts with the normative case and the deviations are categorised as outliers.

4 Experimental Setup

To establish a way to conduct the research, we define a research methodology that form the guide to solving the problem at hand. The quantitative research approach was chosen as it allows for the statistical analysis of the data and maintains an objective standpoint.

The data that was used for the study is the Medicare payments data between 2012–2015. The dataset contained payments and utilization healthcare claims data as well as the details about the procedures rendered to individuals. The data from the List of Excluded Individual or Entities (LEIE) was used to create the ground truth labels.

5 Model

Applying a machine learning algorithm to derive knowledge is just a piece in the puzzle of creating an effective data mining model. The data mining model consists of different processes and each of these processes play a major role in deriving knowledge from the data. In this section, we unpack these individual processes as well as the methods that were used for each individual process.

5.1 The Data Collection Phase

Data is the raw material to be processed in a data mining system. Data can be structured, unstructured or semi-structured. The Medicare dataset used, was in a structured format. Both datasets were loaded and stored in a database for further analysis.

5.2 The Data Pre-processing and Transformation Phase

Normally, data in the real world is dirty, contains missing values and can be incorrect. Therefore, a lot of work to be done cleaning up the data to get it to the form that will be suitable for use. The Medicare data used was already pre-processed by the Centre for Medical Services (CMS). The work first task we performed was filtering the Medicare dataset for only non-prescription data. The Medicare dataset did not contain any label to be used to differentiate between fraudulent and non-fraudulent claims, therefore we used the LEIE dataset to flag the claims that were detected as fraudulent. We made use of fuzzy matching on the practitioner's first name, last name and ZIP code to link the Medicare payment data to a practitioner in the LEIE dataset, as there was no explicit join between the two datasets.

After labelling the dataset and identifying ground truth labels, the next task performed was indexing categorical string features. We based the initial selection of features on the work done by [10] using the same Medicare dataset as shown in Table 1. We also used feedback gotten from the model as calculated by the feature importance to adjust the features used for the model.

Finally, we applied a 70:30 split to the dataset. 70 % of the data was used for training the machine learning model while the remaining 30 % was used for testing the outcome of the model.

5.3 Machine Learning Application

Now that we have completed the pre-processing and transformation of the data, we apply machine learning algorithm to derive insights from the data. We used an exploratory approach in creating the machine learning model as several methods were used and results were collected on the performance of the different methods used.

Table 1. Description of medicare features.

Feature	Description
NPI	Unique provider identification number
last_name	Provider's last name
First_name	Provider's first name
Zip	Provider's 5-digit zip code
provider_type	Medical provider's specialty (or practice)
line_srvc_cnt	Number of procedures performed per provider
bene_unique_cnt	Number of distinct beneficiaries per day services
average_submitted_chrg_amt	Average of the charges that the provider submitted
average_medicare_payment_amt	Amount paid to the provider for services performed

1. The Bayesian classifier is a simple classifier based on statistical methods and it assigns probabilities to each member of the class in such a way that a given sample can be categorised into one class. It makes a strong assumption on the independence of features.
2. Random forest classifier, an ensemble learning method made up of a sequential construct of several decision trees in the training phase and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
3. Logistic regression is another machine learning method based on statistical principles. It is highly suited for predicting categorical features. In predicting a binary outcome, the logistic regression model makes use of the binomial logistic regression and for multiple outcomes it makes use of multinomial logistic regression.
4. Gradient Boosted tree classifier is yet another powerful classification method. The classification method uses ensembles of decision trees and applies the technique known as boosting to improve performance. The idea of boosting emanates from the attempt to combine weaker learners to become better learners. The gradient boosted tree classifier is a combination of a loss function, a weak learner and the additive function responsible for combining the weak learners and reducing the loss function.
5. The artificial neural network is a machine learning method based off the functioning of the neurons in the brain of a biological system. It is made up of a network of interconnected nodes. The nodes do not contain any computation but rather, they function as a group of linear functions. The nodes in the neural network are grouped into layers. The behaviour of each node is defined by an activation function.

6 Results

Once the data has been passed through to the machine learning process, we evaluate how well the model performed by using the following pre-determined benchmarks. We used the following metrics to evaluate the different machine learning models created: Weighted Precision, Weighted Recall, AUC, Test Error, Sensitivity, Specificity, F1. In this section, we unpack how each of the implementations performed against the defined metrics.

Table 2. Result for the performance metric of the different machine learning methods.

ML model	Weighted precision	Weighted recall	AUC	Test error	Sensitivity	Specificity	F1
Naive Bayes	73.7	82.1	47.0	17.9	11.6	99.4	74.6
GBT	**93.3**	**93.3**	**97.0**	**6.7**	**73**	98	**93**
Random forest	88.6	88.61	92	11.4	41.3	**98.8**	86.9
Logistic regression	63.5	82.3	63.5	17.7	45.2	93.0	74.3
Neural network	89	90	93.8	10	71.2	94.8	90

6.1 Naive Bayes

We implemented the Bayesian model using the Apache SparkML library using multinomial distribution of features and a smoothing of 1.0. The model performed poorly with a ROC curve of 47.0 as seen in the figure below, which entails that the predictions of the model was just as good as random guesses. The Naive Bayes model scored well in the other metrics and also recorded a low test error of 17.9%.

6.2 Logistic Regressions

The logistic regression model was a slight improvement to the Naive Bayes model. The logistic regression model was created with a regularization parameter of 0.3 and the elastic net parameter was set at 0.8 based on the recommendation of the SparkML documentation. The logistic regression model was run over several iterations and achieved the best result at 10 iterations. An improved AUC of 63.5% was achieved as seen in Fig. 2, which is better than the Bayesian model but not good enough for predictions. The model achieved an accuracy of 83.7% (Fig. 1).

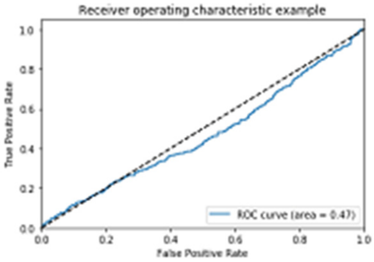

Fig. 1. ROC curve for the Naive Bayes classification model

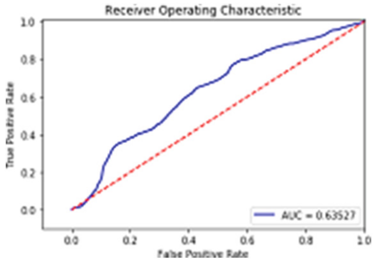

Fig. 2. ROC curve for the logistic regression classification model

6.3 Random Forest Classifier

The random forest classifier was made up of 10 decision trees and a maximum of one hundred bins. The random forest offered a slight improvement to the logistic regression with an AUC of 63.5%. The random forest classifier had a low specificity score of 41.3% and an accuracy of 89.6%. The AUC score of 92.0% makes the random forest classifier suitable for the Medicare dataset as the score implies a great improvement to random guess noticed in the previous models.

6.4 Gradient Boosted Tree Classifier

The gradient boosted tree classifier model presented a great improvement to the previous models. The model was run through 10 iterations. The gradient boosted tree classifier had a weighed precision and recall of 93.3% each. It had an accuracy of 93.3% and an F1 score of 93 %. It recorded a sensitivity score of 73% and a specificity of 98%. The most important improvement was the noticeable increase in the AUC score of 97.0%. Based on the results the gradient boosted tree classifier will be ideal for detecting possible fraud in healthcare claim using the Medicare data (Figs. 3 and 4).

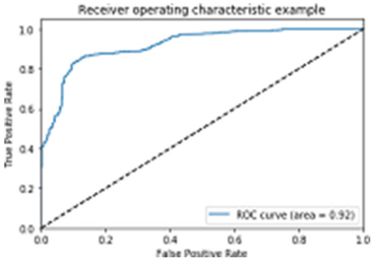

Fig. 3. ROC curve for the random forest classification model

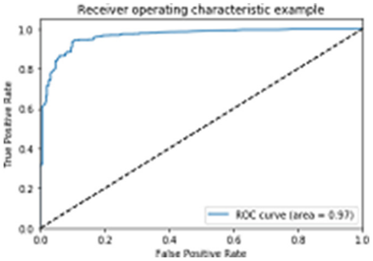

Fig. 4. ROC curve for the gradient boosted tree classification model

6.5 Artificial Neural Net

The artificial neural network slightly underperformed the gradient boosted tree classifier as can be seen in the metrics in Table 2. The model made use of the binary cross-entropy as the loss function. The model had a weighted precision and a weighted recall of 89% and 90% respectively. It also had an accuracy of 90% with specificity and sensitivity measuring 71.2% and 94.8% respectively. The F1 score was 90% and it also had an AUC of 93.8%.

7 Discussion

We have shown how the proposed model functions as well the results that were derived from the model. There are several limitations associated with the model which can act as a hindrance in the implementation of the model. The following limitations were evident through the implementation of the model.

7.1 Critique

Lack of Labelled Data. There was difficulty in obtaining openly available data with ground truth label. We made use of data from the LEIE database for ground truth label. The availability and use of a dataset that has been labelled initially would be a better validation of the model (Fig. 5).

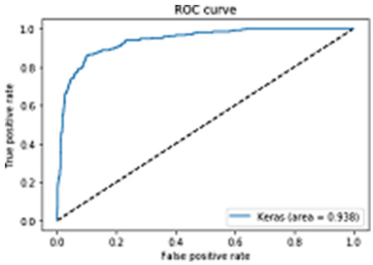

Fig. 5. ROC curve for the artificial neural network classification model

Sample Dataset. The data that was used in the model was only concerned with the healthcare provider. Although fraud mainly occurs through the healthcare provider, it doesn't imply that fraud cannot occur through the other entities in the health insurance ecosystem. The dataset we used constrain the research to only consider fraud from the medical practitioner.

7.2 Support

Applications of Alternative Machine Learning Methods. The choice of the machine learning methods to be used in a data mining system is important. The approach taken in the study is an exploratory approach, applying different types of machine learning methods and the assessing them against the benchmark. The use of multiple machine learning methods allowed for a better analysis of how each method performs with the Medicare dataset as well as the different advantages and disadvantages of each model.

Flexibility of Solution. The model is structured in a modular manner which allows for a flexible implementation of a healthcare claims fraud detection system. The modular implementation of the system means that the different components of the system can be easily replaced thereby reducing the burdens of future system upgrades.

8 Conclusion

The task of building a system that enables the identification of possible fraudulent healthcare claims has become very important as the demand for healthcare increases. The medical billing process, due to its complexities and the volume of claims to be processed has been exploited by fraudsters looking to gain illegal advantage from the system. Machine learning methods have enabled the identification of these fraudulent claims and have improved the effectiveness of the medical billing process.

The study explored the application of different machine learning methods to detect fraudulent claims in the medical billing process. The result generated from the application of these machine learning methods were collected. The analysis of the result showed that the ensemble methods and the artificial neural network performed best with the Medicare dataset.

Through the insights gained from the study, we were able to identify the strengths and weaknesses of the model. The Medicare dataset limited the system to only identify fraud that occurs through the medical practitioner as the Medicare payment data only contained data regarding the medical practitioner. Notwithstanding the limitations, the success of the model in identifying the fraudulent healthcare claims as well as the explorative approach taken to determine the which machine learning method makes this a viable solution.

References

1. Yu, H.: Impacts of rising health care costs on families with employment-based private insurance: a national analysis with state fixed effects. Health Serv. Res. **47**(5), 2012–2030 (2012)
2. Singh, A.: Fraud in insurance on rise. Technical report, Ernst & Young (2011)
3. Davis, L.E.: Growing health care fraud drastically affects all of us, October 2017
4. Rabiul, J., Nabeel, M., Ahsan, H., Sifat, M.: An evaluation of data processing solutions considering preprocessing and “special” features. In: 11th International Conference on Signal-Image Technology & Internet-Based Systems (2015)
5. McLeod, S.: Maslow’s hierarchy of needs. Simply Psychol. **1** (2007)
6. Bush, J., Sandridge, L., Treadway, C., Vance, K., Coustasse, A.: Medicare fraud, waste and abuse. In: Business and Health Administration Association Annual Conference (2017)
7. Thornton, D., van Capelleveen, G., Poel, M., van Hillegersberg, J., Mueller, R.M.: Outlier-based health insurance fraud detection for U.S. medicaid data. In: 16th International Conference on Enterprise Information Systems (2014)
8. Branting, L.K., Reeder, F., Gold, J., Champney, T.: Graph analytics for healthcare fraud risk estimation. In: Advances in Social Networks Analysis and Mining (ASONAM) (2016)
9. Bauder, R.A., Khoshgoftaar, T.M.: A probabilistic programming approach for outlier detection in healthcare claims. 15th IEEE International Conference on Machine Learning and Applications (ICMLA) (2016)
10. Bauder, R.A., Khoshgoftaar, T.M.: Medicare fraud detection using machine learning methods. In: 16th IEEE International Conference on Machine Learning and Applications (2017)