International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

Effective Fraud Detection in Healthcare Domain using Popular Classification Modeling Techniques

Sheffali Suri, Deepa V Jose

Abstract: ***Fraud is any activity with malicious intentions resulting in personal gain. In the Present Day scenario, every sector is polluted by such fraudulent activities to fetch unauthorized benefits. In HealthCare, an increase in fraudulent insurance claims has been observed over the years which may constitute around 3-5% of the total cost. Increasing healthcare costs along with the hike in fraud cases have made it difficult for people to approach these services when required. To avoid such situations, we must understand and identify such illegal acts and prepare our systems to combat such cases. Thus, there is a need to have a powerful mechanism to detect and avoid fraudulent activities. Many Data mining approaches are applied to identify, analyze and categorized fraud claims from the genuine ones. In this paper, various frauds existing in the Health Care sector have been discussed along with analyzing the effect of frauds in the health care domain with existing data mining models. Furthermost, a comparative analysis is performed on two existing approaches to extract relevant patterns related to fraudulent claims.***

Keywords: Anomaly Detection, Data Mining, Fraud Detection, Health Care, Machine Learning.

I. INTRODUCTION

Health plays an eminent role in everyone's life. In this world full of uncertainties, every individual wants to live a healthy and disease-free life. To ensure a healthy future, we all depend on the Healthcare services available. With the rapid growth in Health Care sectors, the risk of falling for jagged practices has also increased. According to research, the USA spends almost one-third of the healthcare expenses on fraud, waste, and abuse which roughly estimates up to 19 to 65 billion[1][2][3]. These frauds involve the participation of many factors like:

- Service Providers which include doctors, hospitals, labs etc.
- Insurance subscribers which includes patients
- Insurance carriers who make the premium transactions on the behalf of patients[10]

Recently, CEO Dr Prem Reddy consented to pay $65 Million to Settle False Claims Act Allegations. Healthcare-related programs such as Medicaid and Medicare are primarily targeted for such FAW acts. India spends overall 600-800 crs on fraudulent claims yearly [12].
Figure 1 and 2 represent popular frauds in health care and methods currently being used to detect frauds.

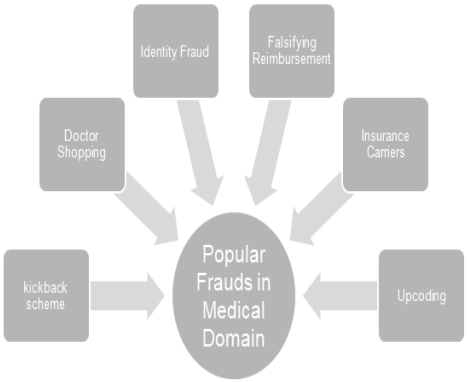

Fig 1-Popular Frauds in Medical Domain

Based on an analysis performed, Health care frauds can be predominantly classified in 67 categories. Out of which, the most common ones are listed below[1][11].

- kickback scheme
- Doctor Shopping
- Identity Fraud
- Falsifying Reimbursement
- Insurance Carriers
- Upcoding

It is quite evident how these frauds are a threat to all and how essential it is to eliminate them. Various algorithmic approaches are proposed to detect doubtful claims. With the recent advancement in claim policies, there is a need of having sophisticated methods that are capable to automatically learn the fraud pattern.

Manuscript published on 30 September 2019.
*Correspondence Author(s)
Sheffali Suri Perusing Master in Computer Science Department of Computer Science, CHRIST
Deepa.V.Jose. FacultyDepartment of Computer Science, CHRIST

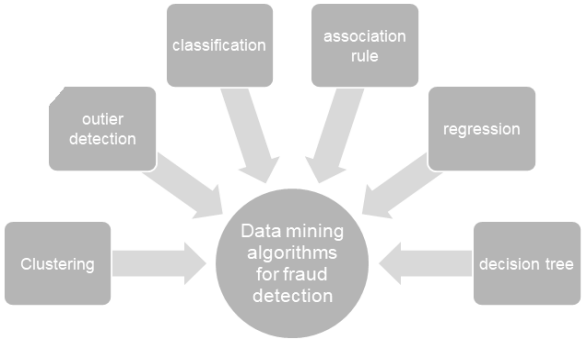

Fig-2 Data Mining Algorithms for Fraud Detection

Some of the most popular data mining methods are;

- Clustering
- outlier detection
- classification
- association rule
- regression
- decision tree[1]

Further researches are being carried to filter objectionable claims from the legit one**.** Again many researches are being carried with the aim to filter objectionable claims from the legit one.

II. RELATED WORKS

In this research, various frauds existing in the Health Care sector is discussed along with analyzing the effect of frauds in that domain. Therefore, the intention here is to focus on any related works on fraud detection in Healthcare. Also, various Data mining techniques are reviewed which have been applied successfully to detect bogus claims.

In [1], Systematic literature review of already published work in health care frauds is discussed. They have discussed the types of frauds detected along with methods used. The article acknowledged how various data mining algorithms are used to combat frauds.In [9], popular health care frauds were discussed present in Medicare data. Upcoding was introduced as one of the most significant ways of filing Fraud claims.In[2], Data mining approaches are discussed along with their pros and cons. The authors propose a hybrid model approach which implements support vector machine (SVM) and expectation conditional maximization(ECM); stating classification and clustering gives a better result and can help reducing frauds.

According to [3], the authors provide a brief introduction to health care .The principle goal of this paper is to think about and examine the works in healthcare fraud, specifically upcoding fraud analysis and detection in order to help alleviate fraud and reduce financial losses in the healthcare domain.The authors in [5] have addressed the need for semi-supervised since data readily available is unstructured and structured is expensive and limited due to legal and privacy issues.

Hao peng[4] proposes an improved feed-forward neural network algorithm to detect medical insurance frauds which combines 3 layers MPL neural networks. The algorithm showed 86% accuracy. They suggested the Use other neural network methods such as adaptive resonance method, self-organizing map networks to get better results.

Verma[6] talks about various Sophisticated statistical methods that are capable of automatic learning from data using data mining and ML models. The authors propose an effective fraud claim detection method along with identifying frequently occurring fraud pattern. This approach is an integration of three mining approaches-Association rule, Clustering, Outlier detection. Richard A. carried out a series of research on the same domain.

In[7], the author performs Empirical study on unsupervised learning algorithms for detecting frauds focusing on isolation and unsupervised random forest. In [8], Fraud detection is evaluated on an ensemble learner, Random Forest, and Logistic Regression. Their results indicated RF ensemble learner to be the best performer with AUC scores significantly higher than Logistic Regression. We reviewed studies that performed for detecting health care fraud and abuse, using various data mining approaches. Most of the studies suggested going ahead with hybrid approaches since dataset available are highly unstructured and there is vast scope to explore and come up with something more efficient.

III. METHODOLOGY USED

Here in this section, details related to datasets, existing algorithms and the performance metrics have been discussed. Figure 3 mentions the methodology followed in this research work.

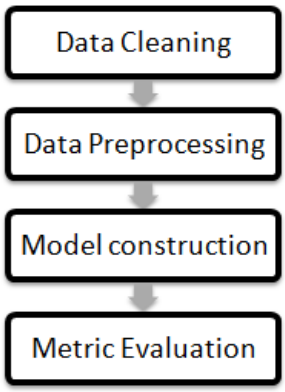

Fig 3. Block Diagram representing Implementation Flow

Additionally, we explain feature selection strategies along with finding principal variables.

These variables are helpful in tracing the pattern of fraud providers. Further, this would help in tracking fraudulent behaviour in the provider's claims to understand the pattern of similar frauds.

A. Data Source

The datasets used here has been taken from Kaggle[13]. The dataset is already partitioned into Test and Training samples which are suitable for both Supervised as well as unsupervised modeling. Table 1 describes the data source in detail.

Table 1 Available Datasets

TRAIN DATA SET		TEST DATA SET	
Dataset	**Shape**	**Dataset**	**Shape**
Inpatient data	40474, 30	Inpatient data	9551,30
Outpatient data	517737, 27	Outpatient data	125841,27
Beneficiary data	138556, 25	Beneficiary data	63968,25
Train Result	5410, 2	Test Result	1353,1

These datasets are categorized as Inpatient claims, Outpatient claims and Beneficiary details of each provider.

A) Inpatient Data-The data provides information about patients admitted in the hospital. Some of the features of these data sets are attending physician, insurance provider, admit date, and discharge date.

B) Outpatient Data-This data provides details about the claims filed for those patients who visit hospitals and not admitted in it.

C) Beneficiary Details Data-This data contains details about patients like gender, country, health conditions etc[13].

The data source also contains datasets named Train and Test data set stating whether the claim filled is fraud or not

B. Data Preprocessing

The datasets available are well structured and is suitable for implementation. The initial step of preprocessing of data is handling missing values. Table 2 has the count of all the missing values in data sources.

TABLE -2 Missing Values

TRAIN DATA SET		TEST DATA SET	
Dataset	**Missing values**	**Dataset**	**Missing Values**
Inpatient data	358960	Inpatient data	72305
Beneficiary data	137135	Beneficiary data	63394
Train Result	0	Test Result	0

The missing values can be substituted using an appropriate imputation idea depending upon the type of column having missing values. The replacing value can be calculated using central tendency statistics of each of the columns or by relevant constant, the next step involves feature selection. It is essential to identify the chief features of the dataset which can contribute to a constructive model building for distinguishing claims. Primary exploratory data analysis is performed based on which a set of features is selected[14].

Age(Figure-3) is picked out as an important feature from the dataset as the frequency of fraud claims being filed is high for age ranging from 30-70 years (Mid years) when compared to 1-30(early years) and 70+(late years)[13].

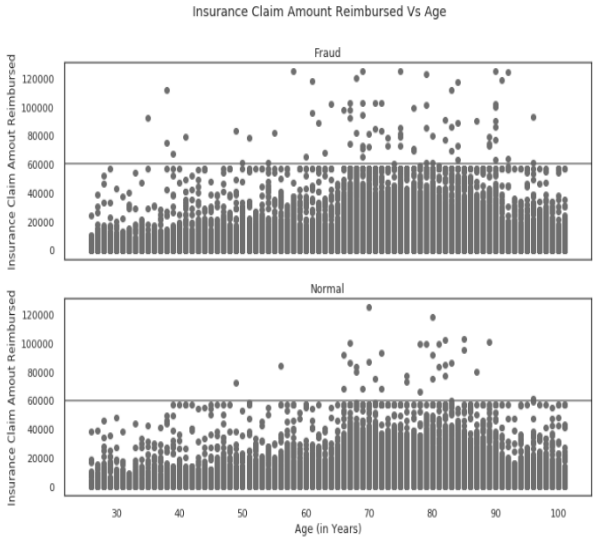

Fig-4 Insurance Claim Amount Reimbursed Vs Age Graph

Also, Race (Figure-4) as an important feature for the modeling as maximum of claimers benefitted are from a particular race (Race 1).

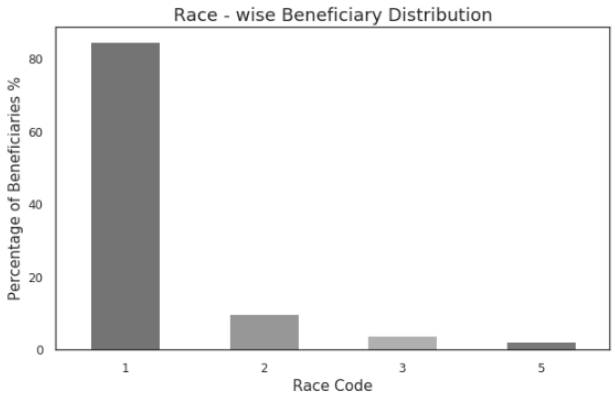

Fig-5 Insurance Claim Vs race-wise beneficiaries Graph

Based on the EDA, 26 features are selected from the test dataset for further modeling and evaluation.

C. Model Construction

In order to assess the performance of various existing algorithms on datasets, we employ the following modelling approaches. Also, approaches are evaluated and compared to come up with the most optimal idea for detection of maximum frauds from real-time datasets.

Logistic Regression-

Logistic Regression is a Supervised Machine learning algorithm. It is a classification algorithm used to classify variable based on various features associated. It is a statistical model which involves more than one independent variables. In this scenario, there exist two classes i.e. Binomial logistic regression which can be coded as 1 (yes, fraud claim) and 0(no, non-fraud claim). The dependent variable is classified under these two classes on the basis of independent variables[15].

Random Forest-

Random Forest is an ML algorithm which involves a large number of decision trees operating together. Each decision tree (belonging to random forest) produces the result in the form of class prediction for a particular data point[16]. The class with maximum majority becomes the final predicted class for the data point. In the case of Random forest, Accuracy of the model is directly proportional to the number of trees [16].

D. Model Evaluation

Here, Confusion matrix performance is used for measuring the performance of the models. Table 3 represents the confusion matrix for the dataset used.

TABLE -3 Confusion Matrix

	Predicted: Non-Fraud Claim	**Predicted: Fraud Claim**
Actual: Non Fraud Claim	True Negatives(TN)	False Positive(FP)
Actual: Fraud Claim	False Negatives(FN)	True Positives(FP)

The Column of the above matrix represents the predicted classes for the data points by above-mentioned models and rows represent the actual classes for the same.

The above-mentioned categories:

- True Positives (TP)-This represents the category when the model predicts a record as a fraud claim and the claim is actually a fraud.
- True Negatives (TN)-This represents the category when the predicted result states the record as Non-fraud Claim and it is not a fraud in real too.
- False Positive (TP)-This represents the category when the classifier predicted “Fraud claim” but it was a “Non-Fraud Claim”.
- False Negatives (FN)-this represents the category when the classifier predicted “Non-Fraud Claim” but it was a “Fraud Claim”.

Based on Confusion Matrix, Following measures are calculated for evaluation.

- Accuracy-It is calculated as all correctly predicted classes by the total number of the dataset. The best-case accuracy is 1.0 and the worst-case accuracy is 0.0[17].
- Sensitivity-It is calculated as the number of correctly predicted classes by total positive classes. The best-case Sensitivity is 1.0 and worst case 0.0[17].
- Specificity- It is calculated as the number of correctly predicted negative classes by total negative classes. The best-case Sensitivity is 1.0 and worst case 0.0[17].

IV. RESULTS AND DISCUSSION

Confusion matrix depicts the frequency of each class for the dataset used. The following Figures (Fig-6 a) & b)) represents Confusion Matrix obtained for Logistic Regression Modeling.

```
Confusion Matrix Train :
 [[ 270   84]
 [ 210 3223]]
```

```
Confusion Matrix Val:
 [[ 103   49]
 [  93 1378]]
```

Fig 6 a)Confusion Matrix for Train Samples
b)Confusion matix for test Samples

Fig- 7 a) & b) represents the Confusion matrix for Random Forest Classifier.

```
Confusion Matrix Test:
 [[ 124   28]
 [ 181 1290]]
```

```
Confusion Matrix Train :
 [[ 320   34]
 [ 388 3045]]
```

Fig 7 a)Confusion Matrix for Train Samples
b)Confusion matix for test Samples

Following measures are calculated for evaluation and thus a comparative analysis is performed between both the classification techniques i.e. logistic Regression and Random Forest Technique.

Table 4 Comparative Analysis between Logistic Regression and Random Forest

Evaluation	**Logistic Regression**	**Random Forest**
Accuracy (Train)-	0.92	0.88
Accuracy-	0.91	0.87
Sensitivity-	0.67	0.81
Specificity-	0.93	0.87
Kappa Value-	0.54	0.47
AUC-	0.80	0.84
F1-score Train	0.647	0.60
F1-score validation	0.591	0.54

Table 4 contains the resultant values for the performance measure metrics calculated. From the values, It is quite evident to consider Logistic Regression as a better classification model in fraud detection when compared to Random Forest. The implementation of Logistic Regression points towards the linearity between variables whereas Random forest is able to handle large sets of High dimensional data. Both models have their own pros and cons. Also, the selection of the model also depends on the data source.

V. CONCLUSION

Ultimately, frauds act as major pollutants in the field of healthcare. Detection of frauds in the healthcare domain is a tedious job if done manually and thus requires a fast and optimal solution. Therefore, there is a need for effective machine learning models to detect and predict future frauds and prevent loss. This research work is an attempt to solve this issue. To find out patterns and predict the behavior of a particular pattern, many supervised and unsupervised algorithms are being practiced.

The central idea of these approaches is to come up with suspicious data from big data sets. The algorithms work with the idea to find those 50 suspicious records which can constitute to FWA from 5000 data rows. Thus, it reduces the human effort of going through each record manually and allows one to work on those potential fraud records.

For future work, the same approaches can be tested on data from various other sources along with different sampling techniques. Also, an approach towards the Development of Self-evolving fraud detection methods can be proposed as well.

REFERENCES

1. 1 . Dallas Thorntona Michel Brinkhuisb Chintan Amritb Robin Alyc: Categorizing and Describing the Types of Fraud in Healthcare, Procedia Computer Science,Volume 64, 2015, Pages 713-720
2. Vipula rawte,G Anuradha: Fraud detection in Health Insurance using Data Mining Techniques, 2015 International Conference on Communication, Information & Computing Technology (ICCICT) ,January 2015
3. Richard bauder,Taghi M.,Naeem Seliya:A survey on the state healthcare upcoding fraud analysis and detection, Health Services and Outcomes Research Methodology, March 2017, Volume 17, Issue 1, pp 31–55
4. Hao Peng ; Mengzhuo You: The Health Care Fraud Detection Using the Pharmacopoeia Spectrum Tree and Neural Network Analytic Contribution Hierarchy Process, IEEE,2016
5. Dr Ananthi sheshasayee ,Surya susan Thomas: Implementation of data mining techniques in upcoding fraud detection in the monetary domains, International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) 21-23 Feb. 2017
6. Aayushi Verma ; Anu Taneja ; Anuja Arora: Fraud detection and frequent pattern matching in insurance claims using data mining techniques, Tenth International Conference on Contemporary Computing (IC3), 10-12 Aug. 2017
7. Richard A.Bauder,Raquel C.da Rosa,Taghi M.: Identifying Medicare Provider Fraud with Unsupervised Machine Learning, International Conference on Information Reuse and Integration (IRI), 6-9 July 2018
8. Richard Bauder ; Taghi Khoshgoftaar,Amri Napolitano: Fraud Detection with a Limited Number of Known Fraudulent Medicare Providers
9. McGee, J., Sandridge, L., Treadway, C., Vance, K., & Coustasse, A: Strategies for Fighting Medicare Fraud, Health Care Manag (Frederick). 2018 Apr/Jun;37(2):147-154
10. Qi Liu ,Miklos Vasarhelyi:Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information, Published 2013
11. Richard Bauder Taghi M Khosgoftaar Naeem Seliya: A survey on upcoding healthcare fraud Analysis and detection: Health Services and Outcomes Research Methodology, March 2017, Volume 17, Issue 1, pp 31–55
12. Steven porter:prime healthcare ceo to pay part of $65m false claims settlement personally, https://www.healthleadersmedia.com/finance/prime-healthcare-ceo-pay-part-65m-false-claims-settlement-personally
13. Rohit Anand Gupta: Healthcare provider fraud detection analysis,https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis
14. Tanveer Sayyed:The penalty of missing values in Data Science, https://www.freecodecamp.org/news/the-penalty-of-missing-values-in-data-science-91b756f95a32/
15. Jason Brownlee :Supervised and Unsupervised Machine Learning Algorithms,https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
16. 16.Dylan Storey: Random Forests, Decision Trees, and Ensemble Methods Explained, https://www.datascience.com/blog/random-forests-decision-trees-ensemble-methods
17. Basic evaluation measures from the confusion matrix, https://classeval.wordpress.com/introduction/basic-evaluation-measures/

AUTHORS PROFILE

Sheffali Suri is perusing her Master in Computer Science from Department of Computer Science, CHRIST (Deemed to be University. She has keen interest in Data Mining, Machine learning and AI.sheffali.suri@cs.christuniversity.in

Deepa.V.Jose is the faculty in Department of Computer Science, CHRIST (Deemed to be University). Her area of research interest includes network security, AI and Machine Learning. deepa.v.jose@christuniversity.in