Health Care Fraud Detection With Iterative Feature Engineering And Regression Model

Sonu Patel[1], Prof. Anshul Khurana[2]
[1]Dept of CTA
[2]Professor, Dept of CSE
[1, 2] Shri Ram Institute of Technology Jabalpur, Madhya Pradesh, India.

Abstract- *The increment of computer technology use and the continued growth of companies have enabled most financial transactions to be performed through the electronic commerce systems, such as using the credit card system, telecommunication system, healthcare insurance system, etc. Unfortunately, these systems are used by both legitimate users and fraudsters. In addition, fraudsters utilized different approaches to breach the electronic commerce systems. Fraud prevention systems (FPSs) are insufficient to provide adequate security to the electronic commerce systems. However, the collaboration of FDSs with FPSs might be effective to secure electronic commerce systems.*

Nevertheless, there are issues and challenges that hinder the performance of FDSs, such as concept drift, supports real time detection, skewed distribution, large amount of data etc. This thesis aims to provide a solution of these issues and challenges that obstruct the performance of FDSs. We have selected the healthcare insurance systems to detect the frauds using Logistic Regression. The prevalent fraud types in those health systems are introduced and analyzed closely. Then a brief discussion on potential research trends in the near future and conclusion are presented.

Keywords- Health Care Fraud, Medical Claims, Inpatient claims, Outpatient claims, Beneficiary Details Data, Machine Learning, Logistic Regression.

I. INTRODUCTION

Data Analytics Technologies have been applied in Healthcare sector providing significant advantages on both the quality of the healthcare services and the control of the cost [1], [2] According to the international literature, the ability to detect fraud in medical claims seems to be a major asset on the expenditure control [3], [4]. Specific Big Data analytic techniques can lead to effective fraud detection in the health domain [5], [6]. Additionally, the examination of outliers in large amount of data saved in records is a major challenge faced in real world datasets in various research fields [7]. An outlier can be considered as a value that escapes normality and this can cause anomalies and unreal deviations in the results obtained through algorithms and analytical systems. The outlier's examination can lead to valuable and accurate results which will have positive effects on the research in every field. In healthcare systems the outliers of a dataset may lead to an incorrect patient clinical evaluation and diagnosis. At the same time, outliers may have an impact on patient's treatment plan and rehabilitation because they affect the normality. On the other hand, outliers are not related only with clinical data and patient's life, but also with financial and administrative data which may affect the health care expenditures [8].

In order to reduce the expenditures and the total cost of the health care services, many attempts have been done by national organizations to analyse raw data, aiming to extract useful conclusions regarding the pricing and cost by detecting the outliers. For example, United States seems to be an outlier on healthcare expenditures based on the current healthcare policy making strategy combined with price transparency. Also the Great Britain's National Health Service (NHS) assesses the outputs of the public sector using a two-stage Monte Carlo simulation of their data [9]. According to M. Cyganska [10], an examination of the patient's length of stay on a hospital is presented, where the usage of statistical metric to select the outliers was utilized producing some important results.

Healthcare sector is one of the most common in occurrence of fraud. In Healthcare domain, fraud can appeared in many schemes like 'Up-coding: billing for more expensive services or procedures than were actually provided or performed' or 'Charging for a not given treatment' or 'Performing medically unnecessary services solely to generate insurance payments' [11]. All these have a significant impact on the healthcare cost control which affects the quantity and the quality of the healthcare services. In order to reduce fraud, prevention and detection mechanisms can be established in health care sector. Robinson and Vigelette, proposed a system and method for preventing healthcare fraud using biometric signatures for authentication [12]. Regarding the fraud detection, a lot of scientific work has took place the last years presenting valuable applications and data usage against financial fraud [13]. Specifically, on the Healthcare domain, data mining technologies seems to be accepted for fraud detection giving some reliable results [14]. Other actions include the development of information systems to monitoring the billing process of healthcare services [15], the classification and the analysis of fraudulent behaviors [16], and the application of a Bayesian Co-clustering in Healthcare data to detect fraud [17]. At the same time, the Healthcare services in Greece consist of a universal health care system provided through national health insurance, and private health care. The Hellenic National Organization for the Provision of Health Services (EOPYY) is the largest public (national) health insurance organization in Greece which aims to cover the healthcare expenditures for Greek insured citizens. At the same time, EOPYY negotiates contracts and remunerates health professionals on the basis of a Health Benefits Regulation (Greek acronym EKPY) prescribing the benefits basket for the beneficiaries which include among others the medical treatment, diagnostic/laboratory/clinical tests, dental treatment, physiotherapy, occupational therapy, speech therapy, psychotherapy, medication, consumables, dietary supplements, medical devices, hospital treatment, supplementary healthcare (orthopaedics, eyeglasses, hearing aids, prosthetics etc), long-term care, obstetric care and IVF, healthcare abroad, and vaccination programs. The aim of this dissertation is to present the methodological approach of the Hellenic National Organization for the Provision of Health Services in data analysis to detect financial or medical fraud in claims.

1.1 Types of frauds

Healthcare fraud and abuse take many forms. Some of the most common types of frauds by providers are:

a) Billing for services that were not provided.
b) Duplicate submission of a claim for the same service.
c) Misrepresenting the service provided.
d) Charging for a more complex or expensive service than was actually provided.
e) Billing for a covered service when the service actually provided was not covered.

Figure 1.1 below show some common areas of frauds.

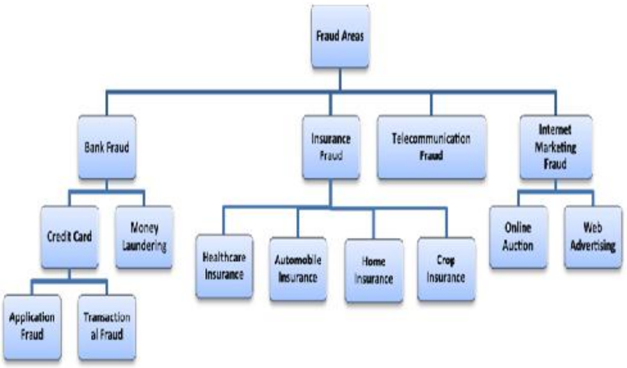


Figure 1.1: Most common areas of frauds [18].

1.2 Motivation

Provider Fraud is one of the biggest problems facing Medicare. According to the government, the total Medicare spending increased exponentially due to frauds in Medicare claims. Healthcare fraud is an organized crime which involves peers of providers, physicians, beneficiaries acting together to make fraud claims.

How Medicare can increase its Health care fraud detection capability, based on the claims provided by the healthcare with Beneficiaries' profiles, health status, Insurance details, diagnostics/procedures carried out on them.

Rigorous analysis of Medicare data has yielded many physicians who indulge in fraud. They adopt ways in which an ambiguous diagnosis code is used to adopt costliest procedures and drugs. Insurance companies are the most vulnerable institutions impacted due to these bad practices. Due to this reason, an insurance company increased their insurance premiums and as result healthcare is becoming costly matter day by day.

II. RELATED WORK

Ameyaa Biwalkar et al. [19] Proposed Study of Data Mining Techniques in the Healthcare Sector. These methods are actively being used in improving line of treatment and diagnosis, customer relationship management, management of healthcare resources and fraud and anomaly detection.

Sandip Vyas and Shilpa Serasiya [20] introducing a blockchain-based framework for enabling secure transactions and data exchange among various interacting agents in the insurance network. Blockchain is a distributed peer-to-peer technology that allows for the safe, immutable, and transparent validation of healthcare claims. Also, discuss how blockchain and smart contracts can be used together to improve organizational operations. In this review authors discuss about types of Fraud Detection in Insurance Claim System and its classification based on different machine learning methods. Also gives the future direction for Fraud Detection in Insurance Claim System.

R. Thomas and J. E. Judith [21] proposed a Hybrid Outlier Detection in Healthcare Datasets using DNN and One Class-SVM. Most of the data mining and machine learning tasks are aimed to detect the general properties of the dataset. Outlier detection plays an important role in various application domains such as malicious activity detection in software applications, fraud detection in financial transactions, intruder detection in communication systems etc. In this paper, the performance of two major outlier detection algorithms is analyzed for healthcare applications. A hybrid model is used for detecting outliers in healthcare systems by using one-class support vector machine model and the autoencoder model is also proposed.

R. A. Bauder et al. [22] Proposed A Medicare Fraud Detection Case Study. In real-world production applications, it is critical to establish a model's usefulness by validating it on completely new input data, and not just using the cross validation results on a single historical dataset. In this paper, we present results for both evaluation methods, to include performance comparisons. In order to provide meaningful comparative analyses between methods, we perform real-world fraud detection.

R. Bauder and T. Khoshgoftaar [23] proposed A Survey of Medicare Data Processing and Integration for Fraud Detection. One aspect contributing to increased costs in healthcare is waste and fraud. In particular, with the rapidly rising elderly population in the United States, programs like Medicare are subject to high losses due to fraud. Therefore, fraud detection approaches are critical in lessening these losses. Even so, many studies using Medicare data do not provide sufficient details regarding data processing and/or integration making it potentially more difficult to understand the experimental results and challenging to reproduce the experiments. In this paper, we present current research using Medicare data to detect fraud, focusing on data processing and/or integration, and assess any gaps in the provided data-related details.

R. A. Bauder and T. M. Khoshgoftaar [24] proposed a Probabilistic Programming Approach for Outlier Detection in Healthcare Claims. Out of the many possible factors for the rising cost of healthcare, claims fraud is a major contributor, but its impact can be lessened through effective fraud detection. We propose a general outlier detection model, based on Bayesian inference, using probabilistic programming. Our model provides probability distributions rather than just point values, as with most common outlier detection methods.

III. PROPOSED WORK

For the purpose of this research, we are considering Inpatient claims, Outpatient claims and Beneficiary details of each provider. Lets s see their details:

A) Inpatient Data

This data provides insights about the claims filed for those patients who are admitted in the hospitals. It also provides additional details like their admission and discharge dates and admit d diagnosis code.

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 beneficiary KYC details like health conditions, regioregion they belong to etc.

D) Train Data:

Data containing provider id's and whether they are fraud or not.

We are given with no details about the health care provider except for the provider id. But in inpatient and outpatient files we have provider id as feature. So we can leverage that and generate information about the provider based on the details of patients.

The General step of proposed methods is shown below:

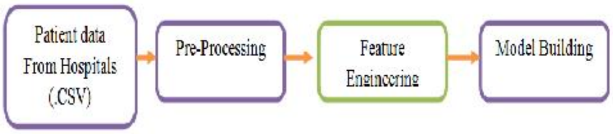


Figure 4.1: General architecture of proposed Model.

The raw data obtained from Kaggle was not clean enough to carry out Exploratory Data Analysis (EDA) or further Machine Learning (ML) building; hence Data wrangling was carried out on them.

4.2 Proposed Algorithm

- Read the dataset.
- Display the dataset like reading all columns and rows.
- Distribution of Fraud and Non-Fraud class.
- Handling Missing Values.
- Finding out Deductible amount.
- Converting Datetime columns to Timestamp.
- Some disease indicator column has listed two values “Y” and 0. The “Y” was assigned to 1 and then the converted to numeric.
- Missing values were dropped from the dataframe.
- The null values were replaced with 0.
- The missing values were replaced with 0.
- The Gender column values were 1 and 2. The entries with 2 were replaced with 0.
- The Potential Fraud column had two values “Yes” and “No”. “Yes” and “No” were replaced with 1 and 0.

After data wrangling and Exploratory Data analysis, Beneficiary dataset was merged with Inpatient and Outpatient datasets. The two newly formed datasets (Beneficiary + Inpatient) and (Beneficiary + Outpatient), now contain four different types of data; Numeric (not categorical), binary categorical, multiple categorical, categorical with unique values in each row.

Finally all four datasets are groups by data frames and are merged together, followed by merging with Train dataset which contains Providers’ information regarding Fraud or not.

The feature engineering on preprocessed data was carried out after splitting the dataset into train and test datasets:

4.3 Feature engineering

Patients wise Claims.

- Max number of days a person was admitted.
- adding a feature signifying whether the patient was inpatient or outpatient
- handling nan values generated due to merging inpatient and outpatient
- since no date is given in the problem statement we are taking last death as the latest day
- whether the person is alive or not
- age of the patient
- number of physician
- medical cases counts
- number of unique attending physicians per provider
- Total number of unique diagnosis.
- Replaced the outliers in each numeric columns assuming that the entries are not mistakenly entered.
- All the numerical columns were standardized.

The data was highly imbalanced with Potential Fraud cases (No): 90%, Potential Fraud cases (Yes): 10%.

Hence SMOTE (Synthetic Minority Oversampling Technique) was applied on the data, made the data balanced, and all the Models were rebuilt. With SMOTE transformed data, the performance of all models improved in terms of recall, with Logistic Regression highest among all.

- Henceforth, further model building was carried out on SMOTE transformed data.
- To improve the model performance, Best Features (i.e most important features) were sorted out and applied to Logistic Regression. During this process, there are many constant and quasi-constant features (columns), which needed to be dropped from the dataset. The dataset is cleaned again dropping the constant and quasi-constant features.
- After this, all machine learning models were again built on cleaned data and proposed algorithm performance was highest.
- To improve the run time and performance of the proposed model, PCA (Principal Component Analysis) was applied.

After pre-processing and Feature Engineering, machine learning models are builder, Evaluated and Optimizated. Machine Learning models were built to predict/classify the Providers as Potential Fraud or not. Different classification algorithms were built on train dataset and evaluated on test datasets.

There are different types of ML classification models available in scikit-learn; Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, Support Vector Classifier, Naive Bayes and few more. We choose to work with the few of them for my problem and considered roc auc score and recall to evaluate the model performance in all cases.

When ML models were built on the feature engineered processed data, all models’ (except Decision Tree classifier) roc auc scores were increased.

The Logistic Regression model performance was highest among all ML classifier models. The LR model performance also remained same when applied to without and with PCA transformed data. In both cases, the roc auc score and run time were almost same, however in PCA case; both recall and precision values were higher. Though the run time in present dataset is same, PCA based model will be faster when applied to large dataset. Taking this also into consideration, the LR model with PCA was finally saved for deployment.

In this work, a mediclaim fraud detection system is designed using predictive analytics, which detects fraudulent medical claims resulting from substantial monetary loss in healthcare systems. An analytical approach to detect medical claim insurance fraud is done using Logistic regression. The fraudulent claims are to be detected using the data from various sectors like insurance company, hospitals, pharmacy, and insurance policyholder. So a multi criteria decision support system is developed to predict if a claim is fraudulent or legitimate.

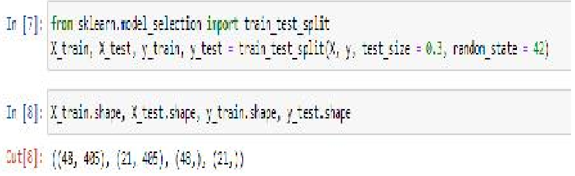
IV. RESULTS WORK

The thesis was built on 4 datasets: Beneficiary, Inpatient, Outpatient and Train. All are large datasets with rows and columns as follows:

Beneficiary: (138556, 25)
Inpatient:(40474, 30)
Outpatient: (517737, 27)
Train: (5410, 2).

To prepare the data for analysis and model building, data wrangling and EDA was carried out in all four datasets separately and finally all four are merged.

In step second, feature engineering is performed. Before applying the feature engineering, the dataframe are divided into features and targets. Figure below shows the code for splitting into train set and test set.



```
In [7]: from sklearn.model_selection import train_test_split
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 42)

In [8]: X_train.shape, X_test.shape, y_train.shape, y_test.shape

Out[8]: ((48, 405), (21, 405), (48,), (21,))
```

After splitting into train and test set, outliers are removed if any from the datasets. As there are some columns where it seems to have outliers, so replace them with appropriate percentile value.

Replace the outliers with 90th percentiles as it seems to be reasonable. After applying the outliers, standard scale the selected numerical data. Two main works are done for feature engineering like Outliers removal and standard scaler conversion. Finally save the database for model building.

The third and final step is the model building that is training and testing the model for final classification. Again read the final database after feature engineering. Again create X_train and Y_train. Then apply the basic machine learning classifiers on the train data.

The final results after all operations are mentioned below in table.

Table 5.1: Final Results.

Algorithm Name	**ROC AUC (%)**
Decision Tree	**78.15**
Random Forest	**84.58**
Support Vector Machine	**82.06**
Gradient Boosting Classifier	**85.13**
Proposed Model	**85.52**

REFERENCE

[1] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: Promise and potential,” Health Information Science and Systems, vol. 2, no. 1, p. 3, 2014.

[2] U. Srinivasan and B. Arunasalam, “Leveraging big data analytics to reduce healthcare costs,” IT professional, vol. 15, no. 6, pp. 21–28, 2013.

[3] L. Morris, “Combating fraud in health care: An essential component of any cost containment strategy,” Health Affairs, vol. 28, no. 5, pp. 1351–1356, 2009.

[4] Y. Shi, C. Sun, Q. Li, L. Cui, H. Yu, and C. Miao, “A fraud resilient medical insurance claim system,” in Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI16, p. 43934394, AAAI-Press, 2016.

[5] P. S. Mathew and A. S. Pillai, “Big data solutions in healthcare: Problems and perspectives,” in 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1–6, March 2015.

[6] R. A. Bauder and T. M. Khoshgoftaar, “Medicare fraud detection using random forest with class imbalanced big data,” 2018 IEEE International Conference on Information Reuse and Integration (IRI), pp. 80–87, 2018.

[7] V. Hodge, Outlier Detection in Big Data, pp. 1762–1771. In: J. Wang and J. Wang, Encyclopedias of Business Analytics and Optimization", PA: IGI Global, 04 2014.

[8] T. Gross and M. J. Laugesen, "The Price of Health Care: Why Is the United States an Outlier?" Journal of Health Politics, Policy and Law, vol. 43, pp. 771–791, 10 2018.

[9] R. Salehnejad, M. Ali, and N. Proudlove, "Combining regression trees and panel regression for exploring and testing the impact of complementary management practices on short-notice elective operation cancellation rates," Health Systems, vol. 0, no. 0, pp. 1–19, 2019.

[10] M. Cyganska, "The impact factors on the hospital high length of stay outliers," Procedia Economics and Finance, vol. 39, pp. 251–255, 12-2016.

[11] S. Peck and L. McKenna, "Fraud in healthcare: A worldwide concern," Health Managements, vol. 17, pp. 124–126, 2017.

[12] R. Robinson and G. Vigelette, 2013. U.S. Patent Application No.13/444,470.

[13] E. Ngai, Y. Hu, Y. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," Decision Support Systems, vol. 50, pp. 559–569, 02 2011.

[14] W.-S. Yang and S.-Y. Hwang, "A process-mining framework for the detection of healthcare fraud and abuse," Expert Systems with Applications, vol. 31, pp. 56–68, 07 2006.

[15] R. Robinson and G. Vigelette, "Health care billing monitor system for detecting health care provider fraud," 2014. U.S. Patent No. 6,826,536. 30 Nov.

[16] J. Li, K.-Y. Huang, J. Jin, and J. Shi, "A survey on statistical methods for health care fraud detection," Health care management science, vol. 11, pp. 275–87, 10 2008.

[17] T. Ekin, "Application of Bayesian methods in detection of healthcare fraud," Chemical Engineering Transactions, vol. 33, 01 2013.

[18] Aisha Abdallah, Mohd Aizaini Maarof, Anazida Zainal, "Fraud detection system: A survey", Journal of Network and Computer Applications 68 (2016) 90–113.

[19] A. Biwalkar, R. Gupta and S. Dharadhar, "An Empirical Study of Data Mining Techniques in the Healthcare Sector," 2021 2nd International Conference for Emerging Technology (INCET), 2021, pp. 1-8, doi: 10.1109/INCET51464.2021.9456157.

[20] S. Vyas and S. Serasiya, "Fraud Detection in Insurance Claim System: A Review," 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), 2022, pp. 922-927, doi: 10.1109/ICAIS53314.2022.9742984.

[21] R. Thomas and J. E. Judith, "Hybrid Outlier Detection in Healthcare Datasets using DNN and One Class-SVM," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1293-1298, doi: 10.1109/ICECA49313.2020.9297401.

[22] R. A. Bauder, M. Herland and T. M. Khoshgoftaar, "Evaluating Model Predictive Performance: A Medicare Fraud Detection Case Study," 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), 2019, pp. 9-14, doi: 10.1109/IRI.2019.00016.

[23] R. Bauder and T. Khoshgoftaar, "A Survey of Medicare Data Processing and Integration for Fraud Detection," 2018 IEEE International Conference on Information Reuse and Integration (IRI), 2018, pp. 9-14, doi: 10.1109/IRI.2018.00010.

[24] R. A. Bauder and T. M. Khoshgoftaar, "A Probabilistic Programming Approach for Outlier Detection in Healthcare Claims," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016, pp. 347-354, doi: 10.1109/ICMLA.2016.0063.