Supervised Learning Based Medicare Hospital Fraud Detection

Prasanna Shivva, Kalpana Kayithi, Sneha Badayyagari

Department of Computer Science and Engineering
Sree Dattha Group of Institutions, Hyderabad, Telangana, India.

ABSTRACT
With the overall increase in the elderly population come additional, necessary medical needs and costs. Medicare is a U.S. healthcare program that provides insurance, primarily to individuals 65 years or older, to offload some of the financial burden associated with medical care. Even so, healthcare costs are high and continue to increase. Fraud is a major contributor to these inflating healthcare expenses. Our paper provides a comprehensive study leveraging machine learning methods to detect fraudulent Medicare providers. We use publicly available Medicare data and provider exclusions for fraud labels to build and assess three different learners. In order to lessen the impact of class imbalance, given so few actual fraud labels, we employ Logistic Regression creating two class distributions. Our results show that the other algorithms have poor performance compared with Logistic Regression. Learners have the best fraud detection performance, particularly for the 80:20 class distributions with average AUC scores, respectively, and low false negative rates. We successfully demonstrate the efficacy of employing machine learning Models to detect Medicare fraud.

Keywords: Medicare hospitals, fraud detection, supervised learning.

1. INTRODUCTION
Health insurers receive millions of claims per year. Given that information asymmetries between the principal (insurer) and the agents (health care providers and the insured) can lead to moral hazard, insurance companies face the choice of either paying out insurance claims immediately without any adjustments or reviewing claims that are suspicious. The most common method for undertaking the latter involves manually auditing claims data, which is a time-consuming and expensive process. Machine learning models can greatly cut auditing costs by automatically screening incoming claims and flagging up those that are deemed to be suspicious – i.e., potentially incorrect – for subsequent manual auditing.

Insurance fraud is a widespread and high-priced problem for each policyholder and insurance businesses in all sectors of the coverage industry [1]. India is one of the quickest developing economies in the international, has a burgeoning middle class, and has witnessed a giant upward push within the demand for medical insurance products [2]. Over the last 10 years, the medical health insurance industry has grown at a capital annual compounded boom rate of round 20%. But, with the exponential growth inside the industry, there has additionally been an extended prevalence of frauds within the us. Health insurance fraud contains a huge range of illicit practices and unlawful acts concerning intentional deception or misrepresentation. Statistics mining has a extraordinary effect in enhancing healthcare fraud detection system. Statistics mining has been implemented to fraud detection in both the way i.e., Supervised and non-supervised way. Information mining strategies and its software for fraud detection in fitness care zone is defined beneath. In latest years, systems for processing digital claims were increasingly carried out to mechanically perform audits and reviews of claims information. These systems are designed for figuring out regions requiring unique interest together with faulty or incomplete data enter, duplicate claims, and medically non-blankedeted services [3]. Even though these structures may be used to locate sure varieties of fraud, their fraud detection
competencies are typically restrained for the reason that detection particularly is predicated on pre-defined easy guidelines special via domain professionals.

2. LITERATURE SURVEY

Herland et. al [4] employed an approach to predict a physician’s expected specialty based on the type and number of procedures performed. From this approach, they generate a baseline model, comparing Logistic Regression and Multinomial Naive Bayes, in order to test and assess several new approaches to improve the detection of U.S. Medicare Part B provider fraud. These results indicate that this proposed improvement strategies (specialty grouping, class removal, and class isolation), applied to different medical specialties, have mixed results over the selected Logistic Regression baseline model’s fraud detection performance. Through this work, they demonstrate that improvements to current detection methods can be effective in identifying potential fraud.

Hancock et. al [5] conducted experiments with three Big Data Medicare Insurance Claims datasets. The experiments are exercises in Medicare fraud detection. They show that for each dataset, they obtain better performance from LightGBM and CatBoost classifiers with tuned hyperparameters. Since some features of the data, they are working with are high cardinality categorical features, they have an opportunity to try different encoding techniques in these experiments. They find that across the different encoding techniques, hyperparameter tuning provides an improvement in the performance of both LightGBM and CatBoost.

Bauder et. al [6] focused on the detection of Medicare Part B provider fraud which involves fraudulent activities, such as patient abuse or neglect and billing for services not rendered, perpetrated by providers and other entities who have been excluded from participating in Federal healthcare programs. Based on the performance and significance testing results, they posit that retaining more of the majority class information leads to better Medicare Part B fraud detection performance over the balanced datasets across the majority of learners.

Herland et. al [7] focused on the detection of Medicare fraud using the following CMS datasets: (1) Medicare Provider Utilization and Payment Data: Physician and Other Supplier (Part B), (2) Medicare Provider Utilization and Payment Data: Part D Prescriber (Part D), and (3) Medicare Provider Utilization and Payment Data: Referring Durable Medical Equipment, Prosthetics, Orthotics and Supplies (DMEPOS). Additionally, they create a fourth dataset which is a combination of the three primary datasets.

Arunkumar et. al [8] provides an extensive study of detecting fraudulent claims in healthcare insurance by leveraging machine learning algorithms. By using the publicly available medicare dataset, they are able to classify as fraud and non-fraud providers. Moreover, synthetically minority oversampling technique is used to avoid the class imbalance problem. Furthermore, a hybrid approach is used which is based on clustering and classification. Additionally, they have used other machine learning algorithms to check the efficiency of the best-suited algorithm.

Chen et. al [9] proposed VAERM coupled with active learning strategy can assist healthcare industry experts to conduct cost-effective fraud investigation. Finally, they propose an online model updating method to reduce the computation and memory requirement while preserving the predictive performance. The proposed framework is tested in a real-world dataset and it empirically outperforms the state-of-the-art methods in both automatic fraud detection and fraud investigation tasks.

Yao et. al [10] proposed the Bagging algorithm based on the weighted threshold method named WTBagging and made ten model combinations using Bagging and WTBagging algorithms. The data
are cleaned and sampled to construct three datasets with different class distributions. The 5-fold cross-validation process was applied to the model training and repeated ten times, and the F1 value was the performance metric to evaluate the model combination. The results show that the model combinations of the WTBagging achieved the highest F1 values under all datasets.

3. PROPOSED SYSTEM

Dataset description

57: Columns: PotentialFraud, BeneID, ClaimID, ClaimStartDt, ClaimEndDt, InscClaimAmtReimbursed, AttendingPhysician, OperatingPhysician, OtherPhysician, ClmDiagnosisCode_1, ClmDiagnosisCode_2, ClmDiagnosisCode_3, ClmDiagnosisCode_4, ClmDiagnosisCode_5, ClmDiagnosisCode_6, ClmDiagnosisCode_7, ClmDiagnosisCode_8, ClmDiagnosisCode_9, ClmDiagnosisCode_10, ClmProcedureCode_1, ClmProcedureCode_2, ClmProcedureCode_3, ClmProcedureCode_4, ClmProcedureCode_5, ClmProcedureCode_6, DeductibleAmtPaid, ClmAdmitDiagnosisCode, AdmissionDt, DischargeDt, DiagnosisGroupCode, AdmitForDays, DOB, DOD, Gender, Race, RenalDiseaseIndicator, State, County, NoOfMonths_PartACov, NoOfMonths_PartBCov, ChronicCond_Alzheimer, ChronicCond_Heartfailure, ChronicCond_KidneyDisease, ChronicCond_Cancer, ChronicCond_ObstrPulmonary, ChronicCond_Depression, ChronicCond_Diabetes, ChronicCond_IschemicHeart, ChronicCond_Osteoporosis, ChronicCond_rheumatoidarthritis, ChronicCond_stroke, IPAnnualReimbursementAmt, IPAnnualDeductibleAmt, OPAnnualReimbursementAmt, OPAnnualDeductibleAmt, Age, WhetherDead

558212-Rows

Data Preprocessing in Machine learning

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task.

Why do we need Data Pre-processing?

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for
cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

**Splitting the Dataset into the Training set and Test set**

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model.

Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models.

If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

![Dataset Diagram](image)

**Training Set**: A subset of dataset to train the machine learning model, and we already know the output.

**Test set**: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

**Support Vector Machine Algorithm**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:
Applications

- Face recognition
- Weather prediction
- Medical diagnosis
- Spam detection
- Age/gender identification
- Language identification
- Sentimental analysis
- Authorship identification
- News classification

Advantages of proposed system

- SVM works relatively well when there is a clear margin of separation between classes.
- SVM is more effective in high dimensional spaces.
- SVM is effective in cases where the number of dimensions is greater than the number of samples.
- SVM is relatively memory efficient.

4. RESULTS AND DISCUSSION

Module implementation

- Medicare Dataset
- Preprocessing
- SVM Training
- Performance estimation
- Trained features
- Test data
- SVM Testing
- Potential Fraud

Sample Training Data
List of attributes

```
PARAMS

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiseaseStatus</td>
<td>String</td>
<td>Disease status</td>
</tr>
<tr>
<td>ProcedureBeforeDischarge</td>
<td>String</td>
<td>Procedure performed before discharge</td>
</tr>
<tr>
<td>DiseaseDiagnosisCode</td>
<td>String</td>
<td>Disease diagnosis code</td>
</tr>
<tr>
<td>AdmissionDiagnosisCode</td>
<td>String</td>
<td>Admission diagnosis code</td>
</tr>
<tr>
<td>AnnualDeductibleAmount</td>
<td>Float</td>
<td>Annual deductible amount</td>
</tr>
<tr>
<td>AnnualReimbursementAmount</td>
<td>Float</td>
<td>Annual reimbursement amount</td>
</tr>
<tr>
<td>Age</td>
<td>Float</td>
<td>Patient age</td>
</tr>
<tr>
<td>WhetherDead</td>
<td>bool</td>
<td>Whether the patient is dead</td>
</tr>
</tbody>
</table>
```

Exploratory Data Analytics
Training and testing data

X_train : (3787, 156)
y_train : (3787,)
X_test : (1623, 156)
y_test : (1623,)

Logistic regression report

```r
print(classification_report(y_test, y_predict))
```

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.97</td>
<td>0.92</td>
<td>1471</td>
</tr>
<tr>
<td>1</td>
<td>0.49</td>
<td>0.74</td>
<td>152</td>
</tr>
<tr>
<td>accuracy</td>
<td>0.90</td>
<td>1623</td>
<td></td>
</tr>
<tr>
<td>macro avg</td>
<td>0.73</td>
<td>0.83</td>
<td>1623</td>
</tr>
<tr>
<td>weighted avg</td>
<td>0.93</td>
<td>0.90</td>
<td>1623</td>
</tr>
</tbody>
</table>

SVM report

```r
print(classification_report(y_test, y_pred))
```

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.94</td>
<td>0.99</td>
<td>1471</td>
</tr>
<tr>
<td>1</td>
<td>0.83</td>
<td>0.83</td>
<td>152</td>
</tr>
<tr>
<td>accuracy</td>
<td>0.94</td>
<td>1623</td>
<td></td>
</tr>
<tr>
<td>macro avg</td>
<td>0.89</td>
<td>0.69</td>
<td>1623</td>
</tr>
<tr>
<td>weighted avg</td>
<td>0.93</td>
<td>0.93</td>
<td>1623</td>
</tr>
</tbody>
</table>

Prediction on test data

<table>
<thead>
<tr>
<th>Provider</th>
<th>PotentialFraud</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 PRV51002</td>
<td>Y</td>
</tr>
<tr>
<td>1 PRV51006</td>
<td>Y</td>
</tr>
<tr>
<td>2 PRV51009</td>
<td>Y</td>
</tr>
<tr>
<td>3 PRV51010</td>
<td>Y</td>
</tr>
<tr>
<td>4 PRV51018</td>
<td>Y</td>
</tr>
</tbody>
</table>

5. CONCLUSION AND FUTURE SCOPE

This work aimed at developing a novel fraud detection model for insurance claims processing based on genetic support vector machines, which hybridizes and draws on the strengths support vector machines. SVMs have been considered preferable to other classification techniques due to several advantages. With other notable advantages, it has a nonlinear dividing hyper plane, which prevails
over the discrimination within the dataset. The generalization ability of any newly arrived data for classification was considered over other classification techniques.

Future Scope

The proposed methodology provides the information that Random Forest performs better than Sequential CNN. The drawback of this methodology is that anyone would expect Sequential CNN can outperform any of the conventional ML methodologies, but it is not happening here. It may happen because the dataset is not enough to train and identify the hidden patterns to predict the future or upcoming data and the initialization of weights was very random that might affect the training process. It can be further improved in two ways. The first way is to tune the hyperparameters through optimization, and the second method is to apply the transfer learning methodology so that the performance of the proposed methodology is improved to detect the fraud transaction through Medicare in the healthcare sector.

REFERENCES