National College of Ireland

Configuration Manual : Detection Healthcare Frauds in Insurance Industry by Healthcare Service Providers

MSc Research Project
Data Analytics

Vinay Reddy Pannala
Student ID: X20138261

School of Computing
National College of Ireland

Supervisor: Bharathi Chakravarthi

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Vinay Reddy Pannala
Student ID:	X20138261
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Bharathi Chakravarthi
Submission Due Date:	31/01/2022
Project Title:	Configuration Manual : Detection Healthcare Frauds in Insurance Industry by Healthcare Service Providers
Word Count:	1488
Page Count:	8

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	30th JAN 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	Q
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	Q
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	Q

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual : Detection Healthcare Frauds in Insurance Industry by Healthcare Service Providers

Vinay Reddy Pannala
X20138261

The health care frauds have been committed by the customers which are intentional, the lot of researches have been done in identifying the fraud claims done by the patients in health care industry and in the field of auto insurance claims, but there are few studies conducted on the healthcare service providers which mainly involves the hospital groups, physicians and peers, who comes together to involve in this kind of fraud. Thus, this project will try to identify the potential fraud claims which help the organization in terms of loss of revenue using machine learning the objectives begin with the literature review in the field of,fraud claims in insurance industry, the second objective was to collect the data set, which was collected from the Kaggle which is publicly available. The next objective begins with the cleaning of data set, such as handling missing values, nan values and identifying the important factors from the data set to make it ready for the model implementation. The final objective was to implementation of various classification models with the cleaned data set and evaluation.

1 System Configuration and Data Exploration

The data analysis phases the relation between the features were evaluated to find out the best suitable features from all the data sets which helped the models to perform the better and some of the graphs were discussed below. The analysis also includes the categorical variables conversion from categorical to numeric so that smooth implementation of the models takes place.

- Objective 1:Extracted data set from kaggle [1] , which is available as open source publicly.
- Objective 2: Imported the data to the jupyter for futher analysis.
- Objective 3: Data pre processing and handling missing values.
- Objective 4: Data exploration.
- Objective 5: Machine learning Model implementation
- Objective 6: Neural network model implementation

These are the important libraries which were imported to execute the the required data analysis and the feature importance for executing the model implementation.

[1]https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis

Device specifications

HP Laptop 15s-gr0xxx

Device name	LAPTOP-ELGSTIS3
Processor	AMD Ryzen 5 3450U with Radeon Vega Mobile Gfx 2.10 GHz
Installed RAM	8.00 GB (5.94 GB usable)
Device ID	B45D80E3-FBA9-46C0-A38F-43EDEED802C2
Product ID	00327-35912-89602-AAOEM
System type	64-bit operating system, x64-based processor
Pen and touch	No pen or touch input is available for this display

Figure 1: system configuration

```
import numpy as np # required libraries for linear algebra
import pandas as pd # for reading the CSV file and data processing
import os
```

```
import scipy as sc
import seaborn as sns
import matplotlib.pyplot as plt
import pandas_profiling as profile   # To check data distributions and correlations
import warnings                      # for supressing a warning when importing large files
warnings.filterwarnings("ignore")

from sklearn.preprocessing import StandardScaler,MinMaxScaler
from sklearn.model_selection import train_test_split
import pickle
import matplotlib.pyplot as plt
from scipy import stats
import tensorflow as tf
from pylab import rcParams
from keras.models import Model, load_model
from keras.layers import Input, Dense
from keras.callbacks import ModelCheckpoint, TensorBoard
from keras import regularizers
%matplotlib inline
sns.set(style='whitegrid', palette='muted', font_scale=1.5)
rcParams['figure.figsize'] = 14, 8
RANDOM_SEED = 42

LABELS = ["Normal", "Fraud"]
```

Figure 2: Required Libraries

2 Data set Structure

From the figure 3, The data set structure was identified, for each of the train data set and outpatient data set, inpatient data set, beneficiary data set.

```
Shape of Train data : (5410, 2)
Shape of Train_Beneficiarydata data : (138556, 27)
Shape of Train_Inpatientdata data : (40474, 31)
Shape of Train_Outpatientdata data : (517737, 27)
Shape of Test data : (1353, 1)
Shape of Test_Beneficiarydata data : (63968, 27)
Shape of Test_Inpatientdata data : (9551, 31)
Shape of Test_Outpatientdata data : (125841, 27)
```

Figure 3: Dataset Structure

From the figure 4, The data set were merged to get the better results, the patients data set merged with the fraudulent providers details with provider as the joining key for inner join.

```
# Merging patient data with fradulent providers details data with "Provider" as joining key for inner join

Train_ProviderWithPatientDetailsdata=pd.merge(Train,Train_AllPatientDetailsdata,on='Provider')

Test_ProviderWithPatientDetailsdata=pd.merge(Test,Test_AllPatientDetailsdata,on='Provider')
```

Figure 4: Merging using inner join

From the figure 5,some feature in the data set were modified such that the categorical variables were converted into numeric for the execution of the models without interruption, the some features were used dummies for execution of models. these conversions helped the models in prediction of healthcare fraud in insurance industry.

Type Conversion

```
## Lets Convert types of gender and race to categorical.

Train_category_removed.Gender=Train_category_removed.Gender.astype('category')
Test_category_removed.Gender=Test_category_removed.Gender.astype('category')

Train_category_removed.Race=Train_category_removed.Race.astype('category')
Test_category_removed.Race=Test_category_removed.Race.astype('category')
```

Dummification

```
# Lets create dummies for categorrical columns.

Train_category_removed=pd.get_dummies(Train_category_removed,columns=['Gender','Race'],drop_first=True)
Test_category_removed=pd.get_dummies(Test_category_removed,columns=['Gender','Race'],drop_first=True)
```

Figure 5: Categorical and dummies for categorical conversion

3 Feature Analysis

From the figure 6,it is observed that the maximum number of claims are from the state number 5, which were converted to numeric while implementing the machine learning models. This analysis can further help the research model, by identifying the states in which most umber of claims are being filed and can focus further more on them can identify the fraud claims. Some of the machine learning models can not allow the variables names in string format, hence the state names were converted to the numeric values such that the models can execute without interruption.

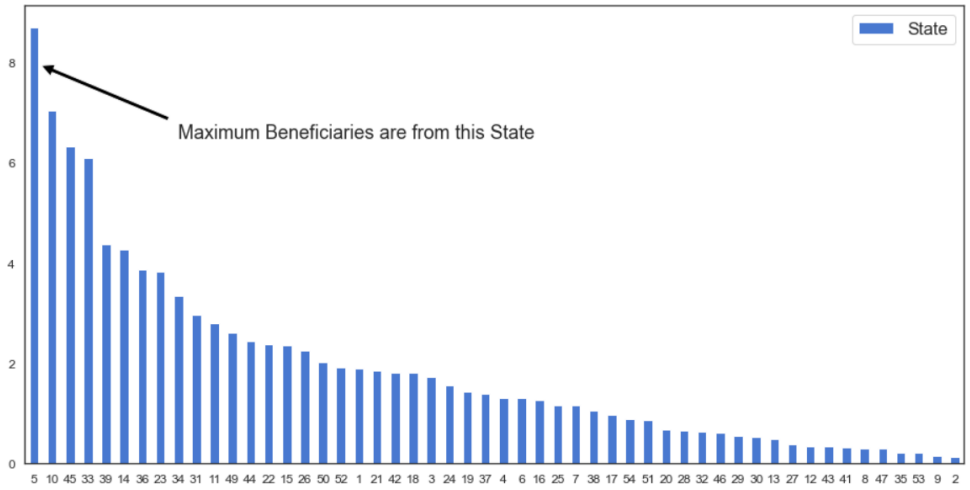

Figure 6: Data Analysis

4 Claim Procedures Diagnosis

From the figure 7, it is observed that the top ten procedures provided in healthcare fraud, the potential frauds number is high for the particular procedure and less for the particular procedure , these can help significantly in finding the frauds when claiming frequently for the same procedure and which has been costlier procedure, which can cost the organization and can cause huge loss , which can be minimised by the adoption these machine learning algorithms to help mitigate the risk and help grow the profits for the organization.

5 Count of fraud claims

From the figure 8, after the both the train and test data analysed together to find the potential fraud distribution in aggregated claim transitional data was described such as fraud claims were almost more than half of the non-frauds claims. The count of the non-fraud claims are around the 350000 then fraud claims were about 200000.

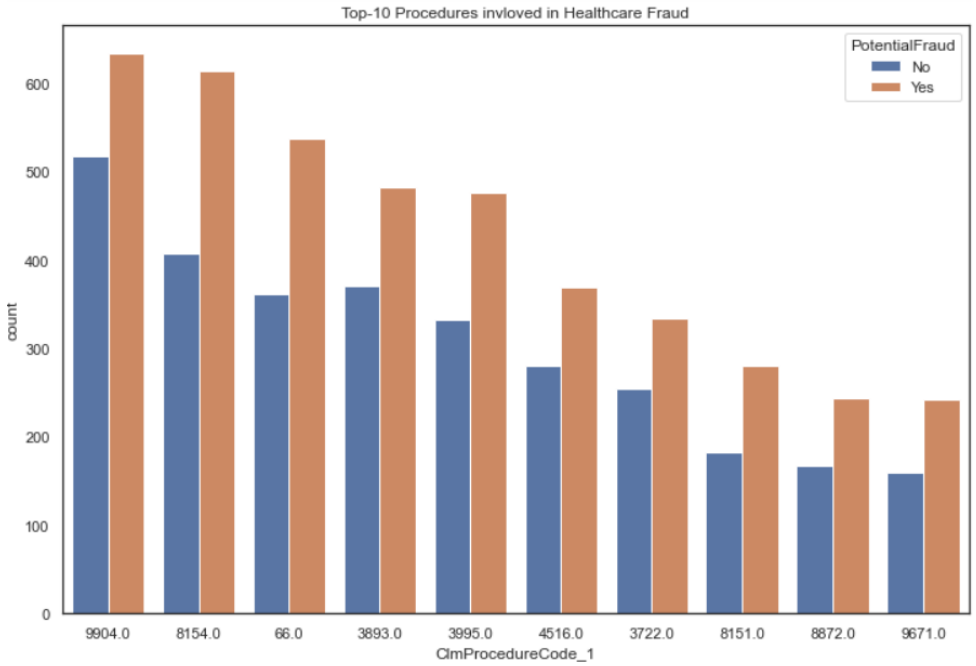

Figure 7: Data Analysis

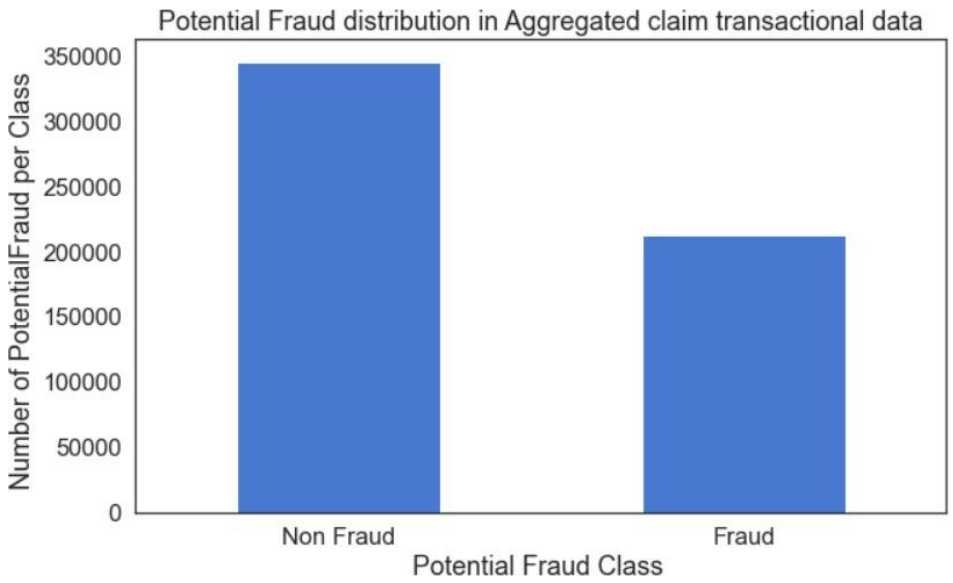

Figure 8: Potential fraud claims

```
Confusion Matrix Train :
 [[ 270   84]
 [ 210 3223]]
Confusion Matrix Val:
 [[ 103   49]
 [  93 1378]]
Accuracy Train:  0.922365988909427
Accuracy Val:  0.9125077017868145
Sensitivity Train :  0.7627118644067796
Sensitivity Val:  0.6776315789473685
Specificity Train:  0.9388290125254879
Specificity Val:  0.9367777022433719
Kappa Value : 0.5438304105142315
AUC         : 0.8072046405953702
F1-Score Train  :  0.6474820143884892
F1-Score Val  :  0.5919540229885056
```

Figure 9: Logistic Regression

6 Model Implementation and Evaluation

From the figure 9,As the research topic was on classification problem, the logistic regression model was one of the efficient model in identifying the classification problems, The data set was divided into the train and test data set and then the model was implemented with the proper data modelling and important features were finalised before implementing the model.The result of the model were discussed, the train and test data was evaluated by using the confusion matrix and the model was successful in identifying the frauds claims about91% in both train and test data

The research problem was to identify the frauds in healthcare sector, as it is the classification problem several classification models were implemented, the most used machine learning model random forest for both classification problems was applied and the result was evaluated by confusion matrix and accuracy. The accuracy of the model was around88% for both the train and test data set, which was less than the logistic model.

At the end the models were implemented together to find out the best performing model for both the data sets and the logistic regression model was outperformed all the models. The results were discussed below. The ADA boost algorithm was significant in predicting the classification binary problem such as fraud or not, and it was the first real time boosting machine models which is helpful in all the fields. For ADA boosting algorithm the f1 score was around 53%. The decision tree classifier algorithm was efficient in identifying the classification model, where it contains binary tree and non-binary tree node structures, the decision starts from the root node which helped to achieve the goal,for the decision tree classifier the f1 score was around 43%.

From the neural network models the auto encoder model was implemented with the different ephocs in multiple implementations, the result was improved with the variation of ephocs count, the auto encoders model was implemented to verify the performance of neural network models compared to machine learning models, and model outperformed

```
Confusion Matrix Train :
 [[ 319   35]
 [ 389 3044]]
Confusion Matrix Test:
 [[ 124   28]
 [ 182 1289]]
Accuracy Train :  0.8880380248217586
Accuracy Test :  0.8706099815157117
Sensitivity :  0.8157894736842105
Specificity :  0.876274643099932
Kappa Value : 0.47589611108548224
AUC         : 0.8460320583920713
F1-Score Train 0.60075329566855
F1-Score Validation :  0.5414847161572053
```

Figure 10: Random Forest

```
{'ada': 0.5309090909090909,
 'dtc': 0.4300341296928327,
 'gbc': 0.5849802371541503,
 'lr': 0.6045340050377834,
 'svm1': 0.5,
 'svm2': 0.5551020408163265,
 'svm3': 0.4100000000000003,
 'xgb': 0.4800000000000004}
```

Figure 11: Models Implemented

over all the machine learning models with the good accuracy in predicting the frauds claims. With 2 different hidden layers the model outperformed. The results were discussed below.The auto encoders model helped the insurance industries in identifying the fraud claims with good accuracy by analysing the outpatient, inpatient and beneficiaries data in right manner, this machine learning models which can help the insurance organization reduce the fraud claim risks by pre identifying the fraud, this research was identified the major threats the insurance organizations were facing.

```
Confusion Matrix Val:
 [[ 61  44]
 [ 54 923]]
Accuracy Val:  0.9094269870609981
Sensitivity Val:  0.580952380952381
Specificity Val:  0.9447287615148413
Kappa Value : 0.5042498480527373
AUC         : 0.7628405712336113
F1-Score Val  :  0.5545454545454545
```

Figure 12: Auto Encoders

7 References

https://www.python.org/

https://jupyter.org/