Report On

Medical Insurance Fraud Detection

Submitted in partial fulfillment of the requirements of the Course project in Semester VII of Fourth Year Computer Engineering

By
Sanika Patil (Roll No. 64)
Hrushikesh Shetty (Roll No. 66)
Rishabh Tripathi (Roll No. 68)

Supervisor
Mrs. Sneha Mhatre

University of Mumbai

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

(2023-24)

Vidyavardhini's College of Engineering & Technology
Department of Computer Engineering

CERTIFICATE

This is to certify that the project entitled “Medical Insurance Fraud Detection” is a bonafide work of " Sanika Patil (Roll No. 64) Hrushikesh Shetty (Roll No. 66) Rishabh Tripathi (Roll No. 68) submitted to the University of Mumbai in partial fulfillment of the requirement for the Course project in semester VII of Fourth Year Computer Engineering.

.

Supervisor

Mrs. Sneha Mhatre

Dr. Megha Trivedi
Head of Department

Abstract

Medicare fraud is a pervasive problem that poses a considerable threat to the integrity and financial sustainability of the Medicare program in the United States. This abstract presents an innovative machine learning solution that combines anomaly analysis and geo-demographic metrics to predict potential instances of Medicare fraud. By leveraging advanced anomaly detection techniques to identify irregular patterns within Medicare claims data and incorporating geo-demographic insights to contextualize claims, our model offers a proactive approach to fraud detection. The project's objective is to create a model that can accurately differentiate between legitimate and fraudulent Medicare claims, leading to substantial cost savings, more efficient resource allocation, improved healthcare services for legitimate beneficiaries, and preventive measures to deter potential fraudsters. This research contributes to the safeguarding of Medicare's financial resources and the provision of high-quality healthcare services to those who rely on the program.

Contents **Pg. No**

Problem Statement

The objective of this project is to develop a pioneering machine learning model that leverages anomaly analysis and geo-demographic metrics to predict Medicare fraud. Medicare fraud is a significant challenge, with fraudulent claims costing the program billions of dollars annually and undermining the healthcare services provided to legitimate beneficiaries. To address this issue, the project aims to create a predictive model that can effectively differentiate between genuine and fraudulent Medicare claims. In order to minimize fraudulent activities, the Centers for Medicare and Medicaid Services (CMS) released a number of “Big Data” datasets for different parts of the Medicare program. This model will incorporate cutting-edge anomaly detection techniques to identify irregular patterns in claims data and integrate geo-demographic insights to assess the appropriateness of claims within specific geographic and demographic contexts, ultimately enhancing the accuracy of fraud detection and safeguarding the program's financial resources while ensuring the delivery of legitimate healthcare services to beneficiaries.

Module Description

Data:

The dataset for medical insurance fraud detection consists of three main components: Beneficiary Data, Inpatient Data, and Outpatient Data. The Beneficiary Data includes information such as unique beneficiary identifiers, dates of birth and, if applicable, dates of death, gender, race, and indicators for various chronic conditions. It also provides details on the duration of Part A and Part B coverage and annual reimbursement and deductible amounts for inpatient and outpatient services. Inpatient Data contains records of inpatient medical claims, including claim identifiers, service providers, admission and discharge dates, attending and operating physicians, and reimbursement amounts. The Outpatient Data similarly records outpatient claims, including diagnosis codes. These datasets collectively offer a comprehensive view of the medical insurance claims landscape, enabling the detection of potential fraudulent activities and patterns within the healthcare insurance domain.

Data Preprocessing:

1) Feature Engineering:

Medicare fraud is categorized as organized crime which involves peers working together to create fraud transactions of claims.Adding features from grouping them helped in improving accuracy of prediction and fraud pattern recognition. Grouping and aggregating numeric features to provider level helped in detecting behavior of their transactions overall.

2) Logistic Regression Classifier:

Features derived from above step are trained using logistic regression and evaluated.My decision of choosing LR is to check linear behavior between dependent and independent variables.Also Logistic model adds explicability to the predictions. Performance of the LR model showcases the linearity between variables.

3) Random Forest Classifier:

One of the benefits of Random forest which excites most is the power of handling large data set with higher dimensionality. It can handle thousands of input variables and identify most significant variables. Further, the model outputs the Importance of variables, which can be a very handy feature.It also checks for non linearity between variables.

4) Auto encoders:

Autoencoders are neural networks that aim to copy their inputs to their outputs. They work by compressing the input into a latent-space representation, and then reconstructing the output from this representation. My aim for the project is to train non fraud data using autoeencoder and reconstructing it back.While reconstructing Fraud data it will create an error,called as reconstruction error.Based on the threshold setting of reconstruction errors, we can easily predict Fraudulent behavior of healthcare providers.

Software and Hardware Used

Languages:

- **Python**: The project is crafted in Python language, from model training to prediction.

Libraries:

- pandas: Data manipulation and analysis in a tabular format.
- numpy: Scientific computing for numerical operations.
- scipy: Scientific and technical computing library for advanced mathematical functions.
- seaborn: Data visualization library for creating attractive statistical graphics.
- matplotlib.pyplot: Data visualization library for creating various types of plots and charts.
- pandas_profiling: A tool for generating data profiling reports to examine data distributions and correlations.
- scikit-learn (sklearn): Machine learning library for data preprocessing and model training.
- pickle: Serialization and deserialization of Python objects.
- tensorflow: Machine learning framework for building and training deep learning models.
- keras: High-level neural networks API running on top of TensorFlow for building and training neural network models.
- rcParams: Setting parameters for Matplotlib figures and plots.
- LABELS: A list of labels used to classify data as "Normal" or "Fraud."

Hardware Required:

Recommended:

- 16gb RAM

Minimum:

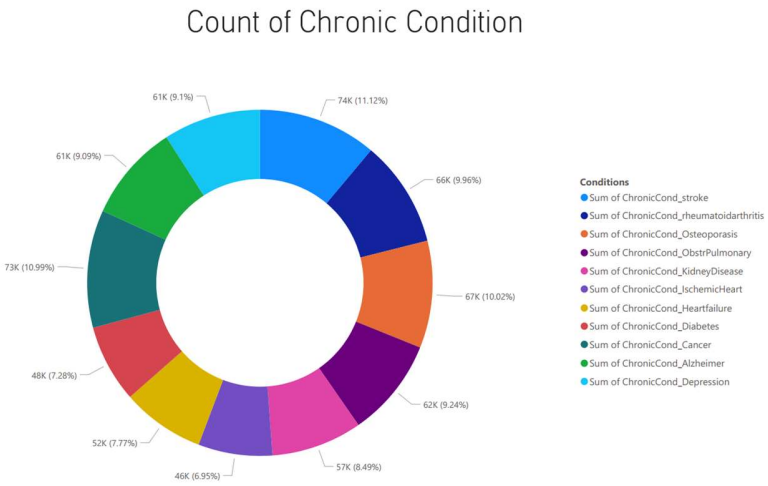
- 8gb RAM

Code

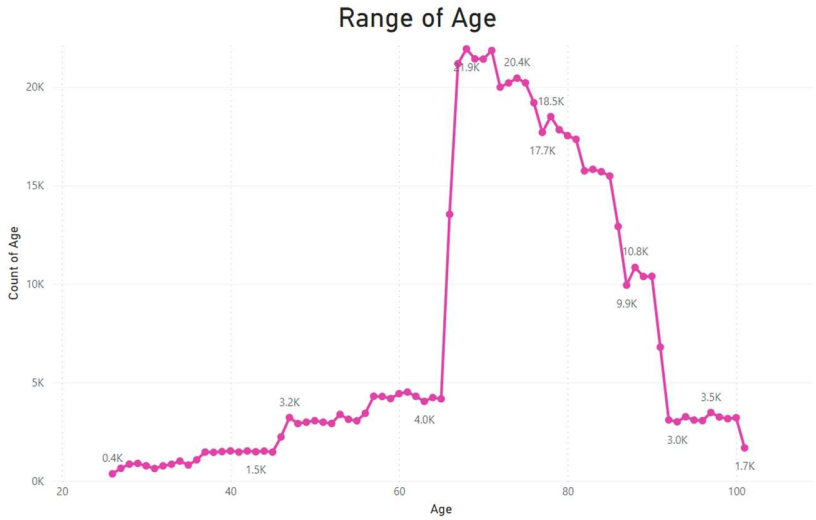
```
import scipy as sc
import seaborn as sns
import matplotlib.pyplot as plt
import pandas_profiling as profile   # To check data distributions and correlations
import warnings     # for supressing a warning when importing large files
warnings.filterwarnings("ignore")
from sklearn.preprocessing import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

Train=pd.read_csv("../input/med-fraud-data/Train-1542865627584.csv")
Train_Beneficiarydata=pd.read_csv("../input/med-fraud-data/Train_Beneficiarydata-
1542865627584.csv")


Train_Beneficiarydata.loc[Train_Beneficiarydata.DOD.isna(),'WhetherDead']=0
Train_Beneficiarydata.loc[Train_Beneficiarydata.DOD.notna(),'WhetherDead']=1
Train_Beneficiarydata.loc[:,'WhetherDead'].head(7)


Train_Inpatientdata['AdmissionDt'] = pd.to_datetime(Train_Inpatientdata['AdmissionDt']
, format = '%Y-%m-%d')
Train_Inpatientdata['DischargeDt'] =
pd.to_datetime(Train_Inpatientdata['DischargeDt'],format = '%Y-%m-%d')
Train_Inpatientdata['AdmitForDays'] = ((Train_Inpatientdata['DischargeDt'] -
Train_Inpatientdata['AdmissionDt']).dt.days)+1

Train_AllPatientDetailsdata=pd.merge(Train_Allpatientdata,Train_Beneficiarydata,left_
on='BeneID',right_on='BeneID',how='inner')
#PLotting the frequencies of fraud and non-fraud Merged transactions in the data

sns.set_style('white',rc={'figure.figsize':(12,8)})
count_classes = pd.value_counts(Train_ProviderWithPatientDetailsdata['PotentialFraud'],
sort = True)
print("Percent Distribution of Potential Fraud class:-
\n",count_classes*100/len(Train_ProviderWithPatientDetailsdata))
LABELS = ["Non Fraud", "Fraud"]
#Drawing a barplot
count_classes.plot(kind = 'bar', rot=0,figsize=(10,6))
```

```
#Giving titles and labels to the plot
plt.title("Potential Fraud distribution in Aggregated claim transactional data")
plt.xticks(range(2), LABELS)
plt.xlabel("Potential Fraud Class ")
plt.ylabel("Number of PotentialFraud per Class ")

plt.savefig('PotentialFraudDistributionInMergedData')

# Lets check the last record of Test_ProviderWithPatientDetailsdata

Test_ProviderWithPatientDetailsdata.iloc[[135391]]
```

Screenshots:



The highest number of identifiable chronic conditions is stroke, accounting for approximately 11.2% or around 74,000 cases, followed by cancer with approximately 7.2%, which is roughly 48,000 cases.



The age distribution of patients encompasses a broad range from 0 to 100, with a noteworthy concentration of individuals falling within the 60 to 80 age bracket, offering valuable insights into the patient demographics.

Results and Conclusion

Adding more fraud data to the training dataset helps in predicting unseen fraudulent behavior from time to time. Above model will help in predicting Provider fraud ,which will be helpful for insurance companies to scrutinize claims thoroughly. Improvement in the model will help in detecting networks of fraud Physicians, Providers and Beneficiaries.

The descriptions of the visualizations highlight important areas of concern in the healthcare sector, such as potential fraud, reimbursement patterns, and beneficiary distribution.

References:

[1] Herland, Matthew & Khoshgoftaar, Taghi & Bauder, Richard. (2018). Big Data fraud detection using multiple medicare data sources. Journal of Big Data. 5. 10.1186/s40537-018-0138-3.

[2]https://data.cms.gov/provider-summary-by-type-of-service/medicare-part-d-prescribers

[3] Rukhsar, L., Bangyal, W. H., Nisar, K., & Nisar, S. (2022). Prediction of insurance fraud detection using machine learning algorithms. Mehran University Research Journal of Engineering & Technology, 41(1), 33-40.

[4] Nalluri, V., Chang, J. R., Chen, L. S., & Chen, J. C. (2023). Building prediction models and discovering important factors of health insurance fraud using machine learning methods. Journal of Ambient Intelligence and Humanized Computing, 14(7), 9607-9619.

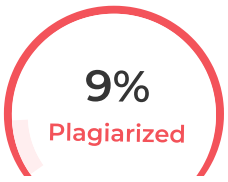
[5] Debener, J., Heinke, V., & Kriebel, J. (2023). Detecting insurance fraud using supervised and unsupervised machine learning. Journal of Risk and Insurance.

[6] Xiao, S., Bai, T., Cui, X., Wu, B., Meng, X., & Wang, B. (2023). A graph-based contrastive learning framework for medicare insurance fraud detection. Frontiers of Computer Science, 17(2), 172341.

Plagiarism Detector .net

Oct 16, 2023

Plagiarism Scan Report



Characters:7151	Words:981
Sentences:46	Speak Time: 8 Min

Excluded URL | None

Content Checked for Plagiarism

Abstract Medicare fraud is a pervasive problem that poses a considerable threat to the integrity and financial sustainability of the Medicare program in the United States. This abstract presents an innovative machine learning solution that combines anomaly analysis and geo-demographic metrics to predict potential instances of Medicare fraud. By leveraging advanced anomaly detection techniques to identify irregular patterns within Medicare claims data and incorporating geo-demographic insights to contextualize claims, our model offers a proactive approach to fraud detection. The project's objective is to create a model that can accurately differentiate between legitimate and fraudulent Medicare claims, leading to substantial cost savings, more efficient resource allocation, improved healthcare services for legitimate beneficiaries, and preventive measures to deter potential fraudsters. This research contributes to the safeguarding of Medicare's financial resources and the provision of high-quality healthcare services to those who rely on the program. Problem Statement The objective of this project is to develop a pioneering machine learning model that leverages anomaly analysis and geo-demographic metrics to predict Medicare fraud. Medicare fraud is a significant challenge, with fraudulent claims costing the program billions of dollars annually and undermining the healthcare services provided to legitimate beneficiaries. To address this issue, the project aims to create a predictive model that can effectively differentiate between genuine and fraudulent Medicare claims. In order to minimize fraudulent activities, the Centers for Medicare and Medicaid Services (CMS) released a number of “Big Data” datasets for different parts of the Medicare program. This model will incorporate cutting-edge anomaly detection techniques to identify irregular patterns in claims data and integrate geo-demographic insights to assess the appropriateness of claims within specific geographic and demographic contexts, ultimately enhancing the accuracy of fraud detection and safeguarding the program's financial resources while ensuring the delivery of legitimate healthcare services to beneficiaries. Module Description Data: The dataset for medical insurance fraud detection consists of three main components: Beneficiary Data, Inpatient Data, and Outpatient Data. The Beneficiary Data includes information such as unique beneficiary identifiers, dates of birth and, if applicable, dates of death, gender, race, and indicators for various chronic conditions. It also provides details on the duration of Part A and Part B coverage and annual reimbursement and deductible amounts for

inpatient and outpatient services. Inpatient Data contains records of inpatient medical claims, including claim identifiers, service providers, admission and discharge dates, attending and operating physicians, and reimbursement amounts. The Outpatient Data similarly records outpatient claims, including diagnosis codes. These datasets collectively offer a comprehensive view of the medical insurance claims landscape, enabling the detection of potential fraudulent activities and patterns within the healthcare insurance domain. Data Preprocessing: 1) Feature Engineering: Medicare fraud is categorized as organized crime which involves peers working together to create fraud transactions of claims.Adding features from grouping them helped in improving accuracy of prediction and fraud pattern recognition. Grouping and aggregating numeric features to provider level helped in detecting behavior of their transactions overall. 2) Logistic Regression Classifier: Features derived from above step are trained using logistic regression and evaluated.My decision of choosing LR is to check linear behavior between dependent and independent variables.Also Logistic model adds explicability to the predictions. Performance of the LR model showcases the linearity between variables. 3) Random Forest Classifier: One of the benefits of Random forest which excites most is the power of handling large data set with higher dimensionality. It can handle thousands of input variables and identify most significant variables. Further, the model outputs the Importance of variables, which can be a very handy feature.It also checks for non linearity between variables. 4) Auto encoders: Autoencoders are neural networks that aim to copy their inputs to their outputs. They work by compressing the input into a latent-space representation, and then reconstructing the output from this representation. My aim for the project is to train non fraud data using autoeencoder and reconstructing it back.While reconstructing Fraud data it will create an error,called as reconstruction error.Based on the threshold setting of reconstruction errors, we can easily predict Fraudulent behavior of healthcare providers. Software and Hardware Used Languages: - Python: The project is crafted in Python language, from model training to prediction. Libraries: - pandas: Data manipulation and analysis in a tabular format. - numpy: Scientific computing for numerical operations. - scipy: Scientific and technical computing library for advanced mathematical functions. - seaborn: Data visualization library for creating attractive statistical graphics. - matplotlib.pyplot: Data visualization library for creating various types of plots and charts. - pandas_profiling: A tool for generating data profiling reports to examine data distributions and correlations. - scikit-learn (sklearn): Machine learning library for data preprocessing and model training. - pickle: Serialization and deserialization of Python objects. - tensorflow: Machine learning framework for building and training deep learning models. - keras: High-level neural networks API running on top of TensorFlow for building and training neural network models. - rcParams: Setting parameters for Matplotlib figures and plots. - LABELS: A list of labels used to classify data as "Normal" or "Fraud." Hardware Required: Recommended: - 16gb RAM Minimum: - 8gb RAM Screenshots: The highest number of identifiable chronic conditions is stroke, accounting for

approximately 11.2% or around 74,000 cases, followed by cancer with approximately 7.2%, which is roughly 48,000 cases. The age distribution of patients encompasses a broad range from 0 to 100, with a noteworthy concentration of individuals falling within the 60 to 80 age bracket, offering valuable insights into the patient demographics. Results and Conclusion Adding more fraud data to the training dataset helps in predicting unseen fraudulent behavior from time to time. Above model will help in predicting Provider fraud ,which will be helpful for insurance companies to scrutinize claims thoroughly. Improvement in the model will help in detecting networks of fraud Physicians, Providers and Beneficiaries. The descriptions of the visualizations highlight important areas of concern in the healthcare sector, such as potential fraud, reimbursement patterns, and beneficiary distribution.

Sources

2% Plagiarized

Tree Based Algorithms | Implementation In Python & R

https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/

2% Plagiarized

machine learning algorithms implementation with dataset

https://github.com/deepak6446/machine-learning

2% Plagiarized

They work by compressing the input into a latent-space representation, and then reconstructing the output from this representation. This kind of network is composed of two parts : Encoder: This is the part of the network that compresses the input into a latent-space representation.

https://towardsdatascience.com/deep-inside-autoencoders-7e41f319999f/

2% Plagiarized

Medical Provider Fraud Detection

https://www.kaggle.com/code/rohitrox/medical-provider-fraud-detection