FEATURE MINIMIZATION IN DETECTION OF CANCER

A PROJECT REPORT

Submitted by

CB.EN.U4CSE16154 Surapaneni Aasritha devi

CB.EN.U4CSE16156 Sushma Shivani Nukala

CB.EN.U4CSE16339 Abhirup Pulla

CB.EN.U4CSE16010 Sarat Chandra

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING

AMRITA SCHOOL OF ENGINEERING, COIMBATORE

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE 641 112

OCTOBER 2019

AMRITA VISHWA VIDYAPEETHAM

AMRITA SCHOOL OF ENGINEERING, COIMBATORE, 641112

BONAFIDE CERTIFICATE

This is to certify that the project report entitled "**Feature minimization in detection Of cancer** " submitted by Surapaneni Aasritha devi (CB.EN.U4CSE16154), Sushma Shivani Nukala (CB.EN.U4CSE16156), Abhirup Pulla(CB.EN.U4CSE16339)and Sarat Chandra (CB.EN.U4CSE16010) in partial fulfillment of the requirements for the award of the Degree **Bachelor of Technology in Computer Science and Engineering** is a bonafide record of the work carried out under our guidance and supervision at Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore

PROJECT GUIDE	CHAIRPERSON
Mr. Arunkumar Chinnasamy	**Dr. (Col) P.N. Kumar**
Assistant Professor	Professor
Dept. of Computer Science and Engg.	Dept. of Computer Science and Engg.

This project report was evaluated by us on :. . . . . .

INTERNAL EXAMINER EXTERNAL EXAMINER

Acknowledgment

We express our gratitude to our beloved Satguru Sri Mata Amritanandamayi Devi for providing a bright academic climate at this university, which has made this entire task appreciable. This acknowledgement is intended to be a thanks giving measure to all those people involved directly or indirectly with our project.

We would like to thank our Vice Chancellor Dr. Venkat Rangan. P and Dr. Sasangan Ramanathan Dean Engineering of Amrita Vishwa Vidyapeetham for providing us the necessary infrastructure required for completion of the project.

We express our thanks to Dr.(Col.P.N.Kumar), Chairperson of Department of Computer Science Engineering, Dr.C.Shunmuga Velayutham and Dr. G. Jeyakumar, Vice Chairpersons of the Department of Computer Science and Engineering for their valuable help and support during our study. We express our gratitude to our guide, Mr. Arunkumar Chinnasamy , for the guidance, support and supervision. We feel extremely grateful to Dr.Prakash.P Ms. Sujee R Mr. Dayanand Ms. Shanmuga Priya for their feedback and encouragement which helped us to complete the project. We also thank the staff of the Department of Computer Science Engineering for their support. We would like to extend our sincere thanks to our family and friends for helping and motivating us during the course of the project.

Abstract

Health insurance works by protecting your assets from the high cost of medical care.Recently,fraudulent activities in medical claims have been prevalent.Detection and prevention of these fraudulent activities have been increasing with various sophisticated tools. But,still there are some lapses in analyzing and finding suspicious activities and mismanagement of system in medical insurance.Our paper provides a comprehensive study leveraging machine learning methods to detect fraudulent Medicare providers. We use the publicly available Medicare data and provider exclusions for fraud labels to build and classify the providers as fraud and non-fraud. In order to lessen the impact of class imbalance, given few actual fraud labels, we deploy oversampling techniques such as SMOTE. We also use hybrid models which involves clustering and classification techniques and many other machine learning algorithms to detect the medical fraud.....

Table of Contents

List of Figures

List of Tables

List of Abbreviations

abbreviation FullForm

Chapter 1

Introduction

1.1 Background

Fraudulent behavior has been prevalent in the healthcare sector which has resulted in huge losses of money every year. Since the number of frauds occurring is minute compared to huge amount of insurance data, it is difficult to detect the fraudulent behaviors.High health insurance fraud rates are a serious detriment not only to insurance companies but also to the broader society which ultimately pays the price in terms of higher fees in a nation wherein health care spending amounts to an astronomical 18 percent of GDP. Incorporating machine learning with data mining and statistics help to anticipate and detect these frauds and minimize costs. Using sophisticated data mining tools, millions of transactions can be searched to spot patterns and detect fraudulent behaviors. This paper gives an insight into the various data mining techniques which are efficient in detecting the frauds especially in the healthcare insurance sector.

1.2 Problem Statement

The goal of this project is to predict the potentially fraudulent providers based on the claims filed by them.We intend to discover important attributes helpful in detecting the behavior of the potentially fraud providers. Further, we will also be studying the fraudulent patterns in the provider's claims to understand the future behavior of providers using data mining techniques such as exploratory data analysis and many machine learning algorithms including hybrid models.

1.3 Specific Objectives

Fraudulent healthcare claims increase the burden to society. Therefore healthcare fraud detection is now becoming more and more important. Generally, healthcare frauds are not obvious and thus difficult to detect. In order to curb these problems of fraud, in our project we will be using the SMOTE (Synthetically Minority Oversampling Technique) for overcoming the problem of class imbalancing. Class imbalance occurs when the class distributions are highly imbalanced where the total number of a class of data (positive) is far less than the total number of another class of data (negative). We will also be focusing on the hybrid model which involves in classification and clustering of the data for the prediction of the potentially fraud providers. We intend in using other machine learning algorithms as well for improving the accuracy of the prediction.

1.4 Findings

Medicare is an insurance program controlled by the US government, it includes various, types of services like prescription drug coverage, hospital insurance, and medical insurance. Medicaid program which is run by state and it has its specific rules for the services. There are three different group of individuals who are

involved in fraud in medical health care commission. 1. 1. Service providers- who are generally doctors, hospitals, ambulance, companies and laboratories. These are various activities involves fraud Ex. Billing, unbundling, upcoding, falsify the patient's treatment, misrepresenting non covered treatment etc., 2. Insurance subscribers- which includes patients and patient's employers. Some examples of their fraudulent behavior include falsifying the records and illegal claims. 3. Insurance carriers- who are mediators from government and private insurance companies will receive the premium amount from the insurance subscribers and pay on behalf of their subscribers to companies. Most of the ways they can involve in fraud falsifying the reimbursements, services and treatments. In addition to these, when a fraudulent activity carried by different parties it is termed as Conspiracy fraud From the existing work, the proposed study incorporates the statistical methods for fraud detection of medical claims . We also incorporate Supervised and unsupervised models along with hybrid approaches. Supervised methods is to be labeled by the experts and requires training dataset for the application domains whereas unsupervised methods is to remove any outliers in the data.The hybrid approach combines both features of prior approaches. We also try to reduce the class imbalance issue caused in the data(Figure 1.1) for better accuracy of the predictions by using oversampling methods.

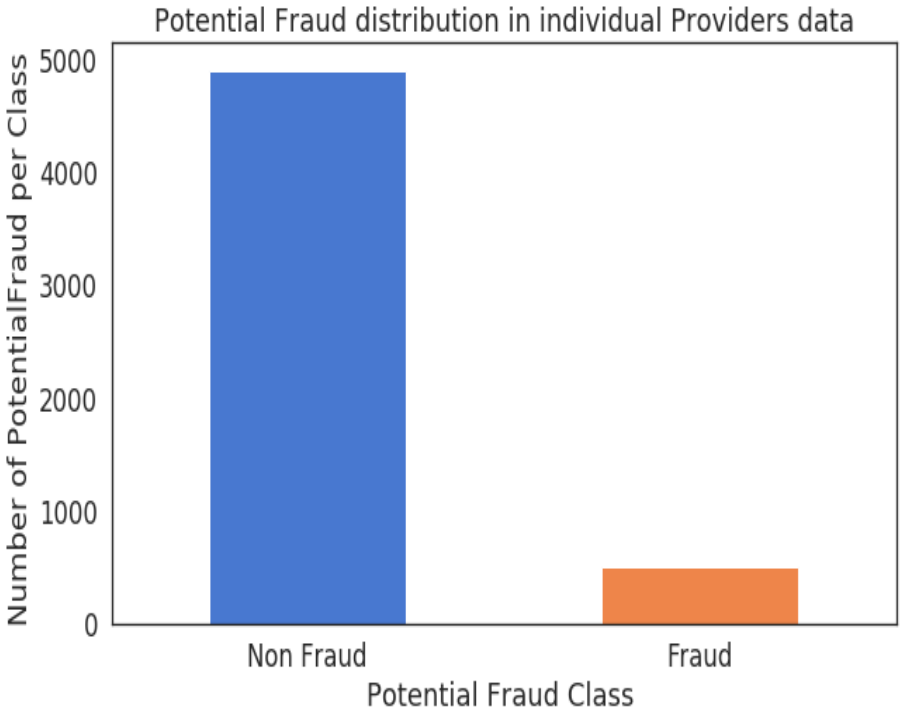

Figure 1.1: Class Imbalance In Fraud detection of Healthcare Insurance

Chapter 2

LiteratureSurvey

The following section provides a review of the literature related to the development of an application in python flask for fraud detection in healthcare using a hybrid of a supervised learning model and unsupervised learning model. The development was based on the methods discussed in "Identifying Medicare Provider Fraud with Unsupervised Machine Learning by Richard Bauder , Raquel da Rosa ,and Taghi Khoshgoftaar which was submitted to the 2018 IEEE International Conference on Information Reuse and Integration (IRI).This paper compares the performance of 5 unsupervised algorithms in health insurance fraud detection.The five algorithms are Isolation Forest , Unsupervised Random Forest ,Local Outlier Factor, autoencodesr, and k-Nearest Neighbors. Overall, based on these results, LOF, for any configuration of neighbors, outperforms all other methods including IF, which was previously shown to outperform LOF.In fact, IF has one of the lower AUC scores, near 0.50, which is almost a random guess. The autoencoders also performed poorly being similar in AUC to IF. On the other hand, URF performs well relative to the other methods.KNN5 has the worst overall performance akin to randomly guessing the fraud or non-fraud labels, which could be due to the high class imbalance.4

The development was based on another paper "Medicare Fraud Detection Using Machine Learning Methods" by Richard A. Bauder and Taghi M. Khoshgof-

taar which was submitted to the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA) This paper talks about some supervised learning, unsupervised learning and hybrid methods. They talk about the following supervised algorithms Gradient Boosting Machines (GBM).Random forest (RF) and The Naive Bayes algorithm .The unsupervised alorithms listed are Local outlier factor (LOF) Autoencoders (which were also used in the previous paper) and hybrid the pre-training autoencoder uses the original unsupervised autoencoder as pre-training input to a supervised model which, in this case, is a neural network. The learner uses the weights from the inputted autoencoder for model fitting. We use the same parameters as with the unsupervised autoencoder, including the threshold for detecting outliers.This learner is more supervised, with unsupervised inputs used only for weighting. The result was that supervised methods being significantly better than the unsupervised and hybrid learners. After viewing both the above papers we have noted the following advantages and disadvantages. In supervised learning,the advantages are that all classes are meaningful to humans and it can be easily used for pattern classification. The disadvantage is the difficulty associated in gathering class labels. Also, when there is bulk input data, it is costly to label all of them, and claims must be identified properly because false positives and true negatives can create a bad impression about the insurance company in the minds of its customers. In unsupervised learning, the advantages are it aims to detect anything which does not abide by the normal behavior and because of the lack of direction, it can find patterns that have not been noticed previously. While the disadvantage being because of lack of direction, there may be times when no interesting knowledge has been discovered in the set of features selected for the training. Considering the advantages and disadvantages of most of the classification and clustering techniques, ECM is chosen as the clustering technique because the data flows in continuously and there is a need to cluster dynamic data and a classification technique used since it provides the scalability and usability that are needed in a good quality data mining.

Chapter 3

Proposed System

3.1 Methodology

In this section, we detail the publicly available Medicare data, and the description of the attributes in the dataset. Additionally, we explain feature selection and sampling as well as the three improvement strategies. We also give a description on the different types of supervised,unsupervised models used along with the hybrid approach as well as the performance metrics.

3.1.1 Data

The Medicare dataset used outlines information based on inpatient claims, outpatient claims and Beneficiary details of each provider. Let us take a closer look on each of these datasets. 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 dates along with diagnosis code.

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

Attributes	Meaning
BeneID	ID of person or entity entitled to receive claim amount and other benefits upon death or on maturity of policy
ClaimID	ID of request that health insurance policyholder submits to the insurance company in order to obtain services covered.
ClaimStart	Start of the claim
ClaimEnd	End of the claim
Provider	ID of company that provides a healthcare service
InsuranceClaimAmtReimbursed	the amount reimbursed in by insurance company when a provider pays for expenses after direct payment.
Attending Physician	ID's of the attending physicians.
Operating Physician	ID's of the operating physicians.
Other Physician	ID's of the other physicians.
Admission Date	Claim Admitting Diagnosis Code (FFS) A diagnosis code on the institutional claim indicating beneficiary's initial diagnosis at admission.
ClmAdmitDiagnosisCode	The amount paid for the covered health care services before the beginning of payment of the insurance plan.
DeductibleAmtPaid	The amount paid for the covered health care services before the beginning of payment of the insurance plan.
DischargeDate	1,2,3,4,5,6,7,8,9,10 & It is a diagnosis code used by physicians and other healthcare providers to classify all diagnoses, symptoms and procedures recorded in conjunction with hospital care.
ClaimDiagonsisCode1,2,3,4,5,6,7,8,9,10	1,2,3,4,5,6,7,8,9,10 & It is a diagnosis code used by physicians and other healthcare providers to classify all diagnoses, symptoms and procedures recorded in conjunction with hospital care
DischargeDate	1,2,3,4,5,6 & Procedure codes are subtypes of medical classification used to identify specific surgical.medical or diagnostic interventions and it indicates the procedure performed during period covered by institutional claim.

Figure 3.1: Inpatient Data

C) Beneficiary Details Data This data contains beneficiary details like health conditions they belong to etc.

3.1.2 Class Imbalance

The Medicare claims data, with fraud labels, is a challenging dataset due to the skewed nature of the provider exclusions. With such class imbalance, the learner will tend to focus on the majority class (i.e. the class with the majority of instances), which is usually not the class of interest. In our case, the non-fraud labels are the majority class.An effective solution to compensate for some of the detrimental effects of severe class imbalance is by changing the class distribution in the training data, to increase the representation of the minority class to help

Attributes	meaning
BeneID	ID of person or entity entitled to receive claim amount and other benefits upon death or on maturity of policy
ClaimID	ID of request that health insurance policyholder submits to the insurance company in order to obtain services covered.
ClaimStart	Start of the claim
ClaimEnd	End of the claim
Provider	ID of company that provides a healthcare service
InsuranceClaimAmtReimbursed	the amount reimbursed in by insurance company when a provider pays for expenses after direct payment.
Attending Physician	ID's of the attending physicians.
Operating Physician	ID's of the operating physicians.
Other Physician	ID's of the other physicians.
Admission Date	Claim Admitting Diagnosis Code (FFS) A diagnosis code on the institutional claim indicating beneficiary's initial diagnosis at admission.
ClmAdmitDiagnosisCode	The amount paid for the covered health care services before the beginning of payment of the insurance plan.
DeductibleAmtPaid	The amount paid for the covered health care services before the beginning of payment of the insurance plan.
ClaimProcedureCode	1,2,3,4,5,6 & Procedure codes are subtypes of medical classification used to identify specific surgical.medical or diagnostic interventions and it indicates the procedure performed during period covered by institutional claim.
ClaimDiagonsisCode1,2,3,4,5,6,7,8,9,10	1,2,3,4,5,6,7,8,9,10 & It is a diagnosis code used by physicians and other healthcare providers to classify all diagnoses, symptoms and procedures recorded in conjunction with hospital care.

Figure 3.2: Outpatient Data

improve model performance. The sampling of data changes the class distribution of the training instances to minimize the effects of these rare events. There are two basic sampling methods: oversampling and under-sampling. Oversampling is a method for balancing classes by adding instances to the minority class, whereas undersampling removes samples from the majority class. Oversampling can increase processing time by increasing the overall size. More critically, oversampling can overfit the data by making identical copies of the minority class. On the contrary, with undersampling, we retain all of the original fraud-labeled instances and randomly sample without replacement from the remaining majority class instances. In this paper, we are using SMOTE(Synthetically Minority Over-Sampling Technique) since oversampling causes more duplicate and undersampling randomly chooses the majority claims to fit with the minority, so there will not be proper learning of the

Attributes	meaning
BeneID	ID of person or entity entitled to receive claim amount and other benefits upon death or on maturity of policy
DOB	Date of Birth.
DOD	Date of Death
Gender	Male or Female
RenalDiseaseIndicator	Indicates whether the person has a Renal Disease or not
NoOfMonthsPartACov	The number of months covered by Medicare Part A which includes Medicare inpatient care, including care received while in a hospital, a skilled nursing facility, and, in limited circumstances, at home
NoOfMonthsPartBCov	The number of months covered by Medicare Part A which includes Medicare outpatient care, preventive services, ambulance services, and durable medical equipment
ChronicCondAlzheimer	Indicator for Alzheimer
ChronicCondHeartfailure	Indicator for Heart failure
ChronicCondKidneyDisease	Indicator for Kidney Disease
ChronicCondCancer	Indicator for Cancer
ChronicCondObstrPulmonary	Indicator for Postpreliminary
ChronicCondDepression	Indicator for Depression
ChronicCondDiabetes	Indicator for Diabetes
ChronicCondIschemicHeart	Indicator for Ischemic Heart
ChronicCondOsteoporasis	Indicator for Osteoporosis
ChronicCondrheumatoidarthritis	Indicator for rheumatoid arthritis
ChronicCondstrokeChronicCondstroke	Indicator for ChronicCondstroke

Figure 3.3: Beneficiary Data

data.(Figure 3.4) SMOTE synthesizes new minority instances between existing minority instances. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen.

3.1.3 Models and Algorithms

For our experiments, we will be using three important machine learning models to predict the Potential Fraud providers and classify them as fraud and non-fraud: In supervised learning algorithms we will be using : Random Forest Classifier, SVMs(Support Vector Machines)]and Logistic Regression. In unsupervised machine

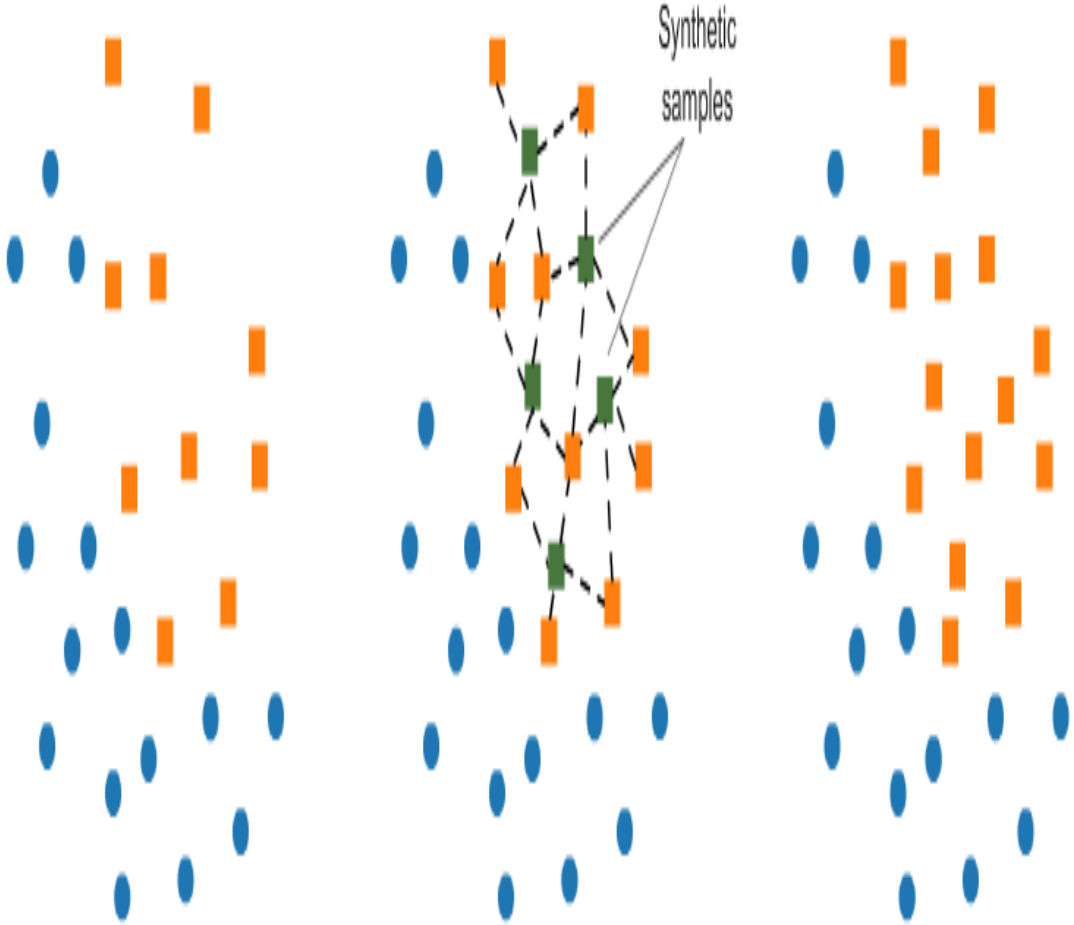

Figure 3.4: SMOTE

learning , we will be using K-means clustering and neural networks. Our main focus is on the hybrid approach we will be deploying which involved in ECM (Evolving Clustering Method) and a classification technique such as SVMs or Ensemble Boost classifiers. Evolving Cluster method is used since because the data is dynamic and new data is generated continuously.ECM clusters them by modifying the position and size of the cluster.

3.1.4 Performance Metrics

In order to gather more detail on learner performance, we also examine false positive rate (FPR) and false negative rate (FNR), with the instances labeled as fraud being the positive class. A classification threshold of 0.5 was used to assess

these metrics for each learner. For the detection of Medicare claims fraud, a low FNR is most important since this indicates a higher detection rate for capturing actual fraudulent claims. Given the current manually intensive process in detecting fraud, we can generally accept a slightly higher FPR (i.e. claims predicted as fraud that are not actual fraud)as long as we obtain the lowest possible FNR. In practice , missing a substantial number of fraudulent events will render any fraud detection system ineffective, but, conversely,having too many false positives will make the system unusable.(Figure 3.5)

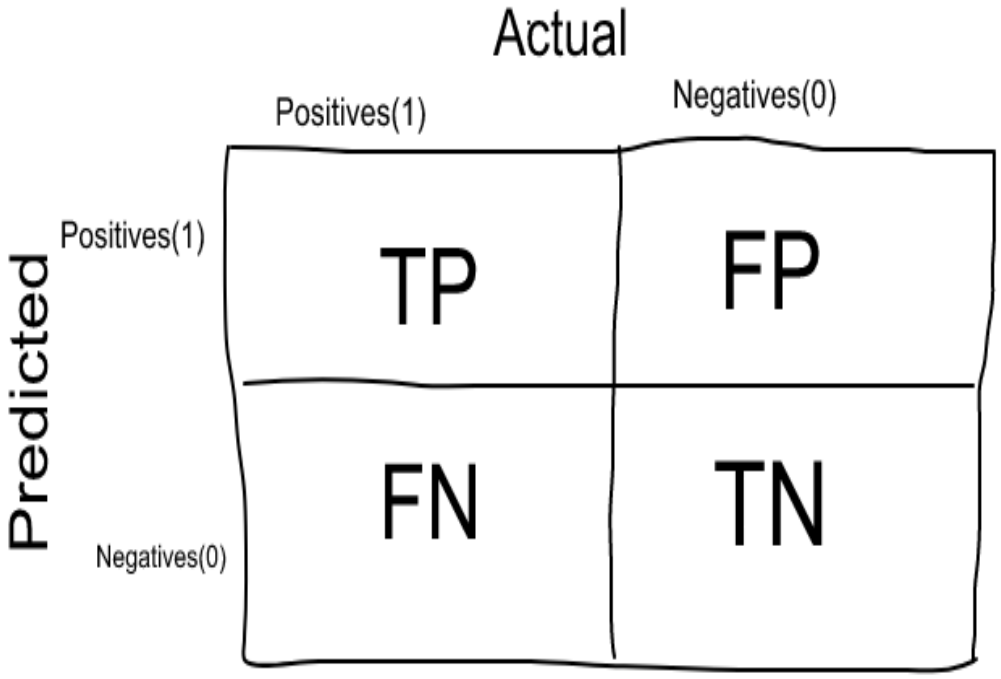

Figure 3.5: Performance Metrics

3.2 Implementation

The following steps are required before implementing the algorithms for the prediction of Potential fraud providers:

- Data Understanding

- Data Preprocessing
- Feature Engineering
- Exploratory Data Analysis
- Model selection
- Prediction and Performance Metrics

3.2.1 Data Understanding and Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. In our project ,we understand how the insurance companies works in the healthcare sector and about the various types of fraud that could occur in that sector. Additionally, understanding the data and each attribute in the dataset was the key step to get started with the process. We load the datasets and check its shape and information of the data. We also check for the null values or duplicates to clean the dataset. We analyze the Inpatient data, the Outpatient data and the Beneficiary data and label encoded the categories for further processing. We merge all the datasets after the understanding and cleaning of the Medicare data.

3.2.2 Feature Engineering and Exporatory Data Analysis

Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work.Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. We add extra

features in the such as age,whetherDead and No.ofAdmitDays for calculating the number of days a person is admitted in the hospital in the Inpatient data. In EDA, we visualize the number of potential frauds and non frauds from the main dataset that was created after the merging of all the data. We also visualize the top 10 procedure codes involved in Healthcare fraud. We append the train data to test data which helps in getting good average scores of new features in Test data,as we see, not all levels of variables are present in test data compared to train data.So our approach here will be to append train data to test data ,derive new average features and take only test data to evaluate results.As we verified that our first record is appended to test data correctly,we are all set to derive Average features grouped according to columns of datasets.Other than basic explorations and visualizations, we can use certain methods to identify clues of fraud and abuse. One such simple method is 'Grouping based on Similarity'. In this method, we basically group all the records by the ProcedureCodes, DiagnosisCodes,Provider.

3.2.3 Models

Supervised Learning This is the most usual learning technique wherein the model is trained using pre-defined class labels. In the context of health insurance fraud detection the class labels may be the "legitimate" and "fraudulent" claims. The training dataset can be used to build the model. Then any new claim can be compared with the already trained model to predict its class. A claim will be classified as a normal claim if it follows a similar pattern to the normal behavior else it will be classified as an fraud.

Random Forest Classifiers: The Random Forest algorithm is an implementation of bootstrap aggregation (bagging) where each tree in an ensemble of decision trees is constructed from a bootstrap sample of feature vectors from the training data.

Support Vector Machines: SVM is fundamentally a classification technique.

The system is trained to determine a decision boundary between classes of "non-fraud" and "fraud" claims. Then each claim is compared with that decision boundary and is placed into either legitimate or fraudulent class.

Unsupervised Learning Unsupervised learning has no class labels. It focuses on finding those instances which show unusual behaviour. Unsupervised learning techniques can discover both old and new types of fraud since they are not restricted to the fraud patterns which already have pre-defined class labels like supervised learning techniques do. Requirement of using hybrid approach. Unsupervised cannot identify duplicate record of medical claims and supervised cannot recognize a new unknown disease that can occur. For example: if SVMs recognize heart disease, diabetes etc. when a new disease such as Parkinson disease occurs it cannot be recognized. Evolving Clustering Method : (ECM) is used for clustering because the data is dynamic and new data is generated continuously and Support Vector Machine (SVM) for classification. ECM used for dynamic data. Dynamic data are those which keep on changing with respect to time. As and when new data point comes in, ECM clusters them by modifying the position and size of the cluster. There is a parameter known as radius associated with each cluster that determines the boundaries of that cluster. Initially, the cluster radius is set to zero. The radius of the cluster increases as more data points are added to that cluster. It has one more parameter known as the distance threshold, which determines the addition of clusters. If the threshold value is small then, there will be more number of small clusters and if the value is large, then there will less number of large clusters.

Chapter 4

Results and Discussion

Chapter 5

Conclusion and Future Work

Chapter 6

Bibliography

[1] Bauder, R. and Khoshgoftaar, T" Medicare Fraud Detection Using Machine Learning Methods" - IEEE Conference Publication(2019)

[2] Bauder, R., da Rosa, R. and Khoshgoftaar, T. "Identifying Medicare Provider Fraud with Unsupervised Machine Learning" - IEEE Conference Publication (2019)

[3] Wang, J., Xu, M., Wang, H. and Jiwu, J." Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding" - IEEE Conference Publication(2019)

[4] Matthew Herland , Richard A. Bauder , Taghi M. Khoshgoftaar "Medical Provider Specialty Predictions for the Detection of Anomalous Medicare Insurance Claims"- IEEE Conference (2017)

[5] Shivani S. Waghade, Prof. Aarti M. Karandikar "A Comprehensive Study of Healthcare Fraud Detection based on Machine Learning "Research India Publication(2018)