INSURANCE CLAIM ANALYSIS AND DETECTING FRAUDS USING MACHINE LEARNING ALGORITHMS

GUJIPINENI VENKATA SAIBABA[1], B.CHARISHMA[2]

[1]PG SCHOLAR (M.Tech., (CSE)), Dept of CSE, Srinivasa Institute of Technology And Science, Kadapa

[2]Associate Professor, HOD, Dept of CSE, Srinivasa Institute of Technology And Science, Kadapa

Abstract

Insurance fraud is an illegal conduct that is done on purpose in order to profit financially. This is currently the most serious issue that numerous insurance companies throughout the world are facing. The majority of the time, one or more gaps in the investigation of false claims has been identified as the primary factor. As a result, the requirement to use computer tools to stop fraud activities increased. Providing customers with a dependable and stable environment while significantly lowering fraud claims. We demonstrated the results of our research by automating the evaluation of insurance claims using a variety of data methodologies, where the detection of erroneous claims would be done automatically using Data Analytics and Machine Learning techniques.

Keywords— — Machine Learning, Digital Era, Client Lifetime Value (CLV).

I. INTRODUCTION

The most important advantage of Machine Learning (ML) to use in Insurance Industry is to facilitate data sets. Machine learning (ML) can be successfully useful across Structured, Semi Structured or Unstructured datasets. Machine learning can be used accurate across the value chain to identify with risk, claims and customer actions, by means of advanced predictive accurateness. The probable applications of machine learning in insurance are plentiful from perceptive risk appetite and premium leakage, to expense administration, subrogation, proceedings and fraud detection. Machine learning is not a novel technology; this technology is following from the last few decades. There are 3 main categories of learning they are supervised learning, Unsupervised Learning and reinforcement learning. The greater part of the insurers arefollowing Supervised Learning fromlast few decades for assessing the risk by means of known parameters in dissimilar combinations to acquire the preferred outcome. Present age insurers are motivated to unsupervised learning, in this predestined goals are clear. If there are any modifications in the variables, the method identifies those modifications and tries to change as per the goals. For example according to traffic the GPS Suggests different routes dynamically based on traffic conditions. In insurance industry also the learning is adopted for usage based insurance. Reinforcement learning is mostly depends on ANN (Artificial Neuron Network), Target/ Goa can be modified dynamically depending on objective. Reinforcement learning is used for IOT applications.

In essence, insurance fraud is deliberate deception that can done by, against, or with the intent to defraud an insurance company or agent. It is a severe and urgent problem that is a threat because fraudulent insurance applications put a greater financial strain on the society through high premium prices. Recent research suggest that there is universal agreement that traditional methods of fraud identification are highly unreliable and imprecise. These worries prompt the machine learning and data analytics community to focus on this issue and seek a solution. Similar to this, our proposed work accurately distinguishes between fraudulent and non-fraudulent claims so that only fraudulent cases need to be investigated and legitimate claims can be made quickly without wasting time or resources. This project aims to suggest the most accurate and simplest way that can be used to fight fraudulent claims. The main problem with detecting fraudulent activities is the massive number of claims that run through the companies systems. This problem can also be used as an advantage if the officials were to take into account that they hold a big enough database if they combined the database of the claims. Which can be used in order to develop better models to flag the suspicious claims.

As we live in a very materialistic world everyone is looking out to protect something they have or own in one way or another. People are willing to pay money as a contingent against the unknown loss that they might face. In the U.S alone the insurance industry is valued at 1.28 trillion dollars and the U.S consumer market losses at least 80 billion to insurance fraud every year. That causes the insurance companies to increase the cost of their policies which puts them in a less competitive position against the competition. This in turn also increased the threshold of the minimal payment for a policy since they can afford to do so while everyone is raising prices. This project will look into the different methods that have been used in solving similar problems to test out the best methods that have been used previously. Searching if examining these methods and trying to enhance and build a predictive model that could flag out the suspicious claims based on the researching and testing out the different models and comparing these models to come up with a simple enough time- efficient and accurate model that can flag out the suspicious claims without stressing the system it runs on.

A. Problem Statement

The main purpose of this project is to come up with a model to be used to find out if a certain insurance claim made is a fraud or not. The model will be designed after testing multiple algorithms to come up with the best model that can detect if a claim is fraud or not. This is aimed at the insurance companies as a pitch to come up with a more tailored model for their liking to their own systems. The model should be simple enough to calculate big datasets, yet complex enough to have a decent successful percentile.

The traditional method for the detecting frauds depends on the event of heuristics around fraud indicators. Supported these, the selection on fraud created is said to occur in either of situations like, uncertain things the principles are shown if the case should be interrogated for extra examination. In numerous cases, an inventory would be prepared with scores for various indicators of the occurred fraud. The factors for deciding measures and additionally the thresholds are tested statistically and periodically recalibrated. Associate aggregation and then price of the claim would verify necessity of case to be sent for extra examination .The challenge with above strategies is that they deliberately believe on manual mediation which might end in the next restrictions

B. Motivation

Basically businesses ought to obtain the responses to prevent fraud from happening or if that is out of the question, to watch it before important damage is finished at intervals 407 the strategy. In most of the companies, fraud is understood entirely once it happens. Measures are then enforced to forestall it from happening over again. At intervals the given time that they can't resist at different time intervals, but Fraud detection is that the most effective suited issue for removing it from the atmosphere and preventing from continuance. Previously frauds related insurance detected or analyzed by manually using this method is not correct way to detect frauds related insurance, therefore increasing accuracy, precision, recall we proposed this project using machine learning algorithm. People are making fool insurance companies by claiming wrong insurance claim so detecting this we did this project.

II LITERATURE SURVEY

Obtaining Dataset: We collected dataset from different sources like haggle , Google. Also we created dummy dataset for analysis and detection of frauds. We used three types of dataset raw dataset which is dummy dataset, processed insurance claim dataset and Integrated dataset.

Loading Dataset: For reading dataset we used panda's python library. We loaded dataset in csv File format. For visualization the features of dataset used python libraries which are matplotlib, seaborn plotly etc. We loaded dataset in jupyter notebook

Preprocessing Dataset: Preprocessing the dataset means data wrangling. In preprocessing dataset we reduced reduced redundancy of data. Under preprocessing step we cleaned the dataset, treated the null values, we did normalization of datta removed the outliers. Encoded variables categorical variables into continuous variables

Comparative Analysis: In comparative analysis we divided dataset into training and testing sets .we Visualized heat map of training and testing dataset. We compared model with Respect to accuracy parameter using different classifier .Finally we generated the ROC curve using the different classifier.

Training and Validation: For Training and Testing we splited dataset using sklearn library into 80-20. For training we feuded or trained model with 100% accuracy for testing we got 65% accuracy. For visualizing training and testing dataset we used heat map.

Calculating Accuracy: Accuracy is defined as percentage of correct prediction for test data .Total accuracy is calculated using confusion matrix. Precision, recall and f1-Score is used for calculating the total accuracy. We got 65% accuracy of our model Total accuracy calculated random forest classifiers

The world is moving towards complete digitization, and many organisations have recently started considering the integration of block chain technology into their operationalworkflows. Numerous sectors have recently invested time and resources learning about the potential of block chain and the impact of its adoption [10]. This technique was primarily introduced for Bitcoin applications to enable peer-to-peer transfer of electronic cash in the absence of a centralised trusted system and prevent the problem of double payment [11, 12]. The Blockchain 1.0 came in to existence inthe year 2009 and the first generation technique featured hardcoded special-purpose protocols that concentrates chiefly on digital currency while also serving potentially dangerous public players. In 2014, the Blockchain 2.0 came in to existence and it focusses on creative application of smart contracts in variety of fields. Hyperledger projects are the base of Blockchain 3.0 and it was introduced in the year 2017. The development of extensive amount of systems in the fields of finance, certification and logistics took place during the second generation [13, 14]. A blockchain based solution is proposed in this work for insurance agencies to ensure digital security and quicker processing. In case of the insurance industry, the Blockchain technique is a relatively newer approach and the several advantages of applying Blockchain technique in the fieldof insurance is given in Fig.1

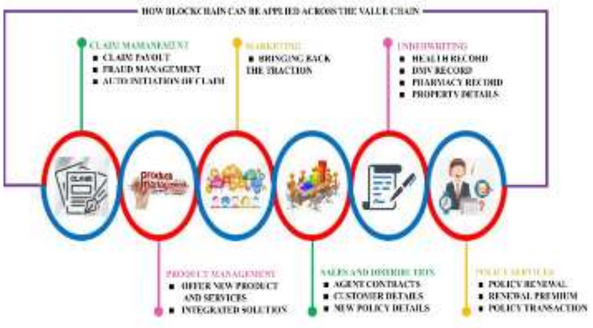


Fig.1.Benefits of blockchain in insurance sector

Blockchainhas the ability to resolve some of the major issues troubling the insurance sector today, including effective fraud detection, product innovation and decreased operational expenses. Additionally, combining the Blockchain with a machine learning technologyincreases its effectiveness and intelligence. So, in order to classify the insurance frauds, the machine learning technique of SVM [15-17] is considered. The working of the SVM is further improved by integrating it with the RF algorithm [18]. For the insurance industry, an innovative Blockchain-based strategy is proposed to achieve the major goal of fraud detection and prevention. Furthermore, the Blockchain technique has the added benefits of cost savings, improved back-end efficiency, better assessment of risks, handling big data in addition to event triggered smart contracts. The remarkable learning ability of hybrid ERFSVM machine learning technique is capable of heightening the efficiency and smartness of the Blockchain approach

III. EXISTING ANALYSIS

Yang and Hwang developed a fraud detection model using the clinical pathways concept and process-mining framework that can detect frauds in the healthcare domain . The method uses a module that works by discovering structural patterns from input positive and negative clinical instances. The most frequent patterns are extracted from every clinical instance using the module. Next, a feature-selection module is used to create a filtered dataset with labeled features. Finally, an inductive model is built on the feature set for evaluating new claims. Their method uses clustering, association analysis, and principal component analysis. The technique was applied on a real-world data set collected from

National Health Insurance (NHI) program in Taiwan. Although the authors constructed different features to generate patterns for both normal and abusive claims, the significance of those features is not discussed. Bayerstadler et al. presented a predictive model to detect fraud and abuse using manually labeled claims as training data. The method is designed to predict the fraud and abuse score using a probability distribution for new claim invoices. Specifically, the authors proposed a Bayesian network to summarize medical claims' representation patterns using latent variables. In the prediction step, a multinomial variable modeling predicts the probability scores for various fraud events. Additionally, they estimated the model parameters using Markov Chain Monte Carlo (MCMC).

Zhang et al. proposed a Medicare fraud detection framework using the concept of anomaly detection . First part of the proposed method consists of a spatial density based algorithm which is claimed to be more suitable compared to local outlier factors in medical insurance data. The second part of the method uses regression analysis to identify the linear dependencies among different variables. Additionally, the authors mentioned that the method has limited application on new incoming data.

IV. PROPOSED WORK

Knowing a risk is that the beginning in bar, associated intensive assessment offers the lightness that want. This is typically usually performed exploitation varied techniques, like interviews, surveys, focus teams, feedback conducted anonymously, detailed study of record and analysis to spot traffic pumpers, service users, and subscription scam which are different fraudulent case. The association of Certified Fraud Examiners offers a detailed guide to follow. This can be usually alleged to be a preventive methodology, fraud analysis and detection is associate certain consequence of associate intensive risk evaluation.

Recognize and classify threats to fraud in knowledge technology and telecommunications sector stereotypically yield the shape of the chances like:

Records showing associate degree inflated rates in calls at associate degree surreal time of day to associate degree uncertain location or far-famed fraud location.

Unusual Dialing patterns showing one variety being referred to as additional of times by external numbers than job out. • Increased calls created in an exceedingly day than the minute's allotted per day, that might indicate an account has been hacked or shared

The following is the proposed method of the model development:

• Different models are tested on the dataset once it is obtained and cleaned.

• On the basis of the initial model performance, different features of the model are engineered and tested again.

• Once all the options area unit designed, the model is made and run victimisation completely different completely different values and victimisation different iteration procedures.

• A predictive model is created that predicts if an insurance claim is fraudulent or not.

• Binary Classification task takes place which gives answer between YES or NO. This report deals with classification algorithm to detect fraudulent transaction. The influence of the feature engineering, feature choice parameter modification area unit explored with an aim of achieving superior prophetic performance with superior accuracy.

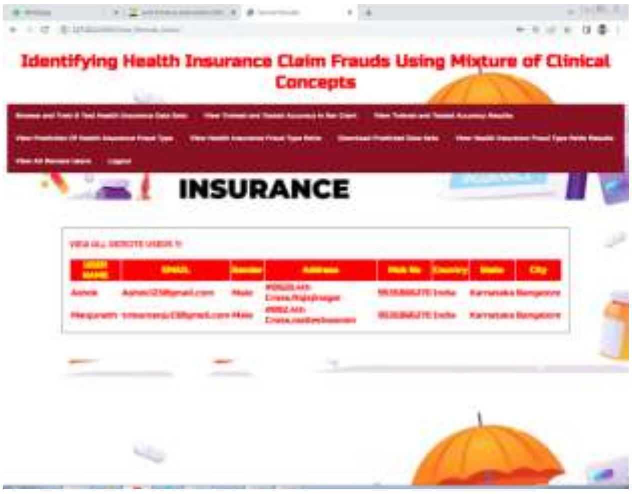
The assorted machine learning techniques area unit utilized in the development of accuracy of detection in unbalanced samples. As a system, the info are divided into 3 completely different segments. These area unit loosely coaching, testing and validation. The algorithmic program is trained on partial set of knowledge and parameters. These area unit later changed on a validation set. This may be studied for evaluation and performance on the particular testing dataset. The high acting models area unit formerly tested with numerous random splits of knowledge.

V. EXPERIMENTAL RESULTS

The effectivenessof the proposed blockchain design in identifying problematic clients and their potential fraudulent claims is the major focus of this section. Then the performance accuracy of the hybridmachine learning technique of ERFSVM in identifying and classifying the variety of fraud types is also discussed in detail in this section. The datasets required for evaluating the performance of the designed blockchain based insurance fraud detection modelis assessed from real world insurance agencie

Identifying Health Insurance Claim Frauds Using Mixture of Clinical Concepts
Home | Remote User | Service Provider
Insurance
Healthcare, Insurance, Fraud, Mixture Model,Clinical Concepts

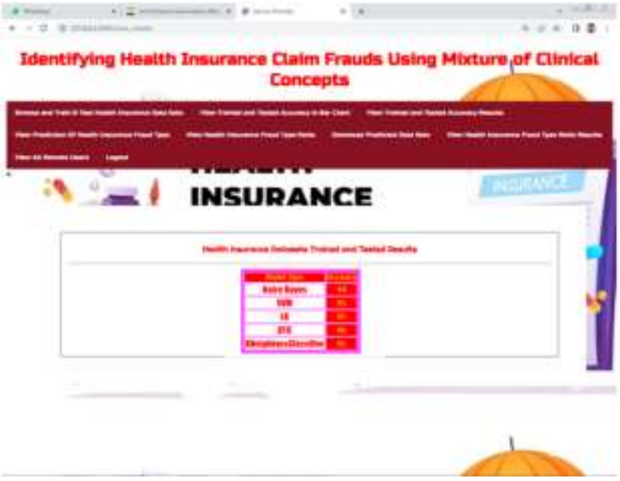
Figure 2: Home Page


Identifying Health Insurance Claim Frauds Using Mixture of Clinical Concepts
Browse and Train & Test Health Insurance Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
View Prediction Of Health Insurance Fraud Type
View Health Insurance Fraud Type Ratio
Download Predicted Data Sets
View Health Insurance Fraud Type Ratio Results
View All Remote Users
Logout
INSURANCE

VIEW ALL REMOTE USERS !!!

USER NAME	EMAIL	Gender	Address	Mob No	Country	State	City
Ashok	Ashok123@gmail.com	Male	[illegible]	[illegible]	India	Karnataka	Bangalore
Manjunath	[illegible]@gmail.com	Male	[illegible]	[illegible]	India	Karnataka	Bangalore

Figure 2: Trained Dataset



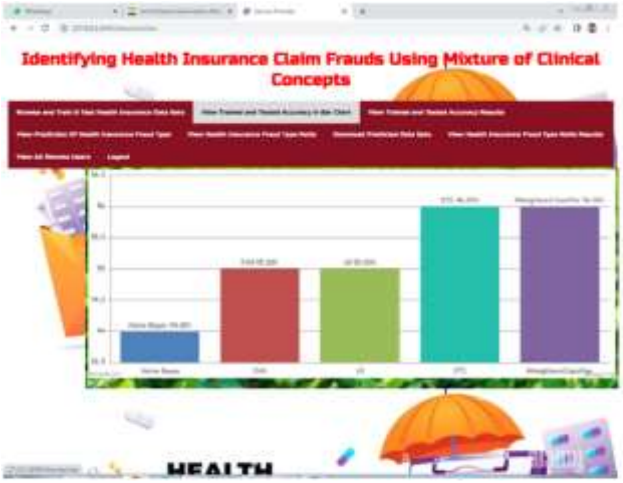
Identifying Health Insurance Claim Frauds Using Mixture of Clinical Concepts

Browse and Train & Test Health Insurance Data Sets | View Trained and Tested Accuracy in Bar Chart | View Trained and Tested Accuracy Results | View Prediction Of Health Insurance Fraud Type | View Health Insurance Fraud Type Ratio | Download Predicted Data Sets | View Health Insurance Fraud Type Ratio Results | View All Remote Users | Logout

Health Insurance Datasets Trained and Tested Results

[illegible]	[illegible]
Naive Bayes	[illegible]
SVM	[illegible]
LR	[illegible]
DTC	[illegible]
KNeighborsClassifier	[illegible]

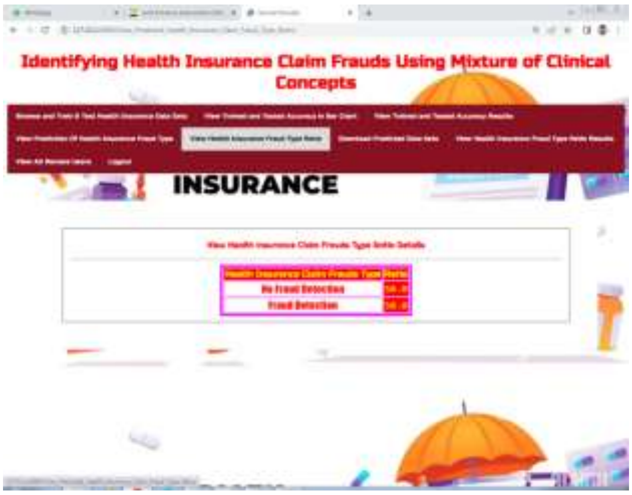
Figure 3: Accuracy Values



Identifying Health Insurance Claim Frauds Using Mixture of Clinical Concepts

Browse and Train & Test Health Insurance Data Sets | View Trained and Tested Accuracy in Bar Chart | View Trained and Tested Accuracy Results | View Prediction Of Health Insurance Fraud Type | View Health Insurance Fraud Type Ratio | Download Predicted Data Sets | View Health Insurance Fraud Type Ratio Results | View All Remote Users | Logout

Figure 4: Accuracy in Charts



Identifying Health Insurance Claim Frauds Using Mixture of Clinical Concepts

Browse and Train & Test Health Insurance Data Sets | View Trained and Tested Accuracy in Bar Chart | View Trained and Tested Accuracy Results | View Prediction Of Health Insurance Fraud Type | View Health Insurance Fraud Type Ratio | Download Predicted Data Sets | View Health Insurance Fraud Type Ratio Results | View All Remote Users | Logout

View Health Insurance Claim Frauds Type Ratio Details

Health Insurance Claim Frauds Type	Ratio
No Fraud Detection	[illegible]
Fraud Detection	[illegible]

Figure 5: Accuracy for Fraud and Non Fraud

CONCLUSION

Modern technologies are moving extremely fast making their ways into various fields of the business. In this respects, the insurance industry does not lack behind the others. The application of statistics in the insurance has a long history. Thus, the fact that insurance companies are actively using data science analytics is not amazing. In essence, the aim of applying data science analytics in the insurance is the same as in the other industries—to optimize marketing strategies, to improve the business, to enhance the income, and to reduce costs. In this paper, we presented several machine learning techniques to analysis the insurance claims efficiently and compare their performances using various metrics

REFERENCE

[1] K. Usage Priya and S. Pushpa, "A Survey on Fraud Analytics Using Predictive Model in Insurance Claims," Int. J. Pure Appl. Math., vol. 114, no. 7, pp. 755–767, 2017.

[2] E. B. Belhadji, G. Dionne, and F. Tarkhani, "A Model for the Detection of Insurance Fraud," Geneva Pap. Risk Insur. Issues Pract., vol. 25, no. 4, pp. 517– 538, 2000, doi: 10.1111/1468-0440.00080.

[3] "Predictive Analysis for Fraud Detection." https://www.wipro.com/analytics/comparativeanalys is-of-machine-learning-techniques-for-

[4] F. C. Li, P. K. Wang, and G. E. Wang, "Comparison of the primitive classifiers with extreme learning machine in credit scoring," IEEM 2009 - IEEE Int. Conf. Ind. Eng. Eng. Manag., vol.1

[5]. Belhadji, E., G. Dionne, and F. Tarkhani, ―A Model for the Detection of Insurance Fraud, Geneva Papers on Risk and Insurance Theory‖, 25: 517-538, may 2012.

[6]. Crocker, K. J., and S. Tennyson,‖ Insurance Fraud and Optimal Claims Settlement Strategies: An Empirical Investigation of Liability Insurance Settlements‖ The Journal of Law and Economics, 45(2), April 2010.

[7]. Kajiamuller, ―The Identification of Insurance Fraud – an Empirical Analysis Working papers on Risk Management and Insurance‖ no: 137, June 2013.

[8]. S. B. Kotsiantis, ―Supervised Machine Learning: A Review of Classification Techniques,‖ Informaticavol 31, pp 249- 268,May 2011.

[9]. Sivarajah U, Kamal M, Irani Z, Weerakkody V (2017) Critical analysis of big data challenges and analytical methods. J Bus Res 70:263–286

[10]. Mishr K (2016) Fundamentals of life insurance theories and applications. In: 2nd ed, Delhi: PHI Learning Pvt Ltd

[11]. The Kaggle Website. [Online]. https://www.kaggle.com/c/ prudential-life-insurance-assessment/data/The Accenture website https://www.accenture.com.