Predicting Health Insurance Claim Frauds Using Supervised Machine Learning Techniques

S.Gayathri[1], Mrs.S.P.Audline Beena[2]
[1]Dept of Computer Science
[2]Assistant professor, Dept of Computer Science
[1, 2] SRI MUTHUKUMARAN INSTITUTE OF TECHNOLOGY, India

Abstract- *The healthcare industry is a complex system and it is expanding at a rapid pace. At the same time fraud in this industry is turning into a critical problem. One of the issues is the misuse of the medical insurance systems. Machine learning and data mining techniques are used for automatically detecting the healthcare frauds. In this paper, we attempt to give a review on frauds in healthcare industry and the techniques for detecting such frauds. With an emphasis on the techniques used, determining the significant sources and the features of the healthcare data we proposed a machine learning model to tackle the issues related to the health insurance claims. The univariate and bivariate analysis are applied on the data to know the features pattern and then proper visualisation of data to know which feature affects the most and a machine learning model is built on the pre-processed data.*

I. INTRODUCTION

Medical coverage in our nation can be a developing segment of India's economy. The Indian wellbeing framework is one in each of the biggest inside the globe. The wellbeing business in Bharat has quickly gotten one in every one of the chief essential parts inside the nation as far as financial addition and employment creation. 100 million Indian family units (500 million individuals) don't get joy from wellbeing inclusion. Arrangements zone unit reachable that supply every person and family. Human services has become a monster use in a large portion of the nations, the tremendous amount of money worried all through this segment had made it as an objective for fakes. To remain with the National Health Care Anti-Fraud Association, human services misrepresentation is respect deliberate trickiness or lie made by somebody, or respect element which can prompt some unapproved benefit to him or his accomplishes.

Social insurance misuse is made once either the provider rehearses are conflicting with sound monetary, business or medicinal practices, partner degrade lead to Associate in Nursing inessential worth or in pay of administrations that don't appear to be therapeutically important or that neglect to fulfil expertly perceived norms for human services. To distinguish the misrepresentation designs different information mining methods are utilized subsequently they are recognized as regular examples.

II. PROPOSED SYSTEM

Healthcare is considered as one of the complex industry and it’s especially need in this pandemic period. Many people are getting infected in this pandemic period and everyone can’t afford payment.In this time insurance comes handy for the people to reduce their financial burden.But there are frauds happening in this also so it is very difficult and some may not get proper response due to this. So there is a need to address this problem. Machine learning is mainly used to tackle these issues so we will be a building a machine learning model where the model is made to train on the previous data and made to learn to find the pattern so that it is capable of analyising and prediction the insurance claim is fraud or not.

III. LIST OFMODULES

- Data Preprocessing:
- Data validation/Cleaning/ Preparing Process:
- Exploring Data Analysis For Visualization:
- Comparing Algorithm With Prediction In The Form
- Prediction Result By Accuracy

DATA PREPROCESSING

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

DATA VALIDATION / CLEANING / PREPARING PROCESS:

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

EXPLORING DATA ANALYSIS FOR VISUALIZATION:

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

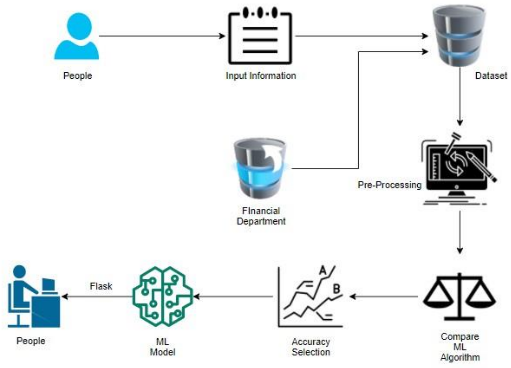
COMPARING ALGORITHM WITH PREDICTION IN THE FORM BEST ACCURACY RESULT

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

PREDICTION RESULT BY ACCURACY

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. It need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.

IV. BLOCK DIAGRAM



V. CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out. This application can help to find the Prediction of Insurance Claim Fraud.

VI. FUTURE WORK

Insurance Claim prediction to connect with cloud. To optimize the work to implement in Artificial Intelligence environment

REFERENCES

[1] Jun Lee, S and Siau, K. (2001), “A review of data mining techniques”, Industrial Management &Data System.

[2] Dallas Thornton, Roland M. Mueller, Paulus Schoutsen, Josvan Hillegersberg, "Predicting Healthcare Fraud in Medicaid: A Multidimensional Data Model and Analysis Techniques for Fraud Detection"

[3] Hossein Joudaki, Arash Rashidian, Behrouz Minaei-Bidgoli, Mahmood Mahmoodi, Bijan Geraili, Mahdi Nasiri & Mohammad Arab, "Using Data Mining to Detect Health Care Fraud and Abuse:.

[4] Pedro A. Ortega, Cristian J. Figueroa, Cristian J. Figueroa, "A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile", URL:https://www.researchgate.net/publication/220704891.

[5] Leonard Wafula Wakoli, Abkul Orto and Stephen Mageto, "Application of The K-Means Clustering Algorithm In Medical Claims Fraud/ Abuse Detection",

[6] Mayank Garg, Akshit Monga, Priyank Bjatt, Anuja Arora, "Android App Behaviour Classification Using Topic Modelling Techniques and Outlier detection using App Permissions"

[7] El Bachir Belhadji, Georges Dionne and Faouzi Tarkhani, "A Model for the Detection of Insurance Fraud",

[8] A. Fiat, and M. Naor, "Broadcast Encryption," in Annual International Cryptology Conference (CRYPTO), 1993, pp. 480–491.

[9] L. Zhang, C. Hu, Q. Wu, J. Domingo-Ferrer, and B. Qin, "Privacy Preserving Vehicular Communication Authentication with Hierarchical Aggregation and Fast Response," IEEE Transactions on Computers, vol. 65, no. 8, pp. 2562–2574, 2016.

[10] L. Zhang, Q. Wu, J. Domingo-Ferrer, B. Qin, and C. Hu, "Distributed Aggregate Privacy-Preserving Authentication in VANETs," IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 3, pp. 516–526, 2017.

[11] J. Liu, J. Li, L. Zhang, F. Dai, Y. Zhang, X. Meng, and J. Shen, "Secure Intelligent Traffic Light Control Using Fog Computing," Future Generation Computer Systems, vol. 78, part 2, pp.

[12] M. Burmester, and Y. G. Desmedt, "A Secure and Efficient Conference Key Distribution System," in Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT), 1995, pp. 275–286.

[13] S. Jiang, "Group key agreement with local connectivity", IEEE Transactions on Dependable and Secure Computing, vol. 13, pp. 143–160.

[14] Q. Wu, B. Qin, L. Zhang, J. Domingo- "Broadcast Encryption and Group Key ", Information Security (ASIACRYPT), 2011, no. 3, pp.326–339, 2016.