A Real time Application for Insurance fraud detection using Machine learning

Kishor Kumar R, M.Tech, PES College of Engineering, Mandya, India, kishukish4741@gmail.com

Prasanna P, Associate Professor, PES College of Engineering, Mandya, India, shivz3@yahoo.com

Abstract – Insurance is something people buy to protect themselves. Insurance fraud is an act of receiving the benefits to which they are not entitled or when a person makes false claims to receive the benefits. The type of insurance fraud are diverse and it occurs in all areas of insurance, among them Auto/vehicle insurance is the most noticeable one. Detecting insurance fraud is the major issue and it is hard through Traditional method as it is more time consuming and expensive, hence it needs an automatic system, so that one can detect whether the claim is genuine or fraud. Through the use of advanced technology like machine learning insurance fraud detection can be done in an efficient way. Here, we focus on identifying the auto/vehicle fraud through the use of machine learning algorithms. Likewise, the performance will be computed by the confusion matrix. Which in turn calculate the accuracy, precision, and recall.

Keywords — Machine learning (ML), K-nearest neighbor (KNN), Naïve Bias (NB), Decision Tree (C4.5)

I. INTRODUCTION

In the present world violations are expanding quickly. Continuously we have numerous classes of wrong doings, wherein financial fraud or crime is the zone of concern. We have 2 kinds of financial frauds, bank fraud and insurance fraud, here we focuses on insurance fraud recognition. The protection business is right now grasping the powerful misrepresentation the board. A couple of individuals cheat the associations for getting pay, anyway others pay premiums. There are two significant groupings for protection, hard protection extortion and delicate protection extortion. Hard protection extortion is characterized as when individuals purposefully counterfeit a mishap. At the point when an individual has a substantial protection guarantee however misrepresents some portion of the case is known as delicate protection extortion. In the event that an organization has great misrepresentation location and limitation management system at that point expanded consumer loyalty, at that point expanded consumer satisfaction. As indicated by the expanded fulfillment, misfortune change costs will be diminished. Presently we have numerous ways for distinguishing cheating claims. The most utilized strategy are investigating the information with its own guidance [1]. So they need complex and tedious inspection, where it manages the diverse area of information. Thus utilizing ML method conquer this whole issue.

ML concerns with development and investigation of framework that can gain from data. For example, ML can be used in E-mail message to learn how to distinguish between spam and inbox messages. A computer program is said to be learn from experience E with respect to some task T and some performance P only if the program performance increases with experience E. ML is a branch of AI which contains statistical, probabilistic, optimization technique that can learn from past experience and discover the pattern from large complex data sets.

For example, we can apply ML technique in predicting student performance based on their behaviors. Student performance depends on many factors such as living locality, SSLC result, PUC result, Family income, Parents education, use of internet, use of mobile, use of bike, use of Social Networking and other habits.

We can predict student performance using ML technique before exams so that we can improve student performance by knowing status of student. ML based technique can be applied to classify the employees in an organization either to be class leave or presence based on their behavior.

There is no automation for protection extortion forecast, current procedure is manual, which is troublesome procedure, where it requires additional time and is costly.

II. RELATED WORK

Machine Learning deals with the development, examination and investigation of calculations that can therefore recognize patterns from information and use it to foresee future data or perform decision making [1]. Machine learning does its usefulness by making models out of it [2]. Machine Learning has become widespread and has its applications in the field of bioinformatics, computer vision, robot locomotion, computational finance, search engine etc.

In real time problems, observations are made on entities associated with a problem so as to make inferences on the

 DOI : 10.35291/2454-9150.2020.0445

target value of those entities. This mapping is encompassed in a predictive model with the help of decision trees. This method of learning is referred to as Decision Tree Learning. This is one of the predictive modelling methods that can be found in the fields of data mining [3], machine learning and statistics.

Scikit-learn is an open source machine learning module in python [4] that is comprised of wide range of classification, clustering and regression algorithms in machine learning.

Logistic regression [5] is a probabilistic view of classification. The approach is used when dependent variable is binary (dichotomous) and help in predicting a discrete outcome. Prediction is done using the probability scores. Logistic regression gives knowledge of the relationships among the variables. In our project we have used logistic regression to find the probability of the student getting placed belonging to different departments by using the BIGLM package in R tool. BIGLM processes the data in chunks at a time to perform the regression optimization and does not need larger memory allocations to the computer because it performs calculations on smaller data sets. The number of lines to be processed at any time is specified by chunk size and we have chosen a chunk size of 100. BIGLM and RODBC package are required to perform calculations using database connectivity and BIGLM will identify the SQL Query as a data frame.

Pal and Pal [6] conducted a study on student data that have information on their academic records and proposed a classification model to find an efficient method to predict student placements. They concluded that Naïve Bayes classifier is the best classification method for use in placements in comparison with Multilayer Perceptron and J48 algorithms. Ramanathan, Swarnalatha and Gopal [7] conducted a study using sum of difference method for students' placement prediction. They used the attributes such as age, academic records, achievements etc. for the prediction.

Arora and Badal [8] conducted a study to predict student placements using data mining. They made predictions on MCA students in Ghaziabad in UP, considering parameters such as MCA result, Communication skills, programming skills, co-curricular activity participation, gender, 12th result and graduation result. They concluded that their model based on decision tree algorithm can assist the placement cell and faculties in identifying set of students that are likely to face problem during final placements. Elayidom, Idikkula and Alexander [9] designed a generalized data mining framework for placement chance prediction problems. They considered the students Entrance Rank, Gender, Sector and Reservation Category to predict the branch of study that is Excellent, Good, Average or Poor for him/her using decision trees and neural networks.

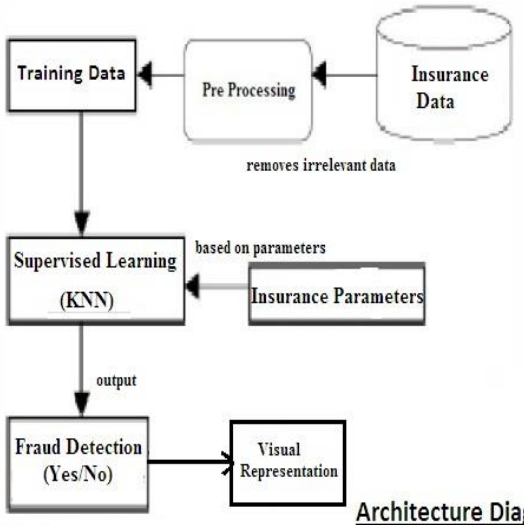
Naik and Purohit [10] made a study to use prediction technique using data mining for producing knowledge about students of MCA course before admitting them.

Hijazi and Naqvi [11] conducted a study to find the factors affecting the academic performance of students. They made use of questionnaires to elicit information from students highlighting factors such as income factor, parents' educational background, Size of the family, regularity of teachers, subject interest created by the teachers and student's interest in co-curricular activities. They used Pearson Correlation Coefficient to highlight the important factors and they found that mother's education and family income played an important role in students' academic performance.

Constraints of existing system

- Difficult to foresee
- Manual investigation
- Less exactness
- Time consuming
- Costly

III. PROPOSED SYSTEM



Architecture Diagram

It's a Real-time application for insurance fraud detection which can be used for the dynamic data in the real world.

Where it is meant for the avoiding the insurance frauds, such that it predicts the insurance fraud claims using the machine learning algorithms. It uses the previous claims data for processing and for claim fraud detection. The data which is been collected is been performed with the necessary data cleansing to generate the valid data which is necessary and remove the irrelevant information.

Insurance parameter which is been listed below is now applied to classification rules. The system uses a classification rules such as Naïve Bayes, Decision tree, K-Nearest Neighbor algorithm for fraud detection and we

 DOI : 10.35291/2454-9150.2020.0445

compare between these algorithms so that we can declare which among these algorithms are efficient and effective and finally the result can be represented graphically for a better view.

We have set standards and peculiarities for making the crude information. The standard and inconsistencies are relied on the arrangement of a trait. In this paper, concentrating on auto/vehicle-related protection misrepresentation claims. With the goal that the crude information which comprise of following subtleties

Parameters Used

- ✧ Guarantee number
- ✧ Premium number
- ✧ Guarantee occurrence start date
- ✧ Guarantee occurrence start time
- ✧ Guarantee open date
- ✧ Guarantee loss date
- ✧ Guarantee event location name
- ✧ Guarantee amount
- ✧ Policy premium
- ✧ Part market cost
- ✧ Guarantee on vehicle
- ✧ Tally of client correspondence
- ✧ Are Guarantee document submitted

At that point change this crude information into the changed informational collection, and took care of into the classification algorithms like decision tree, K-nearest neighbor, and naive bayes. The changed informational index contain following characteristics.

- ✧ Variation between guarantee event date and guarantee report date
- ✧ Variation between guarantee report date and guarantee open date
- ✧ Variation between premium effective and premium occurrence date
- ✧ Are guarantee archive submitted
- ✧ Part cost distinction
- ✧ Credit rating
- ✧ Policy premium
- ✧ Premium event start time
- ✧ Tally of client correspondence
- ✧ Guarantee occasion area name

By using this quality, check the each knowledge concerning the cases. If there should be an occurrence of ordinary cases, the contrast guarantee event date and guarantee report date is under seven. The all report are submitted with confirmation. The contrast between strategy powerful date and guarantee event date is under five days. Moreover check the cases on same vehicle during all course of action periods.

The significant steps in classification is given beneath

- ✓ Recognize and gather information with essential elements.
- ✓ Perform essential information cleansing.
- ✓ Isolate information into preparing set and testing set.
- ✓ Select suitable algorithm.
- ✓ Execute algorithm with preparing set.
- ✓ Assess algorithm with testing set.
- ✓ If acceptable, use the classifier for new dataset.

Functionality

- ✓ Admin gets login to the application by using his Id and password
- ✓ Admin adds different cities, system accessed from different locations.
- ✓ Admin adds insurance branches based on cities, system meant for n insurance branches.
- ✓ Admin creates branch Incharge and sets id and password.
- ✓ Branch Incharge can get login by specifying id and password (given by admin)
- ✓ Branch Incharge can upload the dataset (claims dataset) of previous years.

System predicts whether new insurance claim is genuine or fraud based on old dataset using ML algorithm

IV. Conclusion

This paper simply draws out the property of machine learning. Where it enables the insurance industry to predicts the insurance fraud claims. System predicts fraud claims using machine learning techniques, such that it can be used as a real time application. Proposed system uses previous claims data for processing and for claim fraud prediction. System leads organization effectively & efficiently to deliver key results and ensure high standards. Appropriate parameters used for fraud prediction. Faster decision making the auto\vehicle protection extortion is the most well-known sort of protection misrepresentation, which should be possible by counterfeit mishap guarantee. System predicts fraud claims. Financial frauds can be avoided and reduced. We can differentiate with the algorithms like Decision tree, Naïve Bayes, K-nearest neighbor and it gives the compared result between these algorithms in a graphical representation for a better view.

In future we can work with more algorithms, lastly compute which gives more exactness and also System can be enhanced by adding feedback module, where users can post feedbacks to clarify their doubts.

References

[1] 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

 DOI : 10.35291/2454-9150.2020.0445

[2] 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 2017 International Conference on circuits Power and Computing Technologies [ICCPCT]

[3] Kajia muller, "The Identification of Insurance Fraud – an Empirical Analysis Working papers on Risk Management and Insurance" no: 137, June 2013.

[4] Clifton Phuna damminda, Alahakoon, and Vincent phuaMinority Report in Fraud Detection:Classification of Skewed Data". Sigkdd Explorations, Volume – 6, Issue – 1, sep 2011

[5] Mladeni´c, D & Grobelnik, M. "Feature Selection for Unbalanced Class Distribution and Naive Bayes." In Proceedings of the 16th International Conference on Machine Learning. pp. 258–267.may 2011

[6] Pérez, J. M, Muguerz J, Arbelaitz, O., Gurrutxaga, I., & Martín, J. I., 2005, "Consolidated Tree Classifier Learning in a Car Insurance Fraud Detection Domain with Class Imbalance",Pattern Recognition and Data Mining. Springer-Verlag. S. Singh et al. (Eds.), 381-389

[7] Dionne, G, Giuliano, F. & Picard, "Optimal Auditing for Insurance Fraud", CIRPEE Working Paper No. 03-29, april 2010

[8] Tennyson, S. & Salsas-Forn, P, "Claims Auditing in Automobile Insurance - Fraud Detection and Deterrence Objectives", The Journal of Risk and Insurance, Vol. 69, No. 3, pp. 289-308,sep 2009

[9] Viaene, S, "A Comparison Of State-Of-The-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection", The Journal of Risk and Insurance, Vol. 69, No. 3, pp. 373- 421, march 2011

[10] Bhowmik, R., "Detecting Auto Insurance Fraud by Data Mining Techniques", Journal of Emerging Trends in Computing and Information Sciences, Volume 2 No.4, April 2011

[11] Sokol., B. Garcia, J. Rodriguez, M. West, and K. Johnson, "Using Data Mining to Find Fraud in HCFA Health Care Claims" Topics in Health Information Management,1: 1-13.april 2009.

[12]Barse.E, Kvamstrom.H & Jinson E "Synthesizing test data for fraud detection system". Pro of the 19th annual computer security application, 384- 395, sep2013

[13] Ezawa K & Norton, "Constructing Bayesian network to predict uncollectible telecommunications accounts", IEEE Expert October: 45-51, may2010

[14]Major J& Riedinger, "A Hybrid knowledge/statistical based system for the detection of fraud.", journal of risk and insurance 69(3):309- 324.

[15]He H, Wang J, GRACO W and Hawkin S,"Application of neural networks to detection of medicall fraud,Expert system with applications,13,329-336,jan 2010

[16]S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," Informatica vol 31, pp 249-268,may2011 2017 International Conference on circuits Power and Computing Technologies [ICCPCT]

[17]Viachaslau Sazonau, "Implementation and Evaluation of a Random Forest Machine Learning Algorithm," University of Manchester, Oxford Road, Manchester, M13 9PL, UK Dec 2011.

 DOI : 10.35291/2454-9150.2020.0445