Survey on Fraud Prediction for an Application Using Data Mining

Sudhir Lawand
Dept. Computer Engineering
Vidyalankar Polytechnic MS
India

Dr. Umesh Kulkarni
Dept. of Computer Engineering
Vidyalankar Institute of
Technology.MS India

Abstract— **this research is carried out to propose a model on prediction of fraud in insurance. In the worldwide time, Insurance frameworks had quickly made a considerable measure of gigantic advancement in our general public. Because of the expanded worry in everyday life, the development of interest of protection expanded. Information mining causes protection firms to disclosure helpful examples from the client database. Insurance companies are uncovered different sorts of misrepresentation and every year endure generous misfortunes. The essential objective is extortion location strategies for insurance utilizing information mining approach amid the previous years. Discoveries in this examination show that information mining strategies like SVM, Naïve Bayesian and Random Forest decision tree have been connected most generally to give preparatory answers for the troubles inborn in the recognizable proof and arrange of fake accident protection information. This paper presents fraud prediction method to predict fraud patterns from data. We use Naïve Bayesian Classification, SVM and Random Forest Model algorithms. This enables to build a more robust model with increased accuracy.**

Keywords—***aggregation, Regression, Data Mining.***

I. INTRODUCTION

These days that an assortment of catastrophes undermines individuals' life and riches, insurance is an alluring decision to convey these dangers from guaranteed individuals to insurance companies. Insurance is an agreement between the guarantor and the guaranteed that suggests if misfortunes which are definite in arrangement happen, policyholder will be fiscally remunerated by insurer up to a foreordained bound rather than protection premium that have been paid. Through an insurance policy, the policyholder and the insurance company agree to the terms on which certain risks will be covered. If both parties behave honestly and share the same information, the premium paid by the insured, as well as the compensation due by the insurer, will adequately reflect the probability of the loss event occurring and the implicit estimated loss (plus the insurer's profit margin). However, the behavior of the insured is not always honest.

Insurance companies show an assortment of administrations each of which can cover a piece of misfortunes. Next to quick advancement of data innovation, the measure of put away information in insurance companies' databases is developing quickly. These huge databases contain basically helpful business data. In the other hand, finding advantageous concealed data in these databases and furthermore distinguishing appropriate models are not all that clear. One of the productive techniques for finding shrouded information and discovering designs on enormous databases is data mining. In late two decades many looks into are exhibited around extortion in accident protection.

Belhadji and Dionne in their examination researched on key factors about insurance fraud. They counseled with space specialists here. They figured contingent likelihood of misrepresentation for each factor and afterward they decided the most critical variables and furthermore false misfortunes. Cummins and Tennyson considered on a comparable point. They right off the bat refined the expenses of insurance services and the rate of cost expanding with comparing administration change. At that point they distinguish proficient factors in insurance inflation and uncommonly the variables that develop with cost expanding. In the cases with insusceptible guidelines and ensuring the simplicity of obligation payback, their proposed strategy was effective.

The moral issue behind people's view on insurance fraud provides an attempt at modeling consumer behavior through a consumer's utility function, by assessing the expected value of committing fraud. Most types of fraudulent behavior can be construed as information asymmetry leaned towards the policy holder. In other words, fraud occurs when the insured possesses more information than the insurer.

II. EXISTING SYSTEM

In foreign countries, researchers have established the suitable insurance fraud mining model for different types of insurance, and proposed the anti-fraud recommendations. M Bajec Proposed an expert detection system, which used a new evaluation-iterative evaluation algorithm (IAA) and effectively detected the vehicle insurance fraud.[2] Adrian Gepp provided a comparative analysis on the prediction of a series of data mining technology in analysis of the real data of the insurance fraud Capelleveen introduced the health insurance fraud detection method based on the outlier detection of data mining technology and used the method to detect the suspicious behavior of medical service providers, and a case study was conducted on the data of dental medical claims.[2] Lin K C used data mining technology for manual inspection, repeatedly verified a large number of periodic data and feedback them to the data mining model. The integration of data mining technology can be fed back to the different stages of business to set up warning system as soon as possible. G G Sundarkumar combined the K reverse nearest neighbor and one class support vector machine (OCSVM) to correct the data imbalance problem, and used the data to verified the validity of the model.[2]

At present, many researches on insurance fraud use the existing models, such as Logistic regression model, expert system, game theory and so on. But these models are all established on the assumption that the distribution of claims data is roughly balanced. There widely exists unbalanced data in the real world and insurance claims data is not balanced. This survey paper use the Random Forest Model for mining and prediction of fraud.[2]

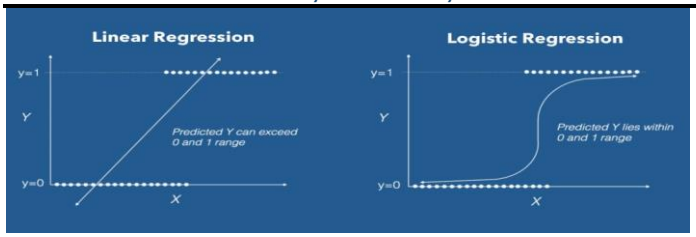

Fig 2.1 Linear regression Vs. Logistic Regression Graph

III. *PROBLEM DEFINATION*

Insurance fraud is gradually spreading in the global scope, and mining insurance fraud is more and more concerned by the society. Concerning that the number of samples in the actual insurance claims data is not balance and the amount of data is large, the real data of a insurance company were selected to establish the random forest fraud mining model based on the theory of insurance fraud mining. The data were processed to screen the index and the importance analysis of each input variable to the output variable was obtained. The error of the model was analyzed. Finally the method has been verified by empirical analysis. The empirical results show that: compared with the traditional model, the insurance fraud mining model introducing Random Forest is suitable for large data sets and unbalanced data. It can be better used for the classification and prediction of the insurance claims data and mining fraud rules. And it has the better accuracy and robustness.[2] In order to reduce the fraud in insurance industry, we propose prediction model using data mining techniques which contains Random Forest algorithm along with Naïve Bayesian and Support vector machine (SVM), it provides better accuracy. Developing Fraud predicting system will help of insurance industry to some extent.

IV. *PROPOSED SYSTEM*

Hence, to increase the accuracy, we are proposing new model for the Insurance Fraud Prediction. Proposed system consists of different modules working together to achieve robust and more accurate system than its predecessors.Proposed system contains following algorithms which will be applied on training data set as well as on testing data set

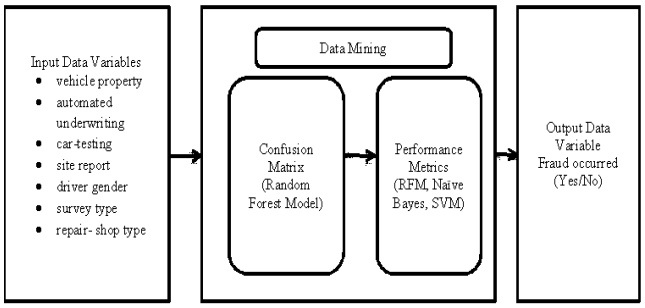

Fig.4.1 System Architecture of Fraud Prediction System

Phases of System Development:

Step 1: Load Training dataset which is in the form of ARFF Format that is Attribute-Relation File Format. (ARRF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software)

Step 2: Data Preprocessing is done where it checks whether we have all attributes available in our dataset or not? If not then the record gets deleted.

Step 3: Data set Training (which gets tested with by our Random Forest Model algorithm.)

Step 4: Load Testing Data set which is in the form of ARFF Format that is Attribute-Relation File Format. (ARRF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software)

Step 5: Calculation of prediction using Random Forest, Naïve Bayes and Support Vector Machine.

Step 6: Compare Output of all algorithms.

Step 7: Find more efficient algorithm.

Following Categorical variables are targeted to predict fraud has occurred or not:

Input Variables	Layered
Vehicle property	Business car is set to 1; The private car is set to 2; Agency car is set to 3.
Automated Underwriting	Automated Underwriting
Car-testing	Not test is set to 1; Have been test is set to 2; The exemption is set to 3
Site Report	Yes is set to 1; No is set to 0.
Driver Gender	Man is set to 1; Woman is set to 0
Survey type	Not survey is set to 1; The first scene is set to 2; Fill the survey site set to 3
Repair Shop	The first kind of factory is set to 1; The second kind of factory is set to 2; The third kind of factory is set to 3; The special service station is set to 4
Fraud	Yes is set to 1, No is set to 0.

Table 4.2 Categorical variables

V.*DATA MINING ALGORITHMS*

RANDOM FOREST

The derived feature vectors for the second subset of training data are used to construct a Random Forest classifier. 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. Each bootstrap sample of feature vectors is obtained by repeated random sampling with replacement until the size of the bootstrap sample matches the size of the original training subset. This helps to reduce the variance of the classifier (reducing the classifier's ability to over-fit the training data). When constructing each decision tree, only a randomly selected subset of features is considered for constructing each decision node. Of the k randomly selected features to consider for constructing each decision node, the yes/no condition that best reduces

the Gini impurity measure g of the data is selected for the next node in the tree:

$$g = 1 - P(Fraud)^2 - P(NotFraud)^2$$

-Equation (1)

The Gini impurity measure is largest when the classifier is most uncertain about whether a feature vector belongs to the fraud class. To support cost sensitive learning, we used a balanced stratified sampling approach to generate bootstrap samples for training the classifier. For training each tree, a bootstrap sample is drawn from the minority class and a sample of the same size is drawn (with replacement) from the majority class. This effectively under-samples the majority class. Each tree in the Random Forest classification model casts its vote for a class label: fraud or not fraud. The proportion of votes for the fraud class is the probability that a randomly selected tree would classify the feature vector as belonging to the fraud class. This is interpreted as the probability of a feature vector belonging to the fraud class.

The parameter Mean Decrease Accuracy is used to measure the degree of prediction accuracy reduction of random forest, when the value of a variable is changed to a random number.[2] The larger the value is, the greater the importance of the variable. The parameter Decrease Gini Mean used the Gini index to calculate the heterogeneity influence of each variable observation at each node of the classification tree, thus the importance of the variables can be compared. Similarly, the larger the value is, the variable is more important.[2]

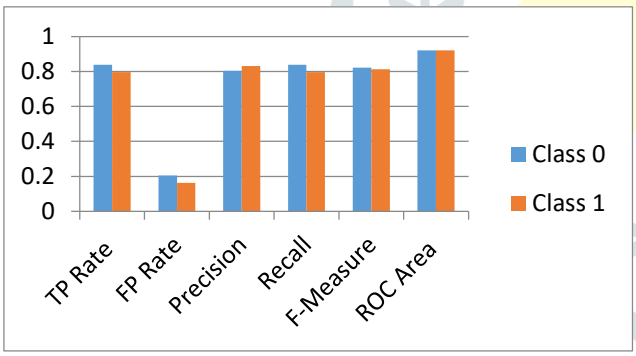

Fig.5.1Accuracy by RFM

NAÏVE BAYESIAN

Naïve Bayes (NB) is a supervised machine learning method that uses a training dataset with known target classes to predict the future or any incoming instance's class value. Naïve Bayes classifier is noted as a powerful probabilistic method that exploits class information from training dataset to predict the class of future instances. Naïve Bayes method assumes that the presence or absence of any attribute of a class variable is not related to the presence or absence of any other attributes. This technique is named "naïve" because it naïvely assumes independence of the attribute. The classification is done by applying "Bayes" rule to calculate the probability of the correct class. Despite their naïve design and oversimplified assumptions, Naïve Bayes classifiers have good performance in many complex real world datasets.

Naïve Bayes rule calculate the probability of the correct class which is the particular attributes of the transactions. Naive Bayes theory is calculated as:

$$Prior\ Probability\ Z = \frac{Number\ of\ Z\ instances}{Total\ number\ of\ instances}$$

$$Likelihood\ of\ Y\ given\ Z = \frac{Number\ of\ Z\ in\ vicinity\ of\ Y}{Total\ number\ of\ Z}$$

-Equation (2)

In the theory the final classification is produced by combining both information (likelihood, priori), to form a posterior probability which is called Bayes rule.

$$Posterior = (Prior * Likelihood)/(Evidence)$$

-Equation (3)

It has a good performance with small amount of training data. It is used to solve both binary classification problem and multi class classification problem

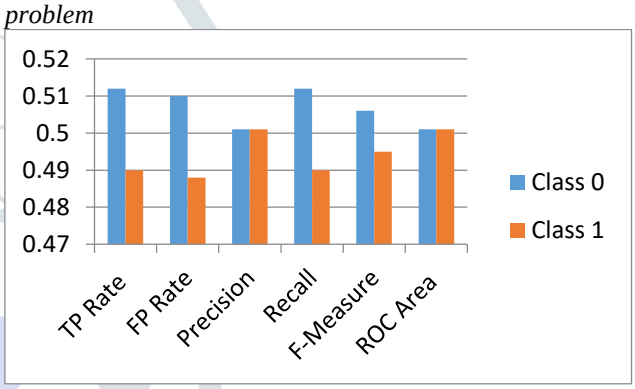

Fig. 5.2 Accuracy by SVM

SUPPORT VECTOR MACHINE (SVM)

SVM models built in this study use four different kernel functions: polynomial, sigmoid, radial basis and linear kernel functions. There are two parameters associated with RBF kernel: C and Ɣ. Regularization parameter (C) controls the trade-off between maximizing the margin and minimizing the training error term. Increasing the value improves the classification accuracy (or reduces the regression error) for the training data, but this can also lead to overfitting. RBF gamma should normally take a value between 3/k and 6/k, where k is the number of input fields.

For example, if there are 12 input fields, values between 0.25 and 0.5 would be worth trying. Increasing the value improves the classification accuracy (or reduces the regression error) for the training data, but this can also lead to overfitting. Gamma is used in Polynomial or Sigmoid kernels. Increasing the value improves the classification accuracy (or reduces the regression error) for the training data, but this can also lead to over fitting. Bias is enabled only if the kernel type is also set to Polynomial or Sigmoid. It sets the constant coefficient value in the kernel function. The default value 0 is suitable in most cases. Degree is enabled only if Kernel type is set to Polynomial. It controls the complexity (dimension) of the mapping space.

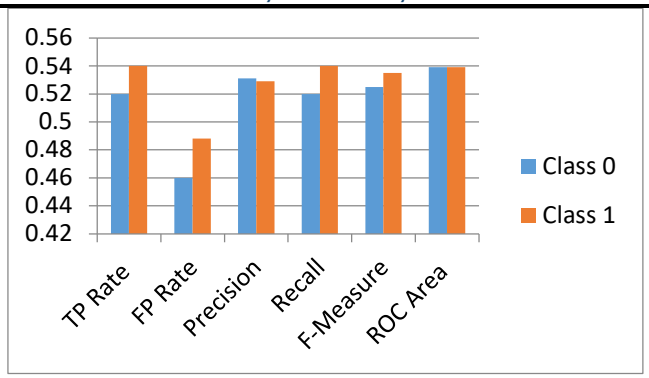

Fig. 5.2 Accuracy by NB

Comparison of all algorithms can be seen as follows

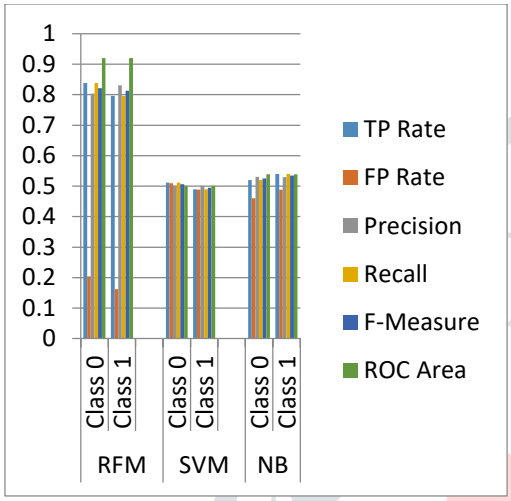

Fig. 5.3 Overall Results

VI. *ADVANTAGES AND DISADVANTAGES*

A. Advantages

- *This Random Forest Model data mining algorithm calculates accuracy at 81%.*
- *Enhances the efficiency of the mining process, by designing Random Forest Model framework.*

B. Disadvantages

- *Random Forest Model algorithm encounters highest computational cost.*
- *Support Vector Machine algorithm calculates accuracy at 51%.*

VII. *CONCLUSION*

We studied problem of fraud prediction in automobile insurance industry. To predict automobile insurance fraud we have used Random Forest Model, Naïve Bayesian and SVM algorithms. Random Forest Decision Model algorithm is applied on training dataset. We have applied Random Forest Model,SVM and Naïve Bayesian on testing dataset to predict the fraud.We have derived performace metrics such as precision,accuracy and recall from the confusion matrix. This perforrmance metrics is reliable in many automobile insurance fraud prediction system.

VIII. *FUTURE SCOPE*

Future research & development may continue to be focused on further improvements of their liability & responsiveness. Since the system focuses only on the improving accuracy with limited dataset further it can improved for larger dataset.

IX. REFERENCES

[1] Subelj L, Stefan Furlan, Bajec M. "An expert system for detecting insurance fraud using social network analysis [J]". Expert Systems with Applications, 2011, 38(1): 1039-1052.[2]

[2] Yaqi Li, Chun Yan, Wei Liu, Maozhen Li. "Research and Application of Random Forest Model in Mining Automobile Insurance Fraud" 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD).

[3] Adrian Gepp, Holton Wilson, Kuldeep Kumar and Sukanto Bhattacharya. "A Comparative Analysis of Decision Trees Vis-à-vis Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection" Journal of Data Science 2012(10), 537-561.[2]

[4] Capelleveen, Guido Cornelis van. "Outlier based predictors for health insurance fraud detection within U.S. Medicaid [D]". [the requirements for the degree of Master of Science in Business Information Technology at the University of Twente], 2013.[2]

[5] Lin K C, Yeh C L, Huang S Y. "Use of Data Mining Techniques to Detect Medical Fraud in Health Insurance [J]". Applied Mechanics & Materials, 2013, 284-287(2): 1574-1578.[2]

[6] Sundarkumar G G, Ravi V. "A novel hybrid under sampling method for mining unbalanced data sets in banking and insurance [J]". Engineering Applications of Artificial Intelligence, 2015: 368-377.[2]

[7] Chen, R., Chiu, M., Huang, Y., Chen, L.:" Detecting Credit Card Fraud by Using Questionnaire-Responded Transaction Model Based on Support Vector Machines". In: IDEAL2004, 800--806(2004).

[8] Brause, R., Langsdorf, T. , Hepp, M.: "Neural Data Mining for Credit Card Fraud Detection". In: 11th IEEE International Conference on Tools with Artificial Intelligence (1999).

[9] SAS, e-Intelligence Data Mining in the Insurance industry: "Solving Business problems using SAS Enterprise Miner Software". White Paper(2000).

[10] Fawcett, T., Flach, P. A.: "A response to web and Ting's on the application of ROC analysis to predict classification performance under varying class distributions". Machine Learning, 58(1), 33—38 (2005).

[11] Ormerod T., Morley N., Ball L., Langley C., Spenser C.: "Using Ethnography To Design a Mass Detection Tool (MDT) for the Early Discovery of Insurance Fraud". Computer Human Interaction, Ft. Lauderdale, Florida (2003).

[12] Viaene S, Derrig R A, Baesens B , Dedene G, "A comparison of state-of-the-art classification techniques for expert insurance claim fraud detection", The Journal of Risk and Insurance ,2002, pp.373–421.

[13] Shaw M J, Subramanian C, Tan G W ,Welge M E, "Knowledge management and data mining for marketing, Decision Support System",2001, pp.127–137.[22]

[14] Turban E, Aronson J E, Liang T P , Sharda "Decision Support and Business Intelligence Systems", Eighth ed, Pearson Education,2007.[22]

[15] Maes, S., Tuyls, K., Vanschoenwinkel, B. & Manderick, B.: "Credit Card Fraud Detection using Bayesian and Neural Networks". Proc. of the 1st International NAISO Congress on Neuro Fuzzy Technologies (2002).

[16] Ahmed S R,(2004). "Applications of data mining in retail business", Information Technology: Coding and Computing, pp. 455-459.

[17] Weatherford, M.: "Mining for Fraud". In: IEEE Intelligent Systems (2002).

[18] Tan P, Steinbach M , Kumar V, "Introduction to Data Mining", First ed.Addison-Wesley Longman Publishing Co., Inc, 2005.[22]

[19] Berry M J , Linoff G S, "Data Mining Techniques: for Marketing, Sales and Customer Relationship Management", Second ed.Wiley, New York, 2004.[22]

[20] Agyemang M , Barker K , Alhajj R, "A comprehensive survey of numeric and symbolic outlier mining techniques", Intelligent Data Analysis, 10,2006, pp.521–538.[22]

[21] Cahill, M., Chen, F., Lambert, D., Pinheiro, J. & Sun, D.: "Detecting Fraud in the Real World". Handbook of Massive Datasets 911-930(2002).

[22] www.ijocit.org

[23] cisjournal.org