International Journal of Recent Development in Computer Technology & Software Applications
Vol. 1, Issue 1- 2017

FRAUD DETECTION IN TELECOMMUNICATION USING DATA MINING

DV BHISE*, PRAJAKTA VIVEK JOSHI*

ABSTRACT

Because of the dramatic increment of Fraud which brings about loss of billions of dollars worldwide every year. Fraud location includes observing the conduct of populaces of clients to assess, distinguish, or keep away from unfortunate conduct. This paper show the idea of data mining and current methods utilized for subscription fraud detection. This paper gives a far reaching comprehensive review of data and distinctive systems to identify Fraud.

KEYWORDS: Data Mining, Fraud Detection, Neural Network Ensemble, Naïve Bayes Classification, Decision Tree.

INTRODUCTION

Fraud has been very common in our society, and affects private enterprises as well as public entities. However, in recent years, the development of new technologies has also provided criminals more sophisticated way to commit fraud and has required more advanced techniques to detect and prevent such events. Telecommunication Company worldwide suffers from customers who use the provided services without paying. The estimated losses amount to several billions of dollars in uncollectible debt per day. Even though this is a small percentage comparing to the Telecom Operators' revenue, it is still a significant loss.

Detection and prevention of frauds is one of the main objectives of the telecommunication industry. However, the volume of data being produced these days is expanding at sensational rate. In this way, extricating valuable information from such information accumulations is an imperative and testing issue. Keeping in mind the end goal to construct such a non-trifling model, numerous inquiries about were done on the plausibility of utilizing the Data Mining (DM) strategies which originates from the need of breaking down high volumes of information gathered by the media transmission organizations (client information, unbilled calls, and so forth.) and identified with various types of exchanges between the organization and its customers.[1] Media transmission segment is a wide division with a great many clients. This segment comprehensively has two sorts of clients – household and business. Associations are given to the local clients at a reasonable rate while the business associations are given at a similarly higher rate as the utilization scale is higher in the last case. There are situations where the associations are purchased under local classification yet utilized are on a business scale. This present reason's significant misfortune to the part, as the associations, when purchased under business class, will yield a more noteworthy wage to the area.

*Department of computer Science and Engineering, Anuradha Engineering College, Chikhali.
Correspondence E-mail Id: editor@eurekajournals.com

There are many methods to identify such fraudulent.

DATAMINING

1. Data mining is a procedure which finds valuable examples from huge measure of information.
2. The objective of this procedure is to discover designs that were beforehand obscure.
3. Once these examples are discovered they can additionally be utilized to settle on specific choices for improvement of their organizations.
4. Four things are required to information mine adequately:
   i. High quality information
   ii. The "right" information
   iii. Adequate test estimate
   iv. Right apparatus
5. There are numerous apparatuses accessible to an information mining specialist. These innovations incorporate artificial neural system, relapse and choice tree.

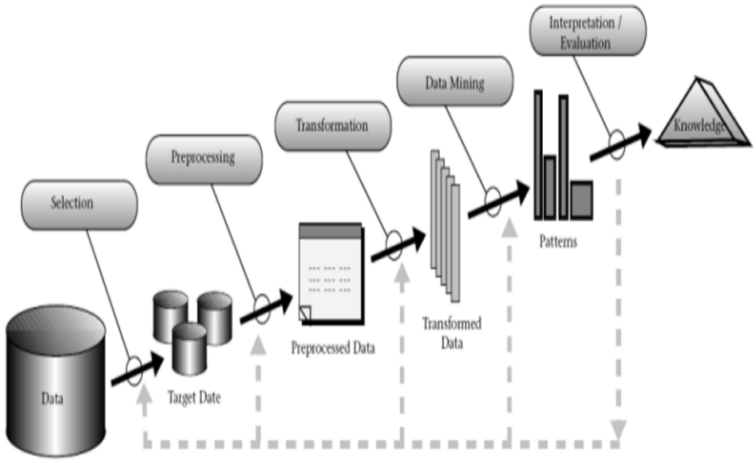

Figure 1.data mining process

FRAUD

The unscrupulous use of telecommunication facilities orservices provided by the telecom operators is termed as telecommunication fraud.

Fraud is costly to the network carrier as network capacity iswasted and also revenue loss occurs.

SUBSCRIPTION FRAUD

1. The fundamental type of telecommunication fraud extortion that has occurred to that is subscription fraud (that is, the inability to pay for administrations used).Detection of such fraud is troublesome in light of the fact that it is effectively conceal as awful obligation.
2. In subscription fraud, guilty parties regularly distort their personality with a specific end goal to stay away from installment. Deception of character is likewise imperative on the grounds that in the most extreme cases, subscription fraud isn't an end in itself yet rather a stage for a few different fakes.
3. Subscription fraud is the demonstration of utilizing telephone utilities without the intension of paying. Client buy in for postpaid administrations, when the time desires client to pay they don't. The telecommunication transmission industry has misfortune billions of dollars to membership fraud.

FRAUD DETECTION TECHNIQUES

There are various techniques to detect the fraud, they are as follows:-

1. Neural Network

2. Decision Tree
3. Probability based method

NEURAL NETWORK

An artificial neural network (ANN), often just called a “neural network”(NN), is a mathematical model or computational model based on biological on neural network.Neural network is a set of connected input/output units and each connection has a weight present with it.During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples.Neural networks have the remarkable abilityto derive meaningfrom complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.

The neural network ensemble technique can be used to detect the fraud in telecom sector.

DATA DESCRIPTION

In a telecommunication company the administrator keeps record of each prepared by the framework. These occasions are recorded in CDRs (Call Details Records), produced naturally and are utilized for charging purposes. Each CDR has data with respect to an arrangement of occasions, voice calls or SMSs, for instance. Normally, the CDR is a content record containing data organized by a predefined set of requested fields isolated by a predefined character. Each line of CDR record is an occasion handled in the administrator framework. In any case, there is an arrangement of fields that, due or their significance, for charging and rating reasons for existing, are generally basic to all CDR structure.

a) Number – identifies the originator of the event.
b) Number – identifies the receiver of the event.
c) Event Date – the date the event stated.
d) Event Type - identifies the type of the event, for example: 1 (Voice), 2 (SMS), 3 (MMS), 4 (Data);
e) Event Amount - measure of the event, for example, in a voice call the event amount is 124 seconds, in a SMS the event amount is 45 characters.
f) Cell ID - identifies the network cell that processed the event. The information contained in the CDRs will be the input for all the future work. The study of the contents of CDRs is not a novelty. They were first created with billing purpose, but know they are used with different purposes of great importance to the operators, for instance, discovering user communities.[2]

FRAUD PREDICTION/DETECTION

An irregular harsh subspace based neural network group strategy is utilized in the advancement of the model to distinguish membership extortion in versatile telecoms. The technique includes making various preparing subsets from the first preparing set. For this investigation, four diverse preparing subsets were utilized to make four classifiers. An anticipated target is acquired by averaging the yields of the four classifiers.

The membership extortion location framework show is accomplished by utilizing successions of call detail records (CDRs), which contain the points of interest of each post-paid clients on the system. The data created for charging likewise contains utilization conduct data significant for misrepresentation location.

Membership misrepresentation anticipation is accomplished by utilizing clients' business precursors which have been displayed in the ANN question. A portion of the presumptions for foreseeing a false membership are as per the following: - Applicant's ID is like that of a fraudster - Applicant's contact telephone number is like a fraudster - Applicant's address, age,

sexual orientation and conjugal are like a fraudster

A fraudster regularly utilizes the principal line he utilized as a part of submitting extortion as the contact telephone number amid application for another line. A specific number is provided a few times.

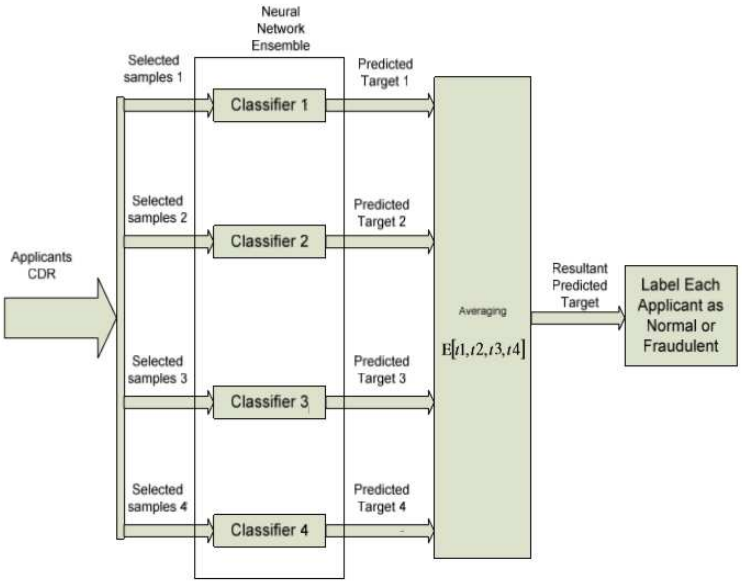

Figure 2.Subscription Fraud Detection Model

The created ANN demonstrate is tried with an alternate 100 examples that are not some portion of the preparation set used to make the ANN show. Out of the testing tests, 80 tests are fake applications while 20 tests are ordinary applications. The contributions to the ANN show are CDR factors 1 to 8. The model is utilized to reenact the contributions to the foresee kind of use (Fraud=1; Normal=0). The anticipated ANN banners are contrasted with the real banners with register:

- True-Positives (TP): extortion tests delegated misrepresentation
- False-Negatives (FN): extortion tests delegated typical
- True-Negatives (TN): ordinary examples delegated typical
- False-Positives (FP): ordinary examples delegated misrepresentation

The discovery rates are figured as takes after:

TP (%) =fraud tests named misrepresentation/ Total number of extortion tests × 100

TN (%) =normal tests named typical/Total number of ordinary examples × 100

FP (%) =normal tests named misrepresentation/ Total number of typical examples × 100

FN (%) =fraud tests named typical/Total number of misrepresentation tests × 100

The precision of the model is registered as:

Precision =Number of right forecast/Total number of test tests × 100 [2]

The exhibitions of the created NN classifiers were tried with another informational collection that was not some portion of the preparation set. The testing set comprises of the statistic information of a few candidates for the post-paid versatile lines. A portion of the candidates with the aim of submitting extortion could supply previous number used to confer misrepresentation, a similar address, financial balance subtle elements or character card like the one provided in the past misrepresentation. The testing information have been marked to recognize the potential

extortion candidates from the typical candidates. Each of the NN classifiers was utilized to mimic the testing information to foresee the mark for every one of the candidates in the information. The anticipated names of the NN outfit classifier are acquired by averaging the names got from the four NN classifiers. The quantity of right orders and wrong characterizations are figured for every one of the classifiers regarding TP, FN, TN and FP.

DECISION TREE

1. Decision trees are used for deciding between several courses of action.
2. Decision trees are predictive model, used to graphically organize information about possible options, consequences and end value.
3. Decision trees are easy to understand and interpret. A decision tree is also a method for expressing algorithms.
4. The goal of a decision tree is to create a model that predicts the value of a target variable based on several input variables.
5. From investigations the telecommunication industry discovered a way for detecting subscription fraud.
6. The procedure is illustrated below;

Two new variables were created; Pay Duration (in days) and ratio.

Pay Duration= Date of checking customer's fraud status (today's date)-last date of payment by customer Ratio= balance/credit limit. Where balance=Outstanding balance + unbilled

Investigations showed that whenever the pay duration is greater than 31 days the customer is likely to be involved in BAD DEBT.

But if it's less than or equal to 31 days, the ratio of the customer is checked, if the ratio is less than 1 then the customer's fraud status is NOT FRAUD, but if the ratio is greater than 1, the pay duration is checked again, if the pay duration is less than or equal to 0days then the customers fraud status is SUBSCRIPTION FRAUD, but if its greater than 0days it is BAD DEBT.

EXAMPLE

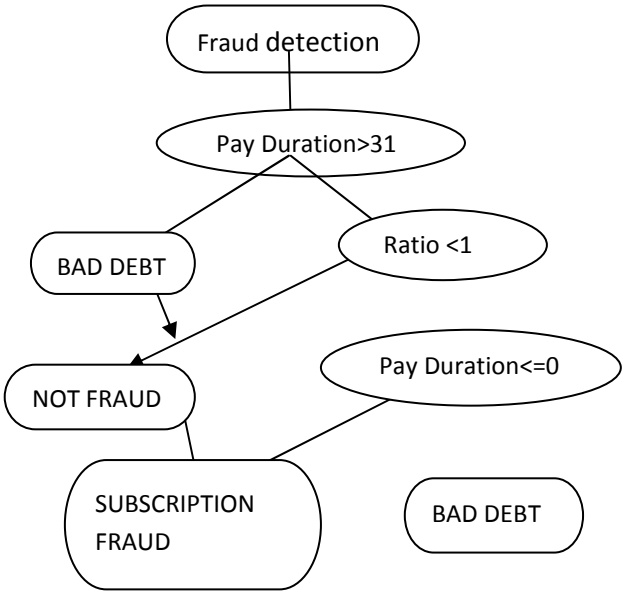

Figure 3.Descision tree for fraud detection

Table 1.CDR for detecting fraud

Name	Phone no.	Customer Segment	Outstanding Balance	Unbilled	Credit Limit	Last Date of payment	Fraud Status
John	12345678	Consumer	$836	$1003	$667	31/8/2011	Subscription fraud
Gate Corporation	12435566	Corporate	$3082	$2027	$6667	8/8/2011	Not fraud
Michael	23767868	Consumer	$70	$300	$500	3/7/2011	Bad Debt
Binary Enterprise	34556578	Enterprise	$10000775	$708000	$6000000	16/8/2011	Bad debt

PROBABILITY BASED METHOD

The probability based method is one of the method for detection of fraud. Is a notable strategy for characterizing records in light of, either, numerical or unmitigated. This technique is utilized for computing likelihood esteems for each record utilizing the ascribed that portray the record. The estimations of each record are numerical. Thus, it executes the equation for nonstop esteems in Naïve-Bayesian technique.

For continuous values probabilities are calculated and difference between the normal and fradulent customer is calculated by using the KL divergence.

PROBABILITY CALCULATION

For continuous numeric values, the probability calculation goes by the following formula,

p(x,Ci) = g(x,μ(Ci),σ(Ci))

x represent thetuple under consideration.

Where μ(Ci) = mean of all attribute in the table.

σ(Ci) =Deviation of all attribute for the class (Ci)

$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$[Mean]

$= \sqrt{\left[\frac{1}{-1}\sum_{=0}(-)^2\right]}$[StandardDeviation]

$() = \frac{1}{\sqrt{2\pi\sigma}} \frac{-(-)^2}{2^2}$[Normal Distributiion] [3]

The mean and standard deviation, calculated for a set of records.

KL DIVERGENCE

After calculating the probability of the user records, reference values are found for calculating reference probability based on which the divergence is calculated. The formula for finding divergence between two probability distributions using KL divergence is

$$\int p(x) * \log\left[\frac{p(x)}{q(x)}\right]$$

Where p(x) represents user probability and q(x) represents reference probability. This equation is straightforwardly connected where the client likelihood is substituted in the place of p(x) and the reference likelihood is substituted in the place of q(x). The disparity, in this manner, computed yields the expected contrast to separate a presumed false client from an ordinary client. The disparity esteem for an example informational index with the distinction in values for ordinary and suspected clients.[3]

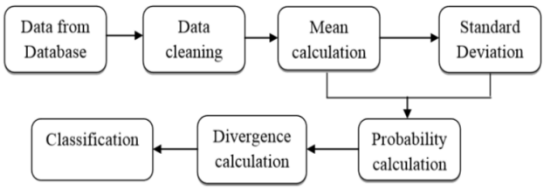

Figure4.Flow diagram of process state

EXPERIMENTAL RESULT

Table 2.Sample Data Set

S.no	Name	Duration	Day time	Night time	Week days	Week ends	Total calls
1.	David	172	58	4	34	28	62
2.	Joseph	567	87	22	67	42	109
3.	Ram	1800	300	0	146	154	300
4.	Lalitha	76	23	12	15	20	35
5.	Gayathri	932	193	172	245	120	365
6.	Suvashini	117	44	16	37	23	60
7.	Varshini	2100	57	24	52	29	81
8.	Kayalvizhi	473	50	17	35	32	67
9.	Sathish	1971	114	59	119	54	173
10	Sriram	396	44	9	26	27	53

RESULT

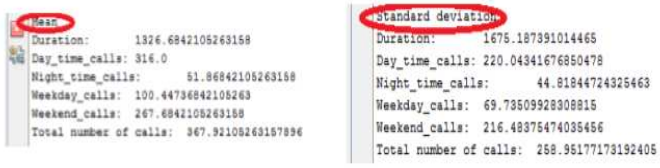

Fig.2 Mean values and standard deviation of the attributes

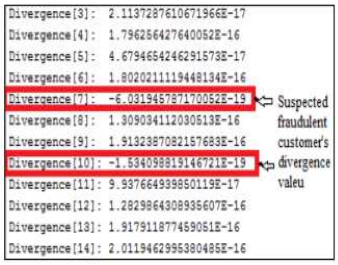

Figure 5.Final Outcome with Fraudulent customer details

CONCLUSION

In this topic we have explained the fraud detection techniques in telecommunication using data mining.It explain about detection of subscription fraud in telecommunication by using various techniques like neural network, probability based method, decision tree.In neural network, we are training the original data set and then ANN objects are created. By averageing the prediction of all classifiers we can train the NN.In probability based method we are calculating the probability of the customer to be fraudulent. In Decision tree creating a model that predicts the value of a target variable based on several input variables.

FUTURE SCOPE

We believe that there is room for further research, especially concerning the design and implementation of new algorithms for rule-discovery for fraud. Among the directions we are pursuing. Optimization methods such as Simulated Annealing and Genetic Programming can be utilized to create a more robust selection process, with a better chance of finding the "best" rule-set.

REFERENCES

[1]. Data mining in Telecommunication by Gray M. Weiss, Fordham University

[2]. Fraud detection in mobile Telecommuni cation (www.IJIRSET.com) Fayemiwo Michel Adebisl and Olasoji Babatunde O.

[3]. Data Mining Approach For Subscription-Fraud, Detection in Telecommunication Sector.

[4]. P. Saravanan, V. Subramaniyaswamy, N. Sivaramakrishnan, M. Arun Prakash and T. Arunkumar. Data Mining: Task, Tools, Techniques and ApplicationsS.D.Gheware, A.S.Kejkar, S.M.Tondare.

[5]. Data mining techniques for Fraud DetectionAnita B. Desai, Dr. Ravindra Deshmukh Sinhgad Institute of Management & Computer Application# Nahre Pune India Ahmednagar College, Ahmednagar, Dist-Pune India.

[6]. Fraud Detection In Communications Networks Using Neural And Probabilistic Methods Michiaki Taniguchi, Michael Haft, Jaakko Hollmén, Volker Tresp Siemens AG, Corporate Technology Department Information and Communications D-81730 Munich, Germany.

[7]. Applications of Neural Networks In Data Mining M. Charles Arockiaraj Head & Asst. Professor, Computer Science Department, Arakkonam Arts and Science College, Arakkonam-631003.

