CYBER FRAUD DETECTION

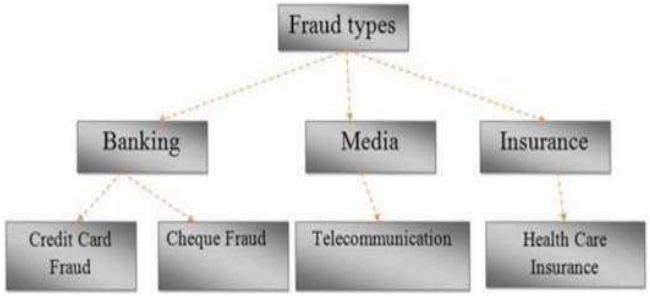
ABSTRACT:

In last decade there's a fast advancement in e-commerce and online banking, the utilization of online transaction has raised. As online transaction become more popular the frauds associated with this are also rising which affects a lot to the financial industry. Cyber fraud is an issue that has wide spreading its consequences in both the finance industry and daily life. To overcome these issues various fraud detection techniques and algorithms are planned, data mining is employed by several corporations related to fraud detection. This paper presents review of various fraud detection techniques and discuss the problems concerning money dataset employed in fraud detection technique, which affects the accuracy of the fraud detection and additionally propose a hybrid approach that uses data processing algorithmic rule at over one stages that contains the info level yet as network level that helps to boost the accuracy of fraud detection. This paper uses three techniques namely-Decision Tree, Random Forrest and at last Multiple Linear Regress. Through these techniques, this paper reveals which algorithm is suitable for detecting Cyber Fraud and which algorithm provides the highest accuracy.

KEYWORDS- cyber fraud, decision tree, Random tree, data mining, linear regression

I. INTRODUCTION

In This World of Advancement, the Cyber Frauds are rapidly increasing at a great speed, every single day we hear about Credit card frauds, Phishing attacks, Spamming, Social Engineering attacks and many more. Cyber fraud is a serious issue that has been wide spreading its consequences in both the finance industry and day to day life. Fraud will cut back confidence in IT trade, and have an effect on people's value of living because of the many increase in fraud it resulted in loss of billions of dollars throughout the world every year; many trendy techniques in detection fraud are evolved and applied to several business fields however Fraudsters(like hackers,etc) are regularly purification their strategies, and per se there's a demand for detection strategies to be ready to evolve consequently.



Various types of fraud

II. BACKGROUND

A helpful and vital structure for applying soft computing to fraud detection is to use them as way for classifying suspicious transactions or samples. Studies show that assessing a pair of transactions might scale back losses up to 1 chronicles of the entire prices of all purchases, with additional

assessments leading to smaller loss however with a rise in auditing prices. A multi-layer pipeline approach are often used with every step applying an additional rigorous methodology to discover fraud. Soft computing are often used to expeditiously separate out additional obvious fraud cases within the initial levels and leave the subtler ones to be reviewed manually.

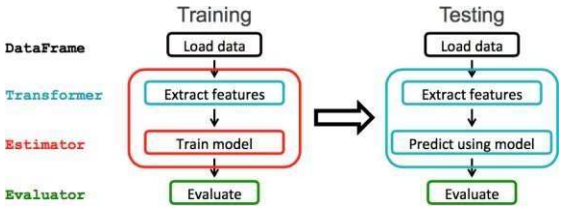
For this paper, a dataset is taken from kaggle.com which is having synthetic datasets generated by the PaySim mobile monesimulator. This dataset comprises of millions of instances out of which 1508 instances were used and simulated in weka software. It is having Nine different attributes out of which first 8 acts as features and 'isFraud' is the output which actually reveals whether the transaction done is fraud or not. If the value of isFraud=0 then no fraud has ocuured or vice versa.

type	amount	nameOrig	oldbalaneOrg	newbalanceOrig	nameDest	oldbalaneDest	newbalanceDest	i
1	9839.64	1231006815	170136	160296.36	1.98E+09	0	0	
1	1864.28	1666544295	21249	19384.72	2.044E+09	0	0	
4	181	1305486145	181	0	553264065	0	0	
3	181	840083671	181	0	38997010	21182	0	
1	11668.1	2048537720	41554	29885.86	1.231E+09	0	0	
1	7817.71	90045638	53860	46042.29	573487274	0	0	
1	7107.77	154988899	183195	176087.23	408069119	0	0	
1	7861.64	1912850431	176087.23	168225.59	633326333	0	0	
1	4024.36	1265012928	2671	0	1.177E+09	0	0	
5	5337.77	712410124	41720	36382.23	195600860	41898	40348.79	
5	9644.94	1900366749	4465	0	997608398	10845	157982.12	
1	3099.97	249177573	20771	17671.03	2.097E+09	0	0	
1	2560.74	1648232591	5070	2509.26	972865270	0	0	
1	11633.8	1716932897	10127	0	801569151	0	0	
1	4098.78	1026483832	503264	499165.22	1.635E+09	0	0	
3	229134	905080434	15325	0	476402209	5083	51513.44	
1	1563.82	761750706	450	0	1.731E+09	0	0	
1	1157.86	1237762639	21156	19998.14	1.877E+09	0	0	
1	671.64	2033524545	15123	14451.36	473053293	0	0	
4	215310	1670993182	705	0	1.1E+09	22425	0	
1	1373.43	20804602	13854	12480.57	1.345E+09	0	0	
5	9302.79	1566511282	11299	1996.21	1.974E+09	29832	16896.7	
5	1065.41	1959239586	1817	751.59	515132998	10330	0	
1	3876.41	504336483	67852	63975.59	1.405E+09	0	0	
4	311686	1984094095	10835	0	932583850	6267	2719172.89	
1	6061.13	1043358826	443	0	1.558E+09	0	0	
1	9478.39	1671590089	116494	107015.61	58488213	0	0	
1	8009.09	1053967012	10968	2958.91	295304806	0	0	

III. WORKING PRINCIPAL

Neural network or fuzzy logic primarly based on fraud detection depends completely on the human brain operating principal. Neural network technology has created a pc capable of suppose. As human brain learn through past expertise i.e. through coaching and useits knowledge or expertise in creating the choice in everyday life downside an equivalent technique is applied with the cyber fraud detection technology.

For example-When a specific person uses his debitcard/credit card, there is a fix pattern in which the credit card is mend to use. By using the previous one or two year data neural network is trained about the specific pattern of using a credit card by a particular consumer. Like the neural network is train on information regarding to various categories about the card holder such as occupation of the card holder, income, occupation may fall in one category, whereas in another category info concerning the big quantity of purchased area unit placed, these info embody the quantity of enormous purchase, frequencies of enormous purchase, location wherever these kind of purchase are take place etc. within a fixed time period.In spite of pattern of debit card/credit card use neural network are trained regarding the various debit card/credit card fraud face by a selected bank antecedently. Based on the pattern of uses of debit card/credit card, neural network create use of prediction formula on these pattern information to classify that weather a selected group action is dishonest or real.



IV. DATA VISUALISATION

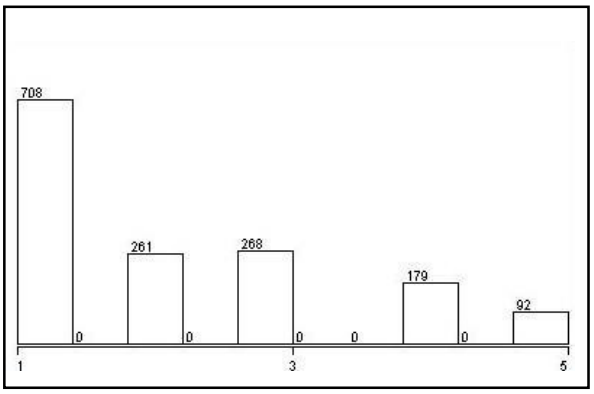
1. Selected Attributes

Name: type Type: Numeric

Missing: 0(0%) Distinct: 5

Unique: 0 (0%)

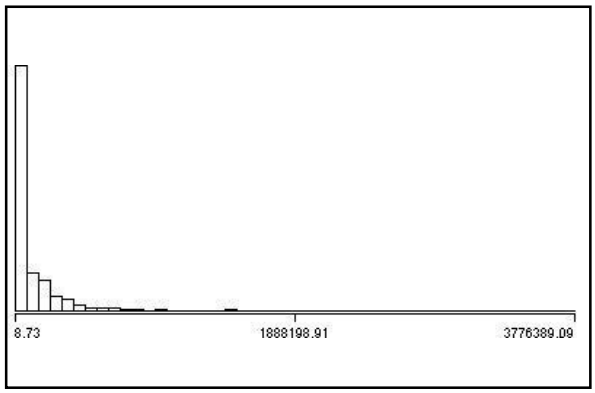
Statistic	Value
Minimum	1
Maximum	5
Mean	2.129
StdDev	1.287



2. Selected Attributes

Name: Amount Type: Numeric
Missing: 0(0%) Distinct:1503
Unique: 1498(99%)

Statistic	Value
Minimum	8.73
Maximum	3776389.09
Mean	115527.786
StdDev	261133.653



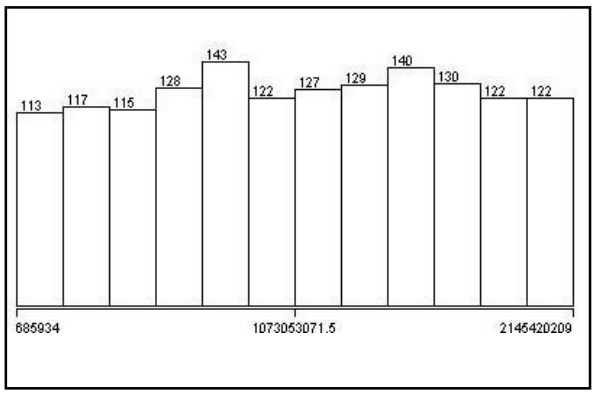
3.Selected Attributes

Name: nameOrig Type: Numeric

Missing: 0(0%) Distinct: 1508

Unique: 1508 (100%)

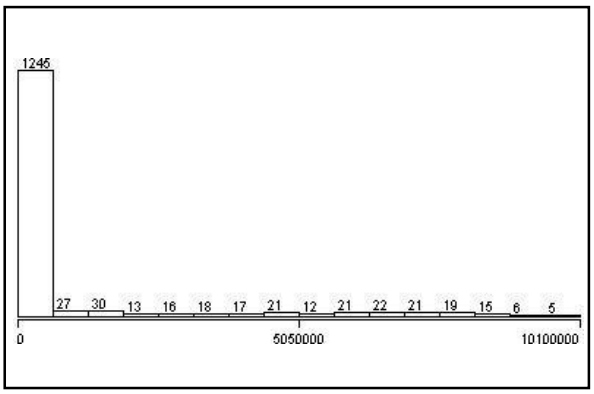
Statistic	Value
Minimum	685934
Maximum	2145420209
Mean	1087711694.629
StdDev	606919762.482



4. Selected Attributes

Name: oldbalanceOrg Type: Numeric
Missing: 0(0%) Distinct: 1073
Uniquw: 1057(70%)

Statistic	Value
Minimum	0
Maximum	10100000
Mean	859633.857
StdDev	2091965.614

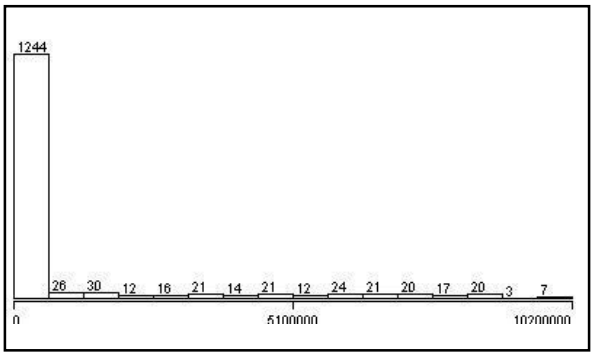


5. Selected Attributes

Name: newbalanceOrig Type: Numeric

Missing: 0(0%) Distinct:870
Unique: 866(57%)

Statistic	Value
Minimum	0
Maximum	10200000
Mean	878100.226
StdDev	2138442.327



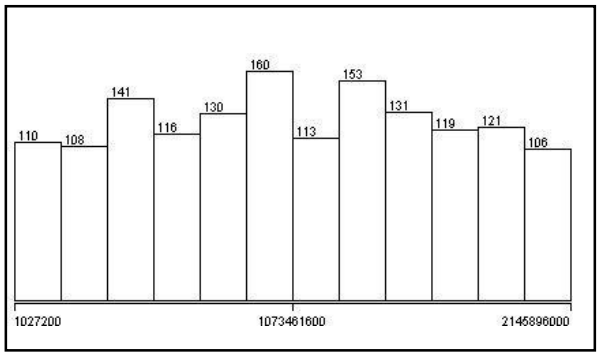
6. Selected Attributes

Name: nameDest Type: Numeric

Missing: 0(0%) Distinct: 900

Unique: 778 (52%)

Statistic	Value
Minimum	1027200
Maximum	2145896000
Mean	1069374977.174
StdDev	595967161.197



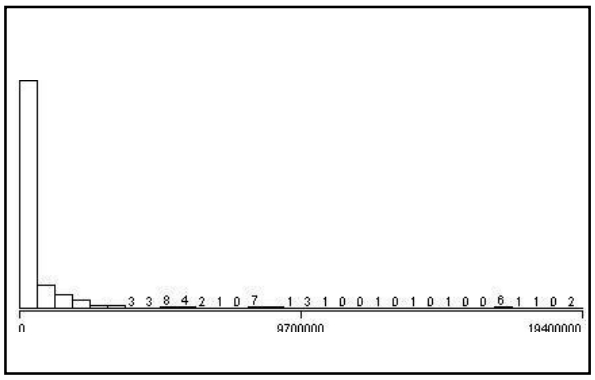
7. Selected Attributes

Name: oldbalanceDest Type: Numeric

Missing: 0(0%) Distinct: 782

Unique: 777 (52%)

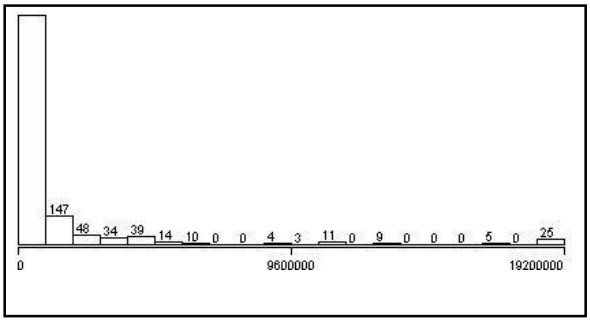
Statistic	Value
Minimum	0
Maximum	19400000
Mean	646206.84
StdDev	1982849.919



8. Selected Attributes

Name: newbalanceDest Type: Numeric
Missing: 0(0%) Distinct: 132
Unique: 34 (2%)

Statistic	Value
Minimum	0
Maximum	19200000
Mean	1071707.834
StdDev	3116895.472



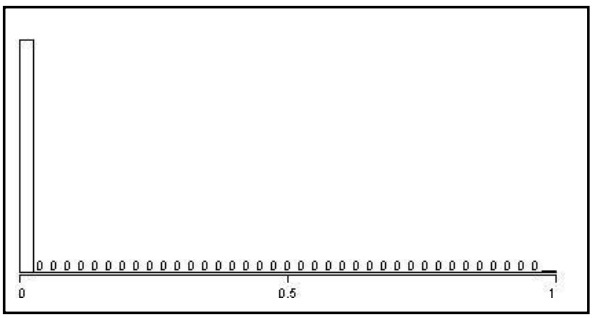
9. Selected Attributes

Name: isFraud Type: Numeric

Missing: 0(0%) Distinct: 2

Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.007
StdDev	0.085



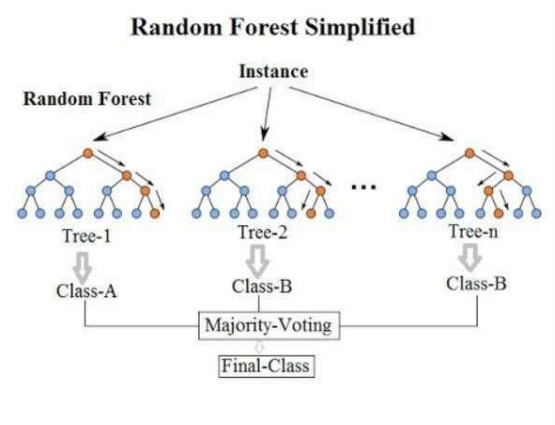
V. ALGORITHM USED:

In order to implement this project we have tried different Algorithms i.e. Decision Tree, Random Forest, Multiple Linear Regression.

ALGORITHM	DECISION TREE	RANDOM FOREST	MULTIPLE LINEAR REGRESSION
ACCURACY	0.98344	0.9984	0.9867

As out of them, all the accuracy of Random Forest is the maximum. So, we have used Random Forest Algorithm in the project

VI. PRACTICAL IMPLEMENTATION:



As shown in the above Figure, Random Forest is a collection of Random Decision Tree. So, in order to implement the Random Forest Algorithm we need to find the Decision Tree first for random values of Datasets.

Now in order to find Decision Tree we need to find,

1. Entropy
2. Gini Index
3. Information Gain

Step1: We have to find the Entropy of every attribute both parent as well as children attribute in the dataset with respect to the output attribute.

Entropy can be calculated by using the formula

$E(X) = \sum - p(x)\log_2 p(x)$ for all $x \in X$

Step2: Now we have to find the information gain for all the Target values with the formula

Information Gain(IG) = E(Target) - E(Target, Attribute)

Step3: Select the Attribute which is having higher Information Gain because,

Attribute Priority(Feature) $\propto$ Information Gain

Hence the Attribute with high Information Gain will come at the top.

In order to implement Random Forest,

Step1: Select some random sets of Data from the available Dataset.

Step2: Apply the Decision Tree Algorithm on it.

Step3: Get the output by input the testing data in all Decision Tree.

Step4: Output with the majority will be the Result.

VII. RESULT

type	Amount	name Orig	oldbalance Orig	newbalance Orig	name Dest	oldbalance Dest	newbalance Dest	isFraud
5	8531.44	1.48E+09	2080	0	1.96E+08	31817.35	40348.79	**0**

Summary:

- Correlation coefficient 0.6144
- Mean absolute error 0.0101
- Root mean squared error 0.068
- Relative absolute error 70.4241%
- Root relative squared error 79.913%
- Total number of Instances 1508

VIII. CONCLUSION

For this paper, we have collected synthetic dataset generated by "PaySim mobile money Simulator" from kaggle.com .By Data Visualization, the irrelevant data has been removed and ranking of the data is done as per their influence in predicting the output. Furthermore, we took three different training algorithms namely- Decision Tree, Random Forest, Multiple Linear Regression for training the dataset and out of them Random Forest predicted with maximum accuracy. Out of the whole dataset , about 80% of data were used for the training purpose and the remaining 20% for the testing purpose.

From our result, we concluded that Higher the deviation in transaction behaviour higher will be the chances of fraud. There are several reasons which may lead to higher deviation in transaction behaviour-

1. Sudden increment in the transaction amount with respect to average transaction behaviour.
2. Information of the recipient is not listed in the database.
3. If the Difference between old Balance Destination and new balance destination is very high.

On the above criteria, the algorithm can predict whether a transaction is fraud or not.

IX. REFERENCES

Madhuvanthi, K & Kailasanathan, Nallakaruppan & Senthilkumar, N.C. & Siva Rama Krishnan, S. (2019). Car sales prediction using machine learning algorithmns. International Journal of Innovative Technology and Exploring Engineering. 8. 1039-1050.

Nallakaruppan M.K , P. Ilango , N. Deepa , Anand Muthukumarappan. Clustering of Wireless Sensor Network Data. Research J. Pharm. and Tech. 2017; 10(1): 73-82.

Kailasanathan, Nallakaruppan & Mohan, Senthilkumar & Thirumalai, Chandra Segar & Suraj, K.A. & Gopu, Magesh. (2014). Accident avoidance in railway tracks using adhoc wireless networks. 9. 9551-9556.

Aleskerov, E., Freisleben, B. & Rao, B. (1997). CARDWATCH: A Neural Network-Based Database Mining System for Credit 11 Card Fraud Detection. Proc. of the IEEE/IAFE on Computational Intelligence for Financial Engineering, 220-226.

Bentley, P. (2000). Evolutionary, my dear Watson: Investigating Committee-based Evolution of Fuzzy Rules for the Detection of Suspicious Insurance Claims. Proc. of GECCO2000.

Tsikerdekis M, Zeadally S. Online deception in social media. Commun ACM. 2014 Sep;57(9):72–80.

Zheng Y, Zhang X, Hou B, Liu G. Using Combined Difference Image and-Means Clustering for SAR Image Change Detection. Geosci Remote Sens Lett IEEE. 2014;11(3):691–5.

Oneto L, Bisio F, Cambria E, Anguita D. Statistical Learning Theory and ELM for Big Social Data Analysis.

Deshmukh, A. & Talluru, T. (1997). A Rule Based Fuzzy Reasoning System for Assessing the Risk of Management Fraud. Journal of Intelligent Systems in Accounting, Finance & Management 7(4): 669-673.

Dorronsoro, J., Ginel, F., Sanchez, C. & Cruz, C. (1997). Neural Fraud Detection in Credit Card Operations. IEEE Transactions on Neural Networks 8(4): 827-834.

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