A SURVEY PAPER ON FRAUD DETECTION USING DATA MINING

INSTITUTE OF COMPUTER SCIENCE AND TECHOLOGY

GANPAT UNIVERSITY – AHMEDABAD

Aniket Roy

KEYWORDS

Data mining, Fraud detection, K-Nearest Neighbour Algorithm , Decision Tree, Data Mining Techniques, association Rule, Neural Network, Bayesian Network, Anomaly Detection, Credit Card Fraud Detection

ABSTRACT

Fraud is a million-dollar business, since the evolution of the internet, many small and large companies have moved their business to the internet, online fraud is most well-known among every other kind of fraud. In online fraud, fraudsters take credit card numbers and purchase online products and vital things from the web. This paper represents various sorts of fraudsters, how information examination helps in extortion investigation, various difficulties, various kinds of misrepresentation, and how to handle those misrepresentations utilizing data mining methods. The two principal kinds of data mining techniques supervised and unsupervised learning strategies utilized in misrepresentation examination have been seriously checked on with the assistance of existing writing.

INTRODUCTION:

Needless to say, frauds have been prevalent since ancient times, and in this digital age, they have evolved into many forms that we counter in our everyday lives, and have become a skilful act of crime. We have either been a part of or witnessed frauds of many kinds from insurance and telecommunication to credit and debit card frauds. With the number of frauds and the people unaware of it are still on the rise, it becomes necessary for the need to counter these problems, especially from the point of view of the computer field. Frauds usually have mainly 3 different types of bodies behind them: Criminal offenders, organized criminal offenders and offenders who commit fraud because of their poor financial situation. Out of the three, the last one remains still hard to be dealt with. To counter these problems, one solution that the computer field has come up with, is the use of the Artificial Intelligence and data mining. The basic theory revolves around training the AI using machine learning to identify whenever there has been a fraud made.

The word “mining” in Data Mining refers to the extraction of “knowledge” from a large data set. This can be done using machine learning and neural networks, and various other methods including regression and clustering.

Data mining can be divided into supervised learning and unsupervised learning. Both have their own merits. With enough training, the AI can learn by identifying a pattern in the regular purchases of a customer. If say a purchase has been made from a location different than the usual, or the product is different than the usual, the AI can detect an irregularity or an “anomaly” and send an alert message to whoever is in charge of the transactions or whoever is connected to it, so that the matter can be further investigated.

Data Mining being the youngest field that it is, is open to many different methods untested and tested alike, and making them turning points for the field. The same goes for its use in fighting frauds.

Data Mining.

[1] A huge measure of information has produced practically every period of everyday exercises. Powerful techniques and instruments for investigating information are significant today as the volume of information is high. It helps analysts in acquiring valuable information from gigantic volumes of information. Data Mining can be presented as the way toward embracing a bunch of information investigation apparatuses and techniques to recognize examples and connections in information for summing up them into valuable data. This interaction is otherwise called Information Disclosure from Information.

Applications of Data Mining.

[1] The information is all over. All the business measures, tests, information today exercises, ventures, and perceptions create billions of information inside a solitary second. It is difficult to acquire the greatest advantages from accepting information. Data Mining comes as the most ideal method of separating information and getting valuable sequences. It has become a critical technique in numerous enterprises.

The retail business is profoundly profited by information mining. The retail business requires constant data of client requests, deals of their item, nature of deals, patterns, and client inclusion, and so on. Data mining can be utilized to create this sort of data utilizing gathered information. Furthermore, it helps in market basket analysis. Market basket analyses in the sense are the hypothesis that if you purchase a specific gathering of things, you are more disposed to purchase another gathering of things This technique is used to understand the buying behaviour of a customer. Product recommendations are also possible for data mining. Can recognize the preferences of the buyer and promote more preferred items. Customer loyalty data can be also processed with data mining [1].

The telecommunication industry is another industry that handles immense measures of information. Telecom organizations use data mining to improve their marketing strategies, distinguish misrepresentation, and enough deal with their organizations as same as the Retail business [1].

Scientific domains such as science, geography, meteorology, and stargazing create gigantic measures of information during investigations and tests. Data mining can be used for examining massive biological data sets for signs about regulatory influences on particular genes, by finding DNA segments negotiating such influence. Environmental data and whether data are also used to predict future environmental changes and disasters [1].

Data mining in sociology and social examinations is well known now because tremendous measures of correspondence information produce while everyday exercises. The world is presently open to utilizing different modes, including www, news, websites, articles, pages, online gatherings, surveys, tweets, messages, commercials, and web-based media [1]

What is Cyber Credit-Card Fraud or no Card Present Fraud?

[2] Recent and current scholars investigating credit-card fraud have divided credit-card fraud into two types; online credit card fraud (or no card-present fraud) and offline credit card fraud (card-present fraud) 1. Online credit-card misrepresentation (in this paper is cyber credit card extortion) is submitted with no presence of a credit card except for all things considered, the utilization of credit card data to make an electronic buy for merchandise and ventures on the web. Offline credit-card extortion is submitted with the presence of a credit card which by and large has been taken or falsified and, in this way, utilized at a nearby store or an actual area for the acquisition of certain products or administrations. Be that as it may, to characterize cyber credit card misrepresentation, it is a situation where the credit card data of a credit-card owner has been taken, or at times, substantial credit-card information has been interestingly created (very much like charge card organizations or guarantors do) and consequently utilized for electronic installment on the web or employing the phone. Much of the time, no I.T or PC ability might be needed to submit online credit card extortion as a result of the various procedures in which charge card data can be taken by cyber fraudsters.

Who are the Cyber Credit-Card Fraudsters?

- Credit-card information buyers: They are fraudsters with practically no expert PC abilities (e.g., PC Programming, Systems administration, and so forth) who purchase hacked (or taken) credit-card information on an unlawful " credit-card deals" site. They purchase this card data intending to make electronic installments for certain merchandise and ventures on the internet.[2]

- Black hat hackers: Late examination on Programmers regarding PC Security characterized a " black hat hacker " (otherwise called a cracker) as a programmer who abuses PC security with malicious aim or for individual addition. They pick their objectives utilizing a two-dimensional interaction known as the "pre-hacking stage"; Focusing on, Exploration and Data Assembling, and Completing the Assault. These kinds of programmers are profoundly gifted in PC Programming and PC Systems administration and such abilities can interrupt an organization of PCs. The principal reason for their demonstration of interruption or hacking is to take individual or private data, (for example, credit card data, financial balance data, and so forth) for their very own benefit (for example making a "credit card deals" site where other cyber credit fraudsters with next to zero PC abilities can purchase credit card information).[2]
- Physical credit-card stealers: They are the sort of fraudsters who truly take credit cards and work out the data on them. They truly take these plastic credit cards (perhaps by pick-stashing in a packed spot) and work out the credit-cards data to utilize this card data to make the electronic installment for some goods and services on the web.[2]

The Role of Data Analytics in Fraud Analysis

Huge information investigation devices and innovations are utilized in combatting dangers. These strategies join text mining, AI, and ontology modelling to help in got danger forecast, discovery, and anticipation at the beginning phase. Intelligence-led examination measures are much quiet with the assistance of these methods and through improved cooperative frameworks dangers can undoubtedly be recognized. Associations are subsequently picking to move away from the regular firewall and endpoint merchant strategies to embracing huge information and cloud answers to keep up security in the association. Data analytic techniques can thus be concluded to have a very important part to play in proactive prediction, identification, and discriminating of fraud. These strategies can permit the association can be allowed to eliminate, examine, decrypt and change business information to recognize potential occasions of extortion and fraud depending on reports generated by these systems. It would thus be possible to realize successful fraud monitoring projects with the help of these hands-on strategies. [3]

In a nutshell using data analytics in fraud prevention would have the following benefits

- Improved efficiency – Repeatable extortion tests that can be run on your information whenever.
- Wider coverage – Full coverage of testing population rather than spot checks on transactions better chance of finding exceptional items.
- Early warning system – Analytics solutions can help you to quickly identify potentially fraudulent behaviour before the fraud becomes materialized.

Main Challenges in Fraud Analytics

1. Changing fraud patterns over time— This one is the hardest to address since the fraudsters are consistently in the post to discover new and imaginative approaches to get around the frameworks to submit the demonstration. In this way it turns into immeasurably significant for the profound learning models to be refreshed with the advanced examples to recognize. This outcomes in a decline in the model's presentation and effectiveness. Subsequently the AI models need to continue to refresh or bomb their destinations.

2. Model Interpretations — This impairment is related with the idea of logic since models ordinarily give a score demonstrating if an exchange is probably going to be deceitful — without clarifying why.

3. Feature generation can be time-consuming — Subject matter experts can require long periods of time to generate a comprehensive feature set which slows down the fraud detection process.

Anomaly Detection

[4]To indicate a process problem, the main objective of anomaly detection is to identify "outliers". The following four modules of task are used in fraud analytics-

1) Classification - During the characterization cycle the information being considered are coordinated into pre-marked gatherings with the utilization of various sorts of information mining calculations. Order is the cycle of blend of information in predefined classes. This is once in a while called directed characterization as it utilizes different class marks to sort the articles in the information bunch. This generally incorporates utilizing a referred to class mark as acquired by past calculations. These sets are called preparing sets and a design is made for this. Distinctive arrangement procedures are utilized for various types of extortion designs. There are two different ways to look at the exhibition of classifiers: I) disarray framework, and ii) to utilize a ROC chart.

2) Clustering – Clustering is similar to classification but this does not use predefined training classes. It is simply meant to cluster similar objects together. Thus it is a type of unsupervised classification. This follows principle of similarity maximization among intra class objects and similarity minimization techniques among inter class objects.

3) Regression – Regression or genetic programming as it is usually called attempts to obtain a Function which models the data of the minimum error.

4) Association rule – Association rules are utilized to discover relationship among information objects by noticing the recurrence sets happening together in value-based data set. Limit esteems known as help and certainty are utilized to discover how successive a thing set is in a specific exchange. [4]

Sorts of an anomaly:

[5] **Point Anomaly:** When we map a situation in a diagram, if a point stands apart distant from the remainder of the information in the perception, at that point that point is called as Point Anomaly.

Contextual Anomaly: When we break down a perception, in view of the setting of the perception, on the off chance that we discover any information occasion inconsistency, we term this as "Contextual Anomaly"

Collective Anomaly: When we analyze an observation, if we find a set of related data instance different than the rest of the data collection, then we term that as “Collective Anomaly”. [5]

[6] **For Fraud Detection, Techniques and Methods**

Ref	**Year**	**System Name**	**Techniques**	**Reasons**
[1]	2010	Subscription Fraud Prevention in Telecommunications using Fuzzy Rules and Neural Networks	Multilayer perceptron neural network	subscription fraud detected
[2]	2008	Temporal Representation in Spike. Detection of Sparse Personal Identity Streams	Spike Analysis Framework	Identity crime
[3]	2009	Fraud Detection Using an Adaptive-Neuro-Fuzzy Inference System in Mobile Telecommunication Networks	Global Services of Mobile Communications	Detect the fraud on the installation phase
[4]	2007	Offline Internet Banking Fraud Detection	Offline internet banking fraud detection system	Detect the Fraud via offline internet bank detection
[5]	2011	Comparison with Parametric Optimization in Credit Card Fraud Detection	Neural Nets(NN),Bayesian Nets(BN),Naive Bayes(NB), Artificial Immune Systems(AIS) and Decision Trees (DT), to detection the credit card fraud	Credit Card Fraud Detection
[6]	2009	Financial Statement Fraud Detection by data mining	Financial statement frauds (FSF)	Detect fraud in financial activities

[7]	2006	Using Identity Credential Usage Logs to Detect Anomalous Service Accesses	anomaly based metric	Prevent the log records
[8]	2011	Holistic Approach to Fraud Management in Health Insurance	(1)Deterrence, (2) prevention, (3) detection, (4) investigation, (5) sanction and redress, and (6) monitoring.	detect fraud on health insurance companies
[9]	2006	Credit card fraud and detection techniques	Genetic algorithms and other algorithms.	Detect fraud on banks or credit card companies
[10]	2008	Utility Based Fraud Detection	Utility-based Rankings	Outlier Ranking`
[11]	2009	Identifying Online Credit Card Fraud using Artificial Immune Systems	Artificial Immune Systems (AIS)	Identifying Online Credit Card Fraud
[12]	2008	Risk-Based Payment Fraud Detection	Risk-Based Payment Fraud Detection system	Identify fraud from one bank account to the other bank account
[13]	2010	Learning Classier Systems for Fraud Detection	Detected Fraud	
[14]	2008	IDENTIFYING BANK FRAUDS USING CRISP-DM AND DECISION TREES	Detecting fraud on bank transactions	Using CRISP-DM AND DECISION TREES
[15]	2011	Detecting Credit Card Fraud by Decision Trees and Support Vector Machines	classification models based on decision trees and support vector machines (SVM)	Detect and stop the fraud in the credit card.
[16]	2010	A Medical Claim Fraud/Abuse Detection System based on Data	multilayer perceptron neural networks (MLP)	To detect fraud in medical domain

		Mining: A Case Study in Chile		
[17]	2010	An Identity Fraud Model Categorizing Perpetrators, Channels, Methods of Attack, Victims and Organizational Impacts	identity about fraud and its category	
[18]	2010	On the communal analysis suspicion scoring for identity crime in streaming credit applications	communal analysis suspicion scoring (CASS)	identity crime in streaming credit applications
[19]	2007	Adaptive Spike Detection for Resilient Data Stream Mining	adaptive spike detection	crime detection domain
[20]	2011	Credit Card Fraud Detection with Artificial Immune System	Artificial Immune Systems(AIS)	credit card fraud detection and compare it to other methods

Credit Card Fraud Variants-

[7] There are various approaches to complete credit card fraud namely:

1) ID theft: When an assailant gets the individual data of a casualty like date of birth, sex, email id, he can easily get access to a new account using victim's details or even a step further by taking hold of the existing account. Wholesale fraud establishes 71% of the most widely recognized kind of extortion.

2) Fake cards: Card which isn't approved or not gave by monetary foundations is named as fake cards. Fake cards are created by skimming the real information of real card which was swiped over an EDC machine. This information is encoded from the attractive strips and later used to make counterfeit cards.

3) Stolen/lost cards: A situation where a card holder inadvertently loses his card or his card has been taken, if the cardholder neglects to report it to the concerned bank there may be chances that the card can be abused by a crook.

4) CNP fraud: Card, not present misrepresentation is a kind of extortion where the criminal requires negligible data, for example, card number and expiry date. In such circumstance, the card need not be available while making the buys on the web.

5) Clean fakes: These cheats are not as spotless as they sound. The buys are made with taken cards and later exchanges are changed hence making it discover a route around the FDS.

6) Friendly fraud: the genuine cardholder himself makes the buys and pays for the administrations utilizing "pull" method of installment with his credit/charge card. Later reports a grievance expressing loss of card and claims for repayment.

7) Affiliate fraud: It is the most generally conveyed misrepresentation where either an individual logs into a site and makes buys utilizing a bogus record or a program is intended to complete misrepresentation exercises.

8) Triangle fraud: Such extortion fundamentally includes 3 stages: (a) Creating a fake or phony site (b) Providing offers, for example, quick conveyance upon MasterCard installment mode (c) Stolen or phony cards are utilized to make the installments and the name acquired at the genuine store is abused by the criminal to later transport the item to the client. [7]

[8] **Credit card fraud**

Classify the credit extortion based into:

I. Offline credit card fraud: Happens when the plastic card is taken by fraudsters, utilizing it in stores as the real proprietor. This is an exceptional kind of extortion (fraud).

II. Online MasterCard extortion: A famous and exceptionally hazardous misrepresentation, charge cards' data are taken by fraudsters to be utilized later in online exchanges by Internet or phone. [8]

Categorized credit card fraud into

We can classified credit card fraud into -

1. Lost cards and stolen cards.

2. Fake cards.

3. Robbery of cards from the mail or non-receipt of issue (NRI).

4. Mail/phone request fraud. [8]

Challenges In Credit Card Fraud Detection

1) Data insufficiency- Basically, CCFD experimentally addresses most concerning issue of ongoing information investigation, because of the privacy of the issue. However, investigators are not discouraged because they can immediately perform scientific work by an industrial partner. Plus, some people advice using synthetic data that mimics the transactions of datasets.

2) Behavioral variation- Fraudulent conduct to try not to recognize sensitivities after some time. [9]

Credit Card Fraud Detection

Methods-

The data mining includes the different method that can be used for detecting credit card fraud.

1. Hidden Markov Model
2. Neural Network
3. Bayesian Network
4. K- nearest neighbor algorithm
5. Decision Tree

Supervised VS Unsupervised Techniques

When it comes to a data set, it is, at its very core, a dummy collection of data mainly used for training. It comes in either of two variations: labelled and unlabelled.

The fraud detection techniques can be broken down to two sub-components.

1) Supervised Machine Learning Technique
2) Unsupervised Machine Learning Technique

The first one comes into the picture when we encounter a set of data which is labelled. [10] For this kind of learning to progress, mapping new examples is necessary. It is done by the algorithm, which takes the earlier mentioned set of data, and infers a function from it.

The implementation of Supervised techniques has its own advantages, when it comes to comparing it with Unsupervised Techniques in terms of possessing a data set which is labelled. However, we might unlabel our data set, taking into consideration that we are dealing with an egregiously unstable data, and then transform it in such a problem which can be handled by unsupervised techniques.

The unsupervised techniques which can handle the aforementioned problem, can also handle a range of problems, like an anomaly or outlier detection problem.

Moving on to the next step, the classification of fraudulent and non-fraudulent records from the data set, come into play. The model, at its very core, works on categorizing the data in one of the two earlier mentioned classifications. But this technique reaches its limits, when it comes to working on future frauds.[11][12]

Hidden Markov Model

Hidden Markov Model is a Statistical Markov Model. It is an easy and the simplest model which can be used for modelling sequential data. The Credit Card fraud detection model is based on the Hidden Markov Model. It can detect frauds without requiring any fraud signatures.

[13] The Hidden Markov Model first processes the expenditure data of a card user after creating clusters of training set. It focuses solely on the total number of products bought, and then uses it for further processing. The data containing various transactions is categorized based on the varying number of transactions, and then it is stored in the form of clusters.

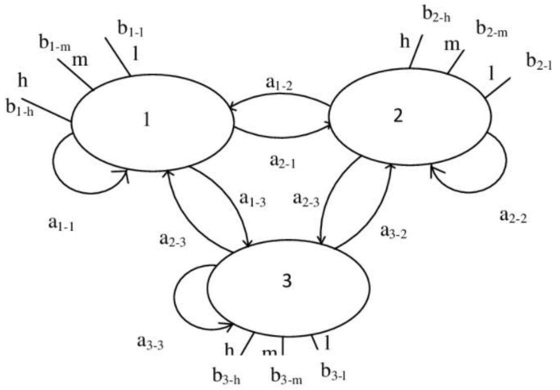
[14-15] HMM has been successfully applied to many applications such as speech recognition, robotics, bio-informatics, data mining etc

[16]Hidden Markov Model is a discrete time stochastic process with a set of states, $S = S0, . . . , SN$ and a constant transitional probability distribution $ai,j = p(qt+1 = Sj \,|qt = Si)$, where $Q = (q0, . . . , qT)$ is a state sequence for the time $t = 0, 1, . . ., T$. The initial state distribution is denoted $\pi = (\pi0, . . ., \pi N)$, where $\pi i = p(q0 = Si)$. The state of the process cannot be directly observed, instead some sequence of observation symbols, $O = (O0, . . . , OT)$ are measured, and the observation probability distribution, $bj (Ot) = bj,t = p(Ot|qt = Sj)$, depends on the current state. The Markov assumption gives that

$p(qt+1 \mid qt, qt-1, . . . , q0) = p(qt+1 \mid qt)$, (1)

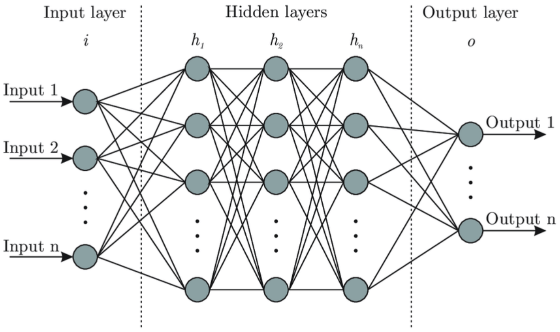
and the probability of the observations satisfies

$p(Ot \mid qt, qt-1, . . . , q0) = p(Ot \mid qt)$. (2)



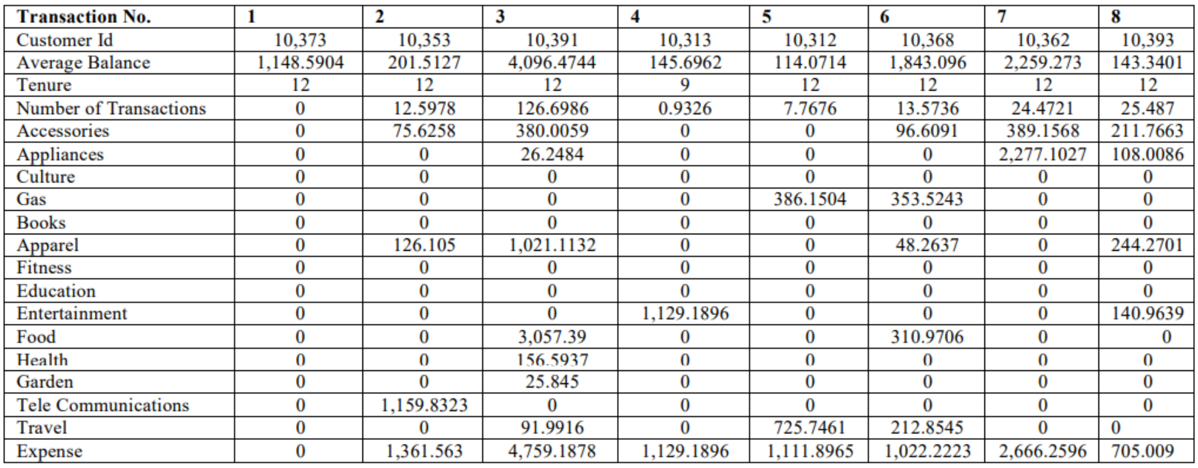
Neural Networks

Neural network is a simple machine learning tool that can be trained to recognize patterns from a data set, and predict an output based on that. Neural Networks are another possible solution for detecting frauds. Neural Networks can be divided into 3 layers: Input layer, Hidden Layer, and the Output layer. The input is taken, and the neurons are trained by breaking down the patterns of the input to its very core at the first layer by assigning “weights” to every neuron, then forwarding it to the next layer and slowly reconstructing the pieces to form a pattern of the input, at each layer, and then finally giving the generated output to the output layer, which determines how much correct the output is, by measuring the error. And then the error is then sent back, via backward propagation.

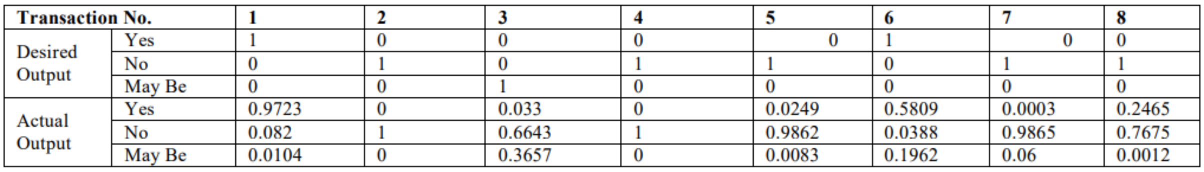


In a study, a data set including a record of 20000 transactions, which contains the amount of average monthly expenditure and average monthly transactions of each credit card owner, was prepared to be fed as an input to the neural network for detecting frauds. The dataset displays various areas of expenditure, each corresponding to its total number of transactions on a monthly average, for each credit card owner. And if there is an irregular amount of expenditure detected, i.e., greater than the average expenditure, and that too, on only one type of item, than it may ne labelled as fraud. For example, if a transaction of $4000 is detected from one category of item, say Television, from a particular card owner, then it can be detected as an abnormal behaviour after comparing it to other monthly average transactions and monthly expenditures. [17] [18]

To identify this kind of pattern, 19 neurons were chosen for the input layer, 15 for the hidden layer and lastly three neurons for the output layer. For the neural network to understand the transactions more efficiently, three principal parameters were provided [17] [18]:



Transaction No.	1	2	3	4	5	6	7	8
Customer Id	10,373	10,353	10,391	10,313	10,312	10,368	10,362	10,393
Average Balance	1,148.5904	201.5127	4,096.4744	145.6962	114.0714	1,843.096	2,259.273	143.3401
Tenure	12	12	12	9	12	12	12	12
Number of Transactions	0	12.5978	126.6986	0.9326	7.7676	13.5736	24.4721	25.487
Accessories	0	75.6258	380.0059	0	0	96.6091	389.1568	211.7663
Appliances	0	0	26.2484	0	0	0	2,277.1027	108.0086
Culture	0	0	0	0	0	0	0	0
Gas	0	0	0	0	386.1504	353.5243	0	0
Books	0	0	0	0	0	0	0	0
Apparel	0	126.105	1,021.1132	0	0	48.2637	0	244.2701
Fitness	0	0	0	0	0	0	0	0
Education	0	0	0	0	0	0	0	0
Entertainment	0	0	0	1,129.1896	0	0	0	140.9639
Food	0	0	3,057.39	0	0	310.9706	0	0
Health	0	0	156.5937	0	0	0	0	0
Garden	0	0	25.845	0	0	0	0	0
Tele Communications	0	1,159.8323	0	0	0	0	0	0
Travel	0	0	91.9916	0	725.7461	212.8545	0	0
Expense	0	1,361.563	4,759.1878	1,129.1896	1,111.8965	1,022.2223	2,666.2596	705.009

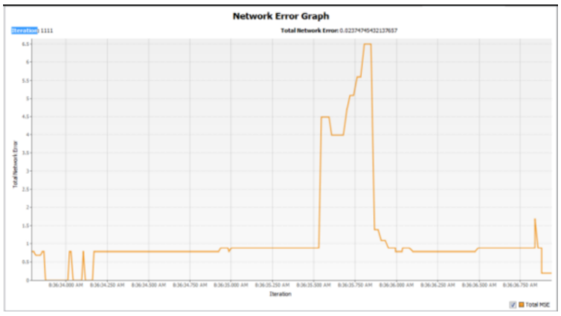


Transaction No.		1	2	3	4	5	6	7	8
Desired Output	Yes	1	0	0	0	0	1	0	0
	No	0	1	0	1	1	0	1	1
	May Be	0	0	1	0	0	0	0	0
Actual Output	Yes	0.9723	0	0.033	0	0.0249	0.5809	0.0003	0.2465
	No	0.082	1	0.6643	1	0.9862	0.0388	0.9865	0.7675
	May Be	0.0104	0	0.3657	0	0.0083	0.1962	0.06	0.0012

1) For controlling the fluctuation in the weight of every neuron, the Learning Rate parameter is used.
2) For preventing the system converging at a local minimum, the Momentum parameter is used. Momentum here is a very crucial factor. If it goes too high, the network will become unstable, because the system will converge very rapidly. If it goes extremely low, the system will surely converge at a local minimum. Therefore, at the time of training the network, choosing the proper amount of momentum is essential.
3) For determining how much error, which is produced at the output layer, is allowed at best, or in other words, the maximum error, while categorizing a transaction, the Max Error parameter is used.

For the current study in discussion, the values of the learning rate, momentum, and max error, were taken as 0.9, 0.2, and 0.03 respectively.

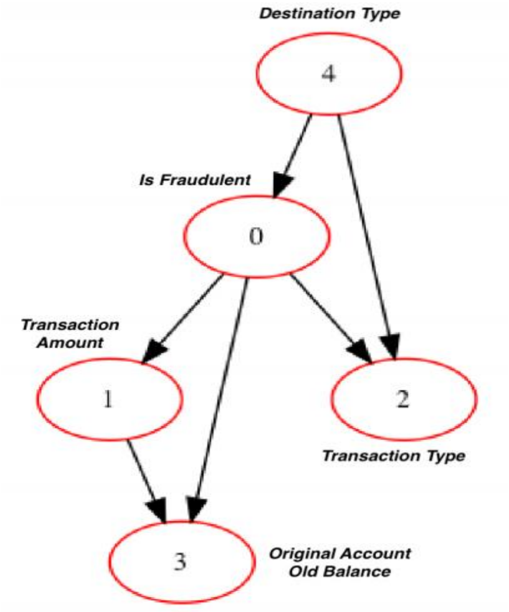
The network was provided a test data set after training it enough. The results are remarkably close to accuracy. They are provided below in the form of a table.



Bayesian Network

[19] It provides a graphic model of casual relationships on which class members are probabilities.

Its aim is to assign a new instance to the class that has to highest posterior probability just for example



[20] Interpretation of the model can be done by associating each of the digits 0 through 4 as follows: 0: Is Fraud (variable to infer)

1: The amount in the transaction

2: The type of transaction that occurred

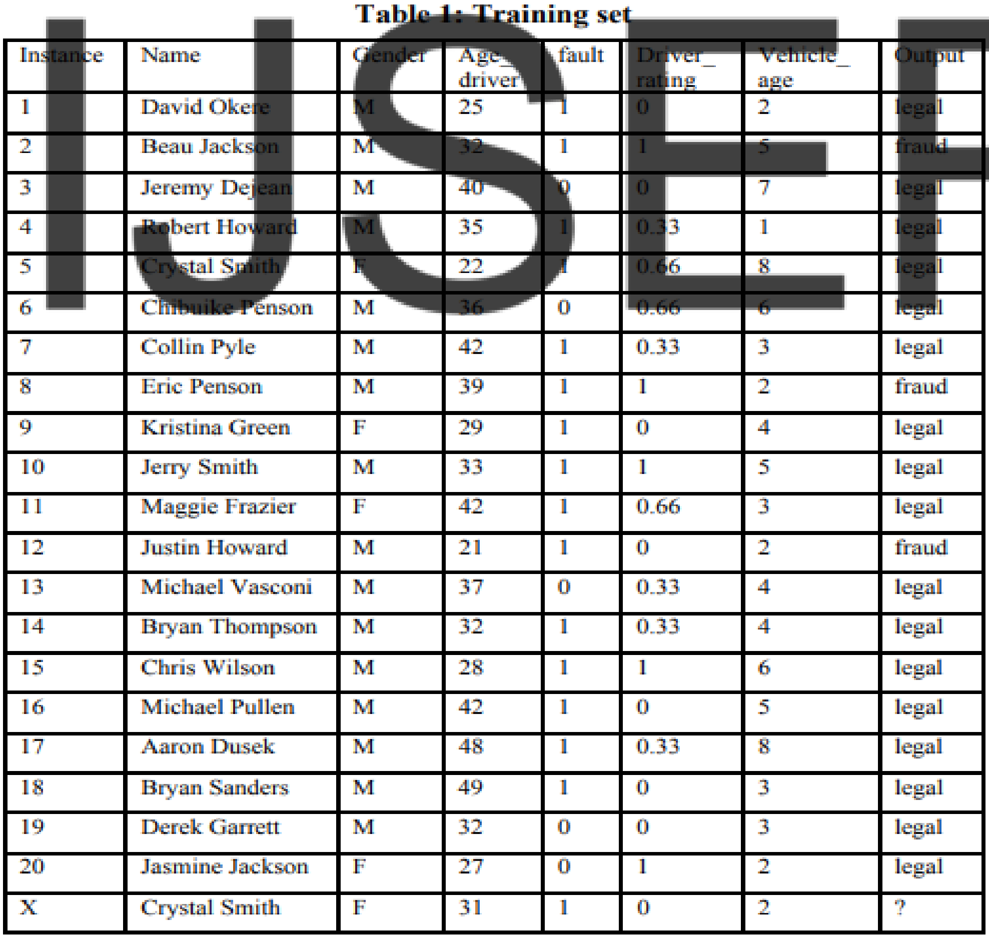
3: Original old balance

4: Destination type - customer or merchant

Taking one more example

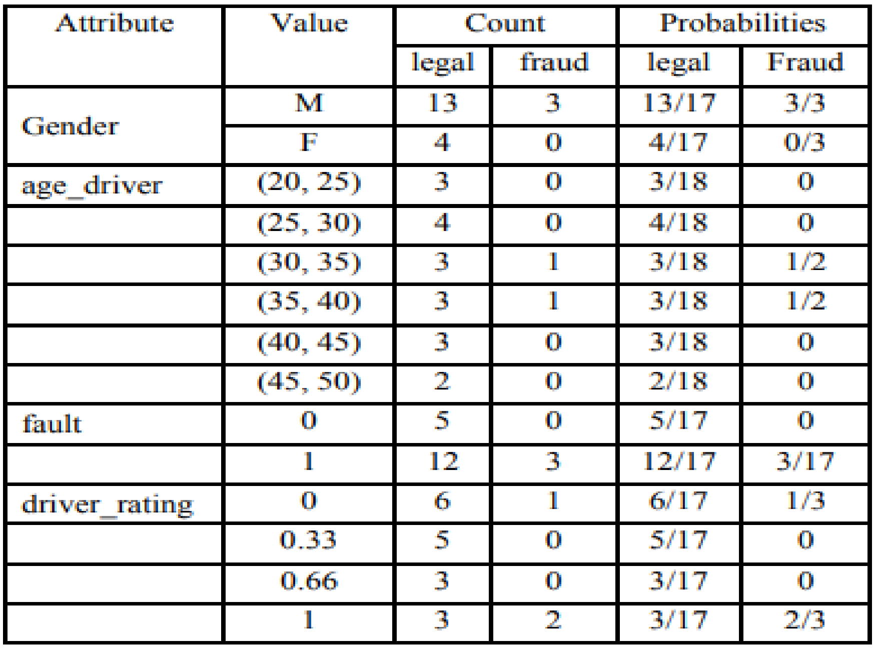
As you can see in there database the is a training data set of driver driving rating and we have to predict that wither this is a fraud or legal in this data base 17 tuples are league and 3 tuples fraud

Table 1: Training set



Instance	Name	Gender	Age_ driver	fault	Driver_ rating	Vehicle_ age	Output
1	David Okere	M	25	1	0	2	legal
2	Beau Jackson	M	32	1	1	5	fraud
3	Jeremy Dejean	M	40	0	0	7	legal
4	Robert Howard	M	35	1	0.33	1	legal
5	Crystal Smith	F	22	1	0.66	8	legal
6	Chibuike Penson	M	36	0	0.66	6	legal
7	Collin Pyle	M	42	1	0.33	3	legal
8	Eric Penson	M	39	1	1	2	fraud
9	Kristina Green	F	29	1	0	4	legal
10	Jerry Smith	M	33	1	1	5	legal
11	Maggie Frazier	F	42	1	0.66	3	legal
12	Justin Howard	M	21	1	0	2	fraud
13	Michael Vasconi	M	37	0	0.33	4	legal
14	Bryan Thompson	M	32	1	0.33	4	legal
15	Chris Wilson	M	28	1	1	6	legal
16	Michael Pullen	M	42	1	0	5	legal
17	Aaron Dusek	M	48	1	0.33	8	legal
18	Bryan Sanders	M	49	1	0	3	legal
19	Derek Garrett	M	32	0	0	3	legal
20	Jasmine Jackson	F	27	0	1	2	legal
X	Crystal Smith	F	31	1	0	2	?

Table 2: Probabilities associated with attributes



Attribute	Value	Count		Probabilities	
		legal	fraud	legal	Fraud
Gender	M	13	3	13/17	3/3
	F	4	0	4/17	0/3
age_driver	(20, 25)	3	0	3/18	0
	(25, 30)	4	0	4/18	0
	(30, 35)	3	1	3/18	1/2
	(35, 40)	3	1	3/18	1/2
	(40, 45)	3	0	3/18	0
	(45, 50)	2	0	2/18	0
fault	0	5	0	5/17	0
	1	12	3	12/17	3/17
driver_rating	0	6	1	6/17	1/3
	0.33	5	0	5/17	0
	0.66	3	0	3/17	0
	1	3	2	3/17	2/3

This table show the car and subsequent probity is associated with the attribute of the database this training data we estimate the prior probabilities .The classifier has to predict the class of instance to be fraud or legal

The classifier has to predict the class of instance

$$P(fraud) = s_i / s = 3/20 = 0.15$$
$$P(legal) = s_i / s = 17/20 = 0.85$$

We assume that classify X = (Crystal Smith, F, 31)

If X is legal then the probity is

$$P(X |legal) = 4/17 * 3/18 = 0.039$$

The submission of x is legal will be around

likelihood of being legal = 0.039 *0.9=0.0351

If X is fraud, then the probity is

$$P(X |fraud) = 3/3 * 1/2 = 0.500$$

The submission of x is fraud will be around

Likelihood of being fraud = 0.500 *0.1= 0.050

On the bases of this we can conclude that

$$P(X) = 0.0351 + 0.050 = 0.0851$$

Finally, we obtain the actual probabilities of each event:

$$P(legal | X) = (0.039 *0.9)/0.0851= 0.412$$
$$P(fraud |X) = (0.500 *0.1) / 0.0851= 0.588$$

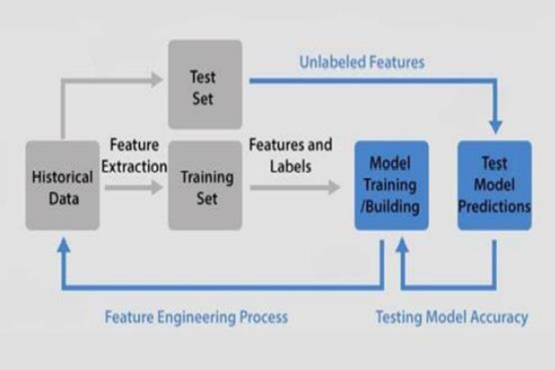
New tuple as fraud because it has highest probity

[21] Since attribute are treated as independent addition of redundant once reduces its predictive power. To relax this conditional independence, add derived attributes which are created from combing of executing attributes

The BAYESIAN NETWORK is easer to use only 1 training data is required this approach can handle missing values by simply omitting that probity when calculating the likelihood of membership in each class the only Disadvantage is it cannot handle continuous data.

K-Nearest Neighbour Algorithm

K-Nearest Neighbour is a non-parametric use classification, it is a lazy learner algorithm where all computation is deferred until classification, we use this algorithm when we have large amount of data in training data set, it is a supervised classification algorithm in which you have some data points or data vectors which is divide in to different number of category's and it tries to predict the new classification of a sample from that particular population set

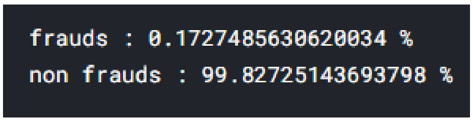


It classifies new points based in the similarity measure

Looking at one example of card fraud detection it contains spending if the money and at which time they have spent it and other contain the value result of pca

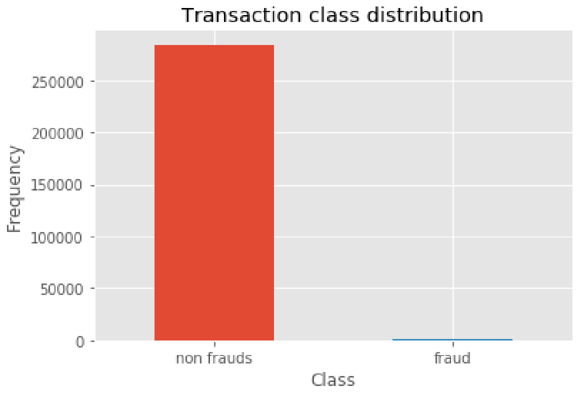
Time	V1	V2	V3	V4	V5	Amount	Class
0	0	-1.359807	-0.072781	2.536347	1.378155	149.62	0
1	0	1.191857	0.266151	0.16648	0.448154	2.69	0
2	1	-1.358354	-1.340163	1.773209	0.37978	378.66	0
3	1	-0.966272	-0.185226	1.792993	-0.863291	123.5	0
4	2	-1.158233	0.877737	1.548718	0.403034	69.99	0

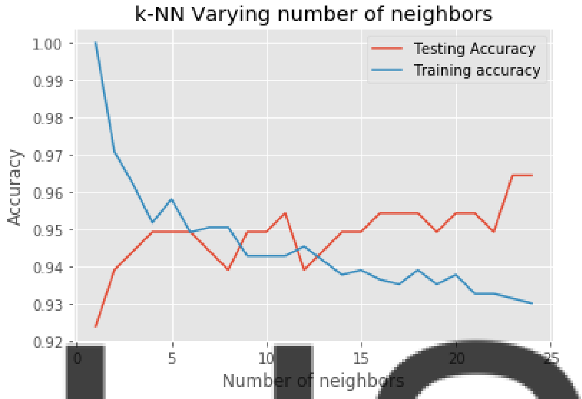
[22]This is the data which has been loaded and in the bellow it there is there the accrues of knn model of this dataset the above there is a graph which shows that if it is a fraud or not in the bases of the class we will find the fraud if class is 1 then it is a fraud transection and if class is 0 then the it is not a fraudulent transection.



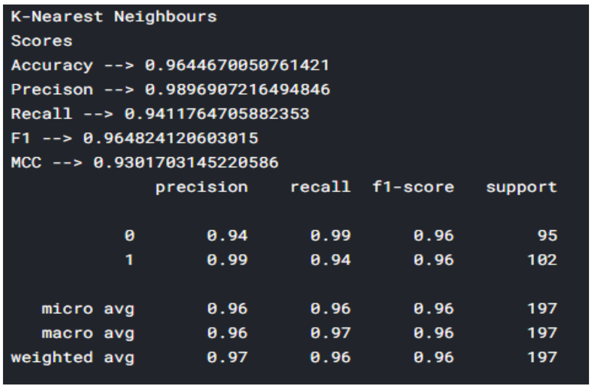
```
frauds : 0.1727485630620034 %
non frauds : 99.82725143693798 %
```

This is the confusion matrix which has been obtained from the card fraud detection model





confusion matrix		actual class	
		positive	negative
predicted class	positive	94	1
	negative	6	96



```
K-Nearest Neighbours
Scores
Accuracy --> 0.9644670050761421
Precison --> 0.9896907216494846
Recall --> 0.9411764705882353
F1 --> 0.964824120603015
MCC --> 0.9301703145220586
                precision    recall  f1-score   support

           0       0.94      0.99      0.96        95
           1       0.99      0.94      0.96       102

   micro avg       0.96      0.96      0.96       197
   macro avg       0.96      0.97      0.96       197
weighted avg       0.97      0.96      0.96       197
```

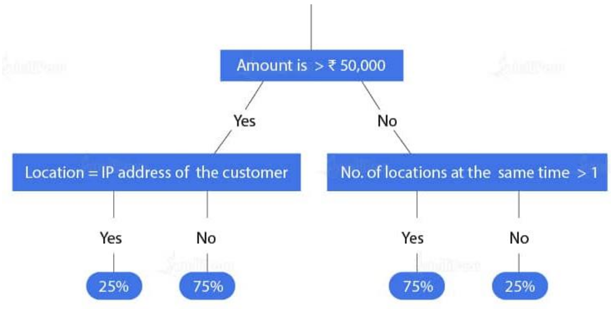
[23]factors:

- The distance metric used to find the closest neighbours.
- The distance rule used to get an arrangement from k nearest neighbour.
- The quantity of neighbours used to group the new example.

Decision Tree

decision tree is mainly used classification (tree structure) and another in the recreation. The classification is a process of dividing the datasets into different categories or in the form of groups by its adding labels on to it. Now coming towards the decision three it is a graphical representation of all the possible solution of a decision given in the dataset decision tree is based in the condition and it is easy to explain comparing the other models not coming towards the structure in the decision tree it has root node and there are leaf node next concept is splitting it means to divide the root node into two or more part on the basis of the condition next concept is data pruning it is basically it is opposite of splitting next concept is parent or child node that are all the concepts in the decision tree

now looking towards a one example assume that


Amount is > ₹ 50,000
Yes
No
Location = IP address of the customer
No. of locations at the same time > 1
Yes
No
Yes
No
25%
75%
75%
25%

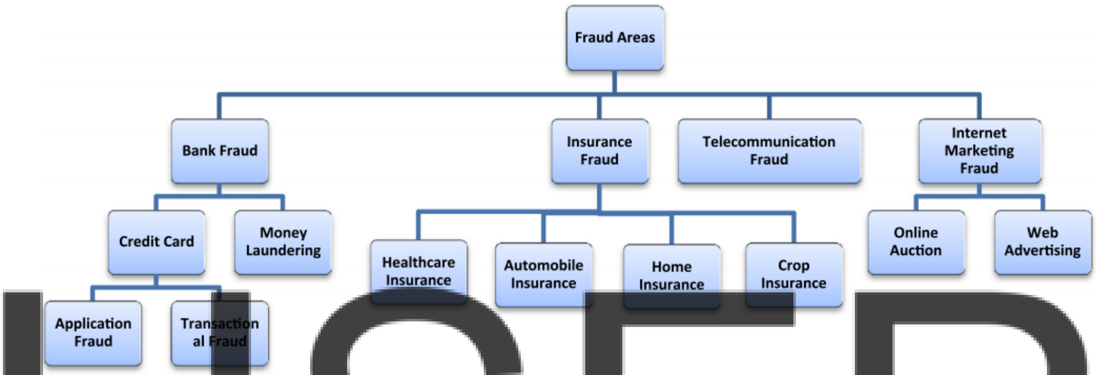
If a user made a transection amount grater, then 50k as per the decision tree we have to verify the location of the transection otherwise then we will check the frequency of the transection

After checking the above conditions (parent node) we can predict the transection is fraud or not fraud if transection made by the user is granter then 50k and the IP Address which we have verified are qual then there is 25% probability of fraud and 75% probability it is not fraud

Similarly, After checking the above conditions (parent node) we can predict the transection is fraud or not fraud if transection made by the user is less then 50k and the number of location is granter then 1 there is 75% probability of fraud and 25% probability it is not fraud

Fraud areas

[24]In the figure it shows is the area where fraud detection is mainly according



CONCLUSION:

Our main goal is to analyze different data mining techniques in a way that they assist us with identifying and predict the credit card fraud. This paper offers techniques and procedures to deal with fraud in several business contexts by citing common scenarios. Data mining techniques are meant to assist and manage these issues with greater fund saving. These techniques can be used alone or combined with an ensemble or meta-learning techniques to build strong detection classifiers. Then, the challenges that obstruct the performance and proficiency of fraud detection is examined in the paper. Each technique have its own pros and cons so we can apply a technique, which would be a best fit for our scenario.

References

[1] Randula Koralage," Data Mining Techniques for Credit Card Fraud Detection", Faculty of Information Technology, University of Moratuwa, July 2019

[2] Vishalakshi,N.S and Deepika ,N," Survey on Data Mining Methodologies For Cyber Credit-Card and Credit-Card Fraud Detection System", Department of Computer Science and Engineering, NHCE, Bangalore, India, 25th December, 2015, https://www.journalcra.com/sites/default/files/issue-pdf/13420.pdf.

[3] Dr. Asoke Nath," Department of Computer Science", St. Xavier's College (Autonomous) Kolkata, India, International Journal of Research Studies in Computer Science and Engineering (IJRSCSE) Volume 3, Issue 4, 2016, PP 1-11, https://www.arcjournals.org/pdfs/ijrscse/v3-i4/1.pdf.

[4] Sreeparna Mukherjee and Triparna Mukherjee, "Fraud Analytics Using Data Mining", 2016.

[5] ALIKA," Survey Paper for Credit Card Fraud Detection Using Data Mining Techniques", 2019.

[6] Muhammad Arif and Amil Roohani Dar,**"**Survey on Fraud Detection Techniques Using Data Mining**",** 2015.

[7] Geeta natrajan," STUDY ON CREDIT CARD FRAUD DETECTION USING DATA MINING TECHNIQUES", 2017.

[8] Aisha Abdallah and Mohd Aizaini Maarof and Anazida Zainal," fraud detection system: A survey", 2016.

[9] Rahul Goyal and Amit Kumar Manjhvar," Review on Credit Card Fraud Detection using Data Mining Classification Techniques & Machine Learning Algorithms", 2020.
[10] Ashay Walke1 . "Comparison of Supervised and Unsupervised Fraud Detection?"

[11] Bolton, R. and D. Hand. Unsupervised profiling methods for fraud detection.

[12] Bolton, R. and D. Hand (2002). Statistical fraud detection: A review. Statistical Science 17 (3), 235–255.

[13] SHAILESH S. DHOK,Credit Card Fraud Detection Using Hidden Markov Model ISSN: 2231-2307, Volume-2, Issue-1, March 2012

[14] Chan, Philip K., Fan, Wei, Prodromidis, Andreas L. & Stolfo, Salvatore J., (1999) "Distributed Data Mining in Credit Card Fraud Detection", IEEE Intelligent Systems, Vol. 14, No. 6, pp. 67-74.

[15] Rabiner, Lawrence R., (1989) "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proc. of IEEE, Vol. 77, No. 2, pp. 257-286.

[16] Fonzo, Valeria De, Aluffi-Pentini, Filippo and Parisi, Valerio, (2007) "Hidden Markov Models in Bioinformatics", Current Bioinformatics, Vol. 2, pp. 49-61.

[17] CREDIT CARD FRAUD DETECTION USING NEURAL NETWORKS

[18] Raghavendra Patidar, Lokesh Sharma, "Credit Card Fraud Detection Using Neural Network", International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-1, Issue-NCAI2011, June 2011, pp. 32-38.

[19] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. Proc. IEEE, 77(2):257–286, 1989.

[20] Kadar D. Qian, Kaushal A. Alate, Nicholas T. Lai, "Identifying Fraudulent Transactions in Mobile Payments" , Stanford University Stanford, CA https://web.stanford.edu/class/aa228/reports/2018/final88.pdf

[21] Rekha Bhowmik ,"Data Mining Techniques in Fraud Detection" , University of Texas at Dallas

https://commons.erau.edu/cgi/viewcontent.cgi?article=1040&context=jdfsl

[22] M. Ummul Safa1 , R. M. Ganga2,” Credit Card Fraud Detection Using Machine Learning” https://www.ijresm.com/Vol.2_2019/Vol2_Iss11_November19/IJRESM_V2_I11_80.pdf

[23] Kalpesh Vishwakarma1 ,“Credit Card Fraudulent Detection using Machine Learning” 1Student, Electronics & Telecommunication Engineering, Mumbai University, Mumbai, Maharashtra, Indiahttps://www.irjet.net/archives/V7/i9/IRJET-V7I9588.pdf

[24] Aisha Abdallahh , Mohd Aizaini Maarof, Anazida Zainal, ”Fraud detection system: A survey” Information Assurance and Security Research Group, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Malays