COMPUTATIONAL INTELLIGENCE AND MACHINE LEARNING
e-ISSN: 2582-7464

Deception Recognition Method Based on Machine Learning

Siddth Kumar Chhajer [1]*, Rudra Bhanu Satpathy [2]

[1]MBA,Marketing, St. Peter's University, Chennai, Tamil Nadu, India.
[2] M.Tech, Electrical, Electronics and Communications Engineering, St. Peter's University, Chennai, Tamil Nadu, India.
*Corresponding Author Email: [1] siddth2011@gmail.com

Abstract

Money extortion is a developing issue with far results in the budgetary business and keeping in mind that numerous procedures have been found. Information removalis effectively functional to back records to computerize the investigation of colossal volumes of multifaceted information. Information removal has additionally assumed a notable job in the location of Visa deception in online exchanges. Deception recognition in credit card is an information mining issue, it gets testing because of two significant reasons–first, the profiles of typical and deceitful practices change much of the time and besides because of the reason that Mastercard extortion informational collections are exceptionally slanted. This paper examines and analyze the presence of the Decision tree, Random Forest, SVM, and strategic regression on exceptionally slanted credit card extortion information. Dataset of Visa exchanges is sourced from European cardholders containing 274,335 exchanges. These function are usedto crude and preprocessed information. The presentation of the strategies is assessed dependent on exactness, affectability, explicitness, accuracy. The outcomes demonstrate the ideal accuracy for logistic regression, decision tree, Random Forest and SVM classifiers are 96.8%, 94.4%,99.5%, and 96.6%.

Keywords

Credit Card, Deception Recognition, Decision Tree, and Support Vector Machine.

INTRODUCTION

Money related extortion is a emergingconcern with broad results in the administration, corporate associations, fund industry, In the reality of extremely dynamic dependency on online innovation, increased visa transactions have been praised, but credit card deception has also escalated as on the network and disconnected trade. When credit card transactions become a much-reaching installment system, the center has been tasked with ongoing technological ideologies to resolve the problem of Visa extortion. There are various extortion discovery and programming agreements that forestall fakes in organizations such as credit card, retail, web-based business, security, and businesses.

The information mining procedure is one prominent and well-known technique utilized in tackling credit extortion discovery issues. It is hard to be completely sure of the true intent and validity behind a request or transaction. In actuality, the best convincing solution is to check for possible verifications of deception from the available knowledge using numerical calculations. Extortion discovery in Visa is the real way to identify certain transactions that are deceptive in two types of legitimate class and deception class transactions, a few methods are designed and implemented to understand credit card deception, such as hereditary recognition calculation, counterfeit neural system visit thing set mining, AI calculations, relocating feathered creatures advancement calculation, near examination of strategic regression, SVM, decision tree and irregular woodland is done. [1]

Mastercard deception recognition is a well-known yet additionally a troublesome issue to unravel. Firstly, due to the question of having only a small measure of knowledge, a credit card attempts to arrange an example for the dataset. In addition, there could be several parts in the database with deceptionster truncations that also match an indication of credible behavior. The problem has a range of specifications, in fact firstly, informational indexes are not effectively open for open and the aftereffects of looks into are frequently protected and managed, rendering the results out of reach and attempting to compare the integrated models for that purpose. Datasets of legitimate knowledge in the documentation in previous inquiries are not referenced. Furthermore, the development of techniques is gradually disturbing in that security forces confinement to the exchange of thoughts and techniques in discovery of deceptions, and particularly in credit and debit cards extortion identification. [2]

In addition, the knowledge databases are constantly developing, and changing process issues of constantly exceptional traditional and deceitful activities that are the legitimate interaction in the past may be manipulation throughout the present or another way around. This paper assesses four propelled information mining draws near, Decision tree, support vector machines, Logistic regression, and arbitrary woodlands, and afterward a collative correlation is made to assess what model performed best. [3]

Mastercard exchange datasets are seldom accessible, profoundly imbalanced, and slanted. Ideal element (factors) decision for the models, reasonable measurement is the most significant piece of information mining to assess the execution of methods on slanted credit card extortion information. Various difficulties are related to Visa discovery, to be specific deceitful conduct profile is dynamic, that is fake exchanges will, in general, appear as though real ones, Credit card extortion location execution is incredibly influenced by the kind of examining approach utilized, the decision of factors and recognition strategy utilized. Toward the finish of this paper, decisions about aftereffects of classifier evaluative testing are made and ordered. From the prosecutions, the result that has already been concluded is that regression analysis has an accuracy of 96.8 percent, while Classifier demonstrates an accuracy of 96.6 percent and the Decision tree demonstrates an accuracy of 94.4 percent, while Random Forest achieves the best results with an accuracy of 99.5 percent. The results obtained in this way suggest Random Forest reveals the most consistent and efficient accuracy of 99.5 percent in the issue of Artificial Intelligence bank card deception identification with data set gave by Artificial Intelligence (AI).

There was also a desire to push forward from a whole new perspective. Attempts were made in the case of deception transactions to enhance alert-feedback interaction. In the case of a deceitful transaction, the authorized program would be informed, and a request would be submitted to repudiate the constantoperation. The Artificial Hereditary Algorithm countered deception from a different direction, unique approaches which shed novelbright in this field.[4]

Methods for detecting deception are constantly being built to protect criminals by responding to their deceitful tactics. These deceptions are categorized as:

- Bank Card Deception: Connected and Disconnected from internet
- Identity Burglary
- Computer Infringement
- Application Deception
- Deception Identity
- Telecommunications Deception

Following are some of the different methodology which is used for detecting the deception:

- Artificial Neural Network,
- Logistic Regression,
- Bayesian Network,
- K-Nearest Neighbor,
- Genetic Algorithm, and
- Fuzzy Logic.

LITERATURE REVIEW

Scam acts as either an illegitimate or unethical deception designed to gain economic or social benefit. It is a premeditated tactic that is clearly illegal, regulation or policy of achieving illegal financial benefit. authors argue related to the detection of anomalies or deception has been already published in this field and seems to be available for public use. The researcher showed that the techniques employed in this field are information accumulation technologies, advanced criminal identification and antagonistic detection.

While in some places these methods and algorithms have created an unforeseen success, they have botched to deliver a enduring and reliable explanation to detect deception. The author presented a related research domain where they used Outlier mining, Outlier detection mining and Distance amount algorithms to reliably predict deception transactions in an emulation experiment of a certain commercial bank's credit card transaction data collection.

Outlier removal is a field of data mining that is used primarily in the monetary and internet sectors. This deals with the identification of items disconnected from the main network, i.e. transactions not genuine. They took consumer behavior attributes and based on the value of those attributes they measured the difference between that attribute's observed value and its predicted value.[5]

Unconventional techniques such as hybrid data mining / complex network classification algorithm may perceive illegal instances in an actual data set of card transactions, based on network reconstruction algorithm.This paper proposes a new collative comparison measure which reflects fairly the gains and losses resulting from deception detection. Using the proposed cost estimate, a cost-sensitive approach based on the Bayes minimal risk is introduced. Improvements of up to 22 percent are made as compared with this approach and other state-of-the-art algorithms. The data collection for this paper is focused on a large European company's real-life transactional data and personal data in data is kept secret, an algorithm's accuracy is around 49.9%.

The purpose of this paper[6] was to find an algorithm and lower the estimate of costs. The result obtained was 22 percent and Bayes' minimal risk algorithm was the one they noticed.Efforts have also been made to advance from a whole new perspective. In the case of deception lent transactions, attempts were made to enhance alert-feedback interaction.The approved system would be notified in case of deception lent transactions, and feedback would be sent to reject the ongoing transaction. One of the approaches which shed new light in this area, the Artificial Genetic Algorithm countered deception from a different direction.

This paper[7], such as the Naive Bayesian Classifier and the model based on Bayesian Networks, the clustering model, offers a contrast between models based on artificial intelligence and a general overview of the evolved deception detection system. And in the end conclusions are based on the findings of the evaluative testing of the models. The number of legal truncations was estimated to be greater than or equal to 0.64, which is their accuracy using Bayesian Network was 64 percent. This paper aims to compare models based on artificial intelligence with a general description of the system developed and to state the accuracy of each model.

METHODOLOGY

The author does experiments by using four types of classifiers. These are:

- Logical Regression,
- SVM (Support Vector Machine)
- Decision Tree Classifier, and Random Forest.

Logistic Regression:

Logistic Regression is a supervised identification process that calculates the likelihood of binary predictor variables predicted from the independent dataset variable that is logistic regression predicts the likelihood of an effect that has two values either zero or one, yes or no and false or actual.Logistic regression has parallels to linear regression but, as a straight line is obtained in linear regression, a gradient is seen in logistic regression. Use one or more determinants or an individual variable is based on what prediction, logistic regression generates logistic curves that show values around 0 & 1.

In another type of problems known as classification tasks, logistic regression is used. The goal here is to determine the category to which the actual object under investigation belongs. Classification is about dividing the data into classes with us, depending on some characteristics.

Let's look at the most widely used example: a tumor must be categorized as malignant or benign based on different characteristics such as size, position, etc. So, the Logistic Regression is a regression framework that has attribute variables such as accurate / inaccurate or 0/1 in the response variable (dependent variable). This basically calculates the likelihood of a binary answer as the answer variable value based on computational equation which relates this to the predictors.

The expression which is used for logistic regression is:

$$\pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x))$$

Where,

- alpha and beta are the parameters that are constant numerical values,
- Y – response variable, and
- x- predictor.

The logistic regression is of two types such as:

1. Binary,
2. Multinomial, and
3. Ordinal

Binary:

Under such a category a variable of interest can only have two possible forms either 1 or 0. Such variables, for example, may indicate final outcome, yes or no, win and lose etc.

Multinomial:

In such a category, variable of interest can have 3 or more potential unsorted types, or forms without quantitative sense. Such factors, for example, may reflect "Type A" or "Type B" or "Type C."

Ordinal:

Under these categories, variable of interest can have 3 or more potential ordered types, or forms with a numerical value. Such parameters, for example, may reflect "bad" or "nice," "very nice," "outstanding," and each category may have values such as 0,1,2,3.

Regression Models of Binary, Multinomial:

The basic example of logistic regression is binary or binomial logistical regression for which the objective or explanatory variables should only have two viable categories, either 1 or 0. A further important feature of logistic regression is multivariate regression logistic regression wherein the objective or explanatory variables may have 3 or maybe more feasible unsorted types, i.e. the types that have no quantitative significance.

Support Vector Machine (SVM):

Support Vector Machine "(SVM) is a administered algorithm that may be utilized to address bothclassification and deterioration. It's mainly utilized in identification issues. In this system, the author plots respectively element of data as a point in i-dimensional space (in which i is the number of options) with the value of each function becoming the value of a unique coordinate. Then they perform segregation by finding the hyperplane which very well differentiates the two groups (see below figure 1).

The parts of support vector machine are:

Support Vector:

The points which are nearest to the plane is called support vectors. The line are separated by the data point for finding the best result.

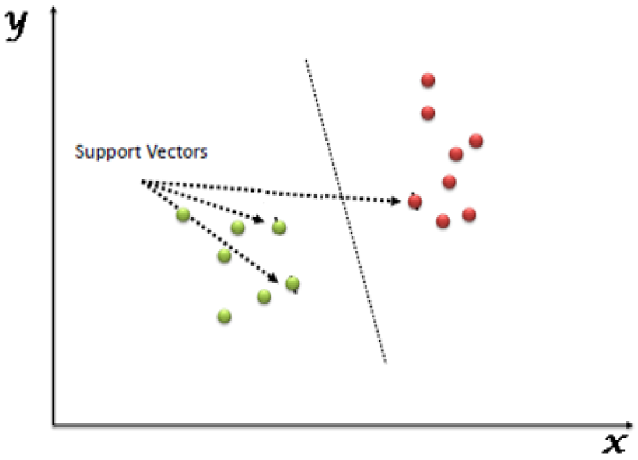

Figure 1: Support Vector Machine classifier

Hyperplane:

It is a plane which is knows as decision plane because it takes decision for giving the outcomes. It is separated between a set of multiple class objects.

Margin:

It can be described as the interface between two lines at the various classes' closet datasets. This could be computed as the distance of the object between both the line and the testing set. Wide margins are called good margins, and low margins are known as poor margins.

Support Vectors Machine are merely the coordinates of distinctcomment. Support Vector Machine is used for segregating of two classes (hyper-plane/ line).SVM can produce the hyperplane iteratively, so the error can be reduced. SVM 's objective is to split the data into groups in order to find a maximum marginal hyperplane (MMH).SVM classifiers have moderately successful and function well enough with high - dimensional feature space. Basically, SVM classifiers are using a set of training samples and thus use far less memory in the end.

Decision Tree:

The decision tree builds a tree structure in the context of classification or regression models. It breaks down a collection of data into smaller and smaller subsets while at the same time incrementally creating a related decision tree. The result is a tree with nodes for decision and nodes for leaves. [8]–[11]A decision node has two or more branches, and a classification or decision is represented by a leaf node. The top decision node in a tree that coincides with the strongest predictor called the root node. Decision trees are capable of handling both categorical and numerical data.

Decision Tree Classifier repetitively divides the working area(plot) into subpart by identifying lines. (repetitively because there may be two distant regions of the same class divided by other as shown in figure 2 below).

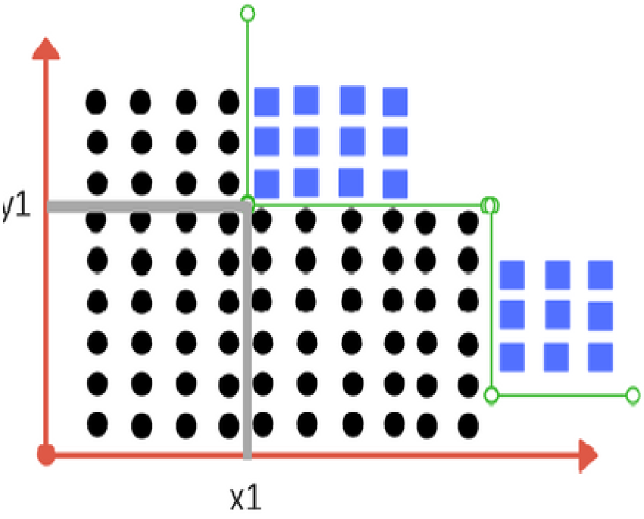

Figure 2: Decision Tree Classifier

Construction of Decision Tree:

A tree can be "learned" by splitting the source set into subsets based on a check of the value of the attribute. This cycle is replicated in a recursive manner, called recursive partitioning on each derived subset. The recursion is completed when the subset at a node all has the same target variable value, or when splitting does not add value to the predictions anymore.

The construction of a decision tree classifier involves no domain knowledge or set of parameters and is therefore ideal for the exploration of explorative knowledge. Decision trees can manage data of large dimensions. Tree classifier has excellent accuracy in general decision making. Decision tree inference is a characteristic inductive approach to learn knowledge on classification.

Terminal Nodes:

While developing tree based terminal nodes, one crucial fact is to determine when and how to stop growing node or generate more terminal nodes. It can even be achieved utilizing two parameters, namely maximum tree depth and minimum node records as follows:

Maximum Tree Depth: This is the maximum number of nodes in a tree after root node, as the name implies. So, people avoid adding terminal nodes until a tree reaches a certain depth, i.e. if a tree has a maximum number of terminal nodes.

Minimum Node Records: The total number of training patterns for which a given node is responsible can be specified. This is necessary to avoid adding terminal nodes once tree is reached at or below these minimum node records.

Now that it is cleared that when to construct terminal nodes, and start building the list. Recursive splitting is a tree-construction method. In this process, once a node is created, recurring is created the child nodes (nodes added to an existing node) on each data group, generated by splitting the dataset, by repeatedly calling the same function.

It is necessary to make a guess about that after creating a decision tree. Prediction essentially includes navigating the decision tree with the data row explicitly given. With the assistance of recursive method, make a prediction, same as above. The same prediction routine is reappointed with left or right child node.

Decision Tree Representation:

Decision trees identify instances by sorting them from the root to some leaf vertex down the tree which gives the instance classification. An instance is defined by beginning at the tree's core point, checking the attribute stated by that node, then heading down the tree branch corresponding to the attribution's value as shown in figure 2 above. For the subtree rooted in the new node this cycle is then repeated.

The instance of a tree structure is given below to predict how often an individual is fit or unfit to provide numerous characteristics such as age, food patterns and exercise levels, offering great accuracy and continuing to work well with high - dimensional feature space.

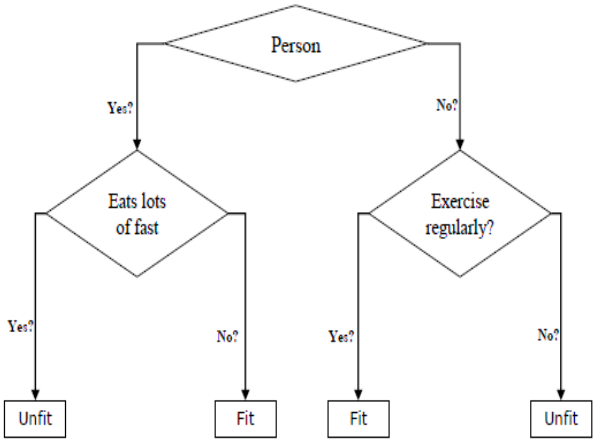

Figure 3: Example of Decesion Tree

According to an above figure 3, the author defined the persons wellness by analyzing many factors. The person has two choices: if he eats lots of fast food or excise regularly. Now, if the individual eats lots of fast food, then again this has two option such as: If yes, then it is sure that an individual is unfit. If no, then an individual is fit.

Now the author came on second option in order to know that an individual is doing exercise regularly or not. If yes, then definitely he is fir or if no, then surely, he is unfit. So, in this way the machine learning decision tree algorithm works.

Random Forest:

Random Forest is a Classification and Regression algorithm. In short, it's a set of classifiers for the decision tree. The random forest has an advantage over the decision tree because it corrects their training collection with the habit of overfitting.

$$\underbrace{P(c|x)}_{\text{Posterior Probability}}=\frac{\overbrace{P(x|c)}^{\text{Likelihood}}\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c|X)=P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

A subset of the training set is sampled randomly so that each tree is trained and then a decision tree is constructed, then each node splits on a feature selected from a random subset of the full feature set. In a random forest, training is extremely fast even for large data sets with many features and data instances, and since each tree is trained independently of the others.It has been found that the Random Forest algorithm provides a good estimate of the generalization error and is prone to overfitting.

Figure 4 shows the random forest algorithm. There are two sets:Training Set, and Test Set. The training set has n samples. It has following steps:

- Firstly, choose the sample from datasets,
- Then develop a decision tree for each sample. It provides the prediction result from each decision tree.
- Then Voting is done for each result, and
- The final prediction result is the most voted prediction result.

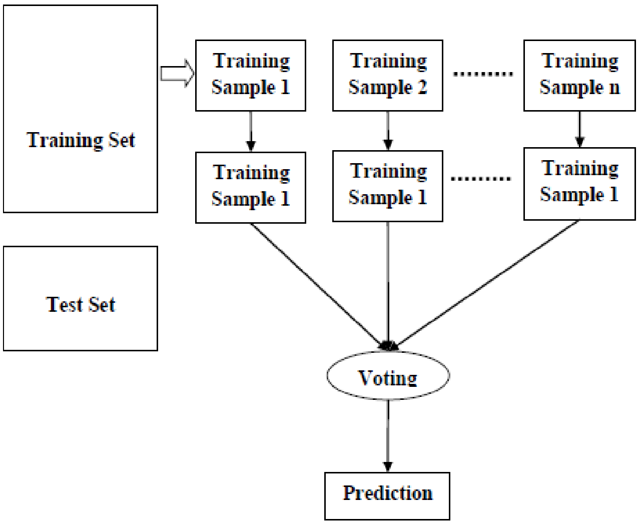

Figure 4: Random Forest Algorithm

The below diagram (Figure 5) helps in understanding the basic rough architecture of this system. Firstly, the incoming transaction is done by the user. This transaction is associated with the bank card deception recognition system and decision maker. These two provides result by comparing all the result and provide best output with help of transaction database as shown in below figure.

The dataset is currently being structured and analyzed. The time and distance and the volume column is uniform, and the column is removed to ensure quality equity. The statistics are processed by a selection of modular algorithms. Above four algorithms is used for this purpose. These data fit into a model.

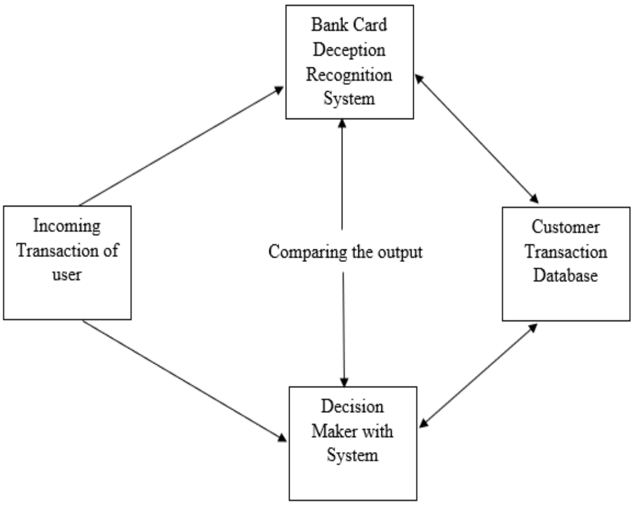
Bank Card Deception Recognition System
Incoming Transaction of user
Comparing the output
Customer Transaction Database
Decision Maker with System

Figure 5: System Architecture

All the people listed in this list have their cards closed in order to evite some risk because of their high-risk profile. Prerequisite is more complicated to the other half. The list is saved only in the restricted data to be properly audited on a case-by - case basis.Credit and collection officers judged half the cases. This list may be regarded as suspect fraudulent comportment.

RESULT& DISCUSSION

This idea is hard to execute in real life, since it needs bank cooperation that is reluctant to share knowledge because of their competitiveness in the market, and for legal reasons, and for the protection of their users ' data.So the authorhas looked up some research papers that is similar methods followed, and findings gathered to make this system useable in a practical manner.

The code shows the number of false positives that it has found and then equate it to real values. Using this it is necessary to calculate the algorithm accuracy, sensitivity and precision.The fraction of the data which the used used for quicker testing is 10% of the full dataset. The complete dataset is also used at the end and the reports are also written out.

These results, along with the report for each classification in the output, the algorithm is as follows, where class 0 is given that means the transaction was legitimate, and 1 means the transaction was assessed as fraud.

Table 1 shows the performance analysis of different classifiers. The accuracy, sensitivity, and precision of all classifiers.In this paper four algorithms are developed for machine learning for identifying credit card fraud. To appraise the algorithms, 70% of the dataset is used for training and 30% is used for testing. It is used for the validation and checking.

Table 1: Performance Analysis

Metrics	Logistic Regression	SVM (Support Vector Machine)	Decision Tree	Random Forest
Accuracy	**0.968**	**0.966**	**0.944**	**0.995**
Sensitivity	**0.965**	**0.964**	**0.944**	**0.993**
Precision	**0.995**	**0.995**	**0.994**	**0.996**

The author performs statistical process in order to find the accuracy, sensitivity and precision for deception recognition method. When the author perform analysis while using linear regression the accuracy, sensitivity and precision is 0.968, 0,965, and 0,995 respectively.

The support vector machine is another algorithm in machine learning that is used for this purpose. Here, the accuracy, sensitivity and precision is 0.966, 0,944, and 0,995 respectively. When decision tree is used for this purpose then the accuracy, sensitivity and precision is 0.944, 0,944, and 0,994 respectively. After that random forest is used for finding the best result then the accuracy, sensitivity and precision is 0.965, 0,963, and 0,996 respectively.

So, after comparing all the analysis and result the author found that the accuracy of logistic regression is the highest i.e. 96.8%. The sensitivity of the logistic regression is again highest sensitivity, i.e. 96.5% but the precision of random forest is the highest precision, i.e. 99.6%.

For understanding the accuracy, sensitivity and precision, firstly people have to understand what is these factor that affect the system in a various manner.

Accuracy:

It shows the difference between the indicated value and the real value. If the indicated value is A_i and the real value is A_r then the accuracy is:

Accuracy= A_i-A_r

It shows the closeness of the indicated value towards the real value.

Sensitivity: The ratio of change in output response is called sensitivity, S. Out of a system for a specified change in input, anywhere this is to be evaluated. This can be expressed mathematically as

$$S = \frac{\Delta A_{out}}{\Delta A_{in}}$$

ΔA_{in} and ΔA_{out} are the change in input and output respectively.

The word sensitivity implies the slightest amount in quantifiable input forced to answer to an instrument. If the standard curve is linear, then the instrument's responsiveness is a perpetual and the standard curve is equal to the slope.

When the standard curve is variable, then the instrument's responsiveness will not be a fixed and will differ with the data.

Precision:

When a device consistently shows a certain value and is often used to calculate a certain quantity for any couple of iterations in the same conditions, so it is assumed that the system has high precision.

CONCLUSION

Misusing the Credit Card is a criminal offense in society. This research has drilled down the most widely recognized strategies for deception alongside their recognition techniques and assessed late discoveries in this field. This paper has additionally clarified in detail, how AI can be applied to show signs of improvement brings about extortion discovery alongside the calculation, pseudocode, clarification its execution, and experimentation results. From all the results the author concluded that the accuracy of the random forest is the highest i.e. 99.5 % and precision is also highest i.e. 99.6%. The sensitivity of the logistic regression is the highest among all the classifiers, i.e. 96.5%.

With a greater number of training data, the Random Forest Algorithm will work better, but speed will suffer during testing and implementation. This will also help to incorporate more pre-treatment procedures. The SVM algorithm still suffers from the problem of the imbalanced dataset and needs more preprocessing to provide better results in the results shown by SVM is good but it could have been better if more preprocessing was done on the data.

REFERENCES

[1] N. Khare and S. Yunus Sait, “Credit Card Deception Detection Using Machine Learning Models and Collating Machine Learning Models,” Int. J. Pure Appl. Math., vol. 118, no. 20, pp. 825–838, 2018.

[2] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, “Credit card deception detection using machine learning techniques: A comparative analysis,” in Proceedings of the IEEE International Conference on Computing, Networking and Informatics, ICCNI 2017, 2017, vol. 2017-January, pp. 1–9, doi: 10.1109/ICCNI.2017.8123782.

[3] L. S. V S S and S. Deepthi Kavila, “Machine Learning For Credit Card Deception Detection System,” 2018. Accessed: 06-May-2020. [Online]. Available: http://www.ripublication.com

[4] A. Oza, “Deception Detection using Machine Learning.”

[5] “(PDF) A Review On Credit Card Deception Detection Using Machine Learning.” https://www.researchgate.net/publication/336552027_A_Review_On_Credit_Card_Deception_Detection_Using_Machine_Learning (accessed May 06, 2020).

[6] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, “Credit Card Deception Detection - Machine Learning methods,” in 2019 18th International Symposium INFOTEH-JAHORINA, INFOTEH 2019 - Proceedings, 2019, doi: 10.1109/INFOTEH.2019.8717766.

[7] S P Maniraj, Aditya Saini, Shadab Ahmed, and Swarna Deep Sarkar, “Credit Card Deception Detection using Machine Learning and Data Science,” Int. J. Eng. Res., vol. 08, no. 09, Sep. 2019, doi: 10.17577/ijertv8is090031.

[8] S. Yaram, “Machine learning algorithms for document clustering and deception detection,” in Proceedings of the 2016 International Conference on Data Science and Engineering, ICDSE 2016, 2017, doi: 10.1109/ICDSE.2016.7823950.

[9] R. A. Bauder and T. M. Khoshgoftaar, “Medicare deception detection using machine learning methods,” in Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, 2017, vol. 2017-December, pp. 858–865, doi: 10.1109/ICMLA.2017.00-48.

[10] P. Raghavan and N. El Gayar, “Deception Detection using Machine Learning and Deep Learning,” in Proceedings of 2019 International Conference on Computational Intelligence and Knowledge Economy, ICCIKE 2019, 2019, pp. 334–339, doi: 10.1109/ICCIKE47802.2019.9004231.

[11] O. S. Yee, S. Sagadevan, and N. H. A. H. Malim, “Credit card deception detection using machine learning as data mining technique,” J. Telecommun. Electron. Comput. Eng., vol. 10, no. 1–4, pp. 23–27, 2018.