Credit Card Fraud Detection using Machine Learning

Gautam Kumar, Shivanesh Kumar, A Arul Prakash

Abstract: ***Now a days credit card plays a very important role in the lives of the human being. It becomes an important part of the businessman, global activities and many more. Even using credit cards give us a most widely used of benefits when it used with the responsibility and carefully, and very small credit and financial harm is also caused by fraudulent activities or transactions. There are a lot of techniques are given to encounter the scope in credit. In spite of, whatever the methods are used they have the same goal of clog the card fraud and each one has its own advantage, drawback and the characteristics too. The deficiency and the good of the credit card detection-methodologies are description and dissimilarity. Moreover, a taxonomy of reference techniques are classified in two fraud-detection perspective, as misuse (supervised) and absurdity (unsupervised) is given. Again, a taxonomy of methods is presented supported caliber to process the categorical and numerical datasets. Other kind of datasets are made in the literature then mentioned and sorted in real and club into the group of the data and therefore the dominant and customary attributes are removed for prosecute application. Consequently, for the new researches, the issues for credit card fraud-detection are described as per the recommendations.***

Keywords **: *Purchasing, Clustering, Datasets, Random Forest, Naïve Bayes Classifie.r.***

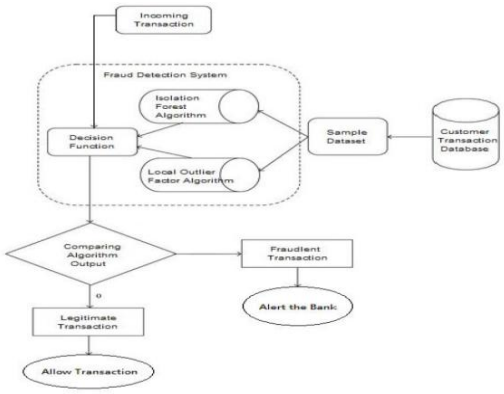
I. INTRODUCTION

Credit card fraud is casual use of someone's account without the proprietor monitoring it. Credit cards gives the benefits to cardholder of the time, that is , it proposed some of the best facilities like the customer can took cash in advance and can be use to buy goods and having services in the limit of the credit card and the customer can repay it latter. Nowadays credit card fraud are very easy to target. in a small period, without knowing to the proprietor a convincing amount can be withdrawn. In the upcoming transactions, necessary steps to be taken against such kind of fraudulent practices by doing analysis and by studying these kinds of fraudulent.

Manuscript received on March 22, 2021.
Revised Manuscript received on April 20, 2021.
Revised Manuscript Received on April 30, 2021.

* Corresponding Author
Gautam Kumar*, Student, Department of School of Computer Science and Engineering, Galgotias University, Greater Noida, India
Shivanesh Kumar, Student, Department of School of Computer Science and Engineering, Galgotias University, Greater Noida, India
A Arul Prakash, Assistant Professor, Department of School of Computer Science and Engineering, Galgotias University, Greater Noida, India

From the point of view of learning this is a very testing issue as it is recognized by numerous components for example, class dissimilarity. Typically the genuine exchanges are more than the deceitful onesWebpage Cloning and False Merchant Sites on the Internet are getting a well known strategy for misrepresentation for some lawbreakers with a gifted capacity for hacking. ML Algorithms calculations are utilized to break down all the approved exchanges and report the dubious ones. These reports are researched by experts who contact the cardholders to affirm if the exchange was certified or fake. The examiners give a criticism to the computerized framework which is utilized to prepare and refresh the calculation to in the long run improve the misrepresentation discovery execution over the long haul.



II. PROPOSED METHOD

The proposed strategies are utilized in this paper, for distinguishing the fakes in charge card framework.. The examination are made for various ML calculations like Logistic Regression, Decision Trees, Gradient Boosting(GSM), to figure out which calculation gives suits best and can be adjusted with Mastercard traders for recognizing extortion exchanges. Given below diagram shows the building outline for addressing the general framework system.

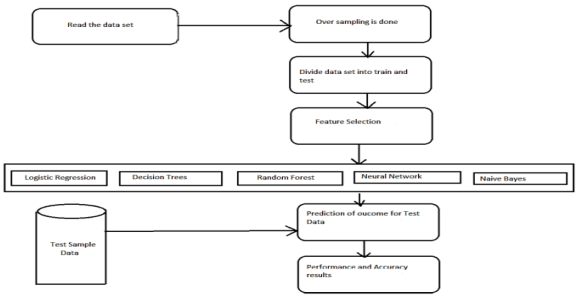


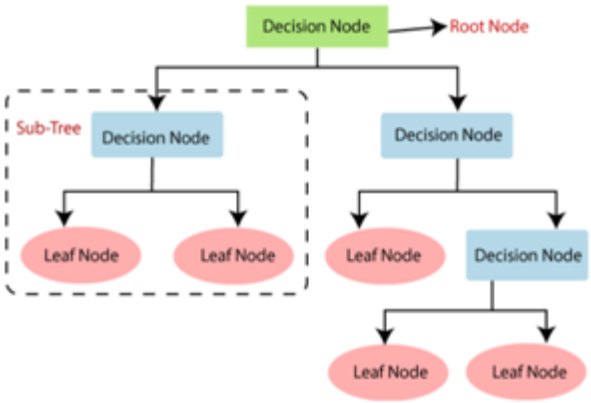
Fig: System Architecture

2.1 Logistic Regression Model

In this segment of the project, we are going to fit very first model, Logistic Regression. A logistic regression is utilized for displaying the result likelihood of a class like pass/come up, + / - and in this case – fraud/not fraud. To address twofold/absolute qualities, chump variables are utilized.
Logistic Regression is a stats model that in its essential structure utilizes a strategic capacity to show a double needy variable, albeit a lot more intricate augmentations exist.

2.2 Decision Tree Model

- Decision tree follows to the **Supervised learning method** that can be used for classification problems and Regression problems as well, but mostly it is go for solving Classification problems. It is a tree-organized classifier, where inward hubs/nodes address the highlights of a dataset, branches address the rule of decision and each leaf hub addresses the result for every input.
- In a Decision tree, two types of nodes are there, one of them is **decision node** and another one is **leaf node.** To make any decision and having branches multiple then decision nodes are utilized, where leaf-nodes are the result for those actions and they don't have any other branches.
- It is known as a choice/decision tree on the grounds that, because it like a tree, it begins with the root hub/node, which develops further branches and builds a tree-like construction.
- To fabricate a tree, we utilize the CART (Classification and Regression Tree Algorithm).
- A choice tree just poses an inquiry, and dependent on the appropriate response (Yes/No), it divides to the tree to sub-trees further.
- In the following below diagram defines the common form to the decision-tree:



(Fig: decision tree)

Terminologies:
Root Node: Where form the decision tree starts, it is said to Root node.
Parent/Child node: Root node is called parent and the other than root node are called child nodes.
Splitting: According to given situations the root node divides into further nodes, that's process is called splitting.
Pruning: The process of removing the unwanted branches from the tree.

Last/Leaf Node: The last node of the tree and also called the final outcome. And after getting the last node, the tree cannot be going to merge anymore.
Branch/Sub Tree: By splitting the tree, a complete tree is formed.

2.3 Random Forest

Random Forest are algorithm which is based on tree .It involves distinct trees and combining output for improving quality of model. This model is used to join distinct trees known as ensembling. It is a joining of weaker section. i.e. each distinct trees give solid beginner. It is used to describe supervise learning i.e. based on classification and regression.

Random Forest Working

It is used to determine random data. Data x1 is used for row a & column b.New data x2 is generated for a different bags randomly replaced with beginner data. Through data x1,1/3rd data of rows are left and it is known as Bag samples. New data x2 is trained to the recent models in which these Bag samples are accustomed determine not biased approximation . Out of b column, B << b column are used to create at each node through the data. B columns are pick out arbitrary. Normally the random choice of B is b/3 for regression & B is sqrt(b) used to classification. Distnict tree, no pruning takings place in other forest. Decision trees has 0 pruning approach to avoid completed. Pruning determine select a branch of trees in which that clues a with at a very low error test rate. Cross validation are accustomed dominate test at a false rate of sub tree. Each individual trees are grow and so final guess is generated through average.

Algorithm:

Step 1: Import the dataset

Step 2: Convert the data into data frames format

Step3: Do random oversampling using ROSE package

Step4: Decide the amount of data for training data and testing data

Step5: Give 70% data for training and remaining data for testing.

Step6: Assign train dataset to the models

Step7: Choose the algorithm among 3 different algorithms and create the model

Step8: Make predictions for test dataset for each algorithm

Step9: Calculate accuracy for each algorithm

Step10: Apply confusion matrix for each variable

Step11: Compare the algorithms for all the variables and find out the best algorithm.

III. PERFORMANCE METRICS AND EXPERIMENTAL RESULTS

3.1 Performance metrics:

It is used to check performance of a metrics. Matrix 2X2 table gives 4 outcome output. Different method like accuracy, sensitivity are described from confusion matrix.Three machine learning algorithm are buid for detecting fraudulent.

for evaluation 75 percent of data are used for training and 25 percent used for accuracy. Accurate data are for logistic regression are 93.5,decision tree are 95.6 and random forest 97.6.our goal here is to fraud transaction while minimizing ratio of in correct cheating classification:

The performance analysis for three different algorithms:

Feature Selection	Logistic regression	Decision tree	Random Forest
For 5 variables	87.2	89	90.1
For 10 variables	88.6	92.1	93.6
For all Variables	90.0	94.3	95.5

IV. OBSERVATIONS

The data set is incredibly favouritism carries with a very less amount of fraud found. The outcome is just 0.172% fraud are found. So this skewed set of data can be judge by the less number of crooked transactions. The dataset contains principle component analysis values to 28, which are namely from V1 to V28.• The 'Time' and 'Amount' attributes are not mutate data.

• In the data there is not any lost change within the data.

V. CONCLUSION

In this research paper, Different machine learning algorithm are used to detect to fraudulent in master card. All data are based on accuracy and specificity.We will use various machine learning algorithm i.e. random forest, decision tree and logistic regression for detecting fraud in master card system. Use of accuracy and error rate is to measure performance report.if we compare all three algorithm, we found that random forest is far better than decision tree and logistic regression.

REFERENCES

1. R. J. Bolton and D. J. Hand. Unsupervised profiling methods for fraud detection. In conference of Credit Scoring and Credit Connol VII, Edinburgh. UK, Sept 5-7,2001.
2. Khyati Chaudhary, Jyoti Yadav, Bhawna Mallick, ―A review of Fraud Detection Techniques: Credit Card‖, International Journal of Computer Applications (0975 – 8887) Volume 45– No.1, May 2012.
3. K. C. Cox, S. G. Eick, G. J. Wills, and R. J. Brachman. Visual data mining: Recognizing telephone calling fraud.J data processing and Knowledge Discover, 1(2):22>231, 1997.
4. Hollmn and Jaakko. Pmbabilistic Appmaches to Fraud Detecrion, Licentiate's ntesis. Helsinki University of Technology, Department of engineering science and Engineering, 1999.
5. K. Chaudhary, B. Mallick, "Credit Card Fraud: The study of its impact and detection techniques", International Journal of Computer Science and Network (IJCSN), vol. 1, no. 4, pp. 31-35, 2012, ISSN ISSN: 2277-5420.

AUTHORS PROFILE

Gautam Kumar, is an aspiring learner. He has Good knowledge of Machine..He is currently pursuing his Bachelor of Technology in Computer Science Engineering from Galgotias University. He has worked on multiple projects related to Machine Learning in his college Academics. He has also attended multiple workshops on Machine Learning during his college. His field of research is in machine learning. Email: gautamkumar7320@gmail.com

Shivanesh Kumar, is an aspiring learner.He has Good knowledge of Machine Learning with a strong inclination towards problem-solving and propelling data-driven decisions .He is currently pursuing his Bachelor of Technology in Computer Science Engineering from Galgotias University.He has worked on multiple projects related to Machine Learning in his college Academics .He has also attended multiple workshops on Machine Learning during his college.His field of research is machine learning. Email: shivaneshkumar97@gmail.com

Mr A Arulprakash, has many years of teaching and industry experience. He received his Ph.D. in Computer Science and researcher, practitioner, and educator. He is presently serving as Associate Professor at Galgotias University Greater Noida. He is an active in machine learning and regularly teach courses. He is a member of the technical program committees for several technical conferences and editorial member of reputed journals. He has chaired, participated in and presented at conferences and seminars in India. Apart from academic pursuits, he has shouldered many administrative responsibilities in various capacities. Email: arulprakash@galgotiasuniversity.edu.in