Credit Card Fraud Detection using Machine Learning

Priti Jadhav[1], Rutuja Ghadge[2], Utkarsha Halpatrao[3], Prof. Vilas Jadhav[4]

[1,2,3]Student, Dept. of Computer Engineering, M.G.M. College of Engineering and Technology, Kamothe, Maharashtra, India

[4]Prof. Dept. of Computer Engineering, M.G.M College of Engineering and Technology, Kamothe, Maharashtra, India

---***---

Abstract - *The use of credit cards has increased, due to a rapid growth in the E-commerce technology. Credit card is the most popular mode of payment, but number of fraud cases associated with it is also rising. Nowadays credit card fraud is a major concern. We have heard of huge amounts of cases where there were credit card frauds and many of the companies had lost fortune and lots of capital from recovering from Credit Card Fraud. Due to huge amount of data, analysing fraud transaction manually is impractical. However, it is possible using machine learning, through informative features. In this project we have worked on modelling of dataset using machine learning for credit card fraud detection. It includes modelling past credit card transactions with the data of the ones that are fraud. In this, the new transaction is fraud or not is checked by this model. In this model, it is expected that using Machine Learning algorithms to perform analysis and find out the number of fraudulent transactions that have taken place by using the given transaction data. The Machine learning algorithm used are Logistic Regression, Decision Tree Model, Artificial Neural Network, and Gradient Boosting Algorithm.*

Key Words: Credit card fraud, applications of machine learning, Logistic Regression, Decision Tree Model, Artificial Neural Network, Gradient Boosting Algorithm.

1. INTRODUCTION

Credit card fraud is referred to as illegal use of credit card or its information without the knowledge of the owner. Now a days there has been a tremendous use of credit cards for online transaction, due to which frequency of transactions is increasing and number of fraudulent transactions are also increasing rapidly. In the period of digitalization, there is necessity to identify credit card frauds. Fraud detection involves monitoring and analysing the behaviour of various users in order to estimate detect or avoid unwanted behaviour. To identify credit card fraud detection successfully, we need to understand the various technologies, algorithms and types involved in detecting credit card frauds. Different machine learning algorithm can differentiate transactions which are fraudulent or not. To find fraud, they need to passed dataset and knowledge of fraudulent transaction.

They analyze the dataset and classify all transactions. Fraud detection involves monitoring the activities of populations of users in order to evaluate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, and defaulting. Machine learning algorithms are employed to analyzes all the authorized transactions and report the suspicious ones.

1.1 Types and techniques of credit card frauds

a. Traditional techniques

Financial fraud

This occurs when an individual seeks to gain more credit than he or she is entitled to. An individual will apply for a credit card under his or her own name. The individual in this scenario will give wrong information with regards his or her financial status. Most commonly an individual exaggerates income or under values his or her outgoings. Banks try to safeguard themselves from the sort of a fraud by requiring the provision of documents to support an individual financial claims. For example, a card is issuer may ask an individual to provide three months of up- to- date account statements, or may be asked to see mortgage statements. Banks have also been known to telephone employers of the individual to confirm their employment. However, the fraudsters have been known to get around all the security procedures. Fraudster have and will be a forge documents and even false telephone numbers. Another security to check that card issuers carry out to safeguard themselves is credit checking. Credit checking reveals and individual's financial status, as well as the individual's current address. It is already plain to see that card issuers are fighting a difficult Battle against the fraudsters.

b. Modern Techniques

Triangulation

Triangulation is also a new phenomenon. Triangulation occurs when merchant offers a product at a very cheap price through a web-site. When a customer seeks to buy the product the merchant tells to customer to pay via e-mail once the item is delivered. The merchant uses a fraud card number to buy the product from a web site and sends the product to the customer, who then sends the merchant his or her credit card detail via e-mail. The merchant goes on operating using the credit card numbers that have been sent from the consumers to purchase products, appearing for a short time to be a legitimate merchant before he or she closes the web site and starts a new one.

Credit card generators

There is also the more sophisticated fraudsters, who use credit card generators computer emulation software that creates valid credit card numbers and expiry dates. These generators are highly reliable at creating valid credit card details and are available for free download off the internet. Making them available to many individuals who run fraudulent operations.

Some of the approaches to detection of such fraud are:

- Logistic Regression
- Artificial Neural Network
- Genetic Algorithm
- Gradient Boosting Algorithm
- Support Vector Machines
- Decision tree
- Fuzzy Logic
- Hidden Markov Model

2. SYSTEM ARCHITECTURE

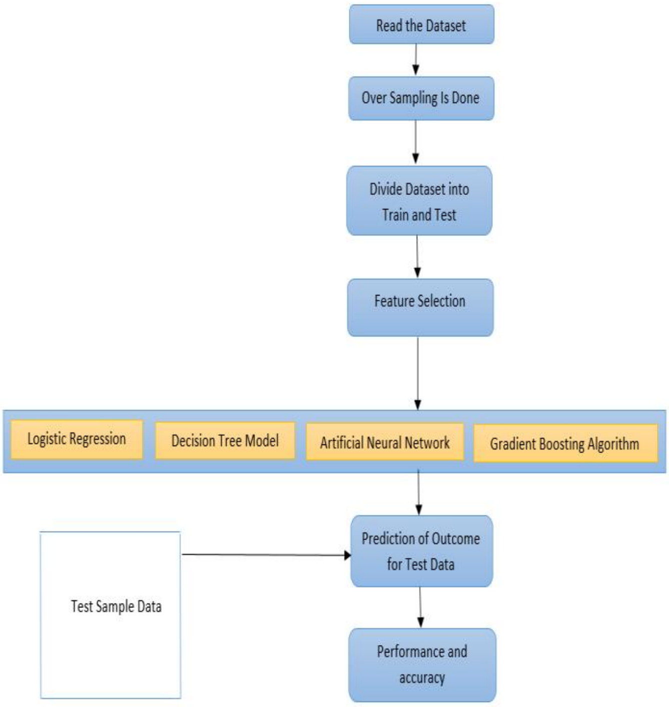

Fig -1: System Architecture

3. PROBLEM STATEMENT AND OBJECTIVE

Credit card frauds are increasing rapidly nowadays because of fraud financial loss is increasing drastically. Billions of amount is lost every year due to fraud. To analyse the fraud there is lack of research. Many machine learning algorithms are implemented to detect credit card fraud. Logistic Regression, Decision Tree Model, Artificial Neural Network, Gradient Boosting Algorithm and Hybrid algorithms are applied. The objective of the project is to detect credit card fraud detection by implementing machine learning algorithm with relevance to time and amount of transaction.

4. LITERATURE REVIEW

This paper presents a survey of various methods used in credit card fraud detection mechanisms. Now a day's credit card has become the most popular mode of payment for online transactions, but cases of fraud associated with it are also rising tremendously. Credit card frauds are increasing day by day despite the prevailing circumstances of the various techniques developed for its detection. Fraudsters are so expert that they find new ways for committing fraudulent transactions each day which demands constant implementation of ideas for its detection techniques. Most of the techniques based on Artificial Intelligence, Fuzzy logic, neural network, logistic regression, Machine learning, logistic regression, decision tree, Bayesian network, meta learning, Genetic Programming etc., these are evolved in detecting various credit card fraudulent transactions. [8]

There are a many number of literature or research papers available on credit card fraud detection in the public platform. A survey study conducted by Clifton Phua and his interns put forward methodologies used in this credit card fraud are data mining applications, Fraud detection, and adversarial detection. In another research paper Suman highlighted on the techniques like supervised and unsupervised learning for fraud detection. These methods were very efficient and effective in some areas of the domain but they are unable to give a permanent solution to the credit card fraud detection. In a similar research paper by Wen-Fang Yu and Na Wang, during which they used outlier mining, outlier detection mining and Distance sum algorithm to estimate fraud in an experiment conducted on the credit card data of some of the commercial banks. Outlier mining is a technique of data mining which is mostly applied in the fields related to finance or internet. It detects the fields which are not authentic. In this technique we take fields of customer's behaviour and on the basis of that we determine the distance between the observed value of the field and the predetermined value. [8]

There are many other literature which suggests a completely new perspective to detect fraud. In case of fraudulent transactions, there have been some research to make the alert feedback interaction more efficient. Whenever a fraudulent transaction is encountered an alert will be generated and sent to the authorised server system which reciprocally will deny the transaction. One of the aspects of Artificial Genetic Algorithm is an advancement to deal with the problem in a totally different aspect. This method resulted in correct fraud detection and less number of false alerts. [8]

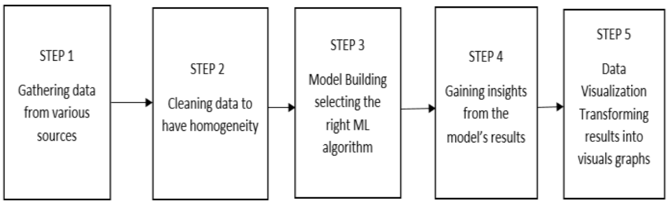
STEP 1
Gathering data from various sources
STEP 2
Cleaning data to have homogeneity
STEP 3
Model Building selecting the right ML algorithm
STEP 4
Gaining insights from the model's results
STEP 5
Data Visualization Transforming results into visuals graphs

Fig -2: Machine Learning Process

5. MACHINE LEARNING ALGORITHM USED FOR CREDIT CARD FRAUD DETECTION

5.1 Logistic Regression

Logistic Regression is a classification algorithm. Logistic regression is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary/categorical outcomes, we use dummy variables. One can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using a log of odds as a dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Logistic Regression is part of a larger class of algorithms known as the Generalized Linear Model (glm).In 1972, Nelder and Wedderburn proposed this model with an effort to provide a means of using linear regression to the problems which were not directly suited for the application of linear regression. In fact, they proposed a class of different models (linear regression, ANOVA, Poisson Regression, etc.) which included logistic regression as a special case.
The fundamental equation of generalized linear model is:

$$g\ (E(y)) = \alpha + \beta x1 + \gamma x2$$

Here, g() is the link function, E(y) is the expectation of target variable and $\alpha + \beta x1 + \gamma x2$ is the linear predictor (α,β,γ to be predicted). The role of link function is to 'link' the expectation of y to linear predictor. [7]

5.2 Decision Tree Model

A decision tree is a graph to represent choices and their results in the form of a tree. The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions. It is mostly used in Machine Learning and Data Mining applications using R. Examples of the use of decision trees are − predicting an email as spam or not spam, predicting of a tumor is cancerous, or predicting a loan as a good or bad credit risk based on the factors in each of these. Generally, a model is created with observed data also called training data. Then a set of validation data is used to verify and improve the model. R has packages that are used to create and visualize decision trees. For a new set of the predictor variable, we use this model to arrive at a decision on the category (yes/No, spam/not spam) of the data. [7]

5.3 Gradient Boost Algorithm

Boosting is famous ensemble learning technique with which we are not concerned with reducing the variance of learners like in Bagging where our aim is to cut back the high variance of learners by averaging lots of models fitted on bootstrapped data samples generated with replacement from training data, so as to avoid overfitting. In Boosting each tree or Model is grown or trained using the hard examples. By hard I mean all the training examples (xi, yi) for which a previous model produced incorrect output Y. Boosting boosts the performance of a simple base-learner by iteratively shifting the focus towards problematic training observations that are difficult to predict. Now that information from the previous model is fed to the next model. And the thing with boosting is that every new tree added to the mix will do better than the previous tree because it will learn from the mistakes of the previous models and try not to repeat them. Hence by this technique, it will eventually convert a weak learner to a strong learner which is better and more accurate in generalization for unseen test examples. [7]

6. STEPS INVOLVED

1. Import the required packages into python environment.
2. Import the data.
3. Process the data and Exploratory Data Analysis.
4. Selection of feature and Data splitting.
5. Six types of classification models are build.
6. using the evaluation metrics classification models are evaluated.

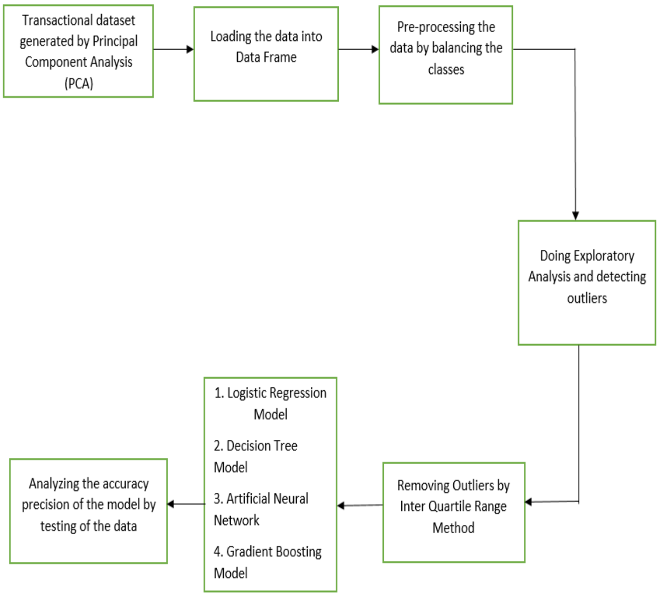
Transactional dataset generated by Principal Component Analysis (PCA)
Loading the data into Data Frame
Pre-processing the data by balancing the classes
Doing Exploratory Analysis and detecting outliers
Removing Outliers by Inter Quartile Range Method
1. Logistic Regression Model
2. Decision Tree Model
3. Artificial Neural Network
4. Gradient Boosting Model
Analyzing the accuracy precision of the model by testing of the data

Fig -3: Block Diagram

7. CONCLUSION

In this paper, we used machine learning techniques like Logistic regression, Decision Tree, Gradient boosting algorithm, Artificial neural network and Random forest to detect the fraud in credit card system. On analyzing the various models, we conclude that the difference in the accuracy of the Logistic Regression Model and the Gradient Boosting Model, for the current data the best model to use will be a Gradient Boosting Model. Such type of data is best fitted using the Gradient Boosting Model method. Thus, this research allowed us to learn how to develop our credit card fraud detection model using machine learning. We used a range of Machine Learning algorithms to implement this model. We learned how data are often analyzed and visualized to find fraudulent transactions from other forms of data.

8. REFERENCES

[1] Credit Card Fraud Detection Based on Transaction Behaviour-by John Richard D. Kho, Larry A. Vea" published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 20107.

[2] CLIFTON PHUA1, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2 " A Comprehensive Survey of Data Mining-based Fraud Detection Research" published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia.

[3] "Survey Paper on Credit Card Fraud Detection by Suman", Research Scholar, GJUS&T Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014.

[4] "Research on Credit Card Fraud Detection Model Based on Distance Sum –by Wen-Fang YU and Na Wang" published by 2009 International Joint Conference on Artificial Intelligence.

[5] "Credit Card Fraud Detection through Parenclitic Network Analysis-By Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral" published by Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages .

[6] "Credit Card Fraud Detection: A Realistic Modelling and a Novel Learning Strategy" published by IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, AUGUST 2018.

[7] "Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models by Navanshu Khare and Saad Yunus Sait" published by International Journal of Pure and Applied Mathematics, Volume 118 No. 20, 2018.

[8] "Fraud Detection in Credit Card using Machine Learning Techniques by Mr.Manohar.s, Arvind Bedi, Shashank kumar and Shounak kr" Singh published by International Research Journal of Engineering and Technology (IRJET) Volume: 07 Issue: 04, Apr 2020.