IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor 6.252

Volume 2, Issue 1, July 2022

Detection of Credit Card Fraud Transactions using Machine Learning based Algorithm

Swetha T[1], Bellam Narendra Nath[2], Manjunath[3], Govinda N[4], H V Naveen Kumar[5]

Project Guide, Department of Computer Science and Engineering[1]

Projecties, Department of Computer Science and Engineering[2,3,4,5]

Sri Jagadguru Chandrashekaranatha Swamiji Institute of Technology, Chikkaballapura, Karnataka, India

Abstract: *The rapid growth in E-Commerce industry has lead to anexponential increase in the use of credit cards for online purchases and consequently they has been surge in the fraud related to it .In recent years, For banks has become verydifficult for detecting the fraud in credit card system. Machinelearning plays a vital role for detecting the credit card fraud in the transactions. For predicting these transactions banks make use of various machine learning methodologies, past data has been collected and new features are been used for enhancing the predictive power. The performance of fraud detecting in credit card transactions is greatly affected by the sampling approach on data-set, selection of variables and detection techniques used. This paper investigates the performance of SVM, decision tree and random forest for credit card fraud detection. Dataset of credit card transactions iscollected from kaggle and it contains a total of 2,84,808 credit card transactions of a European bank data set. It considers fraud transactions as the "positive class" and genuine ones as the "negative class" .The data set is highly imbalanced, it has about 0.172% of fraud transactions and the rest are genuine transactions. The author has been done oversampling to balance the data set, which resulted in 60% offraud transactions and 40% genuine ones. The three techniques are applied for the dataset and work is implemented in R language. The performance of the techniques is evaluated for different variables based on sensitivity, specificity, accuracy and error rate. The result shows of accuracy for SVM, Decision tree and random forest classifier are 90.0, 94.3, 95.5 respectively. The comparative results show that the Random forest performs better than the SVM and decision tree techniques.*

Keywords: Fraud detection, Credit card, SVM, Decision tree, Random forest.

I. INTRODUCTION

Credit card fraud is a huge ranging termfor theft and fraud committed using or involving at the time ofpayment by using this card. The purpose may be to purchase goods without paying, or to transfer unauthorized funds from an account. Credit card fraud is also an add on to identity theft. As per the information from the United States Federal Trade Commission, the theft rate of identity had been holding stable during the mid 2000s, but it was increased by 21 percent in 2008. Even though credit card fraud, that crime which most people associate with ID theft, decreased as a percentage of all ID theft complaints In 2000, out of 13 billiontransactions made annually, approximately 10 million or one out of every 1300 transactions turned out to be fraudulent.

Also, 0.05% (5 out of every 10,000) of all monthly active accounts was fraudulent. Today, fraud detection systems are introduced to control one-twelfth of one percent of all transactions processed which still translates into billions of dollars in losses. Credit Card Fraud is one of the biggestthreats to business establishments today. However, to combat the fraud effectively, it is important to first understand the mechanisms of executing a fraud. Credit card fraudsters employ a large number of ways to commit fraud. In simple terms, Credit Card Fraud is defined as "when an individual uses another individuals' credit card for personal reasons while the owner of the card and the card issuer are not aware of the fact that the card is being used". Card fraud begins either with the theft of the physical card or with the important data associated with the account, including the card account number or other information that necessarily be available to a merchant during a permissible transaction. Card numbers generally the Primary Account Number (PAN) are often reprinted on the card, and a magnetic stripe on the back contains the data in machine-readable format. It contains the following Fields:

DOI 10.48175/IJARSCT-5742

- Name of card holder
- Card number
- Expiration date
- Verification/CVV code
- Type of card

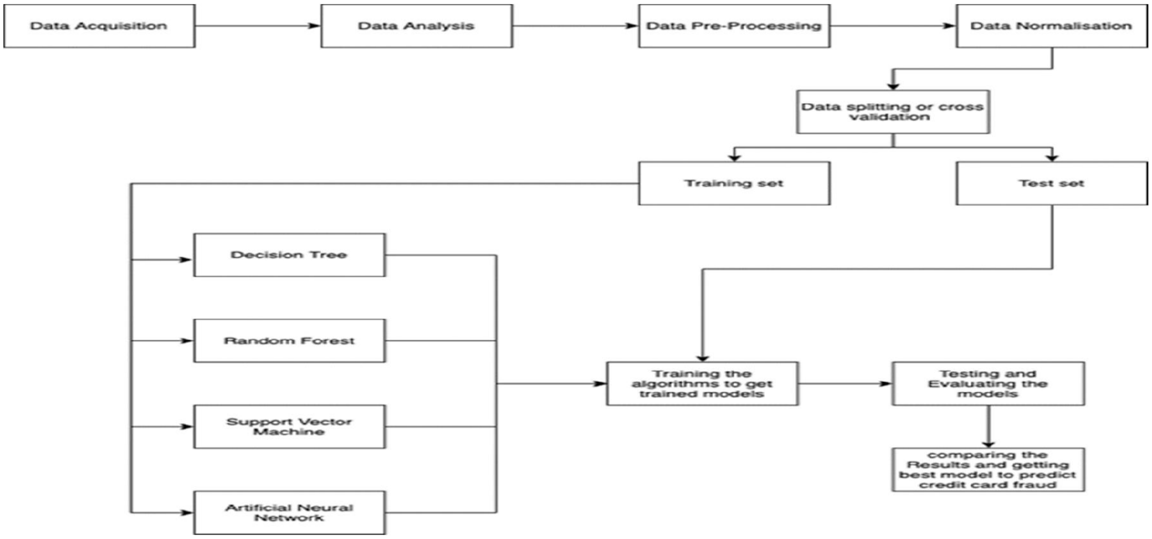
There are more methods to commit credit card fraud. Fraudsters are very talented and fast moving people. In the Traditional approach, to be identified by this paper is Application Fraud, where a person will give the wrong information about himself to get a credit card. There is also the unauthorized use of Lost and Stolen Cards, which makes up a significant area of credit card fraud. There are moreenlightened credit card fraudsters, starting with those whoproduce Fake and Doctored Cards; there are also those who use Skimming to commit fraud. They will get this informationheld on either the magnetic strip on the back of the credit card,or the data stored on the smart chip is copied from one card to another. Site Cloning and False Merchant Sites on the Internetare getting a popular method of fraud for many criminals with a skilled ability for hacking. Such sites are developed to get people to hand over their credit card details without knowing they have been swindled.

Rest of the paper is described as follows: section 2 describes the related work about the credit card system, section 3 described the proposed system architecture and methodology, section 4 shows the performance analysis and results, section5 shows the conclusion.

II. RELATED WORK

A.Shen etal (2017) demonstrate the efficiency of classificationmodels to credit card fraud detection problem and the authors proposed the three classification models ie., decision tree,neural network and logistic regression. Among the three models neural network and logistic regression outperforms than the decision tree. M.J.Islam et al (2017) proposed the probability theory frame work for making decision under uncertainty. After reviewing Bayesian theory, naïve bayes classifier and k-nearest neighbor classifier is implemented andapplied to the dataset for credit card system.Y. Sahin and E. Duman(2019) has cited the research for credit card frauddetection and used seven classification methods took a major role .In this work they have included decision trees and SVMsto decrease the risk of the banks. They have suggested Artificial Neural networks and Logistic Regression classification models are more helpful to improve theperformance in detecting the frauds. Y. Sahin, E. Duman(2020) has cited the research , used Artificial Neural Network and Logistic Regression Classification andexplained ANN classifiers outperform LR classifiers in solving the problem under investigation. Here the trainingdata sets distribution became more biased and the distribution of the training data sets became more biased and the efficiency of all models decreased in catching the fraudulent transactions.

III. PROPOSED TECHNIQUE


Data Acquisition
Data Analysis
Data Pre-Processing
Data Normalisation
Data splitting or cross validation
Training set
Test set
Decision Tree
Random Forest
Support Vector Machine
Artificial Neural Network
Training the algorithms to get trained models
Testing and Evaluating the models
comparing the Results and getting best model to predict credit card fraud

The proposed techniques are used in this paper, for detecting the frauds in credit card system. The comparison are made for different machine learning algorithms such as Logistic Regression, Decision Trees, Random Forest, to determine which algorithm gives suits best and can be adapted by credit card merchants for identifying fraud transactions. The Figure1 shows the architectural diagram for representing the overall system framework

IV. DECISION TREE ALGORITHM

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.

4.1 Types of Decision Tree

1. Categorical Variable Decision Tree: Decision Tree which has categorical target variable then it called as categorical variable decision tree.
2. Continuous Variable Decision Tree: Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree

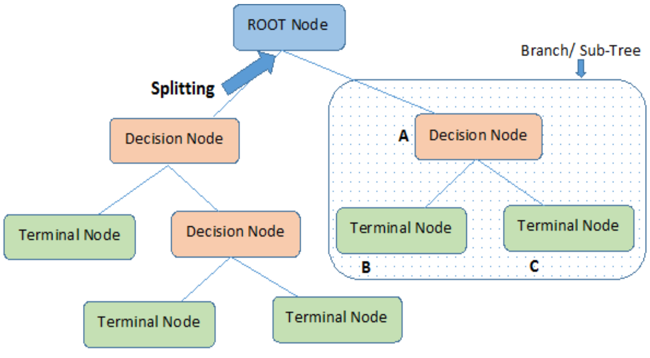
4.2 Terminology of Decision Tree

1. Root Node: It represents entire population or sample and this further gets divided into two or more homogeneous sets.
2. Splitting: It is a process of dividing a node into two or more sub-nodes.
3. Decision Node: When a sub-node splits into further sub- nodes, then it is called decision node.
4. Leaf/ Terminal Node: Nodes do not split is called Leaf or Terminal node.
5. Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.
6. Branch / Sub-Tree: A sub section of entire tree is called branch or sub-tree.
7. Parent and Child Node: A node, which is divided into sub-nodes is called parent node of sub-nodes where as sub-nodes are the child of parent node.

4.3 Working of Decision Tree

Decision trees use multiple algorithms to decide to split a node in two or more sub- nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that purity of the node increases with respect to the target variable. Decision tree splits the nodes on all available variables and then selects the split which resultsin most homogeneous sub-nodes.

1. Gini Index
2. Information Gain
3. Chi Square
4. Reduction of Variance



Note:- A is parent node of B and C.

DOI 10.48175/IJARSCT-5742

V. RANDOM FOREST

Random forest is a tree-based algorithm which involves building several trees and combining with the output to improve generalization ability of the model. This method of combining trees is known as an ensemble method. Ensembling is nothing but a combination of weak learners (individual trees) to produce a strong learner. Random Forest can be used to solve regression and classification problems. In regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical.

5.1 Working of Random Forest

Bagging Algorithm is used to create random samples. Data setD1 is given for n rows and m columns and new data set D2 is created for sampling n cases at random with replacement fromthe original data. From dataset D1,$1/3^{rd}$ of rows are left out and is known as Out of Bag samples. Then, new dataset D2 is trained to this models and Out of Bag samples is used to determine unbiased estimate of the error. Out of m columns,M << m columns are selected at each node in the data set. TheM columns are selected at random. Usually, the default choiceof M, is m/3 for regression tree and M is sqrt(m) for classification tree. Unlike a tree, no pruning takes place in random forest i.e, each tree is grown fully. In decision trees, pruning is a method to avoid over fitting. Pruning meansselecting a sub tree that leads to the lowest test error rate. Cross validation is used to determine the test error rate of a sub tree. Several trees are grown and the final prediction is obtained by averaging or voting.

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 algorithmStep9: Calculate accuracy for each algorithm

Step10: Apply confusion matrix for each variable

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

Table 2: Algorithm steps for finding the best algorithm

VI. SUPPORT VECTOR MACHINES

(SVMs) are a popular machine learning method for classification, regression, & other learning tasks. LIBSVM is a library for Support Vector Machines (SVMs). A typical use of LIBSVM involves two steps: first, training a data set to obtain a model & second, using the model to predict information of a testing data set. For SVC & SVR, LIBSVM can also output probability estimates. Many extensions of LIBSVM are available at libsvm tools. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

1. Set up the training data for model creation
2. Set up SVM's parameters
3. SVM Trainer
4. SVM Predictor

VII. PERFORMANCE METRICS AND EXPERIMENTAL RESULTS

7.1 Performance Metrics

The basic performance measures derived from the confusion matrix. The confusion matrix is a 2 by 2 matrix table contains four outcomes produced by the binary classifier. Various measures such as sensitivity, specificity, accuracy and error rate are derived from the confusion matrix.

Accuracy:

Accuracy is calculated as the total number of two correct predictions(A+B) divided by the total number of the dataset(C+D).It is calculated as (1-error rate).

Accuracy=A+B/C+D (4.1)

Whereas, A=True Positive B=True Negative C=Positive D=Negative

Error rate:

Error rate is calculated as the total number of two incorrect predictions(F+E) divided by the total number of the dataset(C+D).

Error rate=F+E/C+D (4.2)

Whereas, E=False PositiveF=False Negative, C=Positive, D=Negative

Sensitivity:

Sensitivity is calculated as the number of correct positive predictions(A) divided by the total number of positives(C).

Sensitivity=A/C (4.3)

Specificity:

Specificity is calculated as the number of correct negative predictions(B) divided by the total number of negatives(D).

Specificity=B/D. (4.4)

Accuracy, Error-rate, Sensitivity and Specificity are used to report the performance of the system to detect the fraud in the credit card.

In this paper, three machine learning algorithms are developed to detect the fraud in credit card system. To evaluate the algorithms, 80% of the dataset is used for training and 20% is used for testing and validation. Accuracy, error rate, sensitivity and specificity are used to evaluate for different variables for three algorithms as shown in Table 3. The accuracy result is shown for SVM; Decision tree and random forest classifier are 92.7, 95.8, and 97.6 respectively. The comparative results show that the Random forest performs better than the SVM and decision tree techniques.

Table 3: Performance analysis for three different algorithms

Feature Selection	SVM	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

VIII. CONCLUSION

In this paper, Machine learning technique like SVM, Decision Tree and Random Forest were used to detect the fraud in credit card system. Sensitivity, Specificity, accuracy and error rate are used to evaluate the performance for the proposed system. The accuracy for SVM, Decision tree and random forest classifier are 90.0, 94.3, and 95.5 respectively. By comparing all the three methods, found that random forest classifier is better than the and decision tree.

REFERENCES

[1]. Andrew. Y. Ng, Michael. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes", Advances in neural information processing systems, vol. 2, pp. 841-848, 2002.

[2]. A. Shen, R. Tong, Y. Deng, "Application of classification models on credit card fraud detection", Service Systems and Service Management 2007 International Conference, pp. 1-4, 2007.

[3]. A. C. Bahnsen, A. Stojanovic, D. Aouada, B. Ottersten, "Cost sensitive credit card fraud detection using Bayes minimum risk", Machine Learning and Applications (ICMLA). 2013 12th International Conference, vol. 1, pp. 333-338, 2013.

[4]. B.Meena, I.S.L.Sarwani, S.V.S.S.Lakshmi," Web Service mining and its techniques in Web Mining" IJAEGT,Volume 2,Issue 1 , Page No.385-389.

[5]. F. N. Ogwueleka, "Data Mining Application in Credit Card Fraud Detection System", Journal of Engineering Science and Technology, vol. 6, no. 3, pp. 311-322, 2011.

[6]. G. Singh, R. Gupta, A. Rastogi, M. D. S. Chandel, A. Riyaz, "A Machine Learning Approach for Detection of Fraud based on SVM", International Journal of Scientific Engineering and Technology, vol. 1, no. 3, pp. 194-198, 2012, ISSN ISSN: 2277-1581.

[7]. 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.

[8]. M. J. Islam, Q. M. J. Wu, M. Ahmadi, M. A. Sid- Ahmed, "Investigating the Performance of Naive-Bayes Classifiers and KNearestNeighbor Classifiers", IEEE International Conference on Convergence Information Technology, pp. 1541-1546, 2007.

[9]. R. Wheeler, S. Aitken, "Multiple algorithms for fraud detection" in Knowledge-Based Systems, Elsevier, vol. 13, no. 2, pp. 93-99, 2000.

[10]. S. Patil, H. Somavanshi, J. Gaikwad, A. Deshmane, R. Badgujar, "Credit Card Fraud Detection Using Decision Tree Induction Algorithm", International Journal of Computer Science and Mobile Computing (IJCSMC), vol. 4, no. 4, pp. 92-95, 2015, ISSN ISSN: 2320-088X.

[11]. S. Maes, K. Tuyls, B. Vanschoenwinkel, B. Manderick, "Credit card fraud detection using Bayesian and neural networks", Proceedings of the 1st international naiso congress on neuro fuzzy technologies, pp. 261-270, 2002.

[12]. S. Bhattacharyya, S. Jha, K. Tharakunnel, J. C. Westland, "Data mining for credit card fraud: A comparative study", Decision Support Systems, vol. 50, no. 3, pp. 602-613, 2011.

[13]. Y. Sahin, E. Duman, "Detecting credit card fraud by ANN and logistic regression", Innovations in Intelligent Systems and Applications (INISTA) 2011 International Symposium, pp. 315-319, 2011.

[14]. Selvani Deepthi Kavila,LAKSHMI S.V.S.S.,RAJESH B " Automated Essay Scoring using Feature Extraction Method " IJCER ,volume 7,issue 4(L), Page No. 12161-12165.

[15]. S.V.S.S.Lakshmi,K.S.Deepthi,Ch.Suresh "Text Summarization basing on Font and Cue-phrase

DOI 10.48175/IJARSCT-5742