International Journal of Engineering& Scientific Research
Vol.8 Issue 10, October 2020,
ISSN: 2347-6532 Impact Factor: 6.660
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Credit Card Skullduggery Detection System Using Machine Learning Techniques

S.K.SARAVANAN
GNK SURESHBABU

Abstract (10pt)

Frauds in Credit Card are an emerging problem with more consequences in the financial sector and even many techniques have been discovered. The huge volumes of complex data analyzed and automate by applying Data mining techniques successfully. Data mining techniques have also played a vital role in the detection of credit card fraud in online transactions. The main aim of the paper is to design and develop a novel fraud detection method for Transaction Data, with an objective, to analyze the past transaction details of the customers and extract the behavioral patterns. The cardholders are categorized into different groups based on their transaction amount. Banks make use of various machine learning methodologies, past data have been collected and new features are been used for enhancing the predictive power for these transactions. In credit card transactions, the performance of fraud detecting is affected greatly by the sampling approach on every data-set, selection of decision variables and detection techniques used. This proposed work investigates and checks the performance of Support Vector Machines, Decision tree, Random Forest and Logistic Regression on highly skewed credit card fraud data. The European credit card fraud dataset containing around 285,000 transactions have been taken and tested. The above mentioned data mining techniques are applied on the raw and preprocessed data and the performance of these techniques are evaluated based on Accuracy, Sensitivity, Specificity and Precision.

Keywords:

Fraud Detection;
Credit Card;
Logistic Regression;
SVM;
Random Forest.

Author correspondence:

First Author,
Assistant Professor (Sel.G),
SRM Valliammai Engineering College,
Kattangulathur-603 203, Chennai.

Second Author,
Associate Professor,
Acharya Institute of Technology,
Bengaluru

1. Introduction

Generally Credit card refers to a card that is assigned to the customer (cardholder), allowing them to purchase goods and services within the credit limit or withdraw cash in advance. Credit card provides the cardholder an advantage of the time that is it provides time for their customers to repay later in a prescribed time, by carrying this to the next billing cycle. Without having any risks, a significant amount can be withdrawn without the knowledge of the owner, in a short period. The Fraudsters always try to make every fraudulent transaction legitimate, which makes the fraud detection very challenging and difficult task to detect [1].

According to the information from the United States Federal Trade Commission [2], the theft rate of identity had been holding stable during the mid 2000s, but it was increased by 25 percent in 2009. Although the Credit Card fraud, in which most people associate with ID theft, decreased as a percentage of all ID theft complaints in 2000, out of 12 billion transactions made annually, approximately 10 million or one out of every 1200 transactions turned out to be fraudulent.

The 5 out of every 10,000 monthly active accounts was fraudulent. The Fraud detection systems are introduced to control one-twelfth of one percent (1%) of all transactions processed which still translates into billions of dollars in losses. Fraud in Credit Card is one of the biggest and important threats to business establishments today. To control the fraud effectively, it is important to understand the mechanisms of executing a fraud first. The fraudsters apply a large number of ways to commit frauds in Credit card fraud. 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".

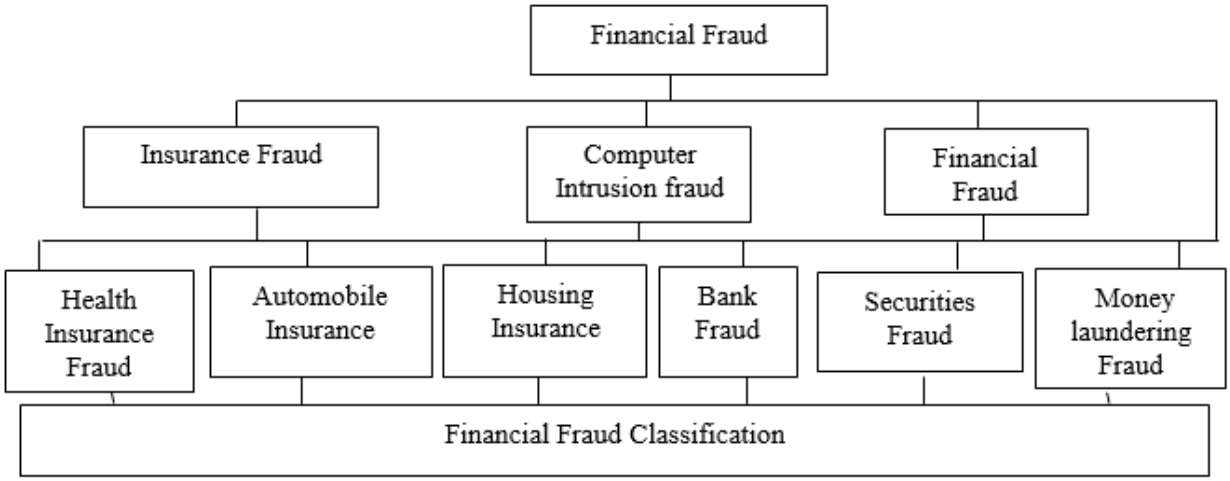


Fig. 1.1 Taxonomy of Frauds

With different frauds, mostly credit cards frauds are often in the news for the past few years, frauds are in the top of mind for most the people's mind. The Credit card dataset is highly imbalanced because there will be more number of legitimate transaction when compared with a fraudulent one.

The datasets of credit card transaction are rarely available, highly imbalanced and skewed. The selection of the optimal feature variables for the models and suitable metric is most important part of data mining to evaluate performance of techniques on skewed credit card fraud data. There are number of difficulties are associated with credit card detection, namely fraudulent behavior profile is dynamic, that is fraudulent transactions are tend to look like legitimate ones, in credit card fraud detection.

2. Literature Survey

There are multiple Supervised and Semi-Supervised machine learning techniques are used for fraud detection, but our aim is to overcome three main challenges with card frauds related dataset i.e., Strong Class imbalance, the Inclusion of Labeled and Unlabelled

samples, and to Increase the ability to process a large number of transactions.

Decision Trees, Naive Bayes Classification, Least Squares Regression, Logistic Regression and Support Vector Machine [3] are different supervised machine learning algorithms used to detect fraudulent transactions in real-time datasets. Two methods Random-tree-based random forest and CART-based under random forests are used to train the behavioral features of normal and abnormal transactions. On small data sets, even though the random forest method obtains good results there are still some problems in case of imbalanced data. On solving the above-mentioned problem and the algorithm of the random forest itself should be improved in the future work.

In [1] this paper represents a research about a case study which is involving credit card fraud detection, has shown that by the clustering attributes neuronal inputs can be minimized where the data normalization is applied before Cluster Analysis and with results obtained from the use of Cluster Analysis and ANN on fraud detection. Results can be obtained by using normalized data and data should be Multi Layer Perception (MLP) trained. The significance of this paper was to find new methods for fraud detection and to increase the accuracy of results.

In [3] various modern techniques based on Artificial Intelligence, Sequence Alignment, Machine learning, Data Mining, Genetic Programming, etc. have been evolved and is still evolving to find fraudulent transactions in credit card. A good, sound and clear understanding on all these approaches are needed will lead to an efficient credit card fraud detection system certainly. The survey of various data mining techniques used in credit card fraud detection mechanisms have been shown in this paper along with the evaluation of each methodology based on certain design criteria. Conducting a survey to compare different credit card fraud detection algorithm and to find the most suitable algorithm to solve the problem is the significance of this paper.

The performances of Logistic Regression, K-Nearest Neighbour, and Naïve Bayes [6] are analyzed on highly skewed credit card fraud data where research is carried out on examining meta-classifiers and meta-learning approaches in handling highly imbalanced credit card fraud data. The supervised learning methods can be used there may fail at certain cases of detecting the fraud cases. A model of deep auto-encoder and Restricted Boltzmann Machine (RBM) that can construct normal transactions to find anomalies from normal patterns.

The expected outcome of this research is to detect the credit card fraud in the dataset obtained from ULB by applying SVM, Decision Tree, Logistic regression, Random Forest [4][5] and to evaluate their Accuracy, Sensitivity, Specificity, Precision using different models and compare and collate them in order to state and find the best possible model to solve the credit card fraud detection problem.

3. Proposed Technique:

In this paper, the techniques used are all for detecting the frauds in credit card system. We made comparison for different Machine Learning Algorithms such as Logistic Regression, Decision Trees and Random Forest, in order to determine which algorithm suits best and can be adapted by the card merchants for identifying fraud transactions. The card transactions are always unfamiliar when compared to previous transactions made the

customer. This unfamiliarity is a very difficult problem in real-world when are called the Concept Drift Problems [7]. A variable which changes over time and in unforeseen ways is called Concept Drift and these variables cause a high imbalance in data. The main aim of our research is to overcome the Concept drift problem to implement on real-world scenario. The table 3.1 shows basic features that are captured when any transaction is made.

Name of Attribute	Meaning
Transaction_ id	Transaction Identification number
Cardholder_ id	Cardholder's Unique Identification number.
Amount	The customer credited or transferred amount in a particular transaction.
Time	To identify, when the transaction was made contains details like time and date.
Label	This specifies whether the transaction is genuine or fraudulent

Table 3.1: Raw features of credit card transactions

Processing or Algorithm Steps:

Step 1: Read the dataset.

Step 2: To make it balanced, Random Sampling is done on the data set.

Step 3: The dataset divides into two parts, that is, Training dataset and Testing dataset.

Step 4: The feature selection techniques are applied for the proposed models.

Step5: To know the efficiency for different algorithms, Accuracy and Performance metrics have been calculated.

Step6: For the given dataset, retrieve the best algorithm based on efficiency.

This system's ability is to automatically learn and improve from experience without being explicitly programmed is called machine learning and it focuses on the development of computer programs that can access data and use it learn for themselves. An algorithm can be stated for the classifier that is used to implement classification especially in concrete implementation, it also refers to a mathematical function implemented by algorithm that will map input data into category[9]. It is an instance of supervised learning i.e. where training set of correctly identified observations is available.

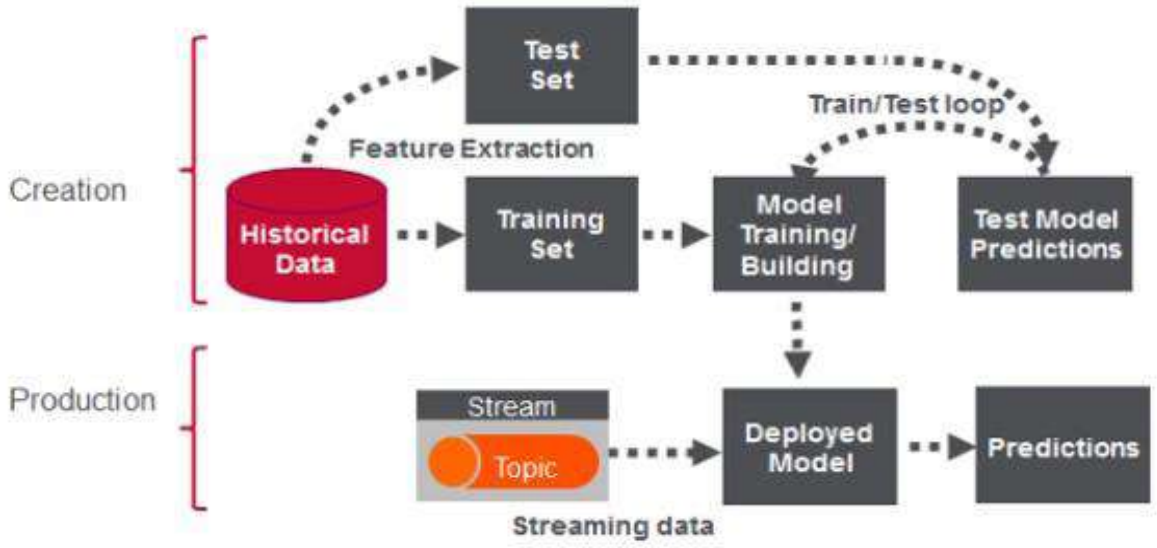


Figure 3.1: Classifier Steps

The dataset contains transactions made by a cardholder in duration in 2 days in the month of September 2013. Where there are total 285,000 transactions among which there are 492 i.e., 0.172% transactions are fraudulent transactions. This dataset is highly unbalanced. Since providing transaction details of a customer is considered to issue related to confidentiality, therefore most of the features in the dataset are transformed using Principal Component Analyses (PCA) [10]. The features V1, V2, V3,..., V28 are PCA applied features and rest i.e., 'Time', 'Amount' and 'Class' are non-PCA applied features, as shown in table 3.2.

S.No	Feature	Description
1	Time	To specify the elapses between the first transaction and current transaction time in seconds
2	Amount	Transaction Amount
3	Class	0 – Non fraud 1 - Fraud

Table 3.2: Attributes of European dataset

3.1 Logistic Regression:

Logistic Regression [11], one of the classification algorithm is used to predict a binary values in a given set of independent variables (1 / 0, Yes / No, True / False). To represent binary / categorical values dummy variables are used. The purpose of special case in the logistic regression is a linear regression, when the resulting variable is categorical then the log of odds are used for dependent variable and also it predicts the probability of occurrence of an event by fitting data to a logistic function.

The Regression is a regression model where the dependent variable is categorical and analyzes the relationship between multiple independent variables. There are many types of logistic regression model such as binary logistic model, multiple logistic model, and

binomial logistic models. To estimate the probability of a binary response based on one or more predictors, Binary Logistic Regression model is used.

3.2 Decision Tree Algorithm:

The Decision Tree is a **supervised learning technique** that can be used for both Regression and Classification problems, but it is mostly preferred for solving Classification problems. DT is a tree-structured classifier model, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.** In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

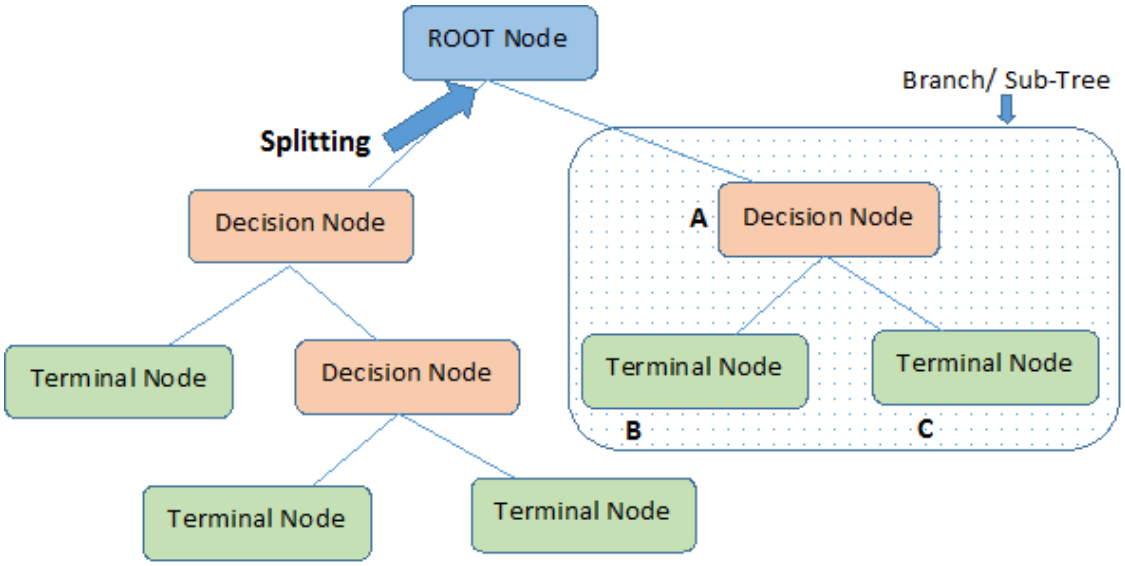
The Decision tree [12] is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. We split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables in this technique.

Decision Tree Types

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

Terminologies used in Decision Trees:

1. Root Node: It represents entire population or samples and further gets divided into two or more homogeneous sets.
2. Splitting: 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.

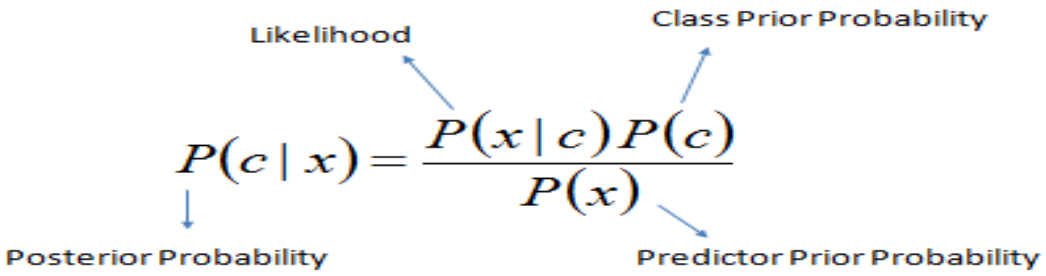


Note:- A is parent node of B and C.

3.2.1 Decision Tree Model

3.3 Random Forest

The best example for regression and classification methods is a Random Forest [14] algorithm. It is a collection of decision tree classifiers. The Random forest algorithm has advantage over decision tree as it corrects the habit of over fitting to their training set. A subset of the training set is sampled randomly so that to train each individual tree and then a decision tree is built. Each node then splits on a feature selected from a random subset of the full feature set. The large data sets with many features and data instances training is extremely fast in random forest and because each tree is trained independently of the others. The Random Forest algorithm has been found to provide a good estimate of the generalization error and to be resistant to over fitting.



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

Random forest algorithms rank the importance of variables in a regression or classification problem in a natural way.

4. Experiments

We have experimented few models on original as well as Synthetic Minority Oversampling Technique (SMOTE) dataset. The results are tabulated, which shows great differences in accuracy, precision and Mathews Correlation Coefficient (MCC) as well. We even used one-class SVM which can be best used for binary class datasets. Since we have 2 classes in our dataset we can use one-class SVM as well.

The data set of credit card is taken from the source, cleaning and validation is performed on it which includes removal of redundancy, empty spaces filling in columns, converting necessary variable into factors or classes then data is divided into 2 parts, one is training dataset and another one is test data set. The original sample is randomly partitioned into k equal sized subsamples by using the K fold cross validation is done. A single subsample is retained among the k subsamples as the validation data for testing the model, and the remaining k −1 subsamples are used as training data. The models are created for Logistic regression, Decision tree, SVM, Random Forest and then accuracy, sensitivity, specificity, precision are calculated and a comparison is made.

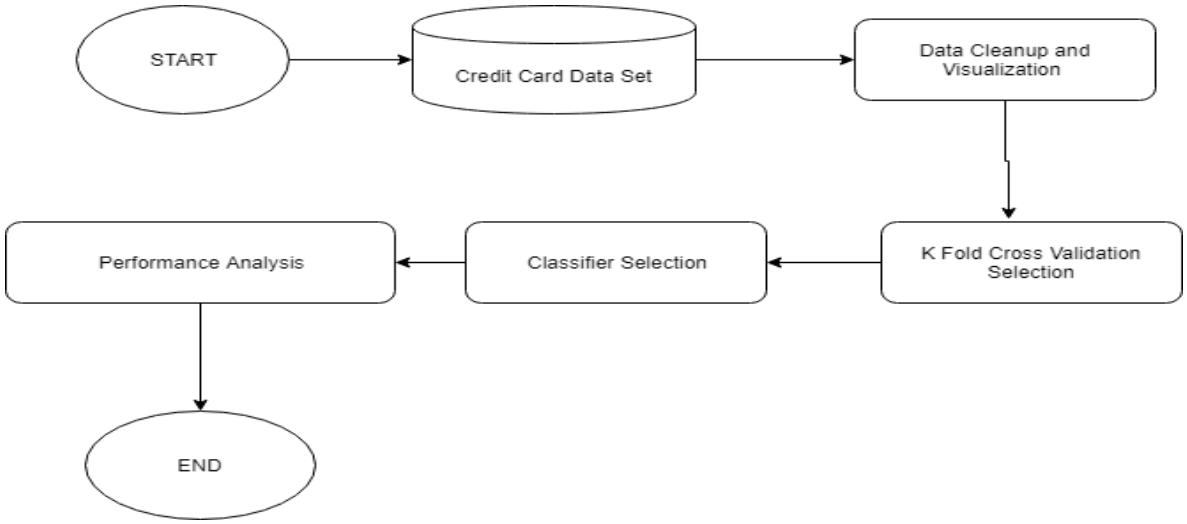


Fig. 4.1: Experiment Model

Credit card transactions made by European cardholders around September 2013 contains the dataset and the occurrence of transactions that happened in two days is presented by this dataset, consisting of 285,000 transactions. The dataset is highly unbalanced and skewed towards the positive class and positive class that is fraud cases make up 0.173% of the transactions data. This contains only numerical (continuous) input variables which are as a result of a Principal Component Analysis (PCA) feature selection transformation resulting to 28 principal components and total of 30 input features are utilized in this study. The behavioral characteristic of the card is shown by a variable of each profile usage representing the spending habits of the customers along with days of the month, hours of the day, geographical locations, or type of the merchant where the transaction takes place. These variables are used to create a model which distinguishes fraudulent activities afterwards. Due to confidentiality issues, the details and background information of the features cannot be presented. The ‘time’ feature stores the seconds that has elapsed between each transaction along with first transaction in the dataset. The 'amount' feature is

the transaction amount. Feature 'class' is the target class for the binary classification and it takes value 1 for positive case (fraud) and 0 for negative case (non fraud).

The basic four basic metrics are used in evaluating the experiments, namely True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) rates metric respectively.

$$\textbf{Accuracy} = TP + TN / TP + TN + FP + FN \quad (1)$$

$$\textbf{Precision} = TP / TP + FP \quad (2)$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (3)$$

The Sensitivity (Recall) gives Accuracy on Positive (fraud) cases classification. The Specificity gives Accuracy on Negative (legitimate) cases classification. The Precision gives Accuracy in cases classified as fraud (positive).

5. Performance and Results

In this paper, we have developed three machine learning algorithms to detect the fraud in credit card system. In order to evaluate these algorithms, 70% of the dataset is used for Training and 30% is used for Testing and Validation. Accuracy, Error rate, Sensitivity and Specificity are used to evaluate for different variables for these three algorithms as shown in Table 5.1. The accuracy result is shown for Logistic Regression, Decision tree and Random Forest classifier are 92.7, 95.8, and 97.6 respectively. The comparative results show that the Random forest algorithm performs better than the logistic regression and decision tree techniques.

Feature Selection	Logistic Regression	Decision Tree	Random Forest
For variables	87.2	89	90.1
For 10 variables	88.6	92.1	93.6
For 5 variables	92.7	95.8	97.6

Table 5.1: Accuracy of three algorithms

Actual / Predicted	Not a Fraud	Fraud
Not a Fraud	True Positive	False Positive
Fraud	False Negative	True Negative

Table 5.2: Confusion Matrix Format

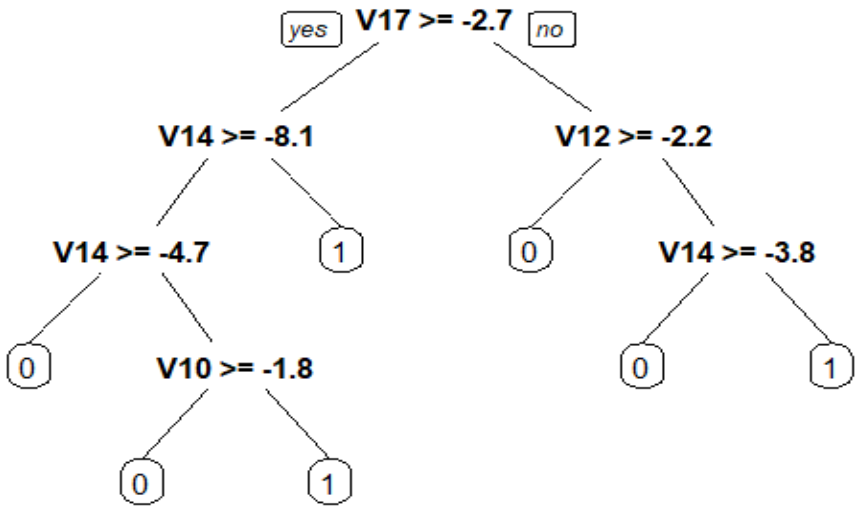


Figure 5.1: Decision Tree References

6. Conclusion

In this paper we have developed a method for fraud detection, where customers are grouped based on their extract behavioral patterns and transactions to develop a profile for every cardholder. Then different classifiers are applied on three different groups later rating scores are generated for every type of classifier. These dynamic changes in parameters lead the system to adapt to new cardholder's transaction behaviors timely. The Machine learning techniques like Logistic regression, Decision Tree and Random forest were used to detect the Credit card fraud system. The performance of the proposed system is evaluated by the parameters Sensitivity, Specificity, Accuracy and Error rate. The accuracy of Logistic regression, Decision tree and Random forest classifier are 92.7, 95.8, and 97.6 respectively. By comparing all the three method, found that Random forest classifier is better than the logistic regression and decision tree.

References

[1] Jiang, Changjun et al. "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism." *IEEE Internet of Things Journal* 5 (2018): 3637-3647.

[2] Pumsirirat, A. and Yan, L. (2018). Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine. International Journal of Advanced Computer Science and Applications, 9(1).

[3] Mohammed, Emad, and Behrouz Far. "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study." IEEE Annals of the History of Computing, IEEE, 1 July 2018, doi.ieeecomputersociety.org/10.1109/IRI.2018.00025.

[4] Randhawa, Kuldeep, et al. "Credit Card Fraud Detection Using AdaBoost and Majority Voting." IEEE Access, vol. 6, 2018, pp. 14277–14284., doi:10.1109/access.2018.2806420.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Jain R., Gour B., Dubey S., A hybrid approach for credit card fraud detection using rough set and decision tree technique, International Journal of Computer Applications 139(10) (2016).\

[10] Hafiz K.T., Aghili S., Zavarsky P., The use of predictive analytics technology to detect credit card fraud in Canada, 11th Iberian Conference on Information Systems and Technologies (CISTI) (2016), 1-6.

[11] "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, 2017

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

[13] "Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya, Mridushi" published by International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016

[14] A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective - Samaneh Sorournejad, Zojah, Atani et.al - November 2016