International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Research Paper Volume-7, Issue-9, September 2020, Regular Edition E-ISSN: 2349-7084

Creditcard Fraud Detection and Classification Using Machine Learning Based Classifiers

[1]Y.Yashasree, [2]Dr.K.Venkatesh Sharma

[1]M.Tech (Pursuing),CVR College of Engineering , Department of Computer Science and Engineering
[2]Professor,Department of Computer Science and Engineering, CVR College of Engineering
Email ID: yashasreeyathipathi@gmail.com

Available online at: http://www.ijcert.org

Received: 13/09/2020 Revised:18/09/2020 Accepted: 28/09/2020 Published: 11/10/2020

Abstract:- Nowadays, most transactions take place online, which means that credit cards and other online payment systems are involved. This method is convenient for the company and the customer. The digital age seems to have provided some very useful features that have changed the way businesses and consumers interact, but for a charge. “Credit card fraud” outlays the card industry literally billions of dollars a year. Financial institutions are constantly striving to improve fraud detection systems, but at the same time, fraudsters are finding new ways to break into systems. Preventing and detecting “Credit card fraud” has become an emergency. Data mining techniques can be very useful in detecting financial fraud, as large and complex financial data processing poses major challenges for financial institutions. In recent years, several studies have used machine learning and data mining techniques to combat this problem. The main aim of this paper is to implement the performance of the machine learning based classifiers on Credit card fraud detection dataset.

Keywords: Machine Learning, Credit card fraud, fraud detection, Classifiers.

---

1. Introduction

A large amount of data is available in the information sector. There is no benefit until this data is transformed into useful information. This extensive data needs to be analyzed and useful information gathered. Data capture is not the only method we need to implement; Data mining also includes alternative processes such as data cleansing, data integration, data transformation, data mining, model evaluation, and data presentation. Once all these processes are complete, we can be ready to use this information in many applications, such as fraudulent detection, market analysis, product control, research, and so on. The aggregated data or knowledge may be used in one of the following cases: Market analysis, Fraud detection, Customer loyalty, Production control, scientific research and Mining applications.

Detecting credit card fraud through machine learning develops a model that provides the best results in the process of finding data and detecting and preventing fraudulent transactions. This is

achieved by combining all relevant card customer transaction functions, such as date, customer territory, product category, size, supplier, and customer behavior. Information flows through a well-designed model that finds patterns and rules for classifying a transaction as fraudulent or legitimate. Now that we know what fraud protection is, let's look at the most common threats.

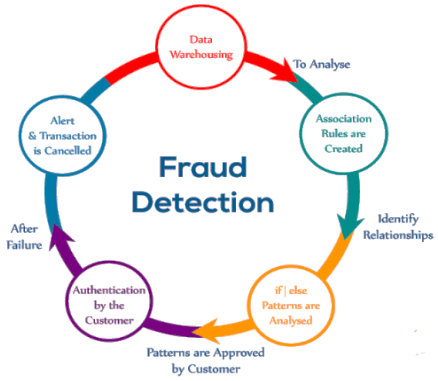

Figure 1.Credit card fraud detection Life cycle

Clone transactions: Cloning transactions have become popular in various credit card scams. This means that the actual transaction is a forgery of the transaction or the actual transaction. This happens when a company tries to pay a partner several times by sending the same invoice to different departments. The traditional rule-based fraud detection algorithm method does not work to separate fraudulent transactions from irregular or irregular transactions.

Creation of false applications.

Credit card fraud is often associated with account / identity / credit card theft. Someone is requesting a new credit account or a new credit card on behalf of another. First, the authors stole documents that could be used as evidence to substantiate their false allegations.

Irregular identification is used to find out if a transaction has unusual patterns, such as date and time or number of items. If the algorithm detects such unusual behavior, the bank's customer will be protected by certain verification methods.

Credit card amortization (electronic or manual).

Credit card counterfeiting or credit card counterfeiting is making an illegal copy of a credit or debit card with a device that reads and copies information from the original card. Credit card fraudsters use machines called "skimmers" to collect and store card numbers and other credit card information and sell them to criminals.

Remainng Paper is organized as Section 2 describes literature review,Section 3 Describes the System study, Section 4 describes result and analysis and Section 5 Concludes the paper

2. Literature Review

Vaishnavi Nath Dornadula et al. (2019) [1]Credit card fraud is a simple and user-friendly goal. E-commerce and many other online sites have increased online payment methods, increasing the risk of online fraud. As fraud rates increase, researchers have begun to use a variety of machine learning techniques to detect and analyse fraud in online transactions.

Venkata Suryanarayana, S et al. (2018) [2] Due to the widespread use of credit cards, fraud seems to be a major problem in the credit card industry. It is very difficult to obtain statistics on the impact of fraud, as companies and banks are reluctant to disclose the amount of losses caused by fraud. At the same time, public data has little to do with privacy issues, leaving many unanswered questions about the best strategy. Another problem with estimating the loss of credit card fraud is that we can only measure the loss of fraud we have seen, not the extent of unreported / unrecognized fraud.

M. A. Al-Shabi et al. (2019) [3]Fraudulent credit card transactions remain one of the problems faced by businesses and banks; It causes billions of dollars in losses every year. One of the main challenges in this area is the efficient development of algorithms. The purpose of this article is to provide an effective method for automatically detecting credit card fraud involving insurance companies using an in-depth training algorithm called Auto Encoders. The proposed solution is based on the training of an auto coder for general data reconstruction. Anomalies can be identified by defining the input of the reconstruction error and treating cases with a higher input as anomalies. The algorithm was able to detect fraudulent transactions between 64% at entry, 79% at the threshold and 91% with a threshold = 91, which is better in terms of performance compared to an unbalanced data set with a logistic regression of 57%.

 https://doi.org/10.22362/ijcert/2020/v7/i09/v7i0901

The literary survey provides a detailed overview of bank fraud and, in particular, credit card fraud. It also looks at the methods proposed to prevent these frauds. Various artificial neural network methods, ant colony optimization techniques, particle flock optimization, hidden Markov model, support vector machines, clustering and outdoor detection methods. In addition to these methods, the formation of multiple clusters and the collective behavior of animals will be considered, which will be used in this study to optimize performance by reducing errors.

3.System Study

3.1 Artificial intelligence techniques

- Fraud detection is a knowledge-based activity. The main AI methods used to detect fraud are:
- Acquisition Data Automatically finds associations and rules in data that represent interesting extraction and fraud patterns to classify, group, and separate research data.
- Fraud detection expert codes in the form of rules.
- An AC class approximation class, a group of suspicious behavior, or samples can be automatically (unnoticed) or associated with given records.
- Machine learning techniques to automatically detect the symptoms of fraud.
- Etc. can independently use neural networks to create taxonomies, groupings, generalizations, and forecasts that can then be compared with internal audit findings or official financial statements.

3.2 Types of Credit card fraud detection Techniques

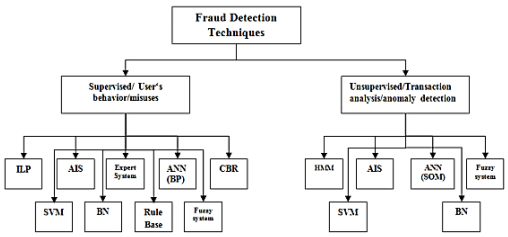

Figure 2. Hirarchical structre for fraud detection techniques

Fraud Detection

Data mining is also used in the field of credit card services and telecommunications to detect fraud. For fraudulent mobile calls, it lets you know the purpose of the call, the duration of the call, the time of day or week, and so on[4]. It also analyzes models that differ from the forthcoming rules.

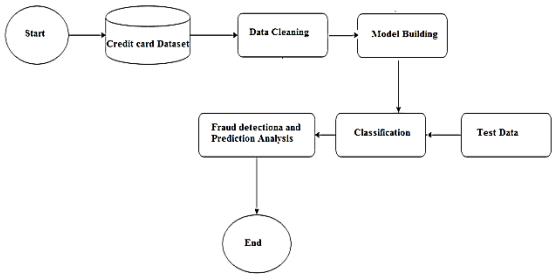

Figure 3. Proposed System Architecture

Dataset and Description:

Source of the dataset for Credit card fraud transactions are collected form the Kaggle Source, which is freely available.

Dataset: Credit card dataset which contains credit card transactions in September 2013by European cardholders occur in two days. Where, we have 492 frauds out of 284,807 transactions. It contains numeric input variables result of PCA transformation .Features V1, V2, V3, and V4……., V28 are the principal components obtained with PCA. The other two features 'time' and 'amount' have not been transformed to PCA. The feature amount' is used as independent variable of transaction amount, and feature 'time' contains seconds elapsed between each transaction and first transaction in the data set. Response variable, as the class variable, takes value 1 in case of fraud and 0 otherwise.

Information gain:

$$\text{Info (D)} = -\sum_{i=1}^{m} pi \log_2 (\text{pi}) \ldots\ldots\ldots (1)$$

In above formula 1, pi is the nonzero probability that an arbitrary tuple in D belongs to class ci, and Info (D) also known as Entropy of D.

3.3 Fraud Detection Methods and Approaches:

Predictive modeling[5] : is used to analyze the data and predict the outcome.Predictive modeling used to predict the unknown event which may occur in the future. In this process, we are going to create, test and validate the

model. There are different methods in predictive modeling. They are learning, artificial intelligence and statistics. Once we create a model, we can use many times, to determine the probability of outcomes. So predict model is reusabl. Historical data is used to train an algorithm. The predictive modeling process is an iterative process and often involves training the model, using multiple models on the same dataset.

The Process of Predictive Modeling:

- Creating model
- To create a model to run one or more algorithms on the data set.
- Testing a model:
- The testing is done on past data to see how the best model predicts.
- Validating a model:
- Using visualization tools to validate the model[6].
- Evaluating a model:
- Evaluating the best fit model from the models used and choosing the model right fitted for the data.

Decision Trees

Decision trees are used to choose between several courses of action.It provides effective structure to investigate the possible outcomes. Decision trees use tree structure to build classification or regression model. A decision tree is a flowchart like tree structure, where non leaf node denotes a test on attribute. In the results, the decision tree will have a decision node and leaf nodes. A decision node is a combination of two or more branches; each branch represents a value for the attribute which is tested. The leaf node holds a class label; the top most node in the decision tree are called as root node. Which corresponds to the best predictor in the data? Decision trees can be used to analyses the categorical data and numerical data. One of the algorithm is used to build a de is decision tree is ID3 which is developed by J. Ross Quinlan. This algorithm uses top down approach and greedy search. The top down approach is recursive divide-and – conquer method. Backtracking is not used in this algorithm[8].

Steps for making a decision tree are that firstly to Calculate the entropy of every :

attribute using the dataset in problem then dataset is divided into subsets using the attribute for which gain is maximum or entropy is minimum after that to make a decision tree node containing that attribute and lastly recursion is performed on subsets using remaining attributes to create a decision tree.

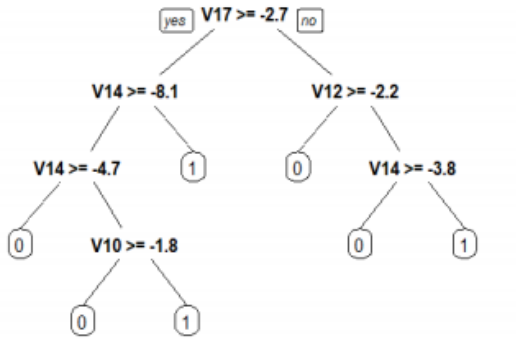

Figure 4.Decision tree

Information gain:

$$\text{Info}(D) = -\sum_{i=1}^{m} pi \log_2(\text{pi}) \ldots\ldots\ldots (1)$$

In above formula 1, pi is the nonzero probability that an arbitrary tuple in D belongs to class ci, and Info (D) also known as Entropy of D. The learning of decision trees from training tuples using ID3 and CART (Classification and Regression Trees) algorithms were invented independently of one another around the same time. The ID3 and CART algorithms are used to generate decision tree induction. These algorithms also follow top down approach in recursive manner. Decision tree is built based on training tuples are recursively partitioned into smaller subsets.

Logistic Regression

Logistic regression is similar to linear regression but interpret curve is using natural logarithm of the "odd" of the target variable which is developed by statistician David cox in 1958[9].To predict the probability of an outcome which has two values either zero or one, yes or no and false or true. The prediction is based on the use of one or several predictors; logistic regression produces logistic curves, which are two values of zero and one. Linear regression

model is used to predict binary target variables. Binary targets variables either 0 or one. The linear regression equation

$$Y=\beta_0+\beta_1+\sum_i \quad \ldots\ldots\ldots\ldots\ldots (2)$$

In equation (2) the actual value of Y is binary variable, then the predicted Y can be less than zero or greater than one. Logistic Regression or logit model is a regression model where the dependent variable is categorical and analyzes the relationship between multiple independent variables. Binary Logistic Regression model is used to estimate the probability of a binary response based on one or more predictors The Logistic Regression can be a binomial, ordinal or multinomial, ordinal Logistic Regression deals with dependent variables that are ordered. In multinomial Logistic Regression where the outcomes can have three or more possible types are not in order. The Logistic Regression[10] is used to determine probability of an event occur over the probability of an event not occurred, and then predicted variable may be continuous or categorical.

KNN classifiers

Let's take a simple case to understand this algorithm. Following is a spread of red circles (RC) and green squares (GS)[11] :

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN[12] is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

First the credit card dataset is taken from the source and cleaning and validation is performed on the dataset which includes removal of redundancy, filling empty spaces in columns, converting necessary variable into factors or classes then data is divided into 2 part, one is training dataset and another one is test data set. Now K fold cross validation is done that is the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k −1 subsamples are used as training data, Models are created for Logistic regression, Decision tree, SVM, Random Forest and then accuracy, sensitivity, specificity, precision are calculated and a comparison is made. The dataset is sourced from ULB Machine Learning Group. The dataset contains credit card transactions made by European cardholders around September 2013 and occurrence of transactions that happened in two days are presented by this dataset, consisting of 284,786 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. It contains only numerical (continuous) input variables which are as a resul t of a Principal Component Analysis (PCA) feature selection transformation resulting to 28 principal components. And total of 30 input features

 https://doi.org/10.22362/ijcert/2020/v7/i09/v7i0901

are utilized in this study. 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. Afterwards these variables are used to create a model which distinguish fraudulent activities. The details and background information of the features cannot be presented due to confidentiality issues. 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). Four basic metrics are used in evaluating the experiments, namely True positive (TPR), True Negative (TNR), False Positive (FPR) and False Negative (FNR) rates metric [13] .

$$TPR = \frac{TP}{P}$$
$$TNR = \frac{TN}{N}$$
$$FPR = \frac{FP}{N}$$
$$FNR = \frac{FN}{P}$$

where FN , FP ,TP,TN, and are the number of false negative false positive ,true positive and true negative test cases classified while total number of positive and negative class cases under test are represented by P and N. Cases classified rightly as negate are termed with true negative and cases classified as positive which are actually positive are termed with True positive .Cases classified as positive but are negative cases are termed as false positive and cases classified as negative but are truly positive are termed as false negative. The performance of Classifiers is evaluated based on accuracy, precision, specificity and sensitivity.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$
$$Sensitivity = \frac{TP}{TP + FN}$$
$$Specificity = \frac{TN}{FP + TN}$$
$$Precision = \frac{TP}{TP + FP}$$

Sensitivity (Recall) gives the accuracy on positive (fraud) cases classification. Specificity gives the accuracy on negative (legitimate) cases classification. Precision gives the accuracy in cases classified as fraud (positive)[14].

4. Result analysis

For implementing the proposed system, we use Python programming with PyCharm IDE[17]. Python is an open-source (free) programming language that is used in web programming, data science, artificial intelligence, and many scientific applications. Learning Python allows the programmer to focus on solving problems, rather than focusing on syntax. Crediticard dataset which contains credit card transactions in September 2013[18] by European cardholders occur in two days. Where, we have 492 frauds out of 284,807 transactions. Load dataset with total of 284807 transactions, Class labeling : 0: Nonfraud and 1 :Fraud among total transactions of 284807.

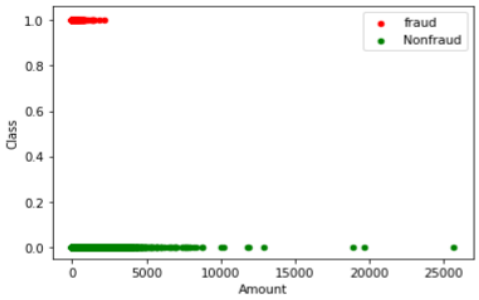

Figure 5. Class labeling on dataset

Data classification based on fraud and non fraud transactions which represent the color as Fraud: red, nongraud: green

After classification it results :

Table 1. Classification on fraud and nonfraud transactions among dataset.

Class name	Result
0 (Nonfraud)	284315
1 (fraud)	492
Total	284807

Comparing the performance on three Classifiers :

1. Logistic Regression
2. Decission Tree
3. KNN Classifier

 https://doi.org/10.22362/ijcert/2020/v7/i09/v7i0901

Logistic Regression:
Prediction Accuracy: 0.9377685548
Congussion matrix:
Table 2. Format of the Confussion matrix

Actual Predicted	Not a fraud	Fraud
Not a fraud	82017	3155
Fraud	7461	77956

Decision Tree :
Prediction Accuracy: 0.99974207
Congussion matrix:
Table 3. Format of the Confussion matrix

Actual Predicted	Not a fraud	Fraud
Not a fraud	85128	44
Fraud	0	85417

KNN Classifier :
Prediction Accuracy: 0.99887448
Congussion matrix:
Table 4. Format of the Confussion matrix

Actual Predicted	Not a fraud	Fraud
Not a fraud	84980	192
Fraud	0	85417

Finally got confusion metrics and high accuracy Overall prediction accuracy among the three machine Learning based classifiers are
Table 5. Performance of Preidction acccuray

Classifier	Prediction Accuracy
Regression Classifier	0.9377
Decision tree Classifier	0.9997
KNN Classifier	0.9988

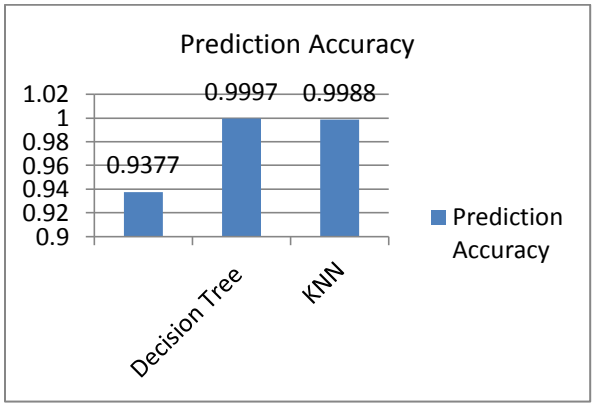

Figure 6. Overall classification performance

From the above results over all prediction accuracy is improved by the Decision tree Classifer rather than the CNN and Regresion Classifiers.

5.Conclussion

In this paper we have implemented three classification algorithms based on machine learning Tehnology and from the experiments the result that has been concluded is that Logistic regression has a accuracy of 93.7% while KNN shows accuracy of 99.88% and Decision tree shows accuracy of 99.97%. The results obtained thus conclude that Decision tree shows the most precise and high accuracy of 99.97% in problem of credit card fraud detection with dataset provided by kaggle source in machine learning. As a future a scope we suggest for the future research , do the experiment on large voume of credict card dataset and compare the performance of various classifiers with prediction accuracy, and efficencuy.

References

[1] Dornadula, V. N., & Geetha, S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. Procedia Computer Science, 165, 631–641. doi:10.1016/j.procs.2020.01.057

[2] Venkata Suryanarayana, S., N. Balaji, G., & Venkateswara Rao, G. (2018). Machine Learning Approaches for Credit Card Fraud Detection. International Journal of Engineering & Technology, 7(2), 917. doi:10.14419/ijet.v7i2.9356

[3] Al-Shabi, M. (2019). Credit Card Fraud Detection Using Auto encoder Model in Unbalanced Datasets. Journal of Advances in Mathematics and Computer Science, 33(5), 1-16. https://doi.org/10.9734/jamcs/2019/v33i530192

[4] SADGALI, I., SAEL, N., & BENABBOU, F. (2019). Performance of machine learning techniques in the detection of financial frauds. Procedia Computer Science, 148, 45–54. doi:10.1016/j.procs.2019.01.007

[5] Leo, M., Sharma, S., & Maddulety, K. (2019). Machine Learning in Banking Risk Management: A Literature Review. Risks, 7(1), 29. doi:10.3390/risks7010029

[6] Sohony, I., Pratap, R., & Nambiar, U. (2018). Ensemble learning for credit card fraud detection. Proceedings of the ACM India Joint

International Conference on Data Science and Management of Data - CoDS-COMAD '18. doi:10.1145/3152494.3156815

[7] Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2019). Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection. Information Sciences. doi:10.1016/j.ins.2019.05.042

[8] John O. Awoyemi Department of Computer Science Federal University of Technology Akure, Nigeria johntobaonline@yahoo.com

[9] Choi, D., & Lee, K. (2018). An Artificial Intelligence Approach to Financial Fraud Detection under IoT Environment: A Survey and Implementation. Security and Communication Networks, 2018, 1–15. doi:10.1155/2018/5483472

[10] Popat, R. R., & Chaudhary, J. (2018). A Survey on Credit Card Fraud Detection Using Machine Learning. 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). doi:10.1109/icoei.2018.8553963.

[11] *Roman Chuprina on April 14, 2020 at 1:30am; Blog, View.* "The In-depth 2020 Guide to E-commerce Fraud Detection". *www.datasciencecentral.com*. Retrieved 2020-05-24.

[12] *Velasco, Rafael B.; Carpanese, Igor; Interian, Ruben; Paulo Neto, Octávio C. G.; Ribeiro, Celso C. (2020-05-28).* "A decision support system for fraud detection in public procurement". *International Transactions in Operational Research: itor.12811*. doi:10.1111/itor.12811. ISSN 0969-6016.

[13] Jump up to:a b c d Bolton, R. and Hand, D. (2002). Statistical fraud detection: A review. Statistical Science 17 (3), pp. 235-255

[14] Jump up to:a b G. K. Palshikar, The Hidden Truth – Frauds and Their Control: A Critical Application for Business Intelligence, Intelligent Enterprise, vol. 5, no. 9, 28 May 2002, pp. 46–51.

[15] Vani, G. K. (February 2018). "How to detect data collection fraud using System properties approach". Multilogic in Science. VII (SPECIAL ISSUE ICAAASTSD-2018). ISSN 2277-7601. Retrieved February 2, 2019.

[16] Michalski, R. S., I. Bratko, and M. Kubat (1998). Machine Learning and Data Mining – Methods and Applications. John Wiley & Sons Ltd.

[17] Salazar, Addisson, et al. "Automatic credit card fraud detection based on non-linear signal processing." Security Technology (ICCST), 2012 IEEE International Carnahan Conference on. IEEE, 2012.

[18] Delamaire, Linda, H. A. H. Abdou, and John Pointon. "Credit card fraud and detection techniques: a review." Banks and Bank system's (2009).

[19] Quinlan, J. Ross. "Induction of decision trees." Machine learning (1986).

[20 Quinlan, J. R. (1987). "Simplifying decision trees". International Journal of Man-Machine Studies.

