IOP Conf. Series: Journal of Physics: Conf. Series **1299** (2019) 012037 doi:10.1088/1742-6596/1299/1/012037

Fraud prediction in bank loan administration using decision tree

I O Eweoya[1], A A Adebiyi[1,2], A A Azeta[1] and Angela E Azeta[3]

[1]Department of Computer and Information Sciences, Covenant University, Nigeria
[2]Department of Computer Science, Landmark University, Nigeria
[3]Department of PTTIM, FIIRO, Nigeria

ibukun.eweoya, ayo.adebiyi, ambrose.azeta{@covenantuniversity.edu.ng}, azetaangela@gmail.com

Abstract. The rate at which banks looses funds to loan beneficiaries due to loan default is alarming. This trend has led to the closure of many banks, potential beneficiaries deprived of access to loan; and many workers losing their jobs in the banks and other sectors. This work uses past loan records based on the employment of machine learning to predict fraud in bank loan administration and subsequently avoid loan default that manual scrutiny by a credit officer would not have discovered. However, such hidden patterns are revealed by machine learning. Statistical and conventional approaches in this direction are restricted in their accuracy capabilities. With a large volume and variety of data, credit history judgement by man is inefficient; case-based, analogy-based reasoning and statistical approaches have been employed but the 21st century fraudulent attempts cannot be discovered by these approaches, hence; the machine learning approach using the decision tree method to predict fraud and it delivers an accuracy of 75.9 percent.

Keywords: Confusion matrix, decision tree, fraud, machine learning, prediction.

1. Introduction

There are unsolved fraudulent practices in financial operations in the society, including bank credit administration, calling for a remedy through intelligent technology [1-4]. Existing fraud detection techniques in bank credit administration have not sufficiently met the desired accuracy, and avoidance of false alarm, and none focused on fraud in bank credit default. Also, fraudulent duplicates, missing data, and undefined fraud scenarios affect prediction accuracy [1-11].

Any unlawful act by human beings or invoked by machines that leads to personal gain at the expense of institutions or the legal human beneficiaries is a financial fraud, but an error must not be taken for a fraud [1],[12-14]. Considering the overall effect of financial frauds, it is referred to as an economic sabotage. The examples of financial fraud are money laundering, bank credit fraud, pension fraud, co-operative society fraud, tax evasion, telecommunications fraud, credit card fraud, inflated

contracts, financial reports fraud, health insurance fraud [15], automobile insurance fraud, and mortgage insurance fraud.

According to [16], there are many types of fraud including, credit card fraud, telecommunication fraud, computer intrusion, bankruptcy fraud, theft fraud or counterfeit fraud, and application fraud. The economy of nations do feel the impact of fraud and many approaches have been employed but with shortcomings. However, machine learning has proved to be more reliable. Machine learning uses data mining techniques to reveal hidden patterns in a large, volatile, and variety of data and make intelligent decisions through the revealed insights.

It is worthy of note that according to [17-19]; a high rate of default has been reported in different nations and this can be reduced using information technology. The rest of this paper is organized as follows: Section 2 is the materials and methods, followed by the results and discussion in Section 3, and section 4 is the conclusion of the paper.

The decision tree classifies data into discrete ones using tree structure algorithms [20-21]. It highlights the structural information contained in the data and classifies from root to the leaf node [22].The advantages of using decision trees include the fact that simplicity and speed of decision trees are second to none; there is no requirement for a domain knowledge or parameter setting; also, it comfortably handles high dimensional data where there are many attributes involved; the way it is represented allows for enhanced comprehensibility; it has a fantastic accuracy though this is dependent on the data in use; it supports incremental learning; they are unvaried, since they are used based on a single feature at each interval node. They work fine on both classification and regression problems; they can handle missing values; trees are plotted graphically, and can be easily interpreted; most interestingly, trees can be easily explained to people [23-25].

Credit default refers to the failure of a client to meet the legal obligations or conditions of a loan according to the promissory note. In other words, loan or credit default is the failure to repay a loan according to the terms agreed to initially before the approval of that loan. Non-performing loan refers to a specific amount of credit taken by a borrower but the debtor has declined in making agreed installment paybacks in 90 days for commercial banking loans and 180 days for consumer loans. Non-payment indicates neither the interest nor the principal gets paid with respect to that credit in 90 to 180 days depending on the type of loan, purpose or industry. Any definition of a non-performing loan is a function of the terms of that loan and the subsisting agreement as definition is not cast in stone but conditional based on promissory notes and agreements.

2. Materials and Methods

A credit dataset of 5000 instances and 9 attributes were employed for this research based on features extraction with the target attribute being the default status, and an effective data pre-processing before subsequent operations. The attributes include age, sex, income, employment status, the track of the last three payments (if any), and balance of loan taken. Python programming language was used for fraud prediction in credit or loan default using spyder 9.0. The classification was executed in Matlab 2017b [26] where cross validation and features extraction were employed. The training and testing was done in Matlab which gave a result of 75.9% accuracy. The scatter plot of the data is in Figure 1.Through the confusion matrix, the true positive rate, and false positive rate are as shown in Figure 2. Decision Tree Positive predictive values and false discovery rates are presented in Figure 3. Also, the ROC, is in Figure 4. Furthermore, 80.04 % was classified rightly, and 19.96% wrongly classified based on weka 3.8 [27] using stratified cross-validation

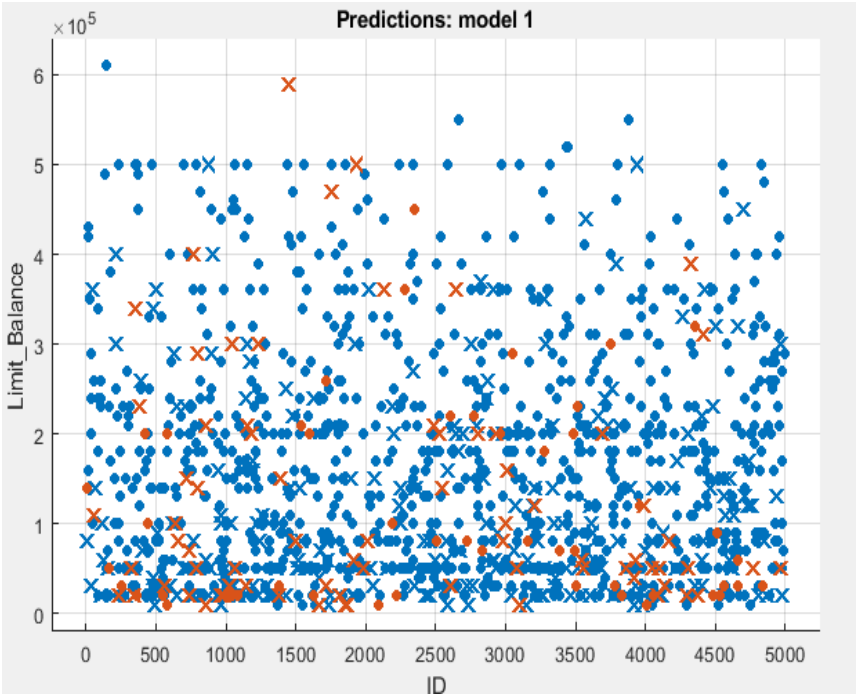

Figure 1: A scatter plot of the ID versus Limit balance

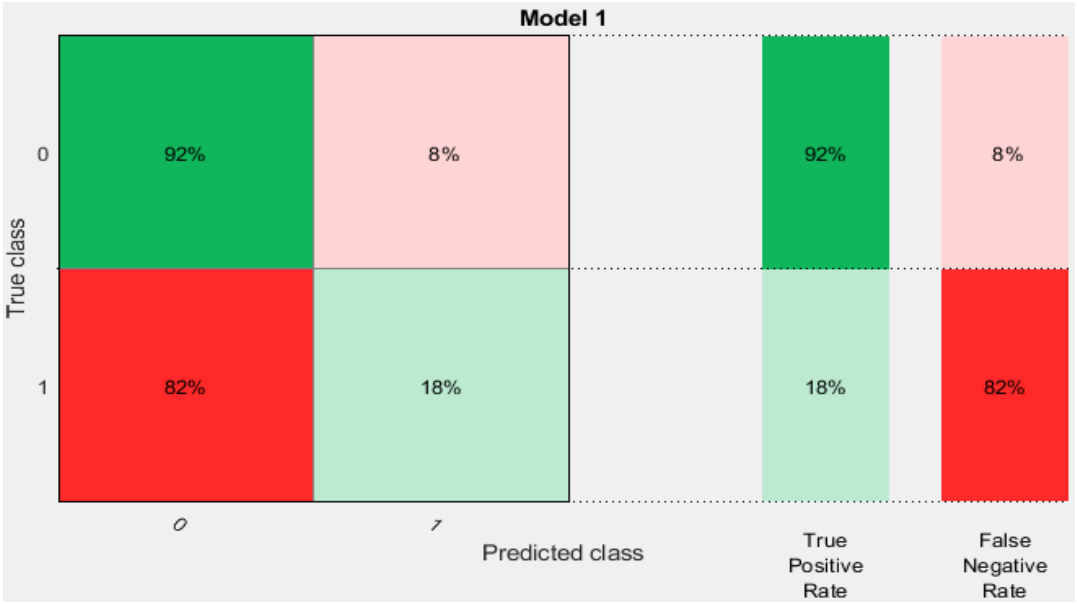

Figure 2: Decision Tree True positive rate and false negative rate

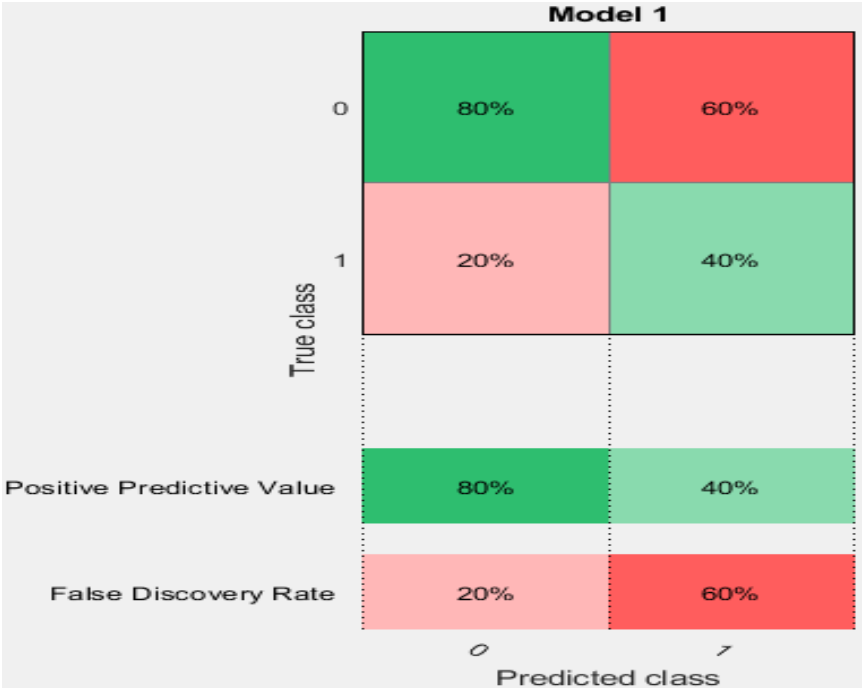

Figure 3: Decision Tree Positive predictive values and false discovery rates

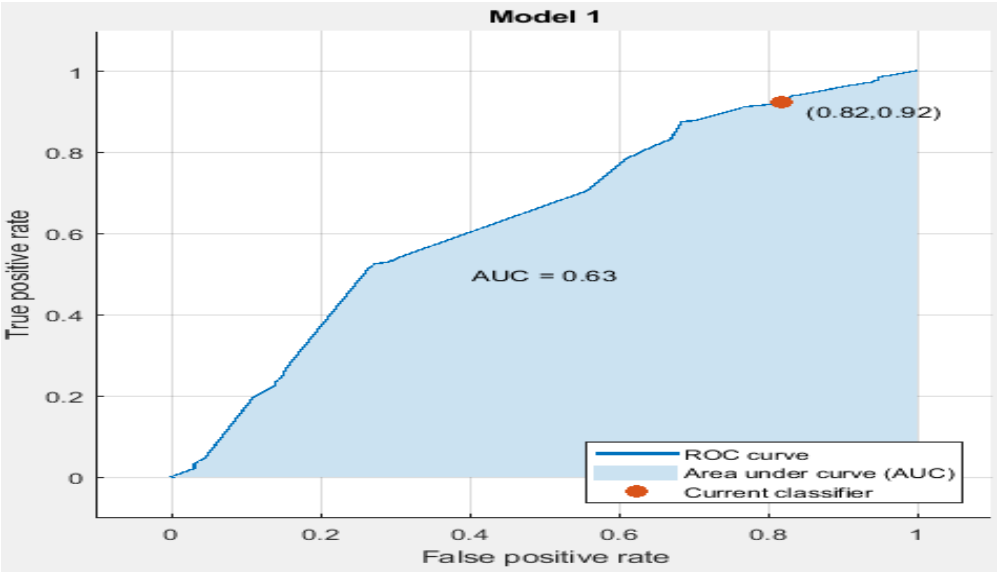

Figure 4: Decision Tree ROC curve

3. Results and Discussion

Credit or loan defaults have led to bank insolvency and nations entering recession, this has an untoward effect on people. For the purpose of extracting relevant features (features engineering), Principal Component Analysis was employed, cross validation was used to avoid overfitting in the model built; data splitting was done to separate testing data from training data. This allows the model to work on a fraction of the data not known before for the testing of the model. The training and testing yielded 75.9 % accuracy with a high true positive ratio. Also, 80.04 % of the instances were

correctly classified and 129 of the testing data identified to be fraudulent based on a python [28] program written for this work. As was presented in [29,30], this current study did not include hypothesis in its model formulation and testing, rather machine learning technique such as decision tree was engaged in the model formulation and prediction.

4. Conclusion

The anomaly of taking credit and ending up in a default to the detriment of the lender has been confirmed to have a remedy in machine learning. Using a real life dataset it has been revealed that false positives can be reduced with an employment of decision tree, thereby getting a highly reliable accuracy that financial institutions can depend on while scrutinizing loan applications.

5. Acknowledgement

In this study, we would like to express our deep appreciation for the support and sponsorship provided by Covenant University Centre for Research, Innovation and Discovery (CUCRID).

6. References

[1] Bagul P D Bojewar S and Sanghavi A 2016 Survey on hybrid approach for fraud detection in health insurance. *International Journal of Innovative Research in Computer and Communication Engineering,* **4**(4): 6918-*6922.*

[2] Hameed A A Karlik B and Salman M S 2016 Backpropagation algorithm with variable adaptive momentum *Knowledge-based Systems*,**114**: 79-87. DOI: https://doi.org/10.1016/j.knosys.2016.10.001

[3] Demla N and Aggarwal A 2016 Credit card fraud detection using svm and reduction of false alarms. *International Journal of Innovations in Engineering and Technology*, **7(**2): 176-182.

[4] Fahmi M Hamdy A and Nagati K 2016 Data mining techniques for credit card fraud detection: empirical study. *Sustainable Vital Technologies in Engineering and Informatics*, pp.1-9, Elsevier.

[5] Vaishali V 2014 Fraud detection in credit card by clustering approach. *International Journal of Computer Applications,* **98(**3)*: 29-32.*

[6] Abid L Masmoudi A and Zouari-Ghorbel S 2016 The consumer loan's payment default predictive model: an application in a Tunisian commercial bank. *Asian Economic and Financial Review,* **6**(1): 27-42.

[7] Sharma S and Choudhury A R 2016 Fraud analytics: A survey on bank fraud and fraud prediction using unsupervised learning based approach. *International Journal of Innovations in Engineering Research and Technology*, **3(**3): 1-9.

[8] Agaskar V Babariya M Chandran S and Giri N 2017 Unsupervised learning for credit card fraud detection. *International Research Journal of Engineering and Technology* (IRJET), **4**(3): 2343-2346.

[9] Rawate K R and Tijare P A 2017 Review on prediction system for bank loan credibility. *International Journal of Advance Engineering and Research Development,* **4(**12): 860-867.

[10] Rimiru R Wa S W and Otienoc C 2017 A hybrid machine learning approach for credit scoring using PCA and logistic regression. *International Journal of Computer,* **27(**1): 84-102.

[11] Boateng E Y and Oduro F T 2018 Predicting microfinance credit default: A study of Nsoatreman rural bank, Ghana. *Journal of Advances in Mathematics and Computer Science (JAMCS),* **26(**1): 1-9.

[12] Rawte V and Anuradha G 2015 Fraud detection in health insurance using data mining techniques. *International Conference on Communication, Information & Computing Technology (ICCICT),* pp. 1-5, Mumbai

[13] Naik J and Laximinarayana J A 2017 Designing hybrid model for fraud detection in insurance. *IOSR Journal of Computer Engineering*, **1**: 24-30.

[14] Akomolafe J A Eluyela D F Ilogho S O Egharevba J W and Aina O 2017 Financial crime in Nigeria public sector: A study of Lagos state ministries. *International Journal of Innovative Research in Social Sciences & Strategic Management Techniques*, **4** (1):13-21.
[15] Kose I Gokturk M and Kilic K 2015 An interactive machine learning-based electronic fraud and abuse detection system in healthcare insurance, *Applied Soft Computing*, **36**:283–299.
[16] Tripathi K K and Pavaskar M A 2012 Survey on credit card fraud detection methods. *International Journal of Emerging Technology and Advanced Engineering,* **2(**11): 721-726.
[17] Central Bank of Nigeria CBN 2016 Financial stability report - December 2016. Available from: https://www.cbn.gov.ng/out/2017/fprd/fsr%20december%202016%20(2).pdf. Retrieved March, 2018.
[18] Central Bank of Nigeria CBN 2017 Financial stability report - June 2017. Available from: https://www.cbn.gov.ng/Out/2018/FPRD/FSR%20June%202017.pdf. Retrieved March, 2018.
[19] World Bank 2018 Economic indicators for over 200 countries, https://www.theglobaleconomy.com/Nigeria/Nonperforming_loans/ Retrieved March, 2018.
[20] Quinlen J R 1986 Introduction of decision trees. *Machine Learning* **1**: pp. 81-106.
[21] Han J and Kamber M 2011 Data mining concepts and techniques. *Elsevier*, p. 744, Morgan Kaufmann.
[22] Kotsiantis S B 2007 Supervised machine learning: A review of classification techniques. *Informatica,* **31**: 249-268.
[23] Williams G J and Huang Z 1997 Mining the knowledge mine: The hot spots methodology for mining large real world databases. *Australian Joint Conference on Artificial Intelligence*, pp. 340-348.
[24] Liou F M Tang Y C and Chen J Y 2008 Detecting hospital fraud and claim abuse through diabetic outpatient services. *Health Care Management Science, **11***(4): 353-358.Available from: http://dx.doi.org/10.1007/s10729-008-9054-y
[25] Shin H Park H Lee J and Jhee W C 2012 A scoring model to detect abusive billing patterns in health insurance claims. *Expert Systems with Applications,* **39(**8), 7441-7450 Available from: http://dx.doi.org/10.1016/j.eswa.2012.01.105
[26] Matlab and Statistics Toolbox Release 2017b The MathWorks Inc Natick Massachusetts United States.
[27] Hall M Frank E Holmes G Pfahringer B Reutemann P and Witten H I 2009 The WEKA data mining software: An update. SIGKDD explorations, Volume 11, Issue 1.
[28]Python Software Foundation. Python language reference, version 2.7. Available at http://www.python.org
[29] Nicholas-Omoregbe O S Azeta A A Chiazor I A and Omoregbe N 2017 Predicting the adoption of e-learning management system: A case of selected private universities in Nigeria. *Turkish Online Journal of Distance Education-TOJDE **18(2)*** 106-121.
[30] Azeta A A Misra S Azeta V I Osamor V C 2019 Determining suitability of speech-enabled examination result management system. *Wireless Networks* 1-8.