IOP Conf. Series: Journal of Physics: Conf. Series **1299** (2019) 012039 doi:10.1088/1742-6596/1299/1/012039

Fraud prediction in loan default using support vector machine

I O Eweoya[1], A A Adebiyi[1,2], A A Azeta[1] and Olufunmilola Amosu[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}, funmiotesh@gmail.com

Abstract. The concept of taking loan has been in existence since inception of the human race but it is now taking diverse dimensions. This spans through personal exchange of loans for repayment based on personal track records, enjoying loans as proceeds of daily contribution without collaterals, except for the banking sector that requests collaterals for official loans. The uniform occurrence of being unable to pay the debts and resulting in a default is evident to the level of bank closures and nations' bankruptcy is experienced across board. With a large volume and variety of data, credit history judgment 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 support vector machine. This work employs a supervised learning approach based on machine learning to predict the possibility of a fraud in a loan application through hidden trends in data instead of giving loans which ordinarily should not be approved; past occurrences discovered through machine learning reveals risky loans and a possible fraud by humans in approvals that can result in a default. Machine learning approaches are able to detect fraudulent financial statements to avert business comatose.

Keywords: Confusion matrix, fraud, machine learning, loan default, support vector machine.

1. Introduction

Loan administration is intended to improve economic buoyancy of loan beneficiaries while the lender gets interest at a specific rate based on diverse circumstances like duration of refund, amount, purpose, and credit track records. A loan enters default when the agreed promissory notes are not fulfilled by the loan beneficiary; precisely when no payment or interest is paid in 90 days. However, intentionally officers that should scrutinize loan applications and decline non- promising applications that are likely

to enter default do approve such loans. Fraudulent practices like overrated collateral, double collateral, fraudulent bank statements are overlooked or judged wrongly.

A fraud has taken place when there is an evidence of intent to mislead, cheat or steal for personal gains [1-4]. To intentionally produce deceptive data is also a fraud. Traditional methods of fraud detection in credit administration are available but limited in capacity to check current sophistication in fraud perpetration and time-consuming [5-7].

The revelation of hidden patterns and features in data through data mining has allowed for machine learning solutions to fraud detection [7]. Machine Learning employs data mining approaches and other learning algorithms in building models of what is happening behind some data to end up in making predictions or detections. There are supervised and unsupervised learning techniques used in detecting fraudulent acts in various domains. Labelled data can be expensive and difficult to get, but this is what a supervised learning approach uses. Grouping bank credit transactions into "legitimate" and "fraudulent" is a sample of labeling. Training and learning of input variables are channeled towards these target variables.

It is high time technology was employed to eliminate the fraudulent practices associated with loans ending up in default. The support vector machine has been employed in many detection and prediction works, for example, telecommunications, pattern recognition, system intrusion detection, age estimation, and facial recognition [8-11]; fraud prediction in bank loan; but not in the context of predicting fraud that might have led to credit or loan default.

This work employs the support vector machine using the kernel mode to predict fraud in loan default. The support vector machine (SVM) is a supervised machine learning approach that employs labeled data for its training and testing to make predictions.

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 initial terms agreed to initially before the approval of that loan.

In a situation where a specific amount of credit is 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; this is called a non-performing loan. 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.

The SVM was developed by [12], based on the structural risk management theory [13]. It uses decision planes to define decision boundaries, separating between a set of objects with diverse class memberships. The SVM creates a hyperplane by using a linear model to implement non-linear class boundaries through some non-linear mapping input vectors into high dimensional feature space [14].

Support Vector Machine (SVM) is a classification and regression prediction tool that employs machine learning theory in the maximization of its predictive accuracy as it automatically avoids overfitting to the data. SVM came to limelight when it outperformed the well-known sophisticated neural networks based on elaborate features in a handwriting recognition problem, using pixel maps as input. It has been employed in many detection and prediction works, for example, telecommunications, pattern recognition, system intrusion detection, age estimation, and facial recognition [8-11].

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. Python programming language was used for fraud prediction in credit or loan default using spyder 9.0. The classification was executed in matlab 2017b [15] using cross validation and features extraction approach was based on the principal component analysis (PCA). Also, an efficient data pre-processing was executed before the subsequent tasks. The training and testing was done using kernel SVM which gave a result of 81.3% accuracy. A

scatter plot of the instances versus limit balance is shown in Figure 1; through the confusion matrix, the true positive rate (96%), and false positive rate (4%) are as shown in Figure 2. Also, the positive predictive values and false discovery rate are in Figure 3 respectively.

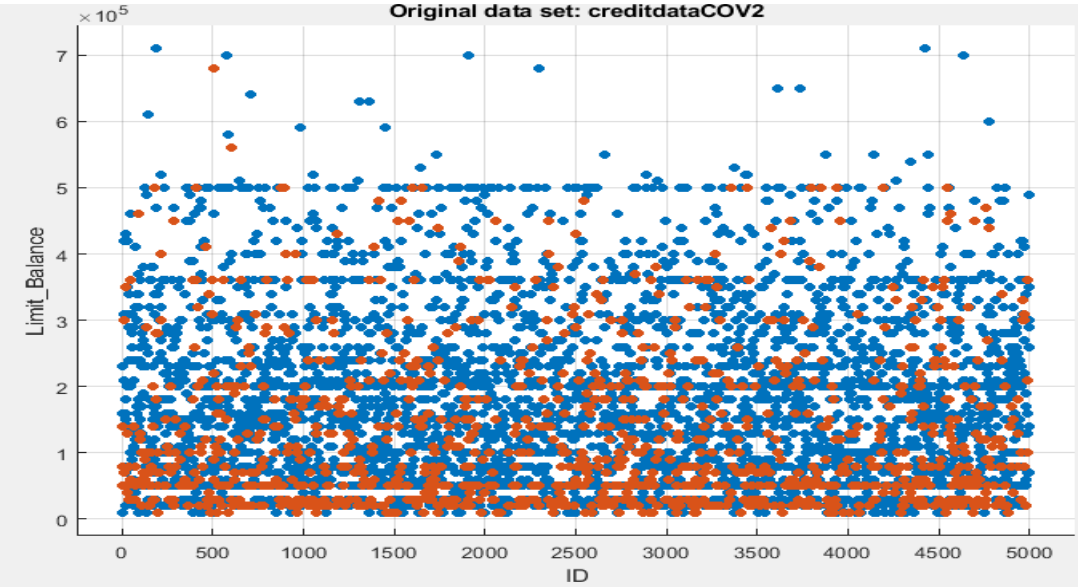

Figure 1: A scatter plot of the ID versus Limit balance

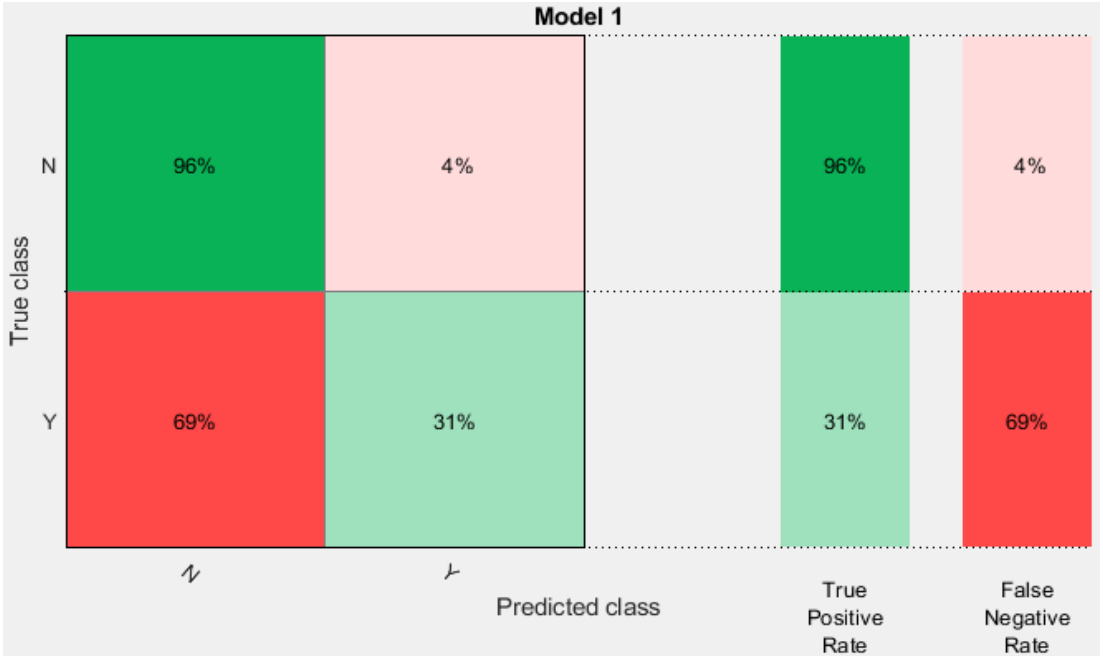

Figure 2: Linear SVM true positive and false negative rates

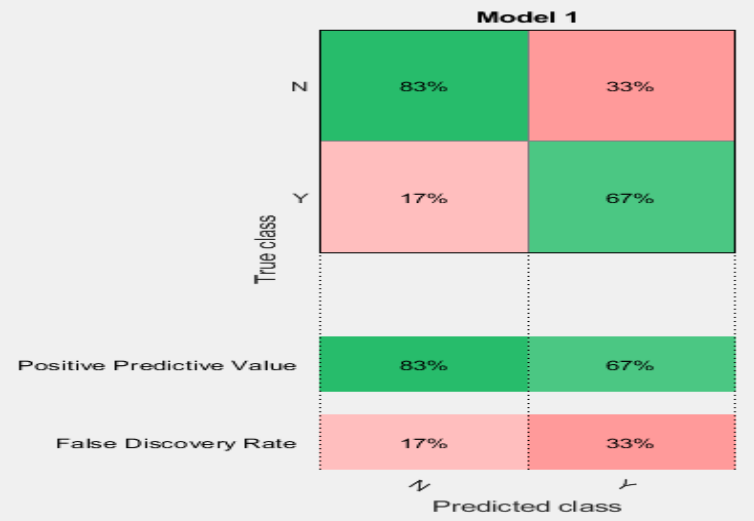

Figure 3: Linear SVM positive predictive values and false discovery rates

3. Results and Discussion

Credit or loan defaults have led to the closure of many banks, also, many nations have experienced recession due to the fraudulent practices in loan administration, and making life tough for people. An approach using SVM in kernel mode yielded 81.3% accuracy with good true positive and false negative rates to confirm its effectiveness if employed loan scrutiny. Based on Python program, a total of 129 fraudulent transactions were identified from the testing data. Some other existing approaches include Naïve Bayes, decision trees, and neural networks among others. However, in comparison the SVM has a higher accuracy than the methods listed above using the dataset employed by this work. As was presented in [16,17], 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 work has confirmed the efficiency of the support vector machine in a new context. A high accuracy is attained with a drastically reduced false positive and high true positives. The usage of technology to salvage loss in loan administration is displayed by this approach. In the future, testing the strength of an ensemble approach to find out the given result is to be pursued to see what it promises to deliver.

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] Oloidi G A and Ajinaja O T 2014. Bank frauds and forgeries in Nigeria: A study of the causes, types, detection and prevention. *IOSR Journal of Economics and Finance*, **4(**2): 41-50.
[2] Rawte V and Anuradha G 2015 Fraud detection in health insurance using data mining techniques. *International Conference on Communication, Information & Computing Technology (ICCICT)*, 1-5, Mumbai.
[3] Naik J and Laximinarayana J A 2017 Designing hybrid model for fraud detection in insurance. *IOSR Journal of Computer Engineering*, **1**: 24-30.

[4] 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.
[5] Abdelhamid D Khaoula S and Atika O 2014 Automatic bank fraud detection using support vector machines. *International Conference on Computing Technology and Information Management*, pp. 10-17, Dubai, UAE.
[6] Rohit K D and Patel D B 2015 Review on detection of suspicious transaction in anti-money laundering using data mining framework. *International Journal for Innovative Research in Science & Technology*,**1(**8): 129-133.
[7] 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*.
[8] Guo G Fu Y Dyer C R and Huang T S 2008 Image-based human age estimation by manifold learning and locally adjusted robust regression. *Transactions on Image Processing, IEEE*, **17(**7): 1178-1188.
[9] Kumar M Ghani R and Mei Z S 2010 Data mining to predict and prevent errors in health insurance claims processing. ACM *16th International Conference on Knowledge Discovery and Data Mining*, pp. 65-74. Available from: http://dx.doi.org/10.1145/1835804.1835816
[10] Kirlidog M and Asuk C 2012 A fraud detection approach with data mining in health insurance. *Procedia-Social and Behavioral Sciences,* **62***:* 989-994. http://dx.doi.org/10.1016/j.sbspro.2012.09.168
[11] Anwar S Zain J M Zolkipli M F Inayat Z Khan S Anthony B and Chang V 2017 From intrusion detection to an intrusion response system: Fundamentals, requirements, and future directions. *MDPI Algorithms*, **10**(2):1-24, DOI: 10.3390/a10020039
[12]Cortes C and Vapnik V 1995 Support vector network. *Machine Learning*, **20(**3): 273- 297. DOI:https://doi.org/10.1023/A:1022627411411
[13]Chiu N H and Guao Y Y 2008 State classification of CBN grinding with support vector machine. *Journal of Material Processing Technology*, **201**:601-605.
[14]Elmi H E Sallehuddin R Ibrahim S and Zain A M 2014 Classification of sim box fraud detection using support vector machine and artificial neural network. *International Journal of Innovative Computing*, **4** (2): 19-27.
[15] Matlab and Statistics Toolbox Release 2017b, The MathWorks, Inc., Natick, Massachusetts, United States.
[16] 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.
[17] 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.