A Naive Bayes approach to fraud prediction in loan default

I O Eweoya[1], A A Adebiyi[1,2], A A Azeta[1], F Chidozie[1], F O Agono[1] and B Guembe[1]

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

ibukun.eweoya, ayo.adebiyi, ambrose.azeta,
felix.chidozie{@covenantuniversity.edu.ng},
frankagono@gmail.com,blessingodede@gmail.com

Abstract. The essence of granting loans to individuals and corporate beneficiaries is to boost the economy while the lenders make profit from the interest that accrues to the lending. However, due to non-compliance to basic rules, fraud is prevalent in credit administration and traditional methods of detecting fraud have failed. Furthermore, they are time-consuming and less accurate. This work uses a supervised machine learning approach, specifically the Naïve Bayes to predict fraudulent practices in loan administration based on training and testing of labeled dataset. Previous works either predict credit worthiness or detect loan fraud but not predicting fraud in credit default. The approach employed in this work yielded 78 % accuracy.

Keywords: Confusion matrix, fraud, machine learning, loan default, Naïve Bayes

1. Introduction

Bank credit administration all over the world has witnessed a high rate of fraud; this is also evident in many other sub-sectors of the economy. It is worthy of note that the traditional ways of detecting frauds in bank credit operations today are unfit because they are inefficient and time-consuming. This is due to the sophistication involved in the 21st century methods of fraud practices. Credit fraud is one of the numerous risks that financial institutions face; it ultimately leads to credit default. This is the highest risk area for financial institutions; it is also a big hole to the treasuries of diverse countries. A host of approaches have been engaged, including statistical methods, knowledge discovery, and case-based reasoning to detect credit fraud in different countries. However, with the ever-increasing large volume of data involved, the application of data mining approaches alongside some sophisticated machine learning algorithms have opened up new ideas towards addressing the problem of financial fraud.

Existing fraud prediction 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. This work applied Naïve Bayes approach to predict fraud in credit default in an attempt to ensure that credits that could enter default were discovered and possible frauds in the transaction predicted based on the data of past transactions.

Globally, fraud perpetration is rising as evident in bank credit administration [1-6].This leads to credit default which is detrimental to economic growth [5-10]. The absolute relevance of a credit officer is forfeited once many loans do enter default. A default means failure to meet the legal obligations of a loan as initially agreed in the promissory notes such that both the interest and the principal are not paid for a continuous 90 days [11]. It results from some of the following: Credit to non-existent borrowers; sham loans with kickbacks and diversion; double pledging of collateral; and linked financing; impersonation, fake documentation, under-valuation of properties, fictitious accounts, unofficial borrowing, fictitious contracts, unauthorized lending, lending to ghost borrowers [7-9]. It is obvious that human judgement of loan approvals with a record of no default is inefficient. 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 Naïve Bayes approach.

Bayesian classifiers are statistical classifiers that predict class membership probability that a given sample belongs to a particular class. This method is simple, elegant, and robust. It is a classification algorithm that has been in existence long ago and despite its simplicity, it is an efficient machine learning approach. It has a wide coverage of applications, for example in spam filtering, text classification, image processing. To enhance its flexibility, it has been modified numerous times in statistics, machine learning, and pattern recognition domains. The advantages of using Naive Bayes include the fact that the training takes a short computational time and the model is easily constructed; it is suitable for large dataset, and the iteration parameter estimation is less complicated. The represented knowledge is easily interpreted. Also, it is not specific to an application in its strength but it is robust and does well across board.

2. Materials and Methods

A credit dataset of 5000 instances with 9 predictors were employed for this research, with default being the target attribute. 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. Using Python 3.6.5 [12], based on IPython 6.4.0 (An enhanced Interactive Python) resident in Anaconda Navigator on a 64 bit (AMD64) system, important packages relevant to the work are: Numpy: 1.14.3; Pandas: 0.23.0; Matplotlib: 2.2.2; Seaborn: 0.8.1; Scipy: 1.1.0.

Weka [13] was used to build the model with cross validation and data splitting employed; WEKA is a Java-based modeling and prediction tool while matlab [14] is a simulation tool that can do a similar task; basic classification, training, testing, and features extraction were carried out. These tools are robust to handle many tasks in machine learning. The testing gave a result of 78% accuracy. The scatter plot of the model is shown in Figure 1. Through the confusion matrix, the true positive rate, false positive rate and other accuracy measures are as shown in Table 1.

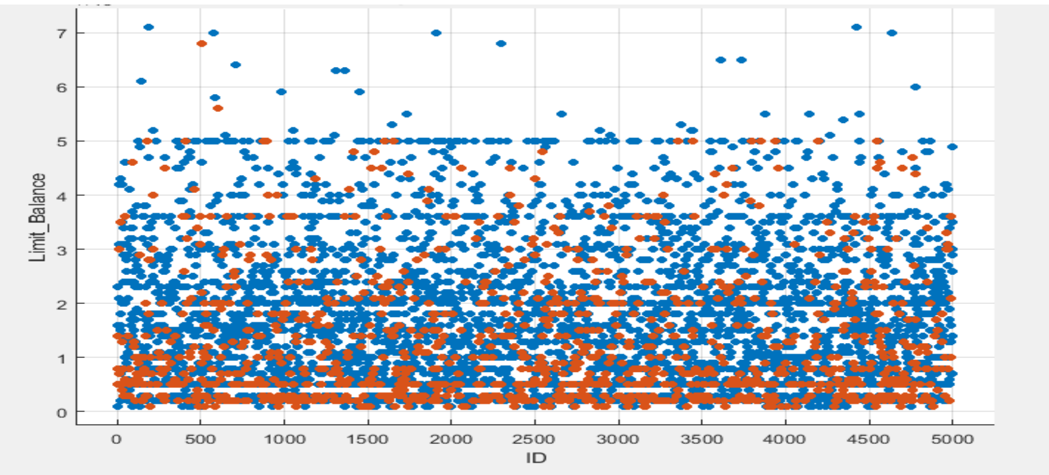

Figure 1: A scatter plot of the ID versus Limit balance

The summary of Naïve Bayes Performance metrics with weka using stratified cross-validation includes: Correctly classified instances which makes 3900 instances that represents 78 %; incorrectly classified instances which makes 1100 instances, representing 22 %; the Kappa statistics is 0.3586; while mean absolute error is 0.3071; root mean squared error is 0.4181; relative absolute error is 88.7218 %; root relative squared error is 100.5106 %. A total of 5000 instances were involved in the process.

Table 1. Detailed accuracy by class.

	True Positive Rate	**False Positive Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.862	0.507	0.856	0.862	0.859	0.735	N
	0.493	0.138	0.506	0.493	0.500	0.735	Y
Weighted Average	0.780	0.425	0.778	0.780	0.779	0.735	

3. Results and Discussion

Credit or loan defaults have led to bank insolvency and nations entering recession, making life unbearable for people. An approach using Naïve Bayes yielded 78% accuracy. Using cross validation and features extraction based on the principal component analysis, the training and testing was done and 25% of the dataset used for testing. The accuracy of the model is good and the extent of false alarm as evident in the false positive rate is minimal. As was presented in [15,16], this current study did not include hypothesis in its model formulation and testing, rather machine learning technique such as Naïve Bayes was engaged in the model formulation and prediction.

4. Conclusion

The study has proffered a technological solution to an age-long challenge to the financial institutions. Using a real life financial dataset and a collection of prediction and simulation tools; basic classification was done that ended in a dependable accuracy. False alarm is reduced to the minimum since the false positives are very few with true positives, precision, and other performance metrics highly favourable. This approach if employed by financial institutions for loan scrutiny will save economic loss, reduce human errors, and eliminate unnecessary bureaucracies in loan administration.

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] 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.

[2] Nigeria Interbank Settlement System, NIBSS 2015 2014-E-payment fraud landscape in Nigeria: A summary and analysis of reported e-payment frauds. Available from: www.nibss-plc.com.ng Retrieved: March, 2018.

[3] 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.

[4] 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.*

[5] 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.

[6] 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.

[7] Pollio G and Obuobie J 2010 Microfinance default rates in Ghana: Evidence from individual-liability credit contracts. *Micro-Banking Bulletin* **20**:8-13.

[8] 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.

[9] 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.

[10]World Bank 2018 Economic indicators for over 200 countries, https://www.theglobaleconomy.com/Nigeria/Nonperforming_loans/.Retrieved March, 2018.

[12] Python Software Foundation Python language reference version 2.7. Available at http://www.python.org

[13]Hall M Frank E Holmes G Pfahringer B Reutemann P and Witten H I 2009 The WEKA data mining software: An update. *SIGKDD explorations*, **11**(1).

[14]Matlab and Statistics Toolbox Release 2017b The MathWorks Inc Natick Massachusetts United States.

[15]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.

[16]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.