An improved Bank Credit Scoring Model A Naïve Bayesian Approach

Olatunji J. Okesola
Dept. of Computer & Information Sciences,
Covenant University,
Ota, Ogun state, Nigeria.
olatunji.okesola@covenantuniversity.edu.ng

Kennedy O. Okokpujie
Dept. of Electrical & Information Engineering
Covenant University,
Ota, Ogun state, Nigeria.
kennedy.okokpujie(@covenantuniversity.edu.ng

Adeyinka A. Adewale
Dept. of Electrical & Information Engineering
Covenant University,
Ota, Ogun state, Nigeria.
ade.adewale@covenantuniversity.edu.ng

Samuel N. John
Dept. of Electrical & Information Engineering
Covenant University,
Ota, Ogun state, Nigeria.
samuel.john(@covenantuniversity.edu.ng

Osemwegie Omoruyi
Dept. of Electrical & Information Engineering
Covenant University,
Ota, Ogun state, Nigeria.
osemwegie.omoruyi@cu.edu.ng

Abstract – Credit scoring is a decision tool used by organizations to grant or reject credit requests from their customers. Series of artificial intelligent and traditional approaches have been used to building credit scoring model and credit risk evaluation. Despite being ranked amongst the top 10 algorithm in Data mining, Naïve Bayesian algorithm has not been extensively used in building credit score cards. Using demographic and material indicators as input variables, this paper investigate the ability of Bayesian classifier towards building credit scoring model in banking sector.

Keywords: Banking sector, Credit score card, Naïve Bayesian, Data Mining, Credit risk evaluation

1 INTRODUCTION

A credit scoring model is an evaluation tool typically used in the decision-making process to reject or accept a loan request. This is indeed a mathematical model meant to estimate the possibility of customers' failure to pay and it is usually measured using a credit score. A higher score indicates a lower default probability.

Credit scoring model comes with a number of credit factors which are the entities of the actual measurement and depends on the loan or credit types. While the credit factors for a credit card loan may include customer's age, payment history, card utilization, and number of accounts, the credit factors for a mortgage loan may comprise of job history, down payment, and loan size. . Hence, this is adjudged as "the result of a statistical model which, based on information (age, number of previous loans, etc.) about the borrower, allows one to distinguish between "good" and "bad" loans and give an estimate of the probability of default"[1] [25].

Fundamentally, Credit scoring models are constructed by two popular statistical tools - logistic regression and Linear Discriminant Analysis [2] [3] [4]**.** However, the advance in Information technology has brought around not only the classical methods (SPSS, Enterprise Miner, etc.) but also new novel predictive modeling and classification techniques such as neural networks, decision tree, k-nearest neighbors, and support vector machine (SVM).

Historically, linear regression and discriminant analysis have been mostly used techniques for building scorecards [5]. Other data mining classifiers such as Artificial Neural Networks, Decision trees, genetic algorithm, and support vector machine are also being extensively used. However, literatures are mostly silent on the use of Naïve Bayern algorithm towards building credit scoring model. Hence, the objective of this paper is to employ Naïve Bayern using historical data (of the already availed loan) and borrowers' characteristics. However, for privacy reasons that make financial data from banks unavailable, this study is restricted to the use of demographic and material indicators as enlisted on Table I.

Table 1: List of Variables Demographic and Material Indicators

Name	*Description*	*Role*
Gender	Gender (Male, Female)	Input
Age	Age (in years)	Input
Tribe	Tribe/Sector	Input
Marrystat	Marital status	Input
Consistuency	Consistuency/Address	Input
YearsAtRes	Years at resident	Input
Status	Credit status (Good, Bad)	Target
ValueOfCars	Value of Cars	Input
ValueOfHouse	Value of House	Input
ValueOfLand	Value of Land	Input
EduLevel	Educational level	Input

978-1-5386-2652-8/17 $31.00 © 2017 IEEE
DOI 10.1109/CSCI.2017.36

2 CREDIT SCORING MODEL

Credit scoring models are developed by credit institutions especially banks to enhance their credit evaluation process and establish the credit worthiness of their intending creditors and assigns credit risks. This is a model-based estimate of the probability that a borrower will show some undesirable behavior in the future [6]. Unlike the judgmental approach commonly used by financial institutions which is based on 3C's, 4C's or 5Cs (capital, character, capacity, collateral and condition), credit scoring will classify a potential credit as either good or bad using demographic characteristics, historical data, and statistical techniques. A credit risk is said to be bad if the scoring model has a high probability of defaulting on the financial obligations, and it is said to be good when there is a high likelihood of repayment [7].

Credit scoring methods are commonly used among financial institutions but they are also applicable to other institutions such as telecommunication, recreational clubs, insurance, and real estate for predicting late payments. The ability of this model in allocating a rating to the credit quality of a loan request makes it easily adaptable to many areas of interests. For instance, the model provides a health score to the possibility that a client is yet to make a payment thereby allows the organization to adjust its risks and monitor its portfolio [1].

3 PARAMETRIC AND NON-PARAMETRIC MODELS

The quantitative models proposed to evaluate consumer loans are grouped as either parametric or non-parametric model [8]. The parametric models include Logistic regression (LR), the Linear Discriminant Analysis (LDA), and Multivariate adaptive regression splines (MARS). However, they are adjudged to be inaccurate as their variables have a linear relationship amongst themselves. For improved accuracy therefore, various non-parametric (data mining) models were built which include the decision trees (DT) [9] k-nearest neighbor [10], artificial neural networks (ANN) [11], genetic programming (GP) [12]; case-based reasoning (CBR) [13], genetic algorithm (GA) [14], Artificial Immune System Algorithm, classification based on association rules [15], rule extraction based on NN [16] and support vector machines (SVM) [17].

a. Use of Classifiers/classification algoritms

Consumer credit risk comes with a lot of prediction tasks. The Basel II standards in particular calls for financial organisations to estimate the exposure at default (EAD), the loss given default (LGD), and the probability of default (POD). These models have recently been well researched and continuously attracting much research interest. Classification and survival analysis are mostly used methods to develop PD models.

As researchers' interest continues growing in the credit risk domain, new techniques of building credit scoring models continue to emerge. The recent ones include: Multivariate adaptive regression splines (MARS), artificial neural networks (ANNs), classification and regression tree (CART), support vector machine (SVM), and case based reasoning (CBR).

Following heavy criticisms on ANNs for its black box' approach and associated interpretative difficulties [7] [17] investigated SVM approach and compare performance in credit rating prediction with back propagation neural networks (BNN). They observed that SVM has just a slight improvement over BNN. Meanwhile, when related with genetic programming, neural networks, and decision tree classifiers, "the SVM classifier achieved identical classification accuracy with relatively few input variables". [7] used MARS and CART to establish the effectiveness of credit scoring. They reported on the basis of scoring accuracy that MARS and CARTS outperform traditional discriminant analysis and other comparative techniques.

b. Naïve Bayesian

Bayes' Theorem is named after Thomas Bayes, and postulates two types of probabilities:

- Posterior Probability of H conditioned on X: [P(H/X)]; and
- Prior Probability of H irrespective of any observation or condition or information

Where:

X is data tuple (evidence) and H is hypothesis.

Hence, the probability that the hypothesis H holds given the "evidence" or observed data tuple X is given as:

$$P(H/X) = P(X/H)\ P(H) / P(X)$$

The Naïve Bayesian (NB) algorithm is centered on the Bayesian theorem with assumed independence amongst the predictors, using a set of training data to estimate the probability terms needed for classification [18]. This performance is measured by the accuracy of the estimated required probability terms, which is always a challenge as the training data is not easily available. Notwithstanding this limitation and despite the availability of numerous classifiers, NB still stands amongst the most popular classifiers and ranked among the top 10 performing data mining algorithms for its simplicity, practicability and effectiveness [19]. NB has been used and proved effective in so many domains including Agriculture [18], medicine [21], and biometrics [22]. However not much is done with it on credit scoring towards enhancing credit evaluation for banks and other financial institutions. This is the motivation for this work.

c. Conceptual framework

Building scorecard calls for historical data on the previously availed loan as score cards are should assign high and low percentage to good and bad borrowers respectively. A lot of variables are also required as indicators for various investigations including gross amount of loan, gender, age, number of dependents, marital status, outstanding mortgage, years lived at residence, monthly salary, spouse take-home, years on the job, bank account type, number of credit references. These are some of the major indicators used by [23] [24] to respectively predict the possible late-paying customers" as well as the credit risk card applicants.

Being guided by the same principles, we came up with a conceptual framework for our Improved Credit Score Model (ICSM) as shown in Figure 1. The value of assets (cars, house, Land) possessed are used as material indicators while the demographic indicators (which are significant especially when capturing various gender and regional differences) are gender, age, tribe/sector, marital status, consistency, years as resident, and educational level.

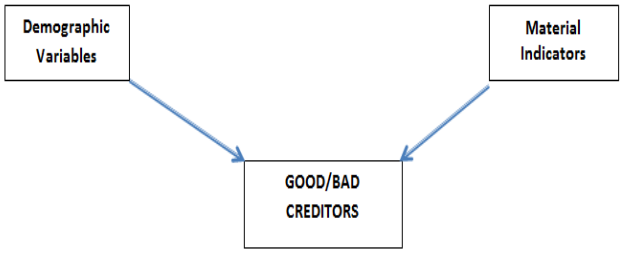

Figure 1: Conceptual framework for Improved Credit Score Model (ICSM)

4 METHODOLOGY

a. *The Variables*

The roles of variables as well as their measurement levels have to be defined and clearly stated when constructing Credit scoring model. Variable is either independent (input) or dependent (target) with respective role as input or target. All variables here other than the target and gender, are nominal. The research target is the payment status which is a binary variable denoted by 1 or 0 respectively and it is either good or bad. A loan is said to be good if the borrower does not default in the monthly payment in three consecutive months. The variable list, roles and descriptions are depicted on Table 1 and Figure 2.

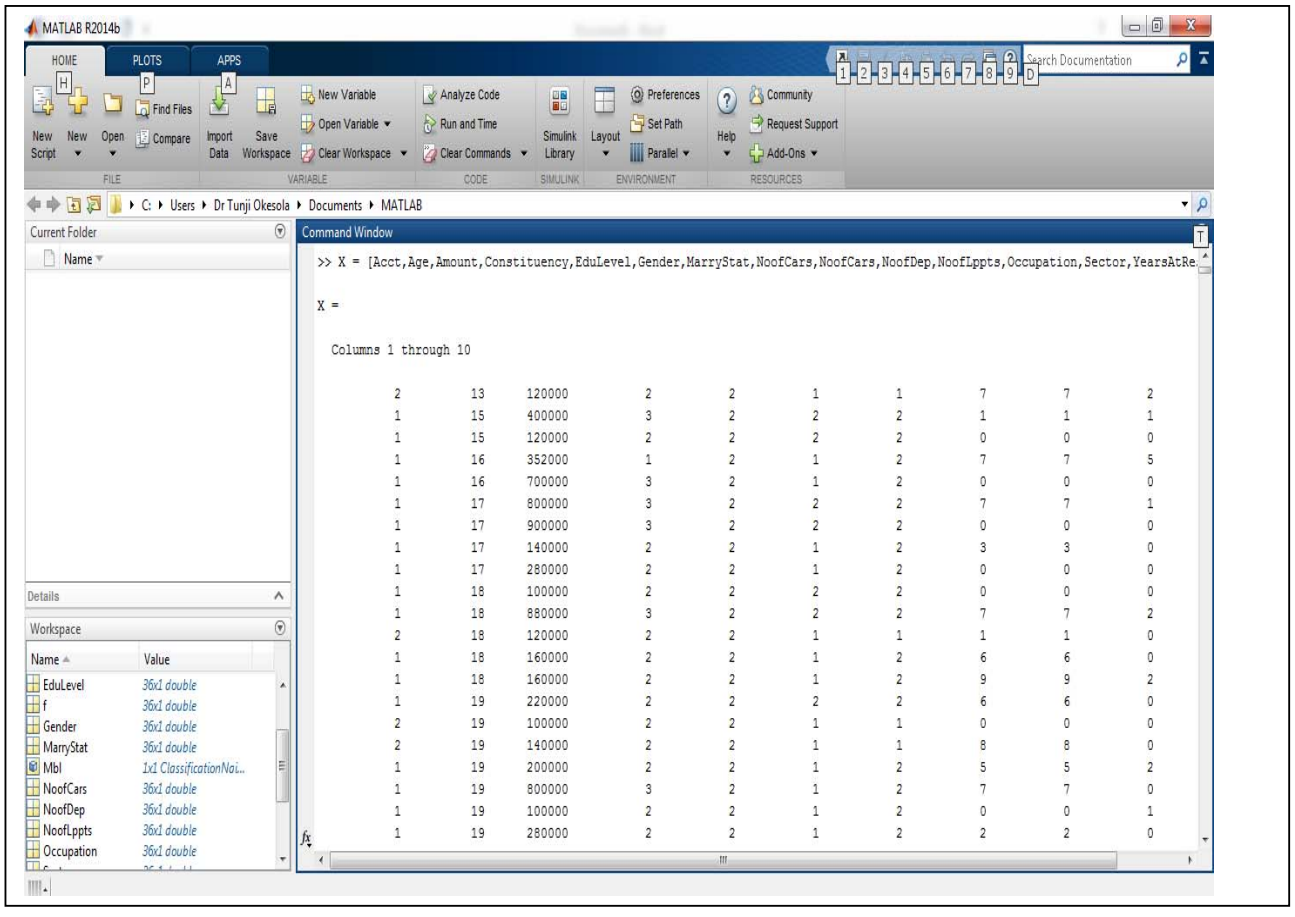

Figure 2: Some variables list, roles and descriptions in the Data set

b. The Construction of the improved credit scoring model (ICSM)

The process flow diagram for ICSM, starting with data retrieval and preparation is represented by Figure 3. Similar to other data mining tools, the Naive Bayesian algorithm provides "a graphical-user-interface (GUI) workspace whereby nodes (tool–icon) can be easily selected from a tools palette and placed into the diagram workspace. Nodes are then connected to form a process flow diagram that structure and document the flow of analytical activities [7].

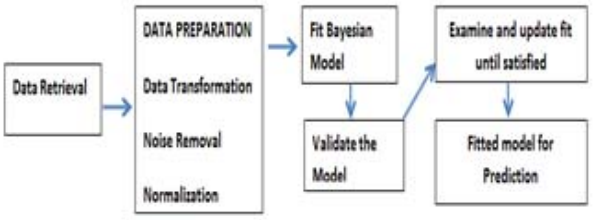

Figure 3: Improved Credit Score Model Process Flow

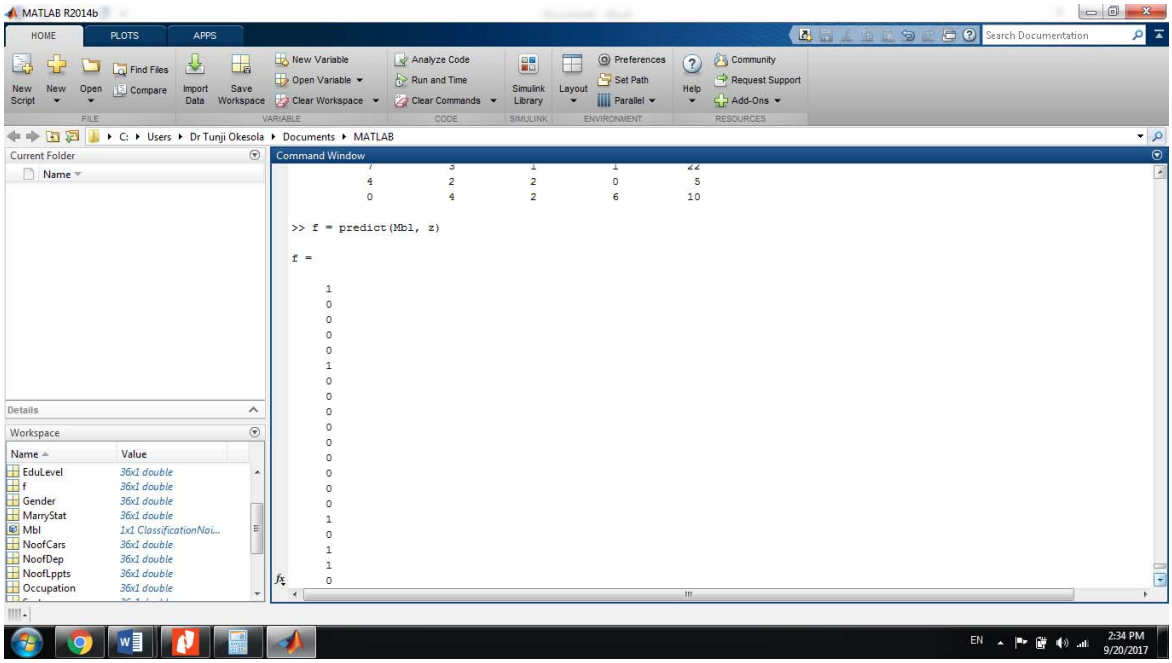

Figure 4: Classified good and bad credit requests

The data set was obtained from a medium size commercial bank and for confidential reasons; the data were analyzed and transformed to suite ICSM required data attributes. Noises like outliers and missing values were adjusted and the data sets were normalized to the permissible range of values. The credit status is the dependent variable and coded as good = 1 or bad = 0. The total population size is 690 consisting of 308 good and 382 bad applicants representing 44.6% and 55.4% respectively.

Using fitcnb fitting function on Matlab, the total population size was partitioned at ratio 34:66 to return a multiclass naïve Bayes model training and test sample data (Table 2). The Naïve Bayesian algorithm was then connected to the partition mode and the predictive performance was assessed.

c. Model Validation and Updates

This work examines the resubstitution error of the Naïve Bayes model to ascertain that all the ICSM data are correctly classified. This is affirmative with resuberror = 0.0933 as reported on table 3. Fitting parameters of the Naïve Bayes model were also tuned to get a more accurate result. Meanwhile, classification tree and K-nearest neighbor were fitted against the data to validate the accuracy of the ICSM.

Table 2: Partitioned Sample Data: Training and Test

Data	***Good***	***Bad***	***Total***
Training	187	266	453 (65.7%)
Test	121	116	237 (34.3%)
Total	**308 (44.6%)**	**382 (55.4%)**	**690 (100%) (100%)**

5. RESULTS

Naïve Bayesian algorithm has been able to predict credit request as good or bad respectively denoted by 1 or 0 (Figure 4).

Naïve Bayesian classifier is good for its speed and memory usage which are really good for simple but not kernel distributions. The model outperforms classification tree (ctree) and K-nearest neighbor (KNN) with 83.3% prediction accuracy as compared to 82% and 82.8% respectively, being that Naïve Bayes can handle categorical predictors almost perfectly compared to the classification tree and KNN.

Table 3: Resubstitution error for Naïve Bayesian

```
resuberror =
    0.0933
Properties for class ClassificationNaiveBayes:
    Y
    X
    W
    ModelParameters
    NumObservations
    PredictorNames
    CategoricalPredictors
    ResponseName
    ClassNames
    Prior
    Cost
    ScoreTransform
    DistributionNames
    DistributionParameters
    CategoricalLevels
    Kernel
    Support
```

6. CONCLUSION

In this work an improved credit scoring model for high speed and effective memory usage was built. Fitcnb fittings from Matlab was used to partition the sample population to training data and test data for our method, Bayesian algorithm to connect to the partition mode and predict loan request as good or bad.

ACKNOWLEDGMENT

This paper is sponsored by Covenant University, Ota, Ogun State, Nigeria as part of research support.

REFERENCES

[1] Deloitte. (2016). Credit scoring Case study in data analytics, (April), 1–18.

[2] Abdou, H., Pointon, J., & El-Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems with Applications*, *35*(3), 1275–1292. https://doi.org/10.1016/j.eswa.2007.08.030

[3] Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, *16*(2), 149–172. https://doi.org/10.1016/S0169-2070(00)00034-0

[4] Vojtek, M., & Kocenda, E. (2006). Credit-Scoring Methods (in English). *Czech Journal of Econimics and Finance*, *56(3-7)*(October 2014), 152–167.

[5] Wah, B., Huat, S., Huselina, N., & Husain, M. (2011). Expert Systems with Applications Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems With Applications*, *38*(10), 13274–13283.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Lessmann, S., Baesens, B., Seow, H., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring : An update of research. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

[7] Wah, B., Huat, S., Huselina, N., & Husain, M. (2011). Expert Systems with Applications Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems With Applications*, *38*(10), 13274–13283.

[8] Dahiya, S., Handa, S. S., Singh, N. P., & Communication, S. (2015). Credit Scoring Using Ensemble of Various Classifiers on Reduced Feature Set, *43*(4), 163–174. https://doi.org/10.5937/industrija43-8211

[9] Zhou, X.Y., Zhang, D.F., & Jiang, Y. (2008). A new credit scoring method based on rough sets and decision tree. *Lecture Notes in Artificial Intelligence,* 5012, 1081-1089.

[10] Henley, W.E., & Hand, D.J. (1996). A k-Nearest-Neighbour Classifier for Assessing Consumer Credit Risk. *Statistician,*45(1), 77. doi:10.2307/2348414

[11] West, D., Dellana, S., & Qian, J.X. (2005). Neural network ensemble strategies for financial decision applications.*Computers and Operations Research,* 32(10), 2543-2559

[12] Abdou, H.A. (2009). Genetic programming for credit scoring: The case of Egyptian public sector banks. *Expert Systems with Applications,* 36(9), 11402-11417.

[13] Chuang, C.L., & Lin, R.H. (2009). Constructing a reassigning credit scoring model, Part 1. *Expert Systems with Applications,* 36(2), 1685-1694.

[14] Zhang, D.F., Huang, H.Y., Chen, Q.S., & Jiang, Y. (2007). A comparison study of credit scoring models. *Natural Computation,* 1(15-18), 24-27.

[15] Yin, X., & Han, J. (2003). CPAR: Classification based on predictive association rule. In: SDM, San Francisco, CA.

[16] Setiono, R., Baesens, B., & Mues, C. (2008). Recursive neural network rule extraction for data with mixed attributes, neural networks. *IEEE Transactions,* 19, 299-307

[17] Huang, C.L., Chen, M.C., & Wang, C.J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications,* 33(4), 847-856.

[18] Diab, D. M., & El Hindi, K. M. (2017). Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification. *Applied Soft Computing Journal*, *54*, 183–199. https://doi.org/10.1016/j.asoc.2016.12.043

[19] Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2008). *Top 10 algorithms in data mining. Knowledge and Information Systems (Vol. 14). https://doi.org/10.1007/s10115-007-0114-2*

[20] Rammal, A., Perrin, E., Vrabie, V., Assaf, R., & Fenniri, H. (2017). Selection of discriminant mid-infrared wavenumbers by combining a naïve Bayesian classifier and a genetic algorithm: Application to the evaluation of lignocellulosic biomass biodegradation. *Mathematical Biosciences*, *289*, 153–161. https://doi.org/10.1016/j.mbs.2017.05.002

[21] Kazmierska, J., & Malicki, J. (2008). Application of the Na??ve Bayesian Classifier to optimize treatment decisions. *Radiotherapy and Oncology*, *86*(2), 211–216. https://doi.org/10.1016/j.radonc.2007.10.019

[22] Dai, Q., Li, J., Wang, J., Chen, Y., & Jiang, Y. G. (2016). A Bayesian Hashing approach and its application to face recognition, *Neurocomputing*, *213*, 5–13. https://doi.org/10.1016/j.neucom.2016.05.097

[23] Ang, J. S., Chua, J. H., & Bowling, C. H. (1979). The Profiles of Late-Paying Consumer Loan Borrowers: An Exploratory Study: Note.

Journal of Money, Credit and Banking, *11*(2), 222. https://doi.org/10.2307/1991836

[24] Chye, K. H., Chin, T. W., & Peng, G. C. (2004). Credit scoring using data mining techniques. *Singapore Management Review*, *26*(2), 25–47.

[25] S. John, C. Anele, O. O. Kennedy, F. Olajide, and C. G. Kennedy, "Realtime Fraud Detection in the Banking Sector Using Data Mining Techniques/Algorithm," in *Computational Science and Computational Intelligence (CSCI), 2016 International Conference on*, 2016, pp. 1186-1191: IEEE.