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 Bayesian algorithm towards building credit scoring model. Hence, the objective of this paper is to employ Naïve Bayesian 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

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Gender (Male, Female)</td>
<td>Input</td>
</tr>
<tr>
<td>Age</td>
<td>Age (in years)</td>
<td>Input</td>
</tr>
<tr>
<td>Tribe</td>
<td>Tribe/Sector</td>
<td>Input</td>
</tr>
<tr>
<td>MarryStat</td>
<td>Marital status</td>
<td>Input</td>
</tr>
<tr>
<td>Constituency</td>
<td>Constituency/Address</td>
<td>Input</td>
</tr>
<tr>
<td>YearsAtRes</td>
<td>Years at resident</td>
<td>Input</td>
</tr>
<tr>
<td>Status</td>
<td>Credit status (Good, Bad)</td>
<td>Target</td>
</tr>
<tr>
<td>ValueOfCars</td>
<td>Value of Cars</td>
<td>Input</td>
</tr>
<tr>
<td>ValueOfHouse</td>
<td>Value of House</td>
<td>Input</td>
</tr>
<tr>
<td>ValueOfLand</td>
<td>Value of Land</td>
<td>Input</td>
</tr>
<tr>
<td>EduLevel</td>
<td>Educational level</td>
<td>Input</td>
</tr>
</tbody>
</table>
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 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 algorithms

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:

\[ P(H/X) = \frac{P(X/H) P(H)}{P(X)} \]

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 Naïve 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].

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.

<table>
<thead>
<tr>
<th>Data</th>
<th>Good</th>
<th>Bad</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>187</td>
<td>266</td>
<td>453</td>
</tr>
<tr>
<td>Test</td>
<td>121</td>
<td>116</td>
<td>237</td>
</tr>
<tr>
<td>Total</td>
<td>308</td>
<td>382</td>
<td>690</td>
</tr>
</tbody>
</table>

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.

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.
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

<table>
<thead>
<tr>
<th>Error</th>
<th>0.9933</th>
</tr>
</thead>
</table>

Properties for classification NaïveBayes:
- Y
- X
- W

Model Parameters
- 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
