*Available online www.jsaer.com*

***Journal of Scientific and Engineering Research*, 2017, 4(12):99-110**

**Research Article**

**ISSN: 2394-2630**
**CODEN(USA): JSERBR**

# Credit Scoring using Principal Components Analysis-based Binary Logistic Regression

**S. Suleiman[1], M.S. Burodo[2], Issa Suleman[1]**

[1]Department of Mathematics, Usmanu Danfodiyo University, Sokoto, Sokoto State, Nigeria, ,
[2]Department of Business Administration and Management, Federal Polytechnic Kaura Namoda, Zamfara State Nigeria

**Abstract** The research paper deals with credit scoring in banking system, using First Bank of Nigeria (FBN), plc as case study. The aim of this study was to improve the predictive power of binary logistic Regression models using principal components as input for predicting applicant status (i.e. Creditworthy or Non-creditworthy) for the new applicant (customer). The developed model was compared with binary logistic regression models. Performance indicator such as Prediction Accuracy (PA), Nagelkerke R Square ($R^2$), Index of Agreement (IA), Normalised Absolute Error (NAE) and Root Mean Square Error (RMSE) were used to measure the accuracy of the models. Results showed that the use of principal component as inputs improved Binary logistic regression models prediction by reducing their complexity and eliminating data co-linearity. Based on eigenvalues over six factors were retained. The factors accounted for 72.4 percent of the variance. The combination of items with loadings greater than 0.40 were considered as separate between important and less important factors.

**Keywords** Principal Component Analysis, Binary Logistics Regression, Principal Component Regression, Performance Indicator, Credit Scoring
**JEL CLASSIFICATION**: C1, C3, C8, G22

## 1. Introduction

Risk is everywhere. May be, risk components have been increased dramatically in the recent years in comparison with the past, especially in the case of health and safety issues, it is also true in the case of financial products, for example, credit risk [1]. And this credit risk develops from the probability that the borrowers may be unwilling or unable to fulfill their contractual obligations [2]. The most important tool for the assessment of credit risk is credit scoring and credit scoring attempts to summarize a borrower's credit history by using credit scoring model [3]. Credit scoring models are decision support systems that take a set of predictor variables as input and provide a score as output and creditors use these models to justify who will get credit and who will not [4].

### 1.1. Credit Scoring

Credit scoring can be formally defined as a statistical (or quantitative) method that is used to predict the probability that a loan applicant or existing borrower will default or become delinquent. This helps to determine whether credit should be granted to a borrower [5]. Credit scoring can also be defined as a systematic method for evaluating credit risk that provides a consistent analysis of the factors that have been determined to cause or affect the level of risk [6]. The objective of credit scoring is to help credit providers quantify and manage the financial risk involved in providing credit so that they can make better lending decisions quickly and more objectively. In the United States, the Circuit Court has found considerable actuarial evidence that credit scores are a good predictor of risk of loss [7]. Similarly, a recent actuarial study has concluded that credit scores are

one of the most powerful predictors of risk; they are also the most accurate predictor of loss seen in a long time [8].

**1.2. Benefits of Credit Scoring**

Credit scoring has many benefits that accrue not only to the lenders but also to the borrowers. For example, credit scores help to reduce discrimination because credit scoring models provide an objective analysis of a consumer's creditworthiness. This enables credit providers to focus on only information that relates to credit risk and avoid the personal subjectivity of a credit analyst or an underwriter [6].

Those are the tips you must know when you are seek for a bank loan in Nigerian such as:

1) You must need to formally apply to a bank for loan
2) Banks charge interest on a per annum basis and they are not fixed
3) Different banks offer different interest rates and terms and conditions
4) Never ignore the terms and conditions
5) Banks always ask for collateral or some form of security
6) Defaulting on repaying your bank loans when their due doesn't mean the bank will take over your business
7) You can always attempt to refinance or restructure your loan
8) You can ask your bank for a moratorium
9) The biggest threat to defaulting is not your interest rate but your debt service coverage ratio
10) Banks have hidden charges

**1.3. Logistic Regression**

Logistic regression is a widely used statistical modeling technique in which the probability of a dichotomous outcome is related to a set of potential independent variables. The logistic regression model does not necessarily require the assumptions of discriminant analysis. However, Harrell and Lee [9] found that logistic regression is as efficient and accurate as discriminant analysis even though the assumptions of discriminant analysis are satisfied. Logistic regression models have been widely discussed in social research, medical research, design, control, bankruptcy prediction, market segmentation, and customer behaviors. Abrahams and Zhang [10] drew attention to the importance of the "Sequence" of the independent variables selection.

**1.4. Principal Component Analysis**

PCA is mostly used for reducing the multiple dimensions associated to Binary Logistics Regression which create new variables called the principal component (PCs) that are orthogonal and uncorrelated to each other. The first PC explains the largest fraction of the original data variability and second PC explains larger fraction than third PC and so on [10-13]. Varimax rotation is mostly used to obtain the rotated factor loadings that represent the contribution of each variable to a specific principal component. Principal Component Logistics Regression (PCLR) is a method that combines Logistics Regression (LR) and PCA. PCD establishes a relationship between the output variable (*i.e.* applicant status) and the selected PC of the input variables ($x$i).

**1.5. Justification for First Bank of Nigeria (FBN) plc**

The First Bank of Nigeria is the largest bank in Nigeria based on branch network, total assets and capital. The bank has a well-established brand name having operated for more than 100 years and reflects sound cooperate governance (Global Credit rating Co., 2012). The creation of loans and management of the risks inherent in the loan portfolio remained a focal point for the First Bank Group, even as the bank continued to contend with a high influx of credit applications due to the general credit squeeze in the market (First Bank of Nigeria Plc Annual Report & Account 2010, page 89). This bank also gives customers loan base on the terms and conditions which had already been stated in form such as: interest rate (charges), type of collateral security and their attribute of applicant such as: Age, sex, ownership of residence, marital status, employment classification, length of service, salary, applicant request, amount request, credit amount, propose tenor in month, other borrowing and applicant status. The model can help the FBN to rate customers and also to reduce the level of default among the customers.

**1.6. Aim and Objectives of the study**

The aim of this study is to classify applicant as credit worthy or non-credit worthy. This aim can be achieved through the following objectives:

1. To build a logistic regression model capable of predicting an applicant credit status
2. To test the homogeneity of variance among the variables using Bartlett's Test.
3. To identify a number of factors that represent the relationship among sets of inter-related variables using principal component and factor analysis.
4. To verify the variables that contributes significantly to the percentage of variance in the components.
4 To build Binary Logistics Regression model capable of predicting an applicant credit status using Principal Component (PC) as an input.

**2. Literature review**

The prescription of credit scoring is to recognize patterns in the population based on the similarities. Fisher [14] introduced the concept in the statistics and Durand [15] identified that it might be applied to recognize good and bad loans, as cited by Thomas, Edelman et al. [16]. Credit scoring was first used in consumer banking in the 1960 among the finance companies, after that gradually retailers and credit card companies started using the concept [17]. At that moment, the tremendously increasing number of applicants for credit cards forced the lenders to automate their credit decisions because of the economic and manpower related reasons, ultimately these organizations found credit scoring system more accurate than judgmental systems (default rate dropped by 50% or more) [16]. Moreover, according to Mays [18] based on opinions of several industry experts, the first one was the Montgomery Ward's scoring system for credit card application, and at present, mortgage industry started adopting the same theory. And this Montgomery Ward was one of best clients of Fair Issacs Company that invented the credit score to help lenders to better analyze applicant's creditworthiness and this company introduced the first credit scoring model in the year of 1958 [19].

A credit scoring model is a complicated set of algorithms that creditors use to evaluate the creditworthiness of a specific customer [20]. These models give unique advantages, for example, they provide a rigorous way of screening credit applications and save huge amount of time and cost (providing salaries to credit analysts) [21]. Among the available models, four approaches are most widely used and those are Linear Discriminant Analysis, Logistic Regression and Probit, K-nearest Neighbor Classifier, Support Vector Machine Classifier [22]. All of these algorithms have one similarity, all of them include parameters that are defined by the variables and the variables can be obtained from a credit report or an application form. The variables can be different types, for example credit history, income, outstanding debt among others, those are explained in detail in the next chapter of this study.

Logistic regression is now widely used in credit scoring and more often than discriminant analysis because of the improvement of the statistical software's for logistic regression [23]. Moreover, logistic regression is based on an estimation algorithm that requires fewer assumptions (assumption of normality, assumption of linearity, assumption of homogeneity of variance) than discriminant analysis [4]. This study is not testing for that assumption.

**3. Data and Methodology**

A real world credit dataset is used in this research. The dataset is extracted from the application forms of First Bank of Nigeria, plc. The dataset is referred to as "Credit Dataset". After preparing the dataset, it is used in the subsequent sections for conducting the analysis with Principal Component and Logistics Regression Analyses. The estimated credit scoring model is based on a binary logistic regression with principal components as exogenous inputs'.

**Table 1:** Credit Dataset Description

| No. | Variable | Type | Scale | Description |
|---|---|---|---|---|
| 1 | Attribute1 | Input Variable | Scale | Age of the Applicant |
| 2 | Attribute2 | Input Variable | Nominal | Sex of the Applicant |
| 3 | Attribute3 | Input Variable | Nominal | Ownership of residence |
| 4 | Attribute4 | Input Variable | Nominal | Marital status |
| 5 | Attribute5 | Input Variable | Nominal | Qualification |
| 6 | Attribute6 | Input Variable | Nominal | Employment status |

| | | | | |
|---|---|---|---|---|
| 7 | Attribute7 | Input Variable | Nominal | Employment classification |
| 8 | Attribute8 | Input Variable | Scale | Length of service |
| 9 | Attribute9 | Input Variable | Scale | Salary |
| 10 | Attribute10 | Input Variable | Nominal | Application Request |
| 11 | Attribute11 | Input Variable | Scale | Amount Request |
| 12 | Attribute12 | Input Variable | Scale | Credit Amount |
| 13 | Attribute13 | Input Variable | Scale | Proposed tenor in month |
| 14 | Attribute14 | Input Variable | Nominal | Other borrowing |
| 15 | Attribute15 | Output Variable | Nominal | Status of the Credit Applicant |

The dataset contains 200 cases, 163 applicants are considered as "Creditworthy" and the rest 37 applicants are treated as "Non-creditworthy". The dataset holds 15 variables altogether. Among the variables, 9 variables are "Categorical" and the rest 6 variables are "Numerical". Moreover, there are 14 independent variables (input variables) and 1 dependent variable (output variable) in the dataset.

**3.1. Logistic Regression Model**

Logistic regression or Logit deals with the binary case, where the response variable consists of just two categorical values. Logistic regression model is mainly used to identify the relationship between two or more explanatory variables $X_i$ and the dependent variable Y. Logistic regression model has been used for prediction and determining the most influential explanatory variables on the dependent variable [24]. The Logistic regression model for the dependence of $p_i$ (response probability) on the values of k explanatory variables $x_1$, $x_2$,……..$x_k$ is given below [25]

$$P_i = \frac{\exp(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)} \quad (1)$$

Which is linear and similar to the expression of multiple linear regressions.

Where $\left(\frac{P_i}{1-P_i}\right)$ is the ratio of the probability of a failure and called odds, $\beta_0$, $\beta_i$ are parameters to be estimated and $P_i$ is the response probability.

In logit model the predicted values of the outcome variable are expected to range between 0 and 1 regardless of the values of the explanatory variables. These are fourteen predictors variables which capable to predict for applicant status

$X_1$ rep. Age of applicant $X_2$ rep. Sex of the applicant, $X_3$ rep. ownership of residence,
$X_4$ rep. marital status $X_5$ rep. Qualification $X_6$ rep. Employment status
$X_7$ rep Employment classification $X_8$ rep. Length of service $X_9$ rep. Salary
$X_{10}$ rep. Application Request $X_{11}$ rep. Amount request $X_{12}$ rep. Credit Amount
$X_{13}$ rep. proposed tenor in month $X_{14}$ rep. other borrowing

**3.1.1. Binary Logistic Model the Analysis of Data**

**Table 2:** Dependent Variable Encoding

| Original value | Internal value |
|---|---|
| Creditworthy | 0 |
| Non-creditworthy | 1 |

The table 2 for the categorical variable coding, above indicates that the creditworthy are coded with 0 while the Non-creditworthy coded 1 as used in the analysis.

**3.2. Omnibus Chi-square Test**

The omnibus Chi-square test is a log-likelihood ratio test for investigating the model statistical significance of the coefficients in logistic regression. The test procedures are as follows:

Hypothesis for Omnibus Chi-square Test:

$H_o$: The model coefficients are not statistically significant

$H_1$ : The model coefficients are statistically significant

Test statistic:

$\chi^2 = 2[\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} \ln(\frac{O_{ij}}{e_{ij}})]$ (2)

Where $O_{ij}$ *is observed value and* $E_{ij}$ *is expected value.* where $E_{ij} = \frac{Row\ total \times Column\ total}{Grand\ total}$

Decision Rule:

Reject $H_o$ if p<0.05 otherwise accept $H_o$ at the 5% level of significance i.e. significance of the logistic model.

**3.2.1. Justification for the use of Omnibus Chi-square test**

It is justifiable and even necessary to investigate the significance of the model coefficient in the logistic model. Hence, the Omnibus test is applied.

**3.3. Wald Test**

The Wald test is used to test the statistical significance of each coefficient $(\beta)$ in the logistic model. A Wald test calculates a $Z$ statistic which is:

$$W = \frac{\hat{\beta}}{SE(\hat{\beta})}.$$

This value is squared which yields a chi-square distribution and is used as a Wald test statistic.

Decision rule: Reject $H_o$ (the null hypothesis that the coefficient is equal to zero) when p-value of that coefficient is less than $\alpha$ level of significance.

**3.4. Principal Component Analysis**

Principal component analysis used to find a small set of linear combinations of the covariates which are uncorrelated with each other. This will avoid the multicollinearity problem. Besides, it can ensure that the linear combinations chosen have maximal variance. Application of principal component analysis (PCA) in regression has long been introduced by Kendall [26] in his book on Multivariate Analysis. Jeffers [27] is suggested for regression model to achieve an easier and more stable computation, a whole new set of uncorrelated ordered variables that is the principal components (PCs) be introduced [28].

Hussain et al. [29]: The steps involved in the analysis of PCA include the method of getting the data, standardizing the data, calculating the covariance matrix, calculating the eigenvectors and eigenvalues of the covariance matrix and visualizing the results. Algebraically, principal components are particular linear combinations of the p random variables.

Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with their development does not require a multivariate normal assumption. On the other hand, principal components derived for multivariate

Normal populations have useful interpretations in terms of the constant density ellipsoids.

Consider the six selected predictor variables which are capable of characterizing applicant. These variables are also believed to vary significantly between Credit worthy ($\pi_1$) and Non-Credit Worthy ($\pi_2$) applicants. These variables are;

PC1 = Principal component 1
PC2 = Principal component 2
PC3 = Principal component 3
PC4 = Principal component 4
PC5 = Principal component 5
PC6 = Principal component 6

### 3.5. Measurement of models performance

Performance indicators were used to evaluate the goodness of fit for the BLR and Principal Component Regression (PCR) to determine which method is appropriate to represent the applicant status (classify customer status in FBN). Performance indicators that are used to determine the best method for predicting applicant status are normalized absolute error (NAE), root mean square error (RMSE), index of agreement (IA), prediction accuracy (PA), and coefficient of determination ($R^2$). The equations used are reported by Lu (2003).

### 3.6. Keiser Meyer Olkin's and Bartlett's test of Sampling Adequacy and measuring the Homogeneity of variance across variables for Credit scoring.

**$H_{01}$:** The sampled data is adequate for the study

**$H_1$:** The sampled data is not adequate for the study.

**$H_{02}$**: $\delta_1 = \delta_2 = \cdots = \delta_k$

H1: $\delta_i \neq \delta_k \; for \; at \; least \; one \; pair(i,j)$

**Test Statistics:** KMO

**Interpretation rule:** 0.5 – 0.7 (mediocre), 0.7 – 0.8 (good), 0.8 – 0.9 (great) and value greater than 0.9 (superb). It can also re-classified as follows: anything in the 0.90s is 'marvelous', in the 0.80s 'meritorious', in the 0.70s 'middling', in the 0.60s 'mediocre', in the 0.50s 'miserable' and below 0.50s 'unacceptable'.

## 4. Result and Discussion

### 4.1. Logistic Regression

The Binary logistic regression models were developed with 200 customers using SPSS version 21.0 and E-view 7.0. Hence the coefficient of determinant is obtained used Nagelkerke $R^2$ (0.535)

**Table 3:** Omnibus Tests of Model Coefficients

| | | Chi-square | Df | Sig. | Nagelkerke R Square |
|---|---|---|---|---|---|
| Step 1 | Step | 163.152 | 14 | 0.000 | 0.536 |
| | Block | 163.152 | 14 | 0.000 | |
| | Model | 163.152 | 14 | 0.000 | |

The significance test for the model chi-square is the statistical evidence of the presence of a relationship between the dependent variable and the combination of the independent variables. In this analysis, the probability of the model chi-square (163.152) is <0.000, less than or equal to the level of significance of .05. The null hypothesis that there is no difference between the model with only a constant and the model with independent variables is rejected. The existence of a relationship between the independent variables and the dependent variable is supported.

**Table 4:** Important Variables Identified By the Logistic Regression Model

| Variables in the Equation | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | B | S.E. | Wald | Df | Sig. | Exp(B) |
| Step 1[a] | Age | 0.535 | 0.233 | 5.259 | 1 | 0.022 | 1.707 |
| | Sex | 0.527 | 1.689 | 0.098 | 1 | 0.755 | 1.695 |
| | Ownership | -1.040 | 1.301 | 0.639 | 1 | 0.424 | 0.354 |
| | Marital Status | -1.147 | 0.899 | 1.628 | 1 | 0.202 | 0.318 |
| | Qualification | -0.277 | 0.543 | 0.259 | 1 | 0.611 | 0.758 |
| | Employment Status | -3.007 | 2.736 | 1.208 | 1 | 0.272 | 0.049 |
| | Employment Classification | -4.006 | 1.946 | 4.236 | 1 | 0.040 | 0.018 |
| | Length Of Service | 0.051 | 0.225 | 0.051 | 1 | 0.821 | 1.052 |
| | Salary | 0.000 | 0.000 | 0.739 | 1 | 0.390 | 1.000 |
| | Application Request | -1.636 | 2.198 | 0.554 | 1 | 0.457 | 0.195 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Amount Request | 0.000 | 0.000 | 9.499 | 1 | 0.002 | 1.000 |
| Credit Amount | 0.000 | 0.000 | 8.538 | 1 | 0.003 | 1.000 |
| Propose Tenure | 0.271 | 0.215 | 1.590 | 1 | 0.207 | 1.311 |
| Other Borrowing | 8.906 | 3.243 | 7.543 | 1 | 0.006 | 7373.903 |
| Constant | -33.936 | 13.974 | 5.898 | 1 | 0.015 | 0.000 |

The independent variables with the probabilities of the Wald statistic less than or equal to the level of significance of 0.05 hold statistically significant relationships with the dependent variable. The statistically significant independent variables are Age of the Applicant, Employment Classification, Amount Request, Credit Amount and Other Borrowing. Here, the insignificant variables have probabilities of Wald statistic greater than the level of significance of 0.05.

### 4.2. Logit Model Result (Build from original data)

The required model for the significant predictor variables as follow:

$$\log_e\left(\frac{\pi}{1-\pi}\right) = -33.936 + 0.535AGE + (-4.006)EC + 0.00AR + 0.000CA + 8.906OB$$

To estimate odds, the model is exponential as

$$\frac{\pi}{1-\pi} = e^{-33.936+0.535AGE+(-4.006)EC+0.00AR+0.000CA+8.906OB}$$

The probability of bad applicant is obtained by applying transformation

$$\pi = \frac{e^{-33.936+0.535AGE+(-4.006)EC+0.00AR+0.000CA+8.906OB}}{1+e^{-33.936+0.535AGE+(-4.006)EC+0.00AR+0.000CA+8.906OB}}$$

Where AGE rep. Age of applicant, EC. Rep. Employment classification, AR rep. Amount request, CA rep. Credit Amount, OB rep. other borrowing.

### 4.3. Principal Component Regression

#### 4.3.1. Test of Principal Component Analysis Assumption (KMO and Bartlett's test result)

Another important test for PCA is the Kaiser Meyer Olkin (KMO) of sample adequacy and Bartlett's test of sphericity. Kaiser (1974) recommends accepting values greater than 0.5, which means the result for this research is acceptant with the value of KMO is 0.566. Bartlett's test is highly significant ($p < 0.001$) and therefore factor analysis is appropriate for this data.

**Table 5:** KMO Statistics for Sample Adequate and Bartlett's test for Homogeneity

| Test | DF | Approx. Chi-Square | P-value |
|---|---|---|---|
| Keiser-Meyer-Olkin Measure of Sampling Adequate | - | - | 0.566 |
| Bartlett's Test of Sphericity | 91 | 1464.453 | 0.000 |

#### 4.3.2. Principal Component Analysis Result (Extracted Components and their factor loadings)

**Table 6:** Total Variance Explained

| Component | Initial Eigen Value | | | Rotated Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|
| | Total | % of variance | Cumulative % | Total | % of variance | Cumulative % |
| 1 | 3.474 | 24.818 | 24.818 | 3.474 | 24.818 | 24.818 |
| 2 | 1.851 | 13.219 | 38.036 | 1.851 | 13.219 | 38.036 |
| 3 | 1.384 | 9.889 | 47.925 | 1.384 | 9.889 | 47.925 |
| 4 | 1.348 | 9.631 | 57.556 | 1.348 | 9.631 | 57.556 |
| 5 | 1.078 | 7.698 | 65.254 | 1.078 | 7.698 | 65.254 |
| 6 | 1.006 | 7.185 | 72.439 | 1.006 | 7.185 | 72.439 |
| 7 | 0.898 | 6.416 | 78.854 | | | |
| 8 | 0.807 | 5.761 | 84.615 | | | |
| 9 | 0.775 | 5.539 | 90.154 | | | |

| | | | |
|---|---|---|---|
| 10 | 0.717 | 5.121 | 95.275 |
| 11 | 0.438 | 3.128 | 98.403 |
| 12 | 0.126 | 0.901 | 99.305 |
| 13 | 0.073 | 0.521 | 99.825 |
| 14 | 0.024 | 0.175 | 100.000 |

Table 6 shows the Eigen values in column two, which are the proportions of total variance in all the variables, which are accounted for by the components. From the output, the first principal component has variance 3.474 (equal to the largest Eigen value) and account for 24.818% of total variance explained followed by second principal component variance 1.851 account for 13.219% of total variance explained and so on. The second component is formed from the variance remaining after those associated with the first component has been extracted, thus this account for the second largest amount of variance. It is worthwhile to note that the principal component coefficient which gives the variance explained for each component gives the values less than 30% of the variance explained. Therefore more than one component is needed to describe the variability of the data. In other to obtain a meaningful interpretation of the principal component analysis, we need to reduce to fewer than fourteen (14) components. In this study, i.e. extraction Eigen Values for the retained components, we observed that six (6) components are retained together with their percentage of variance explained by each component. The cumulative variance gives as well, shows that the first nine components account for about 72.439% of the total variance in the data.

**Table 7**: Rotated Component Matrix

| | **Component** | | | | | |
|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** |
| Age of Applicant | | 0.949 | | | | |
| Sex | | | | | 0.715 | |
| Ownership Residences | | | | | 0.748 | |
| Marital status | | | | 0.705 | | |
| Qualification | | | | -0.716 | | |
| Employment status | | | | | | 0.749 |
| Employment classification | 0.629 | | | | | |
| Length of service | | 0.953 | | | | |
| Salary | 0.922 | | | | | |
| Applicant request | | | | | | -0.747 |
| Amount request | 0.918 | | | | | |
| Credit amount | 0.971 | | | | | |
| Propose tenor in month | | | 0.808 | | | |
| Other borrowing | | | -0.412 | | | |

Table 7 presents rotated component matrix which gives the factor loadings for each variable using varimax rotation method developed by Keiser. This matrix contains the loading of each variable onto each factor where values less than 0.4 are suppressed from the output. We will go across each row and highlight the factor that each variable loads most strongly on. Based on these factor loadings, the factors represent.

- Age of applicant and length of service loaded strongly on factor 2
- Employment classification, Salary, Amount request and credit amount loaded strongly on factor 1
- Sex and ownership Residences loaded strongly on factor 5
- Marital status and Qualification loaded strongly on factor 4
- Employment status and Applicant Request loaded on factor 6
- Proposed tenor in month and other borrowing loaded strongly on factor 3

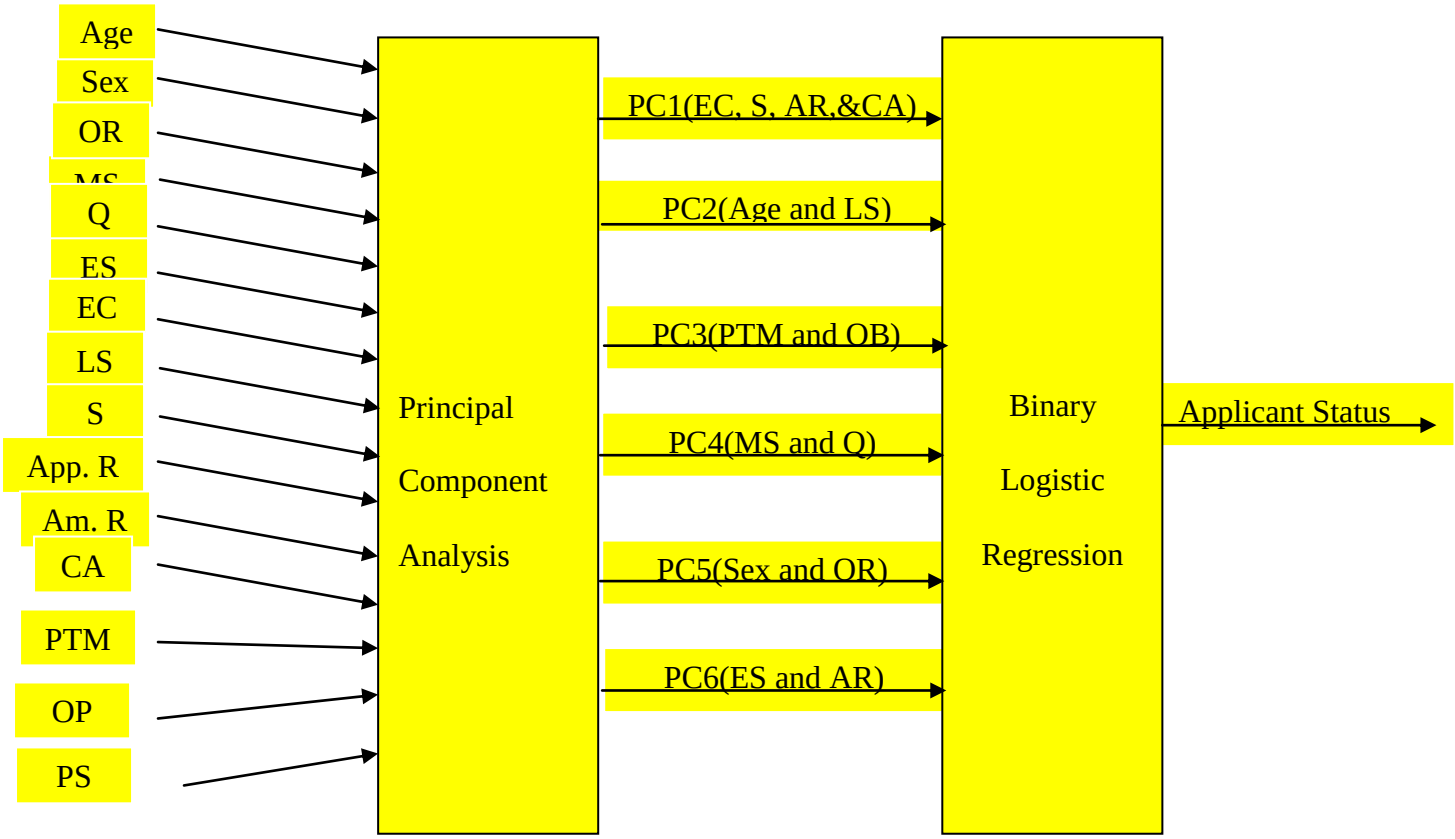


*Figure 1: Architecture of a PCR model for prediction of applicant status for new applicant (customers)*

These new components are shown in Figure 1. The first factor (i.e. Employment Classification (EC), Salary (S), Amount Request (AR), and Credit Amount (AR)) seems to all relate to Credit History parameters. Second factor (Age of applicant and length of service (LS)) is related to applicant background and credit history. Third factor (i.e. propose tenor in month (PTM) and other borrowing (OB)) is label as credit history. Fourth factor (i.e. marital status (MS) and Qualification (Q)) is label as applicant background and academic qualification. Fifth factor (i.e. sex and Ownership Residence (OR)) is label as gender and property of applicant and last factor (i.e. Employment status (ES) and Amount Request (AR)) is label as employment status and credit history.

**Table 8:** Omnibus Tests of Model Coefficients

| | | **Chi-square** | **Df** | **Sig.** | **Nagelkerke R square.** |
|---|---|---|---|---|---|
| Step 1 | Step | 79.954 | 6 | 0.000 | 0.650 |
| | Block | 79.954 | 6 | 0.000 | |
| | Model | 79.954 | 6 | 0.000 | |

The significance test for the model chi-square is the statistical evidence of the presence of a relationship between the dependent variable and the combination of the independent variables. The significant value of the test statistics shows that the there is existence of a significant relationship between the independent variables and the dependent variable. The coefficient of determinant is obtained used Nagelkerke $R^2$ (0.650)

**4.4. Importance of Independent Variables:**

Some independent variables are significantly related with the dependent variable and others are not associated strongly. The significance test is the statistical evidence of the presence of a relationship between the dependent variable and each of the independent variables. The significance test is the Wald Statistic. Here, the null hypothesis is that the b coefficient for the particular independent variable is equal to zero.

**Table 9:** Coefficients of the logistic Regression Model

| | | B | S.E. | Wald | Df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1[a] | Principal component (PC1) | -0.383 | 0.219 | 3.045 | 1 | .081 | .682 |
| | Principal component (PC2) | 1.594 | 0.289 | 30.382 | 1 | .000 | 4.922 |
| | Principal component (PC3) | -0.521 | 0.230 | 5.119 | 1 | .024 | .594 |
| | Principal component (PC4) | -0.290 | 0.220 | 1.738 | 1 | .187 | .749 |
| | Principal component (PC5) | -0.123 | 0.231 | 0.284 | 1 | .594 | .884 |
| | Principal component (PC6) | 0.356 | 0.259 | 1.892 | 1 | .169 | 1.428 |
| | Constant | -2.712 | 0.393 | 47.631 | 1 | .000 | .066 |

The independent variables with the probabilities of the Wald statistic less than or equal to the level of significance of 0.05 and 0.10 respectively hold statistically significant relationships with the dependent variable. The statistically significant independent variables are Principal Component (PC1), Principal Component 2 (PC2) and Principal Component 3(PC3). Here, the insignificant variables have probabilities of Wald statistic greater than the level of significance of 0.05.

### 4.4. Comparison of performance Between PCR and BLR

Performance indicators were used to compare between BLR and PCR for predicting applicant status in First Bank of Nigeria (FNB), plc. Table 10 shows the performance indicator values. The values of the accuracy measure are Prediction Accuracy, Coefficient of Determination, and Index of Agreement. The accuracy measure for PCR is higher than for BLR. The values of the error measures namely Normalized Absolute Error and Root Mean Square Error are smaller for PCR than for BLR. This shows PCR gives better result than MLR based on accuracy measures and error measures. So, PCR should provide a better prediction than MLR.

**Table 10:** Performance Indicator between BLR and PCR models

| Performance Indicator | BLR | PCR |
|---|---|---|
| Normalized Absolute Error (NAE) | 0.121 | 0.112 |
| Prediction Accuracy (PA) | 0.865 | 0.905 |
| Negelkerke R-square | 0.536 | 0.650 |
| Root Mean Square Error (RMSE) | 8.095 | 8.085 |
| Index of Agreement (IA) | 0.927 | 0.956 |

## 5. Findings and Conclusion

Binary logistic regression was used to predict the new applicant status (customer) using predictors' variables. Two different approaches were used, considering original data and principal component as inputs. The result showed the used of principal component as input provides a more accurate result than original data because it reduced the number of inputs and therefore decreased the model complexity. Besides that, the use of PC (principal component) based models was considered more efficient, due to elimination of co-linearity problem and reduction of the number of predictors variables. Based on eigenvalues over six factors were retained. The factors accounted for 72.4 percent of the variance. The combination of items with loadings greater than 0.40 were considered as separate between important and less important factors. However models adequacy checked by various statistical methods showed that the developed principal component binary logistic regression model can also be used for prediction of applicant (customers) status.

The quality and reliability of the developed models were evaluated via performance indicators (NAE, RMSE, PA, IA and $R^2$). Assessment of model performance indicated that principal component regression can predict an input value better than Binary logistic regression. Similar conclusions were found by previous studies [13, 30].

## References

[1]. Culp, C.L. (2001). The Risk Management Process: Business Strategy and Tactics, Wiley.

[2]. Jorion, P. (2000.) Value at Risk: The New Benchmark for Managing Financial Risk,. McGraw-Hill.
[3]. Fabozzi, F.J., Davis, H.A. and Choudhry, M. (2006). Introduction to Structured Finance, Wiley.
[4]. Jentzsch, N. (2007). Financial Privacy: An International Comparison of Credit Reporting Systems (Contributions to Economics), Springer.
[5]. Morrison, J. (2004). Introduction to survival analysis in business. *The Journal of Business Forecasting Methods & Systems* Vol. 23, No. 1, p. 18-22.
[6]. Fensterstock, A. (2005). Credit scoring and the next step, *Business Credit* Vol. 107, No. 3, p. 46-49.
[7]. Johnson-Speck, C. (2005). Abstracts of significant cases bearing on the regulation of Insurance, *Journal of Insurance Regulation* Vol. 23, No. 4, p. 81-8
[8]. Miller, M. (2003). Research confirms value of credit scoring, *National Underwriter*, Vol. 107, No. 42, p. 30.
[9]. Harrell, F. E. and Lee, K. L (1985). A Comparison of the Discrimination of Discriminant Analysis and Logistic Regression, in P. K. Se (ed.) Biostatistics: Statistics in Biomedical, Public Health, and Environmental Sciences, North-Holland, Amsterdam.
[10]. Abraham. C.R. and Zhang. M. (2009). Credit Risk Assessment: The new lending system for borrowers, lenders and investors, Wiley.
[11]. Abdul Wahab S.A., Bakheit C.S. and Al-Alawi S.M., (2005). Principal component and multiple regression analysis in modelling of ground level ozone and factors affecting its concentrations, *Environmental Modelling & Software*, No. 20, Vol. 10, p.1263–1271.
[12]. S. Wang and F. Xiao, (2004). AHU sensor fault diagnosis using principal component analysis Method, *Energy and Buildings*, No. 36, Vol. 2, p. 147–160.
[13]. Sousa, S.I.V., Martins, F.G., Alvim Ferraz M.C.M. and Pereira M.C. (2007). Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations, *Environmental Modeling & Software*, Vol. 22, p. 97–103.
[14]. Fisher, R.A. (1936). The use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenic*, No. 7, p. 179-188.
[15]. Durand, D. (1941). Risk elements in consumer installment financing. (Technical edition) By David Durand. National bureau of economic research [New York].
[16]. Thomas, L., J. Crook, and D. Edelman (2002). Credit Scoring and Its Applications. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.
[17]. Anderloni, L., Braga, M.D. and Carluccio, E.M. (2006). New Frontiers in Banking Services: Emerging Needs and Tailored Products for Untapped Markets, Springer.
[18]. Mays, E. (1998). Credit Risk Modeling: Design and Application, CRC. Mcculloch, W. and W. Pitts (1943). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biology Vol. 5, No. 4, p. 115–133.
[19]. Rosenberger, L. E., Nash, J & Graham A. (2009). The Deciding Factor: The Power of Analytics to Make Every Decision a Winner, Jossey-Bass.
[20]. Burrell, J. (2007). How to Repair Your Credit Score Now: Simple No Cost Methods You. Can Put to Use Today, Atlantic Publishing Company.
[21]. Colquitt, J. (2007). Credit Risk Management, McGraw-Hill.
[22]. Servigny, A. D. & Renault, O. (2004). The Standard & Poor's Guide to Measuring and Managing Credit Risk, McGraw-Hill.
[23]. Greenacre, M. and Blasius, J. (2006) Multiple Correspondence Analysis and Related Methods, Chapman and Hall/CRC.
[24]. Cox, D. R. and Snell, E. J (1994). Analysis of binary data. Chapman & Hall, London
[25]. Collett, D. R. (2003). Modeling Binary data, Chapman & Hall, London.
[26]. Kendall M.G (1957). A course in multivariate Analysis, London, Griffin.
[27]. Jeffers J. N.R (1967). Two case studies in the application of principal component analysis. *Applied Statistics* Vol. 16, p. 225-236.

[28]. Lam, K.C., Tau, T., M.C.K., (2010). A Material supplier selection model for property developers using fuzzy principal component analysis. *Automation in Construction*, Vol. 19, p. 608- 618.

[29]. Hussain, F.; Zubairi, Y. Z.; and Hussin, A. G, (2011). Some application of principal component analysis on Malaysian wind data. *Scientific research and essays.* Vol. 15, p. 3172-3181.

[30]. Ozbay, Bilge, Keskin, Gulsen Aydin, Dogruparnak, Senay Cetin, Ayberk, Savas (2011). Multivariate methods for groundlevel ozone modeling, Atmospheric research doi.10.1016/j.atmosres.2011.06.005