International Journal Of Mathematics And Statistics Invention (IJMSI)
E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759
Www.Ijmsi.Org || Volume 2 Issue 10 || November. 2014 || PP-05-15

Predicting an Applicant Status Using Principal Component, Discriminant and Logistic Regression Analysis

S. Suleiman[1], Issa, Suleman.[2], U.Usman[3], Salami, Y. O.[4]
[1,2&3]Usmanu Danfodiyo University, Sokoto, Nigeria, Department of Mathematics, [4]Kwara State polytechnic, Ilorin, Nigeria, Department of Statistics

ABSTRACT: *The purpose of this study was to improve the predictive power of linear Discriminant and Logistic regression models using principal components as input for predicting applicant status (i.e Creditworthy or Non- creditworthy) for new applicant (customer). The dataset contains 200 applicants and holds 15 variables altogether with 14 independent variables (input variables) and a dependent variable (output variable). Results showed that the use of principal component as inputs improved linear Discriminant and Logistics regression models prediction by reducing their complexity and eliminating data co-linearity. Based on the scree test and eigenvalues over six factors were retained. The factors accounted for 72.4 percent of the variance. The combination of items with loadings greater than 0.30 were considered as separate between important and less important factors.*

KEYWORDS*: Factor Analysis, Principal Component Analysis, Discriminant Analysis, Logistic regression, Credit Scoring*

I. INTRODUCTION

Predictive analysis encompasses a variety of statistical techniques from modeling, machine learning, data mining and game theory that analyzes current and historical facts to make predictions about future events. Predictive analytics is used in actuarial science, financial services, insurance, telecommunications, retail, travel, health care, pharmaceuticals and other fields. One of the most well-known applications is credit scoring, which is used throughout financial services. Scoring models process a customers' credit history, loan application , customers' data, etc, in order to rank-order individuals by their likelihood of making future credit payments on time. Lenders use credits to determine who qualifies for a loan, at what interest rate, and what credit limits. Lenders also use credit scores to determine which customers are likely to bring in the most revenue. One of the most significant banking problems is that of credit scoring. The credit scoring is a method used by the financial institution in order to minimize the number of defaulting customers.

1.1 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 test the homogeneity of variance among the variables using Bartlett's Test.
[2] To identify a number of factors that represent the relationship among sets of inter-related variables using principal component and factor analysis.
[3] To verify the variables that contributes significantly to the percentage of variance in the components.
[4] To build Discriminant model capable of predicting an applicant status using Principal Component (PC) as inputs.
[5] To build Binary Logistics Regression model capable of predicting an applicant status using Principal Component (PC) as inputs.
[6] To compare and contrast the predictive powers of the discriminant model and logistic regression for credit scoring.

II. MATERIAL AND METHODS

The data for this write-up was collected from a sample of 200 applicant on credit scoring, extracted from the application form of First Bank of Nigeria plc. The methods used for credit worthy are Principal Component Analysis and Linear Discriminant Analysis.

2.1 Linear Discriminant Analysis

LDA was first proposed by Fisher (1936) as a classification technique. It has been reported so far as the most commonly used technique in handling classification problems (Lee et al., 1999). In the simplest type of LDA, two-group LDA, a linear discriminant function (LDF) that passes through the centroids (geometric centres) of the two groups can be used to discriminate between the two groups. The LDF is represented by Equation (1):

$$LDF = a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \qquad (1)$$

Where a is a constant, and b_1 to b_p are the regression coefficients for p variables. LDA has been widely applied in a considerable wide range of application areas, such as business investment, bankruptcy prediction, and market segment (Lee et al., 1997; Kim et al., 2000).

2.1 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 (Cox and Snell, 1994). 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 (Collett,2003).

$$Logit(P_i) = Log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k \qquad (2)$$

Or $$P_i = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\right)} \qquad (3)$$

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, β_0, β_i are parameters to be estimated and P_i is the response probability.

In logistic model the predicted values will lies between 0 and 1 regardless of the values of the explanatory variables.

2.2 Principal Component Analysis

Principal component analysis used because 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 (1957) in his book on Multivariate Analysis. Jeffers (1967) 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 (Lam *et al*., 2010).

Hussain et. Al. (2011): 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.

Step 1: Get the data

Consider the linear combinations:

$$Y_1 = a_1 X = a_{11} X_1 + a_{12} X_2 + \cdots + a_{1p} X_p$$

$$Y_2 = a_2 X = a_{21} X_1 + a_{22} X_2 + \cdots + a_{2p} X_p$$

.

.

(3)

$$Y_p = a_p X = a_{p1} X_1 + a_{p2} X_1 + \cdots + a_{pp} X_p$$

Step 2: Standardize the data

Sometimes it makes sense to compute principal components for raw data. This is appropriate when all the variables are in the same units. Standardizing the data is often preferable when the variables are in different units or when the variance of the different columns is substantial. This can be done by subtracting the mean of each column and dividing by its standard deviation namely:

$$Z = \frac{(X - \mu_i}{\sqrt{\sigma_{1i}}}, \qquad i = 1,2, \ldots, p$$

In matrix notation, it is given by:

$Z = (V^{1/2})$-1 $(X - \mu)$

Where

$V^{1/2}$ is the diagonal standard deviation matrix. From this, we obtain mean of Z equals to zero, E(Z) = 0.

2.3 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

Decision Rule: Reject H_0 in favor of H_1 at 0.05 level of significance if p-value $\leq$ 0.05 otherwise do not reject H_0

2.4. Wilks' Lambda Test for significance of canonical correlation Hypothesis canonical correlation:

H_o: There is no linear relationship between the two sets of variables

H_1: There is linear relationship between the two set of variables

Test statistic:

$\lambda = \dfrac{|W|}{|W + H|}$, where W is residual variance

H is the variance due to linear relationship

$W + H$ is the total variance.

Decision rule: Reject H_o if P<0.05 otherwise accept H_o at the 5% level of significance

2.5 Chi-square Test

Hypothesis for Chi-square Test:

H_o: The two variables are independent

H_1: The two variables are not independent

Test statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$, where O_{ij} is the observed value and e_{ij} is the expected value.

Decision Rule:

Reject H_o if P<0.05 otherwise accept H_o at the 5% level of significance

2.6. Omnibus Chi-square Test

The omnibus Chi-square test is a log-likelihood ratio test for investigating the model 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:

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

or

$$\chi^2 = 2\left[\sum_{i=1}^{r}\sum_{j=1}^{c} o_{ij} \ln o_{ij} - \sum_{i=1}^{r} R_i \ln R_i - \sum_{j=1}^{c} C_j \ln C_j - n \ln n\right]$$

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.

2.7. Box M Test for the Equality of Covariance Matrices

Hypothesis for Box's M Test:

H_o: The two covariance matrices are equal

H_1: The two covariance matrices are not equal

Test Statistic:

$M = \frac{|S_L|}{|S_s|}$, where S_L is the larger variance and S_s is the smaller variance.

Decision Rule:

Reject H_o if P<0.05 otherwise accept H_o at the 5% level of significance.

2.8. Wald Test

The Wald test is used to test the statistical significance of each coefficient (β) 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 α level of significance.

III. DATA COLLECTION AND ANALYSIS.

The dataset contains 200 cases, 163 applicants are considered as "Creditworthy" and the rest 37 are treated as "Non-creditworthy".

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 Discriminant Analyses

Table 1: Credit Dataset Description

No.	Variable	Type	Scale	Description
1	Attribute1	Input Variable	Scale	Age of the Applicant
2	Attribut2	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.

Table 2: KMO Statistics for Sampling Adequate and Bartlett's test for Homogeneity

Test	DF	Approx. Chi-Square	P-value
Keiser-Meyer-Olkin Measure of Sampling Adequate	-	-	.566
Bartlett's Test of Sphericity	91	1464.453	0.000

Test Statistics: Bartlett's test $(\chi^2) = 1464.453\ (p - value = 0.000)$

Decision: From table 2, the p-value=0.6 is greater than level of significance (0.05) for KMO measure of sampling adequacy, we therefore fail to reject the null hypothesis. We will reject the null hypothesis for Bartlett's test of Sphericity since p-value = 0.00 < 0.05.

Conclusion: We therefore proceed to conduct factor analysis on the data set since the KMO test revealed that the sample is adequate and the Bartletlett's test revealed that the correlation matrix is not an identity matrix. In what follows we present the factor analysis.

Table 3 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 3: 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	.898	6.416	78.854			
8	.807	5.761	84.615			
9	.775	5.539	90.154			
10	.717	5.121	95.275			
11	.438	3.128	98.403			
12	.126	.901	99.305			
13	.073	.521	99.825			
14	.024	.175	100.000			

Table 4: The Coefficient of Principal Component Score of Variables

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Age	0.248	0.608	0.023	0.137	0.036	-0.032
Sex	0.003	-0.043	0.264	-0.395	0.469	0.255
Residences	-0.079	-0.019	0.370	-0.040	0.629	-0.107
Marital status	0.080	-0.066	-0.167	0.346	0.399	-0.446
Qualification	-0.148	0.107	0.518	0.075	-0.176	0.405
Employment status	-0.075	-0.064	-0.493	0.107	0.178	0.486
Employment classification	0.375	-0.125	-0.183	-0.187	0.216	-0.100
Length of service	0.239	0.606	0.029	0.174	0.044	-0.013
Salary	0.477	-0.180	0.046	-0.143	-0.085	0.077
Applicant request	-0.077	0.076	0.201	-0.432	-0.242	-0.498
Amount request	0.470	-0.153	0.166	0.109	-0.107	0.080
Credit amount	0.476	-0.242	0.185	0.091	-0.078	0.071
Propose tenor	-0.081	-0.156	0.332	0.573	0.066	-0.012
Other borrowing	0.112	0.284	-0.067	-0.258	0.160	0.216

TABLE 4: the first six principal component's scores are computed from the original data using the coefficients listed under PC1, PC2 and PC6 respectively:
PC1= 0.248Age+0.003Sex-0.079Residence+0.080Marital status-0.148Qualification-0.075Emp. Status + **0.375**Emp.classification +0.239Leg.Service+**0.477**Salary-0.077App.req+**0.477**Amt.Req. + **0.476**Credit Amount – 0.081Propose tenor + 0.112 other borrowing.
.
.
PC6= -0.032Age+0.255Sex-0.107Residence-**0.446**Marital status+**0.405**Qualification+**0.486**Emp. Status - 0.100Emp.classification -0.013Leg.Service+0.077Salary-**0.498**App.req+0.080Amt.Req. + **0.071**Credit Amount – 0.012Propose tenor + 0.216 other borrowing.
The interpretation of the principal components is subjective and requires knowledge of the data:

- Employment classification (**0.375**), Salary (**0.477**), Amount Request (**0.476**), and credit Amount (**0.476**) have large positive loadings on component 1, so label this component Employment classification and Credit History
- Employment status (**–0.493**), Sex (**–0.395**) and Marital Status (**–0.446**) have large negative loadings on components 3, 4 and 6, so label this component Applicant background.

- Age (**0.608**), Sex (**0.468**), Residences (**0.629**), Marital Status (**0.399**) and Education qualification (**0.405**) have large positive loadings on Components 2,5 and 6, so label this component Academic and Applicant background.

3.1 Discriminant Model for the data Analysis

Table 5: Test Results of Box's M

Box's M	46.498
Approx.	2.076
df1	21
df2	15565.956
Sig.	0.003

The p-value of the Box's M of 0.003 in table 5 has confirmed the equality of the covariance matrices for the two groups.

3.2 Fisher's Linear Discriminant Function for the Data

Table 6: Fisher's Classification Function Coefficients

Applicant Status	Creditworthy	Non-Creditworthy
Principal component 1(PC1)	0.053	-0.235
Principal component 2 (PC2)	-0.268	1.182
Principal component 3 (PC3)	0.089	-0.394
Principal component 4(PC4)	0.088	-0.388
Principal component 5 (PC5)	0.009	-0.038
Principal component 6 (PC6)	-0.061	0.268
Constant	-0.750	-1.788

Fisher's linear discriminant functions
The Fishers linear discriminant model for each group is computed as follows
Group 1 (Creditworthy)

$$\mathbf{Y_1} = \mathbf{X}'\mathbf{S}^{-1}\left(\overline{\mathbf{X}}_2 - \overline{\mathbf{X}}_1\right)$$

Y_1= (-0.750)+0.053PC1+(-0.268)PC2+0.089PC3+0.088PC4+0.009PC5+(-0.061)3.411PC6

Group 2 (Non-Creditworthy)

$$\mathbf{Y_2} = \mathbf{X}'\mathbf{S}^{-1}\left(\overline{\mathbf{X}}_2 - \overline{\mathbf{X}}_1\right)$$

Y_2= (-1.788)+(-0.235)PC1+1.182PC2+(-0.394)PC3+(-0.388)PC4+(-0.038)PC5+0.268PC6

3.3 Unstandardized Discriminant Function for the Data

Table 7: Unstandardized Classification Function Coefficients

	Function 1
Principal component 1(PC1)	-0.159
Principal component 2 (PC2)	0.799
Principal component 3 (PC3)	-0.266
Principal component 4(PC4)	-0.262
Principal component 5 (PC5)	-0.026
Principal component 6 (PC6)	0.181
Constant	0.000

Unstandardized coefficients

Table 8: Functions at Group Centroids

Applicant status	Function 1
Creditworthy	-0.336
Non-Creditworthy	1.480

Unstandardized canonical discriminant functions evaluated at group means
The Cut-off point ($\hat{\mathbf{M}}$) is computed as follows:

$$\therefore \hat{\mathbf{M}} = \tfrac{1}{2}\left(\hat{\mathbf{I}}_1 + \hat{\mathbf{I}}_2\right) = \tfrac{1}{2}\left(1.480 - 0.336\right) = 0.502$$

Y= (-0.000)+(-0.159)PC1+0.799PC2+(-0.266)PC3+(-0.262)PC4+(-0.026)PC5+0.181PC6

The classification rule is as follows:

Classify as Group 1 (Creditworthy) if $\mathbf{Y} \geq 0.502$

Classify as Group 2 (Non-Creditworthy) if $\mathbf{Y} < 0.502$

Table 9: Prior Probabilities for Groups

Applicant status	Prior probabilities	Cases Used in Analysis
Creditworthy	0.50	163
Non- Creditworthy	0.50	37

The table 9 above indicates the prior probability of misclassifying creditworthy to non-creditworthy is 0.5and prior probability of misclassifying Non-creditworthy to creditworthy is also 0.5

Table 10: SPSS Output: Classification Results:
Predictive Ability of the Discriminant Model

			Predicted Group Membership		
		Applicant status	Creditworthy	Non-creditworthy	Total
Original	Count	Creditworthy	130	33	163
		Non-creditworthy	7	30	37
	%	Creditworthy	79.8	20.2	100.0
		Non-creditworthy	18.9	81.1	100.0
Cross-validated[a]	Count	Creditworthy	127	36	163
		Non-creditworthy	7	30	37
	%	Creditworthy	77.9	22.1	100.0
		Non-creditworthy	18.9	81.1	100.0

a.80.0% of original grouped cases correctly classified.

b.78.5% of cross-validated grouped cases correctly classified.

In the table above shown that the Discriminant model is able to classify 130 good applicants as "Good Group" out of 163 good applicants. Thus, it holds 79.8% classification accuracy for the good group. On the other hand, the same discriminant model is able to classify 30 bad applicants as "Bad Group" out of 37 bad applicants. Thus, it holds 81.1% classification accuracy for the bad group. Thus, the model is able to generate **80.0%**classification accuracy in combined groups. This has justified an acceptable goodness of fit by the linear discriminant model.

3.4 The Logistic Regression Model

The logistic model is constructed for First Bank of Nigeria, plc using the following output results:

3.4.1 Binary Logistic Model the Analysis of Data

3.5 Measurement of Model Performance in Logistic Regression

Logistic Regression is the most important tool in the social science research for the categorical data (binary outcome) analysis and it is also becoming very popular in the business applications, for example, credit scoring (Agresti 2002). The algorithm assumes that a customer's default probability is a function of the variables (income, marital status and others) related with the default behaviour (Blattberg, Kim et al. 2008). 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 (Greenacre and Blasius 2006). 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 (Jentzsch 2007). This study is not testing for that assumption.

After the predictive model development, the most important task is to check the usefulness (utility) of the model. It can be accomplished in two ways. First one is the significance test. 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 79.954 (equivalent to significant value of 0.000), less than the level of significance of .05, which shows that the existence of a relationship between the independent variables and the dependent variable is supported. So, usefulness of the model is confirmed. The table 11 is referred for the test.

Table 11: SPSS Output: Model Test:
Omnibus Tests of Model Coefficients

		Chi-square	Df	Sig.
Step 1	Step	79.954	6	.000
	Block	79.954	6	.000
	Model	79.954	6	.000

Table 12: SPSS Output At Step 0: Logistic Classification Table:

			Predicted		
			Applicant status		Percentage Correct
	Observed		Creditworthy	Non-creditworthy	
Step 0	Applicant status	Creditworthy	163	0	100.0
		Non-creditworthy	37	0	.0
	Overall Percentage				81.5

a. Constant is included in the model.

b. The cut value is .500

Checking Usefulness of the Derived Model

Here, the following table is showing the SPSS generated classification rate that is equivalent to 91.0%. Here, it is noteworthy to mention that, after step 1 (when the independent variables are included in the model), the classification percentage rate is changed to 91.0% from 8150%.

Table 13: SPSS Output At Step 1: Logistic Classification Table:

			Predicted		
			Applicant status		Percentage Correct
	Observed		Creditworthy	Non-creditworthy	
Step 1	Applicant status	Creditworthy	160	3	98.2
		Non-creditworthy	15	22	59.5
	Overall Percentage				91.0

The cut value is .05

Checking Usefulness of the Derived Model

Table 14: Model Summary

	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
Step 1	111.603[a]	0.330	0.535

The table 14 indicates the goodness of fit test for the model.

3.6 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 15: SPSS Output: Significant Variables: Important Variables Identified By the Logistic Regression Model

Variables in the Equation

		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 .05 hold statistically significant relationships with the dependent variable. The statistically significant independent variables are Principal component (PC2) and Principal component 3 (PC3). Here, the insignificant variables have probabilities of Wald statistic greater than the level of significance of 0.05. The fitted model for logistic regression is obtained as follow:

$$p = \frac{e^{\hat{\alpha}+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\hat{\beta}_3 x+\hat{\beta}_4 x_4+\hat{\beta}_5 x_5+\hat{\beta}_6 x_6}}{1+e^{\hat{\alpha}+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\hat{\beta}_3 x+\hat{\beta}_4 x_4+\hat{\beta}_5 x_5+\hat{\beta}_6 x_6}}$$

Where $\hat{\alpha} = -2.712$

$\hat{\beta}_1 = -0.383$

$\hat{\beta}_2 = 1.594$

$\hat{\beta}_3 = -0.521$

$\hat{\beta}_4 = -0.290$

$\hat{\beta}_5 = -0.123$

$\hat{\beta}_6 = 0.356$

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6$$

$$p = \frac{e^{(-2.712)+(-0.383)PC1\ +1.594PC_2+(-0.521)PC3+(-0.290)PC_4+(-0.123)PC_5+0.356PC_6}}{1+e^{(-2.712)+(-0.383)PC1\ +1.594PC_2+(-0.521)PC3+(-0.290)PC_4+(-0.123)PC_5+0.356PC_6}}$$

Alternatively

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = e^{(-2.712)+(-0.383)PC1\ +1.594PC_2+(-0.521)PC3+(-0.290)PC_4+(-0.123)PC_5+0.356PC_6}$$

To compute estimates or forecasts, consider the logistic model as given below:

$$p = \frac{e^{(-2.712)+(-0.383)PC1\ +1.594PC_2+(-0.521)PC3+(-0.290)PC_4+(-0.123)PC_5+0.356PC_6}}{1+e^{(-2.712)+(-0.383)PC1\ +1.594PC_2+(-0.521)PC3+(-0.290)PC_4+(-0.123)PC_5+0.356PC_6}}$$

That will be used to predict the Applicant status using a cut value or threshold probability of 0.5.

IV. FINDINGS AND CONCLUSION

Discriminant model and binary logistic regression were used to classify applicant status(customer) using their characteristics such as age, sex, length of service, salary, credit amount and so on as 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. Among the objectives is to build Discriminant and logistic regression models that are capable to classify applicant status based on their PCs (principal components) and also to compare and contrast the predictive power of the discriminant model and logistic regression for applicant status. Logistic Regression and Discriminant analysis and classification were multivariate techniques employed for the analysis of the work. Box's M test and Wilk's Lambda were used to confirm the equality of the Covariance matrices and to confirm the Significance of the Canonical correlation respectively.

Appropriate predictor variables selection is one of the conditions for successful credit scoring models development. This study reviews several considerations regarding the selection of the predictor variables.The model results are comparable to those obtained using commonly used techniques like Logistic Regression and Discriminant Analysis as described in the following table:

Table 16: **Predictive Models Comparison**

Dataset					
Models	Good Accepted	Good Rejected	Bad Accepted	Bad Rejected	Success Rate
Discriminant Analysis	130	33	30	7	80.0%
Logistic Regression	160	3	22	15	91.0%

There are two noteworthy and interesting points about this table. First of all, it shows the predictive ability of each model. Here, the column 2 and 5 ("Good Accepted" and "Bad Rejected") are the applicants that are classified correctly. Moreover, the column 3 and 4 ("Good Rejected" and "Bad Accepted") are the applicants that are classified incorrectly. Furthermore, it shows that logistic Regression gives slightly better results than discriminant analysis. Secondly, the table 16 gives an idea about the cost of misclassification which assumed that a "Bad Accepted" generates much higher costs than a "Good Rejected", because there is a chance to lose the whole amount of credit while accepting a "Bad" and only losing the interest payments while rejecting a "Good". In this analysis, it is apparent that Discriminant Analysis (equals to 30) acquired much amount of cost "Bad Accepted" than Logistic Regression (equals to 22). So, logistic regression achieves less cost of misclassification.

REFERENCES

[1]. Agresti, A. (2002). Categorical Data Analysis, Wiley- Interscience Blattberg, R.C., Kim, B.D. and Neslin, S.A. (2008). Database Marketing: Analyzing and Managing Customers, Springer.

[2]. Collett, D. R. (2003). *Modeling Binary data*, Chapman & Hall,London.

[3]. Cox, D. R. and Snell, E. J (1994). *Analysis of binary data*. Chapman & Hall, London Fisher, R.A. (1936). The use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenic*, No. 7, p. 179-188.

[4]. Greenacre, M. and Blasius, J. (2006) Multiple Correspondence Analysis and Related Methods, Chapman and Hall/CRC.

[5]. Jentzsch, N. (2007). Financial Privacy: An International Comparison of Credit Reporting Systems (Contributions to Economics), Springer.

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

[7]. Jeffers J. N.R (1967). Two case studies in the applicationof principal component analysis. Applied Statistics, No. 16, p. 225-236.

[8]. Kendall M.G (1957). A course in multivariate Analysis, London, Griffin.

[9]. Kim, J. C., Kim, D. H., Kim, J. J., Ye, J. S., and Lee, H. S. (2000). Segmenting the Korean Housing Market Using Multiple Discriminant Analysis, Construction Management and Economics, p. 45-54.

[10]. 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, No. 19, p. 608-618.

[11]. Lee, T.H., and S.H. Jung (1999/2000). Forecasting creditworthiness: logistic vs artificial neural

[12]. net, *The Journal of Business Forecasting Methods & Systems*. **Vol.** 4, p. 28-30.