Credit Scoring Prediction with Logistic Classifier using Latent Components from Principal Components and Factor Analyses

M. M. Muhammed[1]; A. A. Ibrahim[2]

Department of Mathematical Sciences,
Baze University Abuja, Nigeria.
E-mail: muhammed.muhammed@bazeuniversity.edu.ng[1]

***Abstract* — This research considered developing a credit scoring model using binary logistic regression to discriminate loan applicants based on creditworthiness. The dimension of input variables was reduced to both principal components and factor clusters and the use of principal components analysis was found to reduce collinearity better than factor analysis. Hence, principal components analysis was recommended over factor analysis as a feature selection and dimensionality reduction technique in developing binary logistic regression models for credit scoring.**

Keywords - *Credit scoring, principal component analysis, factor analysis, logistic regression, dimension reduction.*

I. INTRODUCTION

Acredit score is a statistic used to determine the ability of a o repay a loan. Credit scoring is one mechanism used by lenders to evaluate risk before extending credit-to-credit applicants. The method helps to distinguish the creditworthiness of good credit applicants from bad credit applicants [12]. Credit scoring could be defined by breaking the term into two components, credit and scoring. Firstly, the word credit means "buy now, pay later". It is derived from the Latin word "credo", which means, "I believe" or "I trust in". Secondly, the word "scoring" refers to "the use of a numerical tool to rank order cases according to some real or perceived quality in order to discriminate between them and ensure objective and consistent decisions" [2].

Credit scoring is one of the most crucial processes in banks' credit management decisions. This process includes collecting, analyzing, and classifying different credit elements and variables to assess the credit decisions. The quality of bank loans is the key determinant of competition, survival, and profitability [1].

Credit scoring has attracted a lot of attention in both academic and commercial research spheres. The most common techniques used in recent times are statistical-based methods, which largely involve the use of linear discriminant analysis and binary logistic regression.

Research conducted by [13] involved the use of seven well-known feature selection methods namely; t-test, principle component analysis (PCA), factor analysis (FA), stepwise regression, Rough Set (RS), Classification and regression tree (CART), and Multivariate adaptive regression splines (MARS) for credit scoring. Support vector machine (SVM) was a classification technique adopted. The conclusion that CART and MARS methods outperform the other methods by the overall accuracy and type I error and type II error was reached.

In 2014, [10] used a credit applicant's data set to assess the predictive power of linear Discriminant and Logistic regression models using principal components as input for predicting applicant status. The results showed that the use of principal components as inputs improved linear Discriminant and Logistics regression model prediction by reducing their complexity and eliminating data collinearity. The research also showed that the Logistic model 91% performed slightly better than the Discriminant model 80%.

Another study on ways to improve the predictive power of binary logistic Regression models using principal components as input for predicting applicant status was conducted. The study indicated that the use of principal components as inputs improved Binary logistic regression model prediction by reducing their complexity and eliminating data collinearity [11].

Summarily, it has been understood that the accuracy of feature selection methods and

classification techniques depend on the nature of the data. This research therefore seeks to compare the accuracy of principal component analysis and factor analysis as feature selection methods in developing a credit scoring model using Binary Logistic Regression.

II. Research Methodology

The data used in this research was obtained secondarily from a commercial bank in Nigeria (name withheld). The dataset is referred to as the Credit scoring dataset. The data type is a big data set containing 111,107 observations with 9 independent variables and one response variable; hence, Python programming language was used in the analysis of the data. Out of the 111,107 loan applicants, 25,173 were found to be non-creditworthy while 85,934 were recorded as creditworthy. The credit scoring dataset is hereby described below:

Table 1: Credit Scoring Dataset Description

S/N	Variable	Type	Scale
1.	Loan Status	Output	Nominal
2.	Current Loan Amount	Input	Numeric
3.	Credit Score	Input	Numeric
4.	Annual Income	Input	Numeric
5.	Years of credit history	Input	Numeric
6.	Number of open accounts	Input	Numeric
7.	Number of credit problems	Input	Numeric
8.	Current credit balance	Input	Numeric
9.	Bankruptcies	Input	Numeric
10.	Tax Liens	Input	Numeric

The nature of the expected outcome is binary thus; a binary logistic regression will be used to classify loan applicants based on creditworthiness. Since, the predictors are numerous, Principal Components Analysis (PCA) and Factor Analysis (FA) will be used to reduce the dimension and collinearity of the data before the implementation of the Binary Logistic Regression (BLR). The machine learning techniques used in the research are further explained below.

A. *Principal Components Analysis (PCA)*

Principal Components Analysis (PCA) is a data processing technique that allows the data to be transformed from a high-dimensional space to a lower-dimensional space in such a way that information about the data is not lost. Principal components analysis can be used to compress data, visualize high-dimensional data and also speed up machine learning algorithms. So, PCA reduces the dimension of the data by producing the principal components (PCs) which can be used to represent the data more effectively in a machine learning algorithm. The technique achieves this by creating a fewer number of variables which explains most of the variation in the original variables. The new variables created are linear combinations of the original variables. The variance for each principal component is given by the eigenvalue of the corresponding eigenvector. The components are ordered such that the first component (PC1) explains the largest possible amount of variation in the original data, subject to the constraint that the sum of the squared weights is equal to one. The eigenvalues equal to the number of variables in the initial data set. The second component (PC2) is completely uncorrelated with the first component and explains additional but less variation than the first component, subject to the same constraint [9].

B. *Factor Analysis (FA):*

Factor analysis operates on the notion that measurable and observable variables can be reduced to fewer latent variables that share a common variance and are unobservable, which is known as reducing dimensionality [4]. Large datasets that consist of several variables can be reduced by observing 'groups' of variables (i.e., factors) – that is, factor analysis assembles common variables into descriptive categories.

C. *Keiser Meyer Olkin (KMO) and Bartlett Test of Sphericity:*

The KMO Test recommends accepting values greater than 0.5 [10].
H_{01}: The sampled data is adequate for the study
H_1: The sampled data is not adequate for the study.
H_{02}: $\delta 1=\delta 2=\cdots=\delta k$
H_1: $\delta i \neq \delta k \; for \; at \; least \; one \; pair \; (i,j)$
Test Statistics: KMO
Interpretation rule: Values above 0.5 are acceptable and values below 0.5 are unacceptable

D. *Bartlett Test of Sphericity:*

Bartlett's sphericity test [3] is used to test that the null hypothesis of the correlation matrix is an identity matrix (all correlations are zero).
Decision: If the P-value > 0.05, dimension reduction should not be performed on the dataset..

E. *Scree Plot:*

The eigenvalues for successive factors can be displayed in a simple line plot. [5] proposed that this scree plot can be used to graphically determine the

optimal number of factors to retain. The components with eigenvalues greater than or equal to; $\lambda \geq 1$ are retained.

F. *Binary Logistic Regression (BLR)*

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. Consider a data set with binary response variable, where the response default falls into one of two categories, "Yes" or "No", or in the case of likely default or non-default borrower, i.e. Y = (default, Non-default). Rather than modelling this response Y directly, Logistic regression models the probability that Y belongs to a particular category. It is a generalized linear model. As it defaults in probability, it can be directly used for credit scoring and rating.

Logistic Regression Model:

The Logistic regression model for the dependence of Pi(response probability) on the values of *n* explanatory variables, $X_1, X_2, \ldots X_n$ is: [6]

$$logit(Pi) = log\left(\frac{Pi}{1-Pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

or

$$Pi = \frac{exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}{1 + exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}$$

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

Here, $\left(\frac{Pi}{1-Pi}\right)$ is the ratio of the probability of a failure and called odds $\beta_0, \beta_{i's}$ are parameters to be estimated and Pi is the response probability.

Assumptions of Binary Logistic Regression

1. The response variable must be binary
2. The relationship between the response variable and the independent variable does not assume a linear relationship.
3. Large sample size are usually required.
4. There must be little or no multicollinearity.
5. The categories must be mutually exclusive and exhaustive.

G. *Confusion Matrix:*

A confusion matrix contains information about actual and predicted classifications done by a classification system, which in this course of research is the logistic classifier. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two-class classifier. [8]

Table 2: Confusion Matrix Description Table

		Predicted	
		Negative	Positive
Actual	Negative	0 (a)	0 (b)
	Positive	0 (c)	1 (d)

The entries in the confusion matrix have the following meaning in the context of our study:

- a is the number of correct predictions that an instance is negative,
- b is the number of incorrect predictions that an instance is positive,
- c is the number of incorrect of predictions that an instance negative, and
- d is the number of correct predictions that an instance is positive

H. *Binary Logistic Model the Analysis of Data*

Table 3: Response Variable Encoding

Original Value	**Internal Value**
Non-Creditworthy	0
Creditworthy	1

The table above shows that the categorical (response) variable was coded using binary digits. The non-creditworthy and creditworthy are coded with 0 and 1 respectively in the course of analysis.

III. Results And Discussion

A. Results

To develop the model, the dataset was split into a training set of 80% which accounted for 88,885 observations and a test set of 20% which accounted for 22,222 observations. The model was developed from the training set and the prediction was performed on the test set. Binary logistic regression was done on the original data set first before reducing the dimension of the data to enable comparison of logistic regression output of the data with reduced dimension and that of the original dataset.

Principal Components Analysis (PCA): Python Output

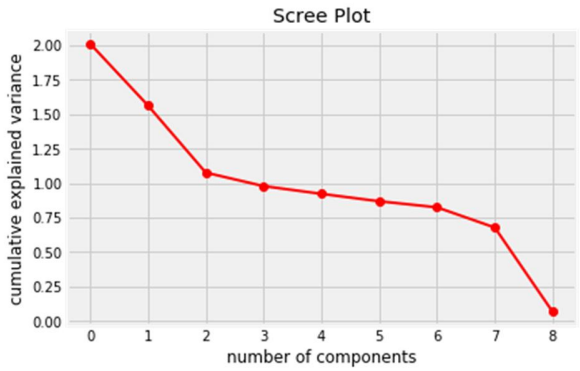

Figure 1: Scree plot showing the Eigenvalues versus the number of components.

The figure above shows the eigenvalues against the number of components explaining the variances in the dataset. The components with eigenvalues $\lambda \geq 1$ were retained. Thus, four (4) components were retained.

Table 4: Logistic Regression Model of the Principal Components

```
Optimization terminated successfully.
         Current function value: 0.440304
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:            Loan_status2   No. Observations:                88885
Model:                           Logit   Df Residuals:                    88881
Method:                            MLE   Df Model:                            3
Date:                 Mon, 17 Jun 2019   Pseudo R-squ.:                 0.1770
Time:                         21:09:11   Log-Likelihood:                -39136.
Converged:                        True   LL-Null:                       -47554.
                                         LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
PC1         6.504e-08   1.93e-09     33.623      0.000    6.12e-08    6.88e-08
PC2         5.984e-06   2.22e-07     26.913      0.000    5.55e-06    6.42e-06
PC3        -3.355e-06   4.95e-07     -6.781      0.000   -4.32e-06   -2.38e-06
PC4           -0.0062   7.32e-05    -84.782      0.000      -0.006      -0.006
==============================================================================
```

The results in Table 1 showed that the Log-Likelihood Ratio test has a p-value < 0.05 which shows that the predictors are significant in explaining the variation in the outcome. The p-values of all the independent variables are statistically significant since the p-values < 0.5. The intercept of the logistic regression model is 1.0462. The coefficients and the intercept can therefore be used to specify the logistic regression model. The insignificant variables will not be used in forming the model.

Logistic Regression Model Result (Build from Principal Components)

The required model for the significant predictor variables as follows:

The probability of a bad applicant is obtained by applying the transformation:

$$Pi = \frac{e^{1.0462+6.5046e^{-8}\,(PC1)+5.984e^{-6}(PC2)-3.355e^{-6}(PC3)-0.0062(PC4)}}{1+e^{1.0462+6.5046e^{-8}\,(PC1)+5.984e^{-6}(PC2)-3.355e^{-6}(PC3)-0.0062(PC4)}}$$

where PC1 = Principal Component 1, PC2 = Principal Component 2, PC3 = Principal Component 3, PC4=Principal Component 4.

Classification Accuracy: Confusion Matrix

The accuracy of a logistic regression model as a classifier can be measured using a confusion matrix. The test set of the data was used to make predictions on the training set of the data, the classification of the 22,222 observations in the test set is represented in a confusion matrix.

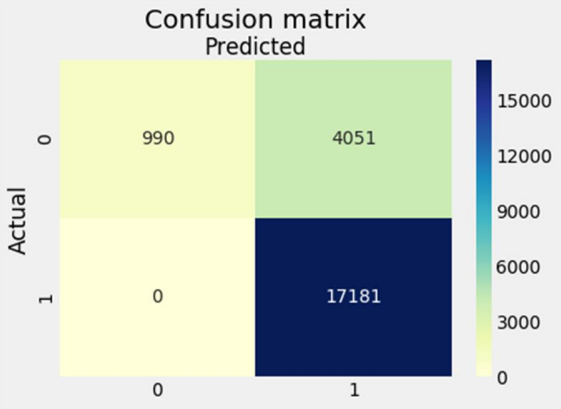

Figure 2: Confusion Matrix

The confusion matrix shows that (990+17,181=18,171) observations were correctly classified and (0+4,051= 4,051) observations were misclassified.

- True Positive (1,1): 17,181 observations were predicted to be creditworthy and turned out to be creditworthy
- True Negative (0,1): 0 observations were predicted to be non-creditworthy and turned out to be non-creditworthy.
- False Positive (1,0): 4051 observation were predicted to be creditworthy but turned out to be non-creditworthy
- False Negative (0,0): 990 observations were predicted to be non-creditworthy and turned out to be non-creditworthy

Table 5: Evaluation Metrics (in weighted average)

Precision	Recall	F1 Score	Accuracy
85%	82%	77%	82%

Receiver Operating Characteristic (ROC)

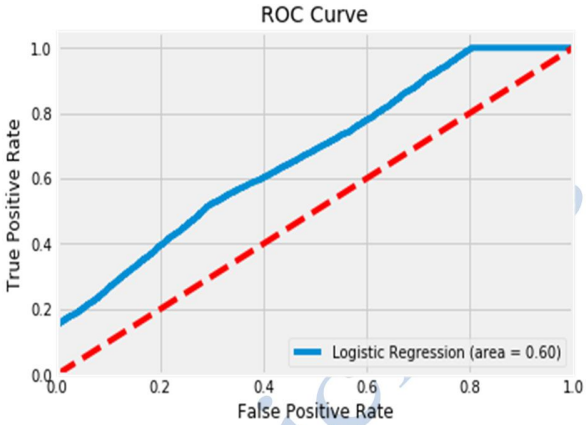

Figure 3: ROC Curve of Logistic Regression (After PCA)

The ROC Curve explains the goodness of a classifier in discriminating binary responses. Since the curve is far left of the center diagonal with an area of 60%, the logistic regression is, therefore, a good classifier for the data.

Factor Analysis (FA): Tests of Data Adequacy

Table 6: Keiser-Meyer-Olkin Measure and Bartletts Test of Sphericity Tests

Test	**P-value**
Keiser-Meyer-Olkin Measure (KMO) of Sampling Adequate	0.535
Bartlett's Test of Sphericity	0.000

The Keiser-Meyer-Olkin Measure (KMO) test of Sampling Adequate recommends accepting a data with p-value>0.5 while the Bartlett's Test of Sphericity shows that the data is appropriate for Factor Analysis since it has a p-value <0.05.

Dimension Reduction Using Factor Analysis

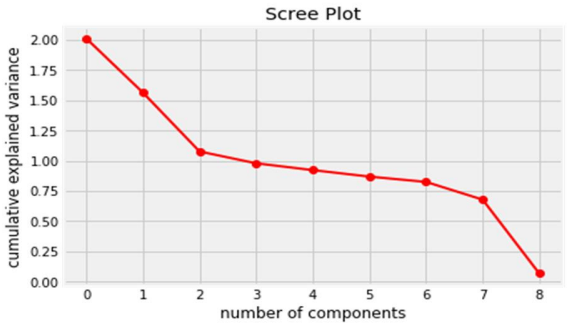

Figure 4: Scree plot showing the Eigenvalues versus the number of components.

The figure above shows the eigen values against the number of components explaining the variances in the dataset. The components with eigenvalues $\lambda \geq 1$ were retained. Thus, four (4) components were retained.

Table 7: Logistic Regression Output of the Factor Clusters

```
Optimization terminated successfully.
         Current function value: 0.439761
         Iterations 11
                           Logit Regression Results
==============================================================================
Dep. Variable:           Loan_status2   No. Observations:                88885
Model:                          Logit   Df Residuals:                    88881
Method:                           MLE   Df Model:                            3
Date:                Tue, 18 Jun 2019   Pseudo R-squ.:                  0.1765
Time:                        00:15:44   Log-Likelihood:                -39088.
Converged:                       True   LL-Null:                       -47468.
                                        LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
FC1            2.0704      0.061     33.739      0.000       1.950       2.191
FC2            0.3057      0.011     27.846      0.000       0.284       0.327
FC3           -0.0551      0.009     -5.868      0.000      -0.074      -0.037
FC4           -8.2585      0.097    -85.194      0.000      -8.448      -8.068
==============================================================================
```

The above shows the Log-Likelihood Ratio test has a p-value<0.05 which shows that the predictors are significant in explaining the variation in the outcome. The p-values of all the independent variables are statistically significant since the p-values < 0.5. The intercept of the logistic regression model is 1.8394. The coefficients and the intercept can therefore be used to specify the logistic regression model. The insignificant variables will not be used in forming the model.

Logistic Regression Model Result

The required model for the significant predictor variables as follows:

$$log_e\left(\frac{Pi}{1-Pi}\right) = 1.8394 + 2.0704\,(FC1) + 0.3057(FC2) - 0.0551(FC3) - 8.2585(FC4)$$

To estimate odds, the model is exponential as:

$$\left(\frac{Pi}{1-Pi}\right) = e^{1.8394+2.0704\,(FC1)+0.3057(FC2)-0.0551(FC3)-8.2585(FC4)}$$

The probability of bad applicant is obtained by applying transformation.

$$Pi = \frac{e^{1.8394+2.0704\,(FC1)+0.3057(FC2)-0.0551(FC3)-8.2585(FC4)}}{1+e^{1.8394+2.0704\,(FC1)+0.3057(FC2)-0.0551(FC3)-8.2585(FC4)}}$$

where FC1 = Factor Cluster 1, FC2 = Factor Cluster 2, FC3 = Factor Cluster 3, FC4 = Factor Cluster 4.

Classification Accuracy: Confusion Matrix

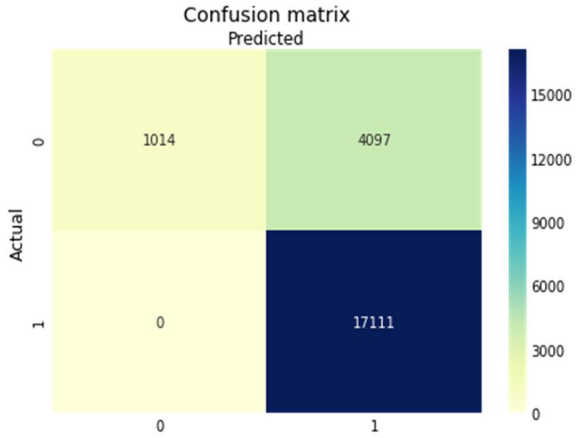

Figure 5: Confusion Matrix

The confusion matrix shows that (1014+17,111=18,125) observations were correctly classified and (0+4,097= 4,097) observations were misclassified.

- True Positive (1,1): 17,111 observations were predicted to be creditworthy and turned out to be creditworthy.
- True Negative (0,1): 0 observations were predicted to be non-creditworthy and turned out to be non-creditworthy.
- False Positive (1,0): 4097 observations were predicted to be creditworthy but turned out to be non-creditworthy.
- False Negative (0,0): 1014 observations were predicted to be non-creditworthy and turned out to be non-creditworthy.

Receiver Operating Characteristic (ROC)

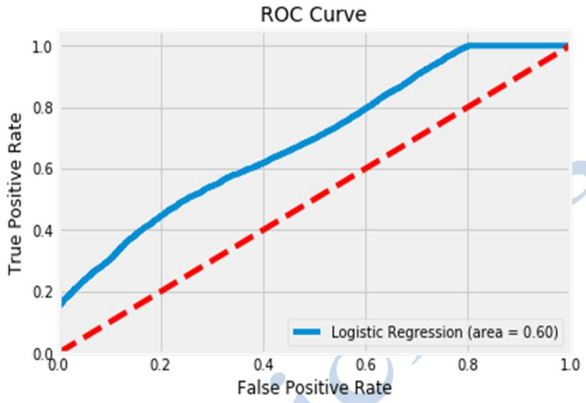

Figure 6: ROC Curve of Logistic Regression (After FA).

The ROC Curve explains the goodness of a classifier in discriminating binary responses. Since the curve is far left of the centre diagonal with an area of 60%, the logistic regression is, therefore, a good classifier for the data.

Table 8: Evaluation Metrics (in weighted average)

Precision	Recall	F1 Score	Accuracy
85%	82%	76%	82%

B. Discussion

Principal Component Analysis (PCA) and Factor Analysis (FA) were used to reduce the dimensionality of the data, each resulting data was applied on a Logistic Regression algorithm. Both PCA and FA outputs retained four components with eigenvalues greater than or equal to one; $\lambda \geq 1$. The PCA and FA both revealed an accuracy of 85% with varying F1 scores of 77% and 76% respectively. The ROC Curves of both PCA and FA reduced-data both had an area of 60% showing that Logistic Regression is a good classifier.

IV. Conclusion

The study revealed that both Principal Components Analysis and Factor Analysis performed well in reducing the data dimension but the PCA reduced-data had 4,051 misclassified observations while the FA reduced-data had 4097 misclassified observations out of the 22,222 test data. This shows that the PCA reduced-data performed more in classifying correctly than the FA reduced-data. Therefore, Principal Components Analysis is recommended for better classification accuracy over Factor Analysis.

Acknowledgment

The authors are grateful to anonymous reviewers for their valuable comments on the original draft of this manuscript.

References

[1] Abdou, H., Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature ', Intelligent Systems in Accounting, Finance & Management, 18 (2-3), pp. 59-88.

[2] Anderson, R. (2007). The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. New York: Oxford University Press.

[3] Bartlett, M. S. Tests of significance in factor analysis. The British Journal of Psychology, 1950, 3 (Part II), 77-85.

[4] Bartholomew, D., Knotts, M., & Moustaki, I. (2011). Latent variable models and factor analysis: A unified approach. (3rd ed.). West Sussex, UK: John Wiley & Sons.

[5] Cattell, R. B. (1966). The Scree Plot Test for the Number of Factors. Multivariate Behavioral Research, 1, 140-161.

[6] D. Collet. (2003). Modelling Binary Data. 2nd edition. New York: Chapman & Hall/RC.

[7] Kaiser, H. F. (1974). An index of factorial simplicity. Psychometrika, 39, 31–36.

[8] Kohavi, R. and Provost, F. (1998) Glossary of terms. Machine Learning—Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. Machine Learning, 30, 271-274

[9] Seema, V. and Lilani, K (2006). Constructing Socio-Economic Status Indices: How to Use Principal Component Analysis. Health Policy Plan, Oxford University Press, 21(6):459-68.

[10] Suleiman, S. Issa, Suleiman, U. Usman (2014). Predicting an Applicant Status Using Principal Component, Discriminant and Logistic Regression Analysis. International Journal of Mathematics and Statistics Invention (IJMSI), 2(10), pp.05-*15*.

[11] Suleiman S., M.S. Burodo, Issa Suleman. (2017). Credit Scoring using Principal Components Analysis-based Binary Logistic Regression**.** Journal of Scientific and Engineering Research. Vol. 4(12):99-110. ISSN: 2394-2630

[12] Sylvester W. W., Richard R., Calvin O. (2017). International Journal of Computer. Vol. 1, pp84-102

[13] Yao, P. (2009). Hybrid Fuzzy SVM Model Using CART and MARS for Credit Scoring. In: Proceedings of the International Conference on Intelligent Human-Machine Systems and Cybernetics, Vol. 2, Hangzhou, China, pp 392–395.

.
.