International Journal of Innovative Mathematics, Statistics & Energy Policies
10(2):1-9, April-June, 2022

© SEAHI PUBLICATIONS, 2022 www.seahipaj.org ISSN: 2467-852X

A Comparison of Logistic Regression and Discriminant Analysis in Predicting Student's Academic Outcome.

F. Meka[1] & R.N. Nwaka[2]

[1]Department of Mathematics and Statistics, University of Delta, Agbor, Nigeria.
Email:fortune.meka@unidel.edu.ng. Phone +2348037421262

[2]Department of Mathematics and Statistics, University of Delta, Agbor, Nigeria.
Email:rita.nwaka@unidel.edu.ng. Phone +2348035397244

ABSTRACT
The need to predict the academic outcome of new intakes in institutions of learning may arise at sometimes. The most widely used statistical methods for prediction when categorical outcome variables are involved includes Linear Discriminant Analysis and Logistic Regression. The question of which classifies better now comes to mind. Data of students who performed academically above average and that of those that performed academically below average was retrieved from their admission forms and analyzed with both Logistic Regression and Discriminant Analysis. The contribution of this work is two-fold. First it compares Linear Discriminant Analysis and Logistic Regression on academic outcome variables in institutions of learning revealing a higher predictive accuracy in Logistic Regression than in Discriminant Analysis. Secondly, the study revealed that the marital status of parents and mother's occupation are key pointers to the academic achievement of learners.
Keywords: Linear Discriminant Analysis, Logistic Regression, Academic outcome, Prediction.

1. INTRODUCTION

The choice of statistical approaches to be employed in researches must be carefully made. This is to avoid misleading conclusions that may be due to wrong choice of approach. Some researches can be simultaneously carried out using more than one statistical approach. In such cases, the best choice to make becomes an issue of concern. In the light of the above, Gareth et al (2013) noted that it is important for researchers to be aware of the major differences between possible statistical modeling approaches that could be applied simultaneously. Azad (2022) stressed that the researcher should have clear idea of the variables that will be used in the research work, whether they are categorical or nominal, ordinal, or rank-ordered, interval, or ratio-level. Logistic regression and linear discriminant analysis can be used to predict the probability of a specified categorical outcome using several explanatory variables. Particularly, logistic regression allows predicting an outcome, which may be discrete, continuous, dichotomous, or a mix. Similarly, discriminant analysis aims to predict membership in two or more mutually exclusive groups from a set of predictors, when there is no necessarily natural ordering on the groups. Discriminant analysis is based on the estimation of orthogonal discriminant functions that are linear combinations of the standardized independent variables, which yield the biggest mean differences between the groups. Thus, it could be suggested that discriminant analysis and logistic regression can be used to address the same types of research question. Based on a set of measurements of a student, a classification model predicts the outcome class of that student. These models are created with a learning set of data where the

outcomes of the students are already known. Often different populations share similar characteristics. This makes it difficult to separate them and a student may be assigned to the wrong class. A good discrimination and classification procedure should result in few misclassifications.

The logistic distribution has many good attributes. It is bounded by zero and one, which is necessary to represent probabilities. Also, the distribution is in the shape of an "S". This indicates that small differences at the extreme values of the predictor variable do not influence the outcome nearly as much as differences around the center (Dey and Astin, 1993). For example, it might not make much of a difference in a student's probability of dropping out if his high school grade point average was a 2.0 or a 2.5, nor if his high school grade point average was a 3.5 or a 4.0. However, there may be a large difference in the probability of a student persisting depending if his high school grade point average was a 2.5 or a 3.0 (Julie,2000). This leads to the logistic distribution's ability to separate and predict binary outcomes. The upper portion of the "S" represents high probabilities of the event occurring and the lower portion of the "S" represents low probabilities of the same event occurring. These two portions determine the two outcomes. The difficulty lies in deciding where to cut the "S" and separate the two outcomes (Dey and Astin, 1993).

Discriminant analysis is a parametric method that works on the assumption that the predictor variables for the different classes are multivariate normal. This implies that the measurements taken on the objects cluster around their class mean vector. When a new observation comes along, the multivariate normal distribution can be used to find the "distance" from the new observation to each of the class mean vectors, or the multivariate normal distribution can be used to find the probability of the new observation belonging to each of the different classes. The new observation is then assigned to a class depending on which class mean vector is the closest or which class yields the highest membership probability.

Both the logistic regression and the discriminant analysis can be used for prediction purposes. The question of which one has a better predictive accuracy, or which one gives a higher percentage in prediction of student's academic outcome is of interest in this study. The rest of this article are structured as follows: First, the extant literatures on relevant pointers to academic outcome are reviewed alongside the implications of logistic regression and discriminant analysis. This is followed by a description of the research materials and methods used in the study. The results of our enquiry are then discussed. Finally, the concluding remarks are presented.

2. Literature Review

2.1 Introduction

Admissions processes in institutions of learning today often banks on the ability to predict student success. It is usually done through the conduction of tests. However, the use of a test to help determine admission has traditionally been problematic and continues to be so. We dispute the notion that merit is identical to performance on standardized tests. Such tests do not fulfill their stated function. They do not reliably identify those applicants who will succeed in college or later in life, nor do they consistently predict those who are most likely to perform well in the jobs they will occupy. As an alternative to standardized tests, some colleges rely on two tests as a means of using multiple criteria, but if the two tests are highly correlated with each other, there is needless duplication in measuring the same aspect of a construct. Because the use of standardized tests has been shown to be problematic, multiple selection methods are being used to predict student success such as the case in most Nigerian universities in recent time whereby students have to take part in the UTME exams and Post-UTME exams and in more recent time have their WAEC examination grades considered. It is therefore pertinent to consider the above, and also other variables that could stand as pointers to academic outcome of students for admission in schools.

2.2 Predictors of Academic Success

Both logistic regression and discriminant analysis builds a predictive model for group membership. The predictor variables provide the best discrimination between groups. For instance, Matthews (1996) revealed that the interaction of learning style, race, and gender could be utilized to predict students' retention in postsecondary institutions. However, there is shortage of data that describes the relationship

between post primary admission criteria and academic performance. By analyzing the admission criteria of groups of students who have been successful against groups who have not, the possibility exists to classify subsequent applicants for retention purposes based upon an analysis of admission criteria. Consequently, what are the best predictors of students' academic performance and retention; and further, are there emerging trends or patterns within such predictor variables? Possessing this knowledge could provide guidance and counsellors with the necessary information that may assist students who are academically below average in order to ensure better performance.

2.2 Logistic Regression and Discriminant Analysis of Academic Success

Researches comparing logistic regression and discriminant analysis has been quite few especially with reference to specific subject matters such as is carried out in this study. Ali et al (1992) used linear discriminant functions in their study to see what admission criteria could help predict student success at Beirut University College (BUC) in Lebanon. BUC had the problem of having far more applicants than space for these aspiring students as is the case in most universities today. Not only had the number of applicants to BUC increased, but also the number of students who were on academic probation had increased. They developed three different linear discriminant models for each of the divisions at the school: business, natural sciences, and humanities. They were satisfied with the predictive ability of all three discriminant models for each academic division. Each model had slightly different predictive variables. Some of the variables considered for the research were: Score on college entrance exam, high school grade point average, type of high school and sex. Another comparison of logistic regression and linear discriminant analysis was carried out by Dey and Astin (1993). They used logistic and linear regression to predict whether first-time, full-time community college freshmen who intend to earn a two-year degree would graduate on time. They also tried predicting student's outcome such as completing two years of college, or being enrolled for a third consecutive semester upon admission. They used predictor variables that "were shown to predict retention among students at four-year colleges and universities". These predictor variables included students' concern about ability to finance their education, their motives for attending college, how many hours they spent per week at various activities their first year, and their high school grade point average. **T**heir results did not reveal any important differences between logistic and linear regression. Each of the techniques had similar classification accuracy as well. In another study done by Meshbane and Morris (1996), the predictive accuracy of logistic regression and linear discriminant analysis were compared. In their presentation, they listed the many conflicting reports about which classification method works better for nonnormal predictors and for small sample sizes. They concluded that there is no specific type of data set that favors logistic regression or linear discriminant analysis. Instead the classification accuracy of both logistic regression and linear discriminant analysis should be carefully compared to determine which may provide a better model. Lim et al (1999) compared thirty-three classification algorithms including logistic regression and discriminant analyses with various data sets. Amongst the algorithms considered in their study, logistic regression and linear discriminant analysis performed exceptionally well at correctly predicting class outcome.

3. MATERIALS AND METHODS

3.1 Data Collection

The data for this study was obtained from secondary schools in Delta State, Nigeria. The cumulative performance of 40 students of which 20 were academically above average and 20 academically below average was traced for their first 3 years in the schools (Junior Secondary). Information about seven outcome variables were sought from the admission application forms of the 40 students that made up the two groups for the study. The variables are as follows:

- FSLC grade
- Type of primary school attended (public or private)
- Age
- Marital status of parents
- Father's occupation

- Mother's occupation and
- Sex

The responses were categorical and were grouped into two using the binary code 0 and 1. This was categorized as follows:

- FSLC grade (below Merit-0, above Merit-1)
- Type of primary school (public-0, private-1)
- Age (below 10years-0, above 10years-1)
- Marital status of parents (separated-0, together-1)
- Father's occupation (Not white cholar-0, white cholar-1)
- Mother's occupation (Not white cholar-0, white cholar-1)
- Sex (Female-0, male-1)

Logistic Regression

The logistic regression model is based upon the assumption that the probability that an object belongs to a given class follows the logistic distribution. Once this assumption has been made all that is left to construct the logistic model is to estimate the parameters using the method of maximum likelihood. The logistic distribution is given by:

$$p(y_i = 1/x_i) = \frac{1}{1+e^{-x_i^T\beta}} \quad (1)$$

Where $x_i^T\beta = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_k$

Thus, the likelihood function for the logistic distribution is:

$$L(X,\hat{\beta}) = \prod_{i=1}^{n} p(y_i = 1/x_i)$$

$$= \prod_{i=1}^{n} \left(\frac{1}{1+e^{-x_i^T\hat{\beta}}}\right) \quad (2)$$

The $\hat{\beta}$ that produces the maximum likelihood becomes the estimate used in the logistic model. In order to make the likelihood function easier to manipulate the natural logarithm of it is taken. This result is called the log likelihood. Since the natural logarithm is a monotonically increasing function, the $\hat{\beta}$ that produces the maximum log likelihood will also be the $\hat{\beta}$ that produces the maximum likelihood. Therefore, finding the estimates for the coefficients for the logistic distribution all boils down to finding $\hat{\beta}$ such that $\log\{L(X,\hat{\beta})\}$is a maximum. This is found by numerical methods. Once $\hat{\beta}$ is found, the logistic distribution is complete, but the classification rule that assigns a student to class 1 or class 0 must still be formulated. This rule is found by determining a "cut-off" probability. Any student whose probability of belonging to class 1 is higher than or equal to the cut-off probability is assigned to class 1, otherwise the student is assigned to class 0. The value that produced the most overall correct predictions in the learning sample is chosen to be the cut-off probability.

3.2 Discriminant Analysis

Linear Discriminant analysis (LDA) was the second classification model used in this study. LDA is one of notable subspace transformation methods for dimensionality reduction (Park and Park 2008). LDA encodes discriminant information by maximizing the between-class scatter, and meanwhile minimizing the within-class scatter in the projected subspace. To illustrate the steps in carrying out an LDA: consider a classification problem involving g groups, each group has n_i m-dimensional samples ($i = 1,2,\ldots,g$).

If X is a matrix whose entries is made up of the samples in each group, the within group scatter matrix is given by

$$S_w = \sum_i^n (x_i - \bar{x}_i)(x_i - \bar{x}_i)^T \quad (3)$$

Also the between group scatter matrix is given by

$$S_b = \sum_i^n (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \quad (4)$$

Linear discriminant analysis (LDA) further assumes that the covariance matrices of the different populations are equal. Linear discriminant analysis can be used to determine which variable discriminates between two or more classes, and to derive a classification model for predicting the group membership of new observations (Worth and Cronin, 2003). The simplest LDA has two groups. To discriminate between them, a linear discriminant function that passes through the centroids of the two groups can be used. When dealing with three or more groups, the linear combination of the discriminating variables, X_s for the ith individual $(i = 1,2,\dots,n_g)$ of group g $(g = 1,2,\dots,G)$ may be written as

$$Z_{ig} = b_1 X_{i1g} + b_2 X_{i2g} + \cdots + b_p X_{ipg} \quad (5)$$

Where Z_{ig} is the LDF score for the ith individual in group g, b_i are the raw weights and X_i are the outcome (or discriminating) variables (Osemwenkhae et al, 2019).

In LDA, a projection matrix $V = (v_1, v_2, \dots, v_r) \in \mathbb{R}^{d \times r}$ that consists of discriminant vectors $V \in \mathbb{R}^d$ is obtainable by solving the following problem

$$\max_V \frac{tr(V^T S_b V)}{V^T S_w V} \quad (6)$$

This is solved through the generalized eigenvalue problem

$$S_b V = \lambda S_w V \quad (7)$$

(See Chun 2019 for clarifications)

4. ANALYSIS OF DATA AND DISCUSSION OF RESULTS

The analysis of data collected for this study is carried out using SPSS. The results are presented below

4.1 Predicting Academic Success with Discriminant Analysis

The following are the results obtained using discriminant analysis

4.1.1 Box's Test of Equality of Covariance Matrices

Table 4.1 below gives the status of the covariance matrices for the two groups

Table 4.1 Box's Test

Results

Box's M		3.946
F	Approx.	3.848
	df1	1
	df2	4.332E3
	Sig.	.050

From the table, the significance value of 0.05 reveals that the covariance matrices for the two groups are equal.

4.1.2 Stepwise Statistics

Table 4.2 Variables Entered/Removed

Step	Entered	Wilks' Lambda							
						Exact F			
		Statistic	df1	df2	df3	Statistic	df1	df2	Sig.
1	Marital status of parents	.677	1	1	38.000	18.102	1	38.000	.000

At each step, the variable that minimizes the overall Wilks' Lambda is entered. From the table, the variable that minimizes the overall Wilks Lambda is "marital status of parents". This indicates that marital status of parents of students is the major discriminatory variable between the two groups of students.

Table 4.3 Structure Matrix

	Function
	1
Marital status of parents	1.000
Mothers Occupation	-.305
Sex	.278
Typy of primary school attended	.150
Father's Occupation	-.144
First school leaving certificate grade	.128
Age	.044

From the structure matrix above, the discriminant function is a measure of marital status of parents. This explains that the academic outcome of students from broken homes is quite different from that of students who are not from broken homes. Other variables do not account much for the difference between the two groups.

Table 4.4 Classification Results

			Predicted Group Membership		
		group	1	2	Total
Original	Count	1	18	2	20
		2	7	13	20
	%	1	90.0	10.0	100.0
		2	35.0	65.0	100.0
Cross-validated	Count	1	18	2	20
		2	7	13	20
	%	1	90.0	10.0	100.0
		2	35.0	65.0	100.0

Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
77.5% of original grouped cases correctly classified.
77.5% of cross-validated grouped cases correctly classified.

The above table presents the classification result of the discriminant analysis. 77.5% of the original grouped cases were correctly classified. Hence, 77.5% is the predictive accuracy of the data under review in this research. Our interest now is to compare this value with the predictive accuracy under logistic regression.

4.2 Predicting Academic Success using Logistic Regression

From table 4.5 below, it is revealed that the predictive accuracy under logistic regression for the same data is 85%.

Table 4.5 Logistic regression Classification Table

Observed		Predicted		
		Group		Percentage Correct
		1	2	
Step 1	Group 1	18	2	90.0
	2	4	16	80.0
	Overall Percentage			85.0

The above results can be tabulated as follows

Table 4.6 comparison of predictive accuracy of discriminant analysis and logistic regression

	Predictive accuracy
Discriminant analysis	77.5%
Logistic regression	85%

It can therefore be concluded that logistic regression is a better predictor of academic performance of students when compared to discriminant analysis.

SUMMARY AND CONCLUSION

The target of this research work is to study the effectiveness of logistic regression and discriminant analysis in forecasting student academic outcomes. Data on fresh intakes into secondary schools in Delta State, Nigeria was gathered and the average score of the students were collected for the period of their junior secondary school. Each of the variables used for the study were dichotomous, and were coded using 0 and 1. Comparison between discriminant analysis and logistic regression where carried out. The result proved that logistic regression was able to more significantly predict student outcome when compared to discriminant analysis. The study identified the factors that have the most significant impact on student academic outcome through the values of standardized coefficients in the discriminant analysis. The results indicate the most influential factor to be the marital status of parents. Another important factor identified is Mother's Occupation. The findings of the study help us to understand the student who needs preliminary assistance from their advisors.

REFERENCES

1. Abdulhafedh, A. (2022) Comparison between Common Statistical Modeling Techniques Used in Research, Including: Discriminant Analysis vs Logistic Regression, Ridge Regression vs LASSO, and Decision Tree vs Random Forest. Open Access Library Journal, 9: e8414. https://doi.org/10.4236/oalib.1108414
2. Ali, Hamdi F., Abdulrazzk Charbaji and Nada Kassim Hajj (1992). A Discriminant Function Model for Admission at Undergraduate University Level, International Review of Education, 38, 505-518.
3. Chun-Na Li, Meng-Qi Shang, Yuan-Hai Shao, Yan Xu, Li-Ming Liu, Zhen Wang (2019).
4. Sparse L1-norm two dimensional linear discriminant analysis via the generalized elastic net regularization. Journal of Neurocomputing 337. 80-96.

5. Dey, Eric L. and Alexander W. Astin. (1993). Statistical Alternatives for Studying College Student Retention: A Comparative Analysis of Logit, Probit, and Linear Regression, Research in Higher Education, 34, 569-581.
6. Gareth, J., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical
7. Learning: With Applications in R. Springer, Berlin, Heidelberg.
8. Julie Luna (2000). Predicting Student Retention and Academic Success at New Mexico Tech New Mexico Institute of Mining and Technology Socorro, New Mexico
9. Lim, Tjen-Sien, Wei-Yin Loh and Yu-Shan Shih (1999). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms, Department of Statistics, University of Wisconson, Madison, (http://www.stat.wisc.edu/p/stat/ftp/pub/loh/treeprogs/quest1.7/mach1317.pdf), [July 20, 1999].
10. Matthews, D. B. (1996). An Investigation of the learning styles of students at selected postsecondary and secondary institutions in South Carolina. (Research Bulletin No. 60.) Washington, DC: U.S. Department of Agriculture.
11. Meshbane, Alice and John D. Morris (1996). Predictive Discriminant Analysis Vs. Logistic Regression in Two-Group Classification Problems, Eric Document ED 400 280.
12. J. E. Osemwenkhae, A. Iduseri and F. Meka (2019). Determinants of the level of stress experienced by teachers at different educational levels: a descriptive discriminant approach. Journal of the Nigerian Statistical Association. Vol. 31
13. Park C. and Park H (2008). A comparison of generalized linear discriminant analysis algorithms, Pattern Recognition, 41: 1083-1097.
14. Worth, A.P. and Cronin, M.T.D. (2003): The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. Theochem, 622, 97-111.