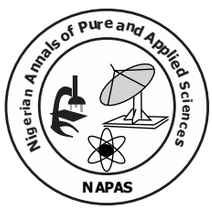


Prediction of Student's Academic Performance Using Linear Regression

Bum, S.[1], Iorliam, I. B.[2], Okube, E. O.[1], and Iorliam, A[1]*.
[1]Department of Mathematics & Computer Science,
Benue State University, Makurdi, Nigeria
[2] CEFTER, BSU, Makurdi, Nigeria
Corresponding author: Iorliam, A.: miorliam@yahoo.com

Abstract

Recently, there has been a tremendous increase of failure rate in higher institutions. Often times, students of tertiary institutions drop-out of school or end up with certain classes of degrees that are far lesser than their ideal intellectual capabilities. As fresh students enroll into higher institutions, some may not even know what a class of degree is, and the level of work that is likely to place them in a particular class of degree. Predicting their academic performance in this wise becomes necessary to ascertain the future of their performances in order to make informed decisions. Traditionally, we find out "who failed in an examination" or who "passed in an examination". However, predictions usually finds out who is likely to fail and who is likely to pass an examination and to what extent the failure or pass could be. Hence, predictions serve as counseling tool for the students to either improve their work rate or maintain their rate to achieve higher performance in subsequent examinations. In this paper, we predicted the academic performance of Benue State University students using linear regression. Linear regression assumes a linear relationship between the dependent variables (x) and an independent variable (y) such that the slope (predicted) can be specifically calculated from a computational linear combination of the variables x and y. Our proposed system correctly predicted the performance of mathematics/computer science students of the Benue State University with an accuracy of up to 100%.

Key words: Prediction, Linear Regression, Performance.

Introduction

Predicting students' academic performance has gained importance due to the increasing rate of failure among students in educational institutions (Mohammadi *et al.*, 2019). Educational systems (schools) pay a good deal of attention to academic performance of students, hence makes research around predicting the future of such performances worth exploring (Mohammadi *et al.*, 2019). In practice, predicting students' performance is generally a challenging task. The primary goals of data mining in practice tend to be prediction and description (Hand *et al*, 2001; Fayyad et al., 1996). Academic prediction involves variables like cumulative grade point average (CGPA) and test or assessment results while description focuses on finding human interpretable patterns that can describe the data. For example, identifying exceptional students for scholarships, identifying average students and identifying weak students who are likely to fail could be classified under description. The main objective of prediction in academic performance is to determine the likelihood of a particular student excelling or failing in a particular discipline. Usually, all educational institutions measure this intellectual ability by assessing the students either through oral tests, written assignments, group work or examinations. Apart from knowing the result of an academic performance after an assessment, it is also possible to also know the future of such performances provided a relationship is established between past and present performances. Reliable academic predictions help admission officials to differentiate between suitable and unsuitable candidates for a particular academic program. It also identifies students that are likely to do well or not in the progress of an academic program (Ayan and Garcia, 2013). The results obtained from the prediction of academic performances may be used in categorizing students – the intelligent, average, poor or weak. This categorization of students may enable lecturers or academic instructors to efficiently carry all students along by giving additional support to those that may be prone to failure. These predictions may also help instructors to identify the most suitable teaching methods for different categories of students and provide them with further assistance tailored to their needs in order to minimize failure. Students on the other hand can be informed from predicted results about the likelihood of either failure or pass and may make adjustments accordingly. For the fact that the students will understand their intellectual abilities from such prediction, they may also develop a suitable learning strategy to cope with their intellectual abilities. Accurate prediction of student academic achievements is one way to enhance the quality of education and provide better educational services (Romeo and Ventura, 2007). A variety of techniques have been employed to predict academic performance. In these techniques, a set of mathematical formulae are used to describe the quantitative relationships between outputs and inputs. The prediction is accurate if there is no error between the predicted and actual values. Basically, linear regression is a machine learning technique based on supervised learning (Fayyad et al., 1996). The linear regression model seeks to establish a relationship between x and y variables along their coefficients. These variables and coefficients can be calculated analytically using linear algebra. Linear regression is a type of predictive analysis model which examines a set of predictor variables (independent variable) in predicting an outcome and determine specific predictors of the outcome variable (dependent variable). These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. Linear regression has been applied in different science disciplines to determine the extent to which there is a linear relationship between a dependent variable and one or more independent variables. In simpler linear regression, a single independent variable is used to predict the value of a dependent variable. Even though this method has been used by different researchers, to the best of our knowledge, this is the first time linear regression will be used to predict student's academic performance paying attention to Benue State University (BSU) students.

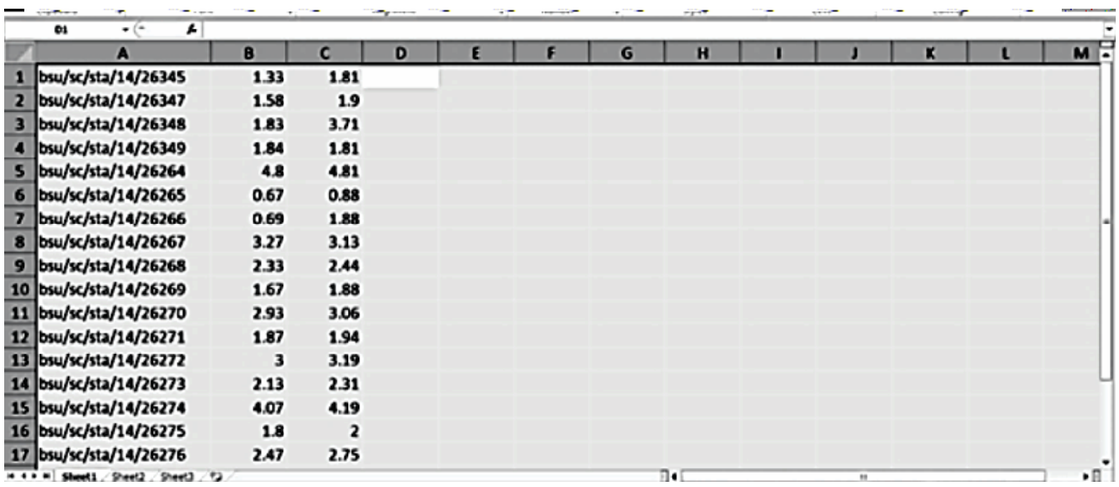
It is observed that the increasing rate of school drop-outs due to academic incompetence and the observed poor academic performance of students in higher institutions can be traced to misinformation of the relevant university authorities handling results to the students. The students have little or no knowledge about the

future of their academics and turn to concentrate on "passing" without necessarily paying attention to how well the passing should be. Most students are withdrawn from higher institutions because of their inability to cope after the completion of their first two academic sessions. BSU Makurdi for example withdraws a good number of students yearly due to poor academic performances. These students may possibly meet this fate of withdrawal as a shock since they have little or no knowledge about the future of their performances. Although poor academic performance cannot be traced to future unawareness alone, awareness can be core to adjustments in students' performances.

It should also be noted that failure rate as a result of porous traditional admission process has been an age long and global problem. Again, Iorliam and Ode (2014) noted that social media usage could negatively affect the performance of students in tertiary institutions. Other researchers such as Mingle and Adams, (2015), Marker, *et al.*, (2018), and Dadgarmehr *et al*, (2018) all agreed that social media usage by students could negatively affect their performance. Hence, this study takes a scientific approach to tackling the problem of uninformed failure among BSU students by exploring the possibility of using linear regression to predict the performance of students of the institutions taking the grade point average (GPA) as input.

Materials and Methods

The proposed system uses linear regression to predict the academic performance. The system will take a dataset as input containing the students' GPA of two consecutive semesters and make prediction of the next semester. This system will accept input data in form of Microsoft excel sheet. It is expected to produce the predicted result also on an excel sheet specifying the class of degree each student is likely to come out with. It should however be noted that data in other formats can also be easily used by the proposed system to predict students academic performance. Figure 1 shows a sample of the input data used for this experiment. Columns A, B and C indicate student's matriculation number, student's first semester and second semester GPAs respectively.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	bsu/sc/sta/14/26345	1.33	1.81										
2	bsu/sc/sta/14/26347	1.58	1.9										
3	bsu/sc/sta/14/26348	1.83	3.71										
4	bsu/sc/sta/14/26349	1.84	1.81										
5	bsu/sc/sta/14/26264	4.8	4.81										
6	bsu/sc/sta/14/26265	0.67	0.88										
7	bsu/sc/sta/14/26266	0.69	1.88										
8	bsu/sc/sta/14/26267	3.27	3.13										
9	bsu/sc/sta/14/26268	2.33	2.44										
10	bsu/sc/sta/14/26269	1.67	1.88										
11	bsu/sc/sta/14/26270	2.93	3.06										
12	bsu/sc/sta/14/26271	1.87	1.94										
13	bsu/sc/sta/14/26272	3	3.19										
14	bsu/sc/sta/14/26273	2.13	2.31										
15	bsu/sc/sta/14/26274	4.07	4.19										
16	bsu/sc/sta/14/26275	1.8	2										
17	bsu/sc/sta/14/26276	2.47	2.75										

Figure 1: Sample Input Data

The Gradient Descent Linear Regression

Here, we fit linear regression parameters θ to our dataset using gradient descent (Harrington, 2012). Basically, linear regression assumes that a dependent variable is related to an independent variable linearly. The data from the variables is used to plot and find the line-of-best fit. Gradient descent is used while training a model for regression. It is an optimization algorithm, based on a convex function, which twists its parameters iteratively to minimize error in a given function to its barest minimum. It simply measures the change in all weights with regard to the change in error. It could be thought of as the slope of a function, which implies that the higher the gradient, the steeper the slope and the faster a model can learn. But if the slope is zero, the model stops learning. In mathematical terms, a gradient is a partial derivative of a function with respect to its inputs.

In linear regression, a training set and the values of its parameters are a straight line gotten from the line of best fit. Making predictions for the values generally implies solving a minimization problem, i.e. we want to minimize the difference between the input and the output.

To achieve that, we use:

$\min \theta_0\theta_1 \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^{(i)})^2$ …....Equation 1

Here, m is the number of training examples. If we put a factor of $\frac{1}{2m}$ it gives us the same value of the process. Conventionally, we will be doing a cost function which is otherwise known as the squared error function.

$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^{(i)})^2$...Equation 2

We minimized J (θ) mostly by a trial an error method. The gradient descent is a better way of implementing the above. Basically, it changes the θ values, or parameters, bit by bit, until we hopefully arrive at a minimum. We start by initializing θ_0 and θ_1 to any two values, say 0 for both, and go from there. Basically, the formula is as follows:

$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_i} J(\theta_0, \theta_1)$ ……..Equation 3

(*for* $j = 0$ and $j = 1$)

Where α (alpha), is the learning rate or how quickly we want to move towards the minimum. We may not converge if α is too large. The main steps to predict student's performance using linear regression are as follows:

1. Prepare input dataset in excel format
2. Perform predictions using linear regression
3. Output the predicted results to an excel file

Results and Discussion

Using data from Figure 1 as input into our proposed system, the system predicted student's performance as shown in Figure 2. The output of the prediction is shown in five columns specifying the students' matriculation number, initial first and second semester GPAs, the predicted next level CGPA and the class of degree respectively.

	A	B	C	D
1	bsu/sc/sta/14/26345	1.41928963	Failed	
2	bsu/sc/sta/14/26347	1.60094812	Third class	
3	bsu/sc/sta/14/26348	2.58874494	Second class lower	
4	bsu/sc/sta/14/26349	1.70382237	Third class	
5	bsu/sc/sta/14/26264	4.76128329	First class	
6	bsu/sc/sta/14/26265	0.61519367	Failed	
7	bsu/sc/sta/14/26266	1.0950369	Failed	
8	bsu/sc/sta/14/26267	3.12029415	Second class lower	
9	bsu/sc/sta/14/26268	2.27246856	Third class	
10	bsu/sc/sta/14/26269	1.64178608	Third class	
11	bsu/sc/sta/14/26270	2.8977977	Second class lower	
12	bsu/sc/sta/14/26271	1.78148865	Third class	
13	bsu/sc/sta/14/26272	2.99778027	Second class lower	
14	bsu/sc/sta/14/26273	2.09995803	Third class	
15	bsu/sc/sta/14/26274	4.06342619	Second class upper	
16	bsu/sc/sta/14/26275	1.77055624	Third class	
17	bsu/sc/sta/14/26276	2.49586796	Third class	

Figure 2: Prediction Results of the Proposed System

Comparison of Actual and Predicted Results

The actual results obtained from the department of mathematics and computer science, BSU, Makurdi was compared with the predicted results as shown in Tables 1.0 and 2.0. Using the Chi-square $x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$, where O_i is the observed and E_i is the expected values, we observed the divergence between the two data sets. A low value for chi square(x_c^2) means there is a high correlation between the two sets of data.

Table 1.0: Predicted results of some sample students

S.NO.	MATRIC NUMBER	ACTUAL GPA 100 L		PREDICTED CGPA
		1st Sem.	2nd Sem.	200 Level
1	BSU/SC/STA/14/26345	1.33	1.24	1.4192
2	BSU/SC/STA/14/26347	1.58	3.71	1.6009
3	BSU/SC/STA/14/26348	1.83	1.81	2.5887
4	BSU/SC/STA/14/26349	1.84	1.81	1.7038
5	BSU/SC/STA/14/26264	4.80	4.81	4.7612
6	BSU/SC/STA/14/26265	0.67	0.88	0.6155
7	BSU/SC/STA/14/26266	0.69	1.88	1.0950
8	BSU/SC/STA/14/26267	3.27	3.13	3.1202
9	BSU/SC/STA/14/26268	2.33	2.44	2.2724
10	BSU/SC/STA/14/26269	1.67	1.88	1.6417

Table 2.0: Comparison Between the predicted CGPA with the actual CGPA of the sample students

S.NO.	MATRIC NUMBER	200 LEVEL		CHI SQUARE
		Predicted (O)	**Actual** (E)	$x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$
1	BSU/SC/STA/14/26345	1.4192	1.40	0.0002
2	BSU/SC/STA/14/26347	1.6009	1.70	0.0057
3	BSU/SC/STA/14/26348	2.5887	2.42	0.0103
4	BSU/SC/STA/14/26349	1.7038	1.75	0.0012
5	BSU/SC/STA/14/26264	4.7612	4.73	0.0002
6	BSU/SC/STA/14/26265	0.6155	0.91	0.0955
7	BSU/SC/STA/14/26266	1.0950	1.73	0.2330
8	BSU/SC/STA/14/26267	3.1202	3.29	0.0087
9	BSU/SC/STA/14/26268	2.2724	2.45	0.0128
10	BSU/SC/STA/14/26269	1.6417	1.64	0.0000

From the prediction results obtained in Table 1.0 and the Chi-square values obtained from Table 2.0, we observed that our proposed method accurately predicted student's next results by up to 100%. If this system is properly used in higher institutions, it will positively improve students' performance.

Conclusion

This work proposed a novel linear regression technique to predict student's performance. With the help of this system, student's performance will be improved. This system when properly deployed in higher institutions will produce graduates with high grades and also reduce the number of higher institution drop-outs. In our future work, we hope to test this proposed system on more student's results in different higher organizations. Furthermore, we hope to use other machine learning techniques to predict student's performance and make a comparative analysis between these machine learning techniques.

References

Ayán, M. N. R., & García, M. T. C. (2008). Prediction of university students' academic achievement by linear and logistic models. *The Spanish journal of psychology*, *11*(1), 275-288.

Dadgarmehr, M., Safari, E., & Gharehkhani, M. (2018). Proposing a Model for Analysing Impact of Social Media on Academic Performance of Students: A Case Study of Allameh Tabatabai University. *Journal of Soft Computing and Decision Support Systems*, *5*(2), 9-15.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996, August). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *KDD* (Vol. 96, pp. 82-88).

Hand, D., Mannila, H., & Smyth, P. (2001). Principles of data mining. 2001. *MIT Press. Sections*, *6*, 2-6.

Harrington, P. (2012). *Machine learning in action* (Vol. 5). Greenwich: Manning.

Iorliam, A., & Ode, E. (2014). The impact of social network usage on university students academic performance: a case study of

Benue state university Makurdi, Nigeria. *International Journal on Computer Science and Engineering*, *6*(7), 275.

Marker, C., Gnambs, T., & Appel, M. (2018). Active on Facebook and failing at school? Meta-analytic findings on the relationship between online social networking activities and academic achievement.

Mingle, J., & Adams, M. (2015). Social media network participation and academic performance in senior high schools in Ghana. *Library Philosophy and Practice*, 1.

Mohammadi, M., Dawodi, M., Tomohisa, W., & Ahmadi, N. (2019, February). Comparative study of supervised learning algorithms for student performance prediction. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 124-127). IEEE.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146.