TOWARDS RELIABLE PREDICTION OF ACADEMIC PERFORMANCE OF ARCHITECTURE STUDENTS: USING DATA MINING TECHNIQUES

ABSTRACT

Purpose
In recent years, there has been a tremendous increase in the number of applicants seeking placement in the undergraduate architecture programme. It is important to identify new intakes who possess the capability to succeed during the selection phase of admission at universities. Admission variable (i.e. prior academic achievement) is one of the most important criteria considered during selection process. The present study investigates the efficacy of using data mining techniques to predict academic performance of architecture student based on information contained in prior academic achievement.

Design/methodology/approach
The input variables, i.e. prior academic achievement, were extracted from students' academic records. Logistic regression and support vector machine (SVM) are the data mining techniques adopted in this study. The collected data was divided into two parts. The first part was used for training the model, while the other part was used to evaluate the predictive accuracy of the developed models.

Findings
The results revealed that SVM model outperformed the logistic regression model in terms of accuracy. Taken together, it is evident that prior academic achievement are good predictors of academic performance of architecture students.
Research limitations

Although the factors affecting academic performance of students are numerous, the present study focuses on the effect of prior academic achievement on academic performance of architecture students.

Originality/value

The developed SVM model can be used a decision-making tool for selecting new intakes into the architecture program at Nigerian universities.

Keywords Academic performance, decision making, logistic regression, modelling, support vector machine.

Paper type Research paper

INTRODUCTION

Student academic performance is an objective metric for assessing the knowledge gained by a student. Poor academic performance and non-completion of undergraduate studies are problems faced by tertiary educational institutions around the world (see Adewale and Adhuze, 2014; Ishitani, 2006). Previous research has established that academic performance is positively related to salary earned upon graduation, job performance, psychological empowerment, resilience and spiritual well-being (Dalessio, 1986; Roth and Clarke, 1998; Kuncel et al., 2004; Beauvais et al., 2014). Furthermore, empirical evidence shows that job performance and job satisfaction are related (Judge et al., 2001). Based on the foregoing, it is reasonable to suggest that academic performance could influence job outcomes and job attrition rates, among others.
Hence, an understanding of factors affecting academic performance of students would help stakeholders to develop intervention strategies to address this problem.

Research into student’s academic performance at universities has a long history. In one of the first published studies on this topic, Wagner and Strabel (1935) found that prior academic achievement at high school, motivation, and homogeneity of study sample contribute to ease of predicting academic performance of university students. It has been reported that teaching approaches and learning styles significantly affect student academic performance (Kvan and Jia 2005; Ling et al., 2010; Herrmann et al., 2016). Oluwatayo et al. (2009) demonstrated that improvements in the quality of learning environments increased the academic performance of students. Similarly, studies have revealed that a positive relationship exist between prior academic attainment and academic performance (Young, 1989; Wait and Gressel, 2009). Information gleaned from literature shows that several factors affect the academic performance of students. A full discussion of factors affecting the academic performance of university students lie beyond the scope of this study. However, the current study focuses on forecasting of academic performance of architecture students based on data relating to prior academic achievement.

Prior academic achievement is one of the main factors used to evaluate and select new intakes for university study. A large and growing body of literature has modelled the relationship between prior academic achievement and academic performance (see Al-Razgan et al., 2014; Dakduk et al., 2016). Two important issues emerge from the studies focused on modelling of the relationship between prior academic achievement and academic performance. First, the results of studies seem to be contrasting. For instance, the strength of the relationship between the two (i.e. prior academic achievement and academic performance) varies from strong (Holt et al., 2006) to
weak (Wao et al., 2016; Adewale and Adhuze, 2014) and no relationship (Suhayda et al., 2008).

Second, there is a conflation between studies targeted at 'explanation' and 'prediction'. According to Shmueli (2010), explanatory studies focuses on testing causal theories while predictive studies are targeted at theory building (such as practical relevance and evaluation of existing theories). For in-depth discussion of the difference between explanation and prediction, interested readers are referred to Shmueli (2010). There is a need to assess the efficacy of using data mining techniques for modelling and forecasting of academic performance of architecture students.

The current study, therefore seeks to provide answers to the following questions: Can prior academic achievement be used as predictors of academic performance of architecture students? What is the effect of prior academic achievement on the predictability of performance of architecture students? Can data mining techniques effectively predict academic performance of architecture students? By providing answers to the research questions, the article will contribute to existing literature on academic performance in several ways. First, it will provide insights on the practical relevance of using prior academic performance as predictors of academic performance. This information is vital for theory building as suggested by Shmueli (2010). Second, the developed model can be used as a decision support tool for effective screening of new intakes into the undergraduate architecture programme.

**LITERATURE REVIEW**

**Prior academic achievement and academic performance of student**

Over the years student academic performance has received much attention from researchers, parents, institutions of learning, government, education reformers and policy makers, among others. For instance, policy makers in the US have moved motions for
improvements in students’ academic performance (Wenglinsky, 2002). This clearly highlights the importance of academic performance to relevant stakeholders. Student academic performance is a metric that quantifies learning outcomes (Steinmayr et al., 2015). Metrics such as examination scores, test scores and Grade Point Average (GPA) are among the indicators used to determine the academic performance of a student (Caro et al., 2014). But William (2016) argues that school grade or score may not always indicate student’s levels of intelligence and knowledge as some student could perform low in class but do well in intelligent quotient test. Also, Levy and Murray (2005) argues that grades are obtained in tertiary entrance examination is not a determinant of academic performance. Despite the difference that exist in literature, GPA remains the most acceptable metric for assessing students’ performance (Richardson et al., 2012; Ogbogu, 2014).

Previous research have modeled the relationship between prior academic achievement and academic performance of students. Based on Shmueli’s (2010) classification of models, these models can be categorized into two broad groups: explanatory and predictive modelling. For explanatory modelling, studies have focused on the following academic discipline: architecture (Adewale and Adhuze, 2014; Roberts 2007; Sutton et al., 2016), real estate management (Allen and Carter, 2007), building technology (Abisuga et al., 2015), construction management (Wao et al. 2016) and dentistry (Ihm et al., 2013), among others. Several modelling techniques have been used to generate forecast of the academic performance of architecture students (Aluko et al., 2016), engineering students (Huang and Fang, 2013), computer science students (Aziz et al., 2015) and mature students (Chong et al., 2016). This article builds on the findings of an earlier study (Aluko et al., 2016) and explores the efficacy of using new data
mining techniques for modelling and forecasting of academic performance of architecture students.

**Application of data mining techniques**

Data mining techniques are useful for exploring the underlying patterns present in large volumes of data. Data mining techniques can be applied to regression, association rule discovery, classification and clustering tasks (see Romero and Ventura, 2007; Cortez et al., 2009). The main strength of data mining techniques lies in its ability to capture nonlinear relationships which are present in real-world data. Data mining techniques have been successfully applied to problems in the field of construction economics (Bala et al., 2014), civil engineering (Chou et al., 2010), food processing (Cortez et al. 2009), medicine (Delen, 2009) and finance (Yeh and Lien, 2009), among others. For additional information on data mining techniques, interested readers are directed to Witten et al. (2011). Hence, the present study examines the predictive performance of data mining techniques when applied to forecasting of academic performance of architecture students’.

**METHOD**

A number of methods have been used to solve research problems in academic disciplines within the built environment. Research methods found in published studies related to built environment include: survey, experiment, literature survey, case-study, modelling, archival research and grounded theory (Laryea and Leiringer, 2012). The multiplicity of methods have resulted in debates on the suitability of a particular approach. Wing et al. (1998) affirmed that nature of the problem under investigation influences the choice and appropriateness of a research
method. Modelling techniques are suitable for solving forecasting problems (see Shmueli, 2010; Fellows and Liu, 2015). Therefore, two data mining modelling techniques (logistic regression and support vector machine) were applied in the present study. These techniques were used for modelling and forecasting of the academic performance of architecture students. In addition, the efficacy of using these techniques for forecasting of the academic performance of architecture students was assessed and compared.

**Student data**

The sample for the present study was drawn from Department of Architecture at Olabisi Onabanjo University Ogun State, Nigeria. The Department runs a 4-year undergraduate program. As stated earlier, prior academic achievement is one of the main criteria that influences selection of new intakes during the admission process. In the current study, data was collected on the grades obtained by each student in Ordinary level (O-level) examinations and unified tertiary examination score (JAMB). In addition, CGPA was used as the metric for quantifying students’ academic performance in the undergraduate program.

Data relating to prior academic achievement (O-level and JAMB) and academic performance (CGPA) was collected from academic records of each student. From the database, 102 architecture students have completed the program in four consecutive cohorts. However, due to incomplete data, the information relating to one student was deleted from the collected data. Thus, data on 101 students (i.e. prior academic achievement and academic performance) was used for model development.

For the forecast models, the input variables are: grades obtained in O-level examinations (11 subjects), mode of entry (JAMB/Direct entry), and JAMB score. This is similar to the inputs
variable used in a previous study (see Young 1989). The output variable is academic performance (i.e. CGPA). The output variable was transformed into binary outcome (“pass” or “fail”). A student with a CGPA of greater than 2.4 on a scale of 5 is classified as “pass”. This is because a minimum CGPA of 2.4 is required for progression to the Master of Science (MSc) Architecture program. Thus, CGPA was used a criterion for classifying students into two groups (pass/fail). Table 1 gives a description of the variable used to develop the forecast models.

*Insert Table 1*

**DATA MINING MODELS**

In the present study, two data mining techniques (i.e. logistic regression and support vector machine [SVM]) were used for modelling and forecasting of academic performance of architecture students. The Logistic regression and SVM algorithm were implemented in R programming software (R Core Team, 2015) and *rminer* R-package (Cortez, 2010).

*Logistic regression*

Logistic regression has its roots in statistics and it is widely used for classification problems (see Delen, 2009; Şen et al., 2012). Traditionally, linear regression is used to investigate the relationship between continuous dependent variable and a set of independent variables. However, Larose (2006) points out that linear regression models (LRM) are not suitable for categorical dependent variable. Logistic regression is a linear function which has been found to be useful when applied to classification tasks. The logistic regression model can be expressed as:

\[
\text{logit}(y) = c_0 + c_1 x_1 + c_2 x_2 + \ldots + c_k x_k
\]  

(1)
where \((x_1, x_2, \ldots, x_k)\) are independent variables, \(y\) is the dependent variable, \((c_0, c_1, \ldots, c_k)\) are coefficients which are adjusted using the maximum likelihood technique and logit \((y) = \ln \left( \frac{y}{1-y} \right)\). The coefficients of this model are easy to interpret and it is widely used for classification tasks. Unlike the LRM that predicts point estimate of an event, logistic regression models predict the odds of its occurrence. For a two-class problem, odds greater that 50% is assigned to predefined "1" or "0" for otherwise.

**Support vector machine**

Support vector machine is a data mining technique that has been applied to classification and regression tasks. In this research, SVM is used for a classification task. Due to the absence of local minima at the learning phase, SVM has theoretical advantage over artificial neural network. Empirical evidence from previous studies shows that SVM outperform neural network (NN) model (Lam et al., 2009; Tinoco et al., 2011). For detailed explanation of the SVM, interested readers are referred to Witten et al. (2011). In this research, the radial basis function (RBF) is utilized as the kernel function for SVM model. The RBF produces optimal solutions when compared to other kernel functions (see Bin et al., 2006; Lam et al., 2009).

Using RBF kernel, there are three hyperparameters \((C, \varepsilon, \gamma)\) that need to be determined. Optimal values for the hyperparameters need to be identified for the SVM model. As suggested in Hsu et al. (2003), a combination of 'grid-search' algorithm with cross-validation was used to identify the best values of the SVM hyperparameters. In this research, the leave-one-out (i.e. 5-fold) cross-validation technique was used to identify the optimal hyperparameters and assess the predictive performance of the SVM model.
The same dataset (input and output variables) was used to develop the two forecast models. This ensured that the predictive performance of both models can be compared. The hold-out approach was used to prevent model over-fitting. The collected data was divided into two subsets: training (70 per cent) and test sets (30 percent). The models were trained using all the training data. Subsequently, the trained model was used for forecasting of academic performance in the test dataset. Forecasting of test data ensures that the generalization capability of the developed model on previously unseen data was assessed. The predictive performance of the models was evaluated by computing confusion matrix and overall accuracy of forecasts (see Tables 3 and 4). Using measures of performance, a comparative analysis of logistic regression and SVM model was carried out. Then, sensitivity analysis as explained in Cortez (2015) was applied to evaluate the relative importance of the input variables used for the development of the models.

RESULTS

As stated previously, the forecast models (logistic regression and SVM) were trained and tested using the collected data. Subsequently, the predictive performance of the developed models was compared. The results from model development and validation phase are summarized and presented in this section

Logistic regression

The logistic regression model is estimated and used to predict performance of architecture students into one of the two class labels (“Pass” or “Fail”). The coefficients of the logistic regression model (under column heading B) and other related statistics are summarized and presented in Table 2.
For the logistic regression model, the value of the log-likelihood ratio test is significant at 5 percent level of significance. Also, the Wald test was conducted to examine the significance of individual variables. Based on results of Wald test, it was found that MATH and BIO are the only significant input variables. Additional tests (i.e. Hosmer-Lemeshow and McFadden’s pseudo $R^2$) were carried out to assess the goodness-of-fit, measure of discrimination and calibration of the developed logistic regression. The $p$-value from the Hosmer-Lemeshow test is below 0.05, which suggest that the logistic regression model does not have a good fit. The value of McFadden’s pseudo ($R^2$) ranges between 0 and 1 (values closer to zero indicate the model has no predictive power). The calculated value of McFadden $R^2$ is 0.35. On the overall, the predictive performance of the logistic regression model is 50 percent (see Table 3).

**Support Vector Machine**

The sequential minimal optimization algorithm training algorithm was used to train the SVM model. The hyperparameters of the SVM model was optimized using an internal grid search (i.e. using only training data), the values of the hyperparameters are: $\gamma = 0.2529$; $C = 4.4375$ and $\varepsilon = 0.0020$. The results of out-of-sample prediction (i.e. the test set) for the SVM model are presented in Table 4. The overall prediction accuracy of the SVM model is 76.67 percent.
Comparison of various models

This section provides the comparative analysis of predictive performance of the developed models. The developed models were used for forecasting of academic performance of architecture students in the test set (unseen data). The output (i.e. forecasts of academic performance of architecture students) of the logistic regression and SVM model were presented in the preceding section. The results shown in Table 3 and 4 revealed that the SVM model provides a better forecast when compared with the logistic regression model. In addition, the findings of the study suggests that prior academic achievement is a good predictor of academic performance of architecture students. It is important to note that the percentage of correctly classified for the ‘fail’ class was significantly lower when compared to the ‘pass’ classification (see Tables 3 and 4). A possible explanation for this disparity may be linked to the small size of study sample that belonged to the ‘fail’ group (only 25.24 per cent of the records). Similarly, the improved performance of the SVM model suggests that a nonlinear relationship exist between the input and output variables (i.e. prior academic achievement and academic performance).

Sensitivity analysis

Compared to the SVM model, the coefficients of the logistic regression model are easier to interpret. The SVM model and other non-linear data mining models are often referred to as ‘black-box’ models. This is because little or nothing is known about the strength of the relationship between input and target variables. However, in recent years, sensitivity analysis has been proposed and utilized in collecting comprehensible information from black box data mining models in previous studies (Hsieh, 2004; Tinoco et al., 2011).
In this study, the SVM model was subjected to sensitivity analysis due to its superior predictive performance. Sensitivity analysis provides insight into the relative importance of each input variable to the prediction of the target variable. As can be seen from Figure 1, MATH, BIO and PHY are the most significant variables which influence academic performance of undergraduate architecture students. Surprisingly, F/MATHS, ECON and TD had the least effect on academic performance of architecture students.

Insert Figure 1

DISCUSSION OF FINDINGS

As stated earlier in this paper, the current study explores the efficacy of using data mining techniques for forecasting of academic performance of architecture student's. In recent years, number of applicants seeking admission placement in Nigerian universities have grown exponentially. As a result, stakeholders in the tertiary education sector are seeking for ways to improve the student selection process.

The current study found that prior academic achievement are valid predictors of academic performance for architecture students. A comparison of the developed model revealed that the SVM model outperformed the logistic regression model in terms of accuracy of prediction. The overall accuracy of the SVM model is 76.67 percent. Grades obtained in the following O-level examinations are good predictors of academic performance for architecture students: mathematics, biology and physics. The superior performance of the SVM model suggest that a nonlinear relationship exist among the variable used to develop the model. It is somewhat surprising that grades obtained in technical drawing/fine art (TD) was the least contributor to academic performance of architecture students. This is because grades obtained in TD could be used as a metric for assessing the ability of a student to draw objects.
The results of the current study support the findings of previous research (see Holt et al. 2006; van Rooyen et al., 2006) which showed that prior academic achievement is positively related to students’ academic performance. This outcome is contrary to the findings reported in previous studies which reported that weak relationship exist between prior academic achievement and academic performance (Abisuga et al., 2015; Adewale and Adhuze, 2014; Wao et al., 2016). A possible explanation for this might be the nonlinear relationship between prior academic achievement and academic performance. On the overall, the SVM model provides reliable forecast of the academic performance of architecture students. While the findings of the present study are promising, it must be acknowledge that several factors affect the academic performance of undergraduate students. For example, it was reported that academic performance of students’ is affected by quality of learning environment, teaching approach and learning styles, among others (Kvan and Jia 2005; Ling et al., 2010; Herrmann et al., 2016; Oluwatayo et al., 2009). The developed SVM model can be used as a decision support tool for selection of new intakes into undergraduate architecture programs in Nigerian Universities.

CONCLUSION

The main goal of the current study was to examine the efficacy of using data mining models for forecasting of the academic performance of architecture students in a Nigerian university. Three main findings emerge from the results presented in the preceding section. First, prior academic achievement is a good predictor of academic performance of architecture students. Second, SVM is a good modelling tool for forecasting of academic performance of architecture students. Third, grades obtained by students in the following O-level subjects: mathematics, biology and physics have the most significant effect on academic performance of
architecture students. Although the SVM model produced good forecasts, the classifications were not 100% accurate. The results suggests that other predictors of academic performance of architecture students were not captured in the developed models.

Taken together, the results suggest that prior academic achievement is a good indicator of the ability of a student to succeed in the study of architecture at the undergraduate level. The present study makes noteworthy contribution to the existing literature in several ways. The results presented here extend the current knowledge relating to the academic performance of architecture students at an undergraduate level of study. In addition, the study provides evidence of the efficacy of using data mining techniques for solving forecasting problems in the real-world.

There are obvious limitations that affect the outcomes of research based on modelling approach. The major weakness of this study is the small size of the available data. However, the use hold-out sample of the original data for validation of the model ensures robustness and reliability of the developed models. This is similar to the approach used in other previous studies (see Edwards et al., 2007). Despite the limitation, the current study offers insights into the relationship between prior academic achievement and academic performance. A future study investigating the predictability of academic performance using prior academic achievement and other information on student's background (such as level of parent's education, economic status, availability of financial support, etc.) would be very interesting.
REFERENCES


