ABSTRACT

Universities have been struggling to meet the demands of growing numbers of students who enter higher education with limited academic skills. Research has established that students change substantially (in terms of their field of study) over the course of their undergraduate academic experience. The most dramatic changes however occur during their first year of study.

The purpose of this paper is to predict the probability for a first-year student to register and pass all required first-year courses using various predictor variables at Technikon Pretoria, South Africa.

A statistical model will be used to find the most important factors/variables distinguishing/discriminating between successful and unsuccessful students at Technikon Pretoria.
1. BACKGROUND AND NEED FOR THE STUDY

Research on retention studies abroad is more than 100 years old. In South Africa, the National Plan for Higher Education (NPHE (2001)) indicated that the reasons for the decline in retention rates in South Africa are not clear and require investigation. Publications on factors influencing retention rates in South Africa are limited and the importance of this paper in a South African context is crucial for institutional planning in Higher Education.

The integration-commitment model of attrition developed by Tinto (1975) and later modified by Pascarella and Terenzini (1983) has been used repeatedly in past research. According to this model, persistence is strongly related to a student’s: (a) level of academic and social integration (‘fit’) with an institution, (b) commitment to earning a degree (goal commitment), and (c) commitment to an institution (institutional commitment).

Liu (2000) stated that commonality between integration and satisfaction is crucial to the success of academic performance and persistence and that student satisfaction is highly related to student retention, and key to academic withdrawal.

Yorke (1999) has identified three primary causes of withdrawal among full-time students: a mismatch between students and their choice of field of study, financial difficulties and poor quality of the student experience which refers to the ‘quality of the teaching, the level of support given by staff and the organization of the program’.

Predicting retention and student performance is an increasing concern for administrators due to the costly effects associated with non-persistence. Being able to predict more accurately which students might potentially drop out or take longer to graduate would enable institutions to focus on intervention strategies.

Descriptive statistics to describe differences in retention rates and student characteristics have been compiled by Technikon Pretoria in order to find out how well it is doing with the students it already serves. However, it has become crucial to make use of inferential statistics to make predictions from institutional data. The purpose of this study is to identify the factors influencing student success at Technikon Pretoria. Specifically, the study attempts to:

Predict the probability of a student being a successful first-year student, i.e. a first-time entering student that registered and passed all required first-year courses, using various predictor variables.

The study will therefore attempt to predict certain retention characteristics of first-year students, using Technikon Pretoria as a South African case study.

2. METHODOLOGY

A variety of information on background, demographic and performance linked to the operational database of Technikon Pretoria has been used to predict the probability of a first-time entering student being successful or not during the first year of study. The statistical techniques, logistic regression and classification tree analysis have been used to predict the probability of a student being classified as a successful student.

A first-time entering student is defined as a student who enrolled for the first time for any qualification at any Higher Education Institution.

A first-time entering student in this study is regarded as a successful first-year student if the degree credits passed (DCP) during the first year of study are equal to or greater than one (unsuccessful if DCP is less than one). DCP, a concept used in this paper, is a variable combining the success of a student (passing or failing a subject) and the course load (number and weight of subjects taken). This variable will be used as the bivariate dependent variable in the analyses.

The credits of all the subjects within the curriculum of a first-year student in one normal year should amount to 1.00 credit. For example, a first-year student will typically take four
subjects, with a credit value of 0.25 for each subject. The credit value of each subject is
determined by the hours of teaching and practical classes involved in that specific subject.

Table 1 gives an example of the DCP for a hypothetical student. It is clear from this table that
the DCP for a student in a given year is calculated by adding the credits of the subjects that the
student passed during that year. The total DCP of the student in Table 1 is equal to 0.75 for the
first year and the student will therefore be classified as an unsuccessful student (DCP<1) for
the purposes of this study. However, if the student passed all four the subjects (DCP=1), the
student would have been classified as successful in this study.

Table 1:  First-time entering National Diploma student

<table>
<thead>
<tr>
<th>Subjects in first year</th>
<th>Pass/Fail</th>
<th>Credit</th>
<th>Degree credit passed (DCP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
<td>0.25</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>P</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>4</td>
<td>P</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>DCP</strong></td>
<td><strong>0.75</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.1 Operational data and Statistical techniques

Operational data (Source: Technikon Pretoria Management Information System) for first-time
entering students at Technikon Pretoria for the period 1994 to 2000 was used to model the
data. Due to missing values, only 4 848 (17.62%) of the original dataset of 27 508 records
could be used in the analysis (distance education students were excluded from the original
dataset). In total, 1 016 (20.96%) of these students were classified as successful students and 3
832 (79.04%) were classified as unsuccessful, using the DCP criteria explained previously.

Logistic regression used in this study is a technique that models the relationship between a
bivariate, discrete, dependent (response) variable and categorical and/or continuous
independent (predictor) variables. The logistic regression model uses the predictor variables to
predict the probability that the response variable takes on a given value.

The following independent variables have been included in the analyses:

Age of student
Province of matriculation
Grade 12 aggregate
Grade 12 English symbol (defined as adequate or inadequate)
Ethnic group
Gender
Campus of study (Pretoria campuses versus satellite campuses)
Method of study (full-time versus part-time)
Financial Aid (Yes/No)
Marital status
Type of accommodation (resident student or not)
Classification of Educational Subject Matter (CESM) category (i.e. major field of study)

The Grade 12 English symbol has been categorized as adequate (English on Higher Grade – D
or better, Standard Grade – C or better, Lower grade – B or better) or inadequate.
The 22 first-level CESM categories (major field of study) as defined by the South African
Department of Education are:

1. Agriculture and Renewable Natural Resources
2. Architecture and Environmental Design
3. Arts: Visual and Performing
4. Business, Commerce and Management Sciences
5. Communication
Therefore, in terms of the present study, logistic regression was used to predict the probability of a student being a successful first-year student making use of the 12 independent variables listed above.

The 12 independent variables listed above were included in the initial logistic regression model and the stepwise regression technique was used to find the predictors in the model. Stepwise selection is an exploratory process of finding the model that is the best of the competing models for the data. The stepwise selection option used combines the backward elimination and forward selection methods to respectively add variables to a model or remove variables from a model as they meet or fail to meet specified significance levels. All variables were required to have a p-value of at least 0.05 to enter a model (in forward selection) or to stay in the model (in backward selection).

All 4848 records were used in the initial stepwise logistic regression model in order to identify the predictor variables. Thereafter, the dataset was subdivided into training/calibration and validation subsets. The logistic regression calibration dataset contained 1700 records of first-time entering students (850 randomly selected successful students and 850 randomly selected unsuccessful students). The remaining 3148 records (166 successful students and 2982 unsuccessful students) were used as the validation dataset.

The adequacy of the logistic regression model was assessed by making use of the following statistical measures: Maximum re-scaled $R^2$ coefficient, Hosmer and Lemeshow goodness-of-fit test, the percentage correctly predicted, the sensitivity, the specificity, the area (c) under the receiver operating characteristic curve (ROC) and a validation dataset.

As a final step the data was explored by using tree analysis, a classification technique for predicting student success. The tree structure is generated by dividing the sample recursively into a number of groups, each division being chosen as to maximize some measure of the difference in the response variable in the resulting two groups (Everitt (1998)). In the present study, classification tree analysis is introduced as an attractive alternative to logistic regression models due to its visual properties. This paper only touches on classification tree analysis and it is used purely as a descriptive and exploratory technique to further support the results obtained from the logistic regression models.

3. RESULTS

3.1 Logistic regression

The Proc Logistic procedure in SAS (SAS Institute Inc. (1995)) was used for conducting the logistic regression.
The stepwise logistic regression technique retained the following eight variables (in order of importance) as significant predictor variables in the model:

- CESM category
- Grade 12 aggregate
- Ethnic group
- Financial aid
- Province of matriculation
- Campus
- Method of study
- Grade 12 English symbol

It was evident that CESM category and Grade 12 aggregate were very strong predictor variables when the p-values for the different predictor variables in the stepwise regression model (not shown here) were compared. These two variables were more important than the remaining six variables.

**Table 2: Evaluating the predictive abilities of the model**

<table>
<thead>
<tr>
<th>Statistical measure</th>
<th>Model 1 8 variable model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum rescaled $R^2$</td>
<td>0.4086</td>
</tr>
<tr>
<td>Hosmer and Lemeshow’s goodness-of-fit test  p=0.1376</td>
<td></td>
</tr>
<tr>
<td>Bias-adjusted % correctly predicted (for prob. threshold set at 0.5)</td>
<td>74.9%</td>
</tr>
<tr>
<td>Bias-adjusted sensitivity (for prob. threshold set at 0.5)</td>
<td>73.4%</td>
</tr>
<tr>
<td>Bias-adjusted specificity (for prob. threshold set at 0.5)</td>
<td>76.5%</td>
</tr>
<tr>
<td>Area under ROC curve (c)</td>
<td>0.831</td>
</tr>
<tr>
<td>Percentage successful students predicted correctly (validation dataset)</td>
<td>76.51%</td>
</tr>
<tr>
<td>Percentage unsuccessful students predicted correctly (validation dataset)</td>
<td>74.68%</td>
</tr>
</tbody>
</table>

The maximum rescaled $R^2$ coefficient of 0.4086 and 0.3725 for compares favourably to similar studies done in Higher Education Institutions in the USA where values of 0.3 were reported (Huey (1999)). Furthermore, the Hosmer and Lemeshow goodness-of-fit statistic suggests that the model was adequate and provide a good fit to the data (p>0.05).

The most rigorous test for determining the accuracy of a logistic model is to apply the model to a validation dataset. If the model is accurate in its predictions when applied to an independent validation dataset that was not used in the initial calibration/training of the model parameters, then the model is of true value. The model performed satisfactory in this regard and the logistic model predicted 76.51% of the successful students correctly in the validation dataset.

### 3.2 Classification Tree Analysis

Classification Tree Analysis, as part of the Multivariate Exploratory Techniques module in STATISTICA, was used for the tree analysis. The discriminant-based univariate splits for categorical and ordered predictors option was used for constructing the tree. Prune on misclassification error was used as stopping rule with the minimum set to five observations per terminal node and the standard error rule kept on the default value of one.
Classification tree analysis was included in the study to illustrate an alternative to logistic regression and to visually and intuitively support the results from the regression model. Therefore, it was deemed sufficient to include a classification tree using only CESM category and Grade 12 aggregate as predictor variables for predicting successful students. Figure 1 gives the results of the classification tree analysis. Note that the splitting criteria shown in Figure 1 always refer to the left branch and that the complement of that criteria holds for the right branch.

For example, the 704 students studying in CESM categories 3, 9 and 11 belong to the left branch and the remaining 4 144 students studying in the other CESM categories belong to the right branch of the tree.

Of the 4 848 students in the original dataset, only 1 016 (21%) were successful first-year students (top box in Figure 1). However, 66.5% of the students studying in CESM categories 3, 9 or 11 and having an aggregate of greater than 1 166.2 were successful. That means that a student studying in Visual and Performing Arts, Health Care/Sciences or Industrial Arts, and having a Grade 12 aggregate in excess of 1 166.2, has a probability of 0.67 to be a successful first-year student (compared to the probability of only 0.21 for the total group). However, if a student enrols for any of the remaining CESM categories and has an aggregate smaller than or equal to 1 526, the probability is only 0.14 that the student will be successful in the first year.

Figure 1: Classification Tree for successful and unsuccessful students, based on DCP
allow the Technikon to intervene and try to prevent students identified as ‘high-risk’ students, from dropping out. It can also have implications for enrollment planning at Technikon Pretoria.

Logistic regression was used again to establish a model to predict the probability of a student dropping out. The same 12 predictor variables used for predicting the probability of a student being a successful first-year student was used in this analysis. Unfortunately, logistic regression did not accurately predict the dropout students. The model only accurately predicted 58.8% of dropout students and 55.56% of non-dropout students in the validation dataset ($R^2 = 0.1409$).

However, even though student dropout could not be modelled with logistic regression, there was a highly significant association (Chi – square = 210.85, p<0.0001) between student dropout and the success of students, as illustrated in Figure 2.

**Figure 2: Relationship between success and dropout classification of first-time entering students at Technikon Pretoria (1994 – 2000)**

Figure 2 illustrates that a student who obtained adequate DCP in the first year of study is unlikely to drop out after the first year (only 5.02% of successful first-year students did not return the next year). This relationship will be useful in follow-up retention studies at Technikon Pretoria.

**4. SUMMARY AND RECOMMENDATIONS**

It is clear from the results that the Grade 12 aggregate and major field of study (CESM) play an important role in terms of the student’s first-year success. This is in agreement with a comprehensive analysis of 175 Higher Education Institutions in the United Kingdom. In this study the Higher Education Funding Council for England demonstrated a clear relationship between entry qualifications, curriculum contents and non-completion (Patrick (2000)).

Further analysis should be undertaken to find out in which major fields of study intervention should take place (in the form of bridging courses or enrollment management, for example). Currently, the admission criteria at Technikon Pretoria differ between the major fields of study
(CESM categories). This could be one of the reasons for the differences in success between students from the different CESM categories.

It is important to note that the statistical analyses used to conduct these studies are a first attempt to determine possible ‘at-risk’ students. Incorporating non-academic factors (not linked to the operational database such as family income or part-time employment) in the analyses would result in more accurate predictions of retention characteristics.

In order to identify these non-academic factors, Technikon Pretoria initiated the following longitudinal study.

Identify a stratified random sample of first-time entering students in 2002, as the initial cohort.

Collect base-line, non-academic information (for example, financial information and living arrangements) by administering questionnaires to the selected cohort.

Link the non-academic information obtained from the questionnaires with the operational data.

Identify dropout and unsuccessful students from the initial 2002 cohort in 2003.

Utilize the expanded dataset (operational and non-academic variables) to re-calibrate the logistic models, predicting the probability of a student being unsuccessful or a dropout student. It is anticipated that the additional non-academic predictor variables will increase the accuracy of these models.

Profile successful versus unsuccessful students and dropout versus non-dropout students using the results from the logistic models.

As part of a qualitative study, the students of the initial cohort who dropped out during or after the first year (2002) will also be targeted to complete a questionnaire in 2003, supplying reasons for them dropping out. The results from the quantitative and qualitative study will be integrated and consolidated to inform the management of Technikon Pretoria on variables influencing student retention and possible intervention strategies to solve the problem.

Enrollment management, recruitment criteria and intervention strategies such as bridging courses in specific fields of study are some of the recommendations that are anticipated.

5. ACKNOWLEDGEMENT

The assistance of the Directorate: Management Information at Technikon Pretoria towards the extraction of the operational data is hereby acknowledged.
6. REFERENCES


