APPLYING PREDICTIVE ANALYTICS IN IDENTIFYING STUDENTS AT RISK: A CASE STUDY

A. Lourens
Unit for Business Mathematics and Informatics
North-West University
Potchefstroom Campus, South Africa
e-mail: Amanda@idsc.co.za

D. Bleazard
Institutional Planning
Cape Peninsula University of Technology
Bellville, South Africa
e-mail: BleazardD@cput.ac.za

ABSTRACT

In this article, a case study is presented of an institutional modelling project whereby the most appropriate learning algorithm for the prediction of students dropping out before or in the second year of study was identified and deployed. This second-year dropout model was applied at programme level using pre-university information and first semester data derived from the Higher Education Data Analyzer (HEDA[1]) management information reporting and decision support environment at the Cape Peninsula University of Technology. An open source platform, namely Konstanz Information Miner (KNIME[2]), was used to perform the predictive modelling. The results from the model were used in HEDA automatically to recognize students with a high probability of dropping out by the second year of study. Being able to identify such students will enable universities, and in particular programme owners, to implement targeted intervention strategies to assist the students at risk and improve success rates.

Keywords: students at risk, predictive learner analytics, retention of students, student dropout, logistic regression, decision trees, Naïve Bayes

INTRODUCTION

Institutions seeking a competitive advantage are increasingly looking to the future where more and more data will broaden the scope of the predictive or learning analytics they can do in order to forecast student behaviour. Predictive analytics is the process of discovering interesting and meaningful patterns in data with a view to predicting likely future behaviour of the phenomena analysed (Abbott 2014). It draws from related disciplines including statistics, machine learning and data mining (Abbott 2014). Delmater and Hancock (2001) indicated that 'the science

underlying predictive modelling is a mixture of mathematics, computer science and domain expertise'.

Data mining and specifically predictive analytics are extensively used in the business world (Luan 2002). The application of predictive analytics in higher education has started to evolve within the last ten years, with institutions focusing on learner analytics to answer questions like: 'What are the characteristics of the students who persist in their studies and graduate?' The establishment of the South African Higher Education Learning Analytics (SAHELA) network and other initiatives from the Council on Higher Education (CHE) have given added impetus to this trend.

Monitoring and supporting first-year students is critical. In 2007, Scott et al. reported that 25 per cent of students in South Africa drop out during their first year of study. South Africa's National Development Plan has characterised the higher education sector as a 'low participation, high attrition system' (NDP 2011). Analysis of higher education performance in South Africa (CHE 2010, 2014) has shown low student throughput rates and a report by the CHE has evaluated the feasibility of restructuring the three- and four-year undergraduate degrees and diplomas (CHE 2013). Predicting student retention is therefore an increasing concern for administrators due to, in part, the costs associated with non-persistence.

A number of benefits in using predictive analytics for higher education have been described by Long and Siemens (2011). The benefits are mainly focused at an administrative level, such as improving decision-making, and assisting institutions in identifying intervention programmes to affect change and improve student success. It is important to note that predictive analytics is only a tool to assist institutions in terms of enhancing decision-making and informing processes and practices. The pedagogy should drive predictive analytics and not the reverse (Greller and Drachsler 2012). Predictive analytics therefore assists in improving human decision making and providing information to alert the institution to matters requiring attention. Adding predictive analytics to institutional management information allows for better informed decisions as indicated in many studies (Verbert et al. 2012; Watters 2012).

However, even the most comprehensive data set cannot take into account all aspects in relation to predicting student retention, such as interpersonal issues or historical identity and contexts. There are a number of factors impacting on students' likelihood to persist of which some cannot be measured, such as being emotionally unprepared (Eduventures 2013). Careful consideration should be given in the deployment and adoption of a predictive analytics model (Stiles 2012), especially from an ethical point of view. Campbell et al. (2007) provide a list of

issues that should be addressed before implementing the outcome of a predictive analytics model.

While important, the ethical issues associated with the implementation of predictive analytics in the context of higher education are beyond the scope of this article. The purpose of this article is to report on the use of predictive modelling algorithms such as Logistic Regression, Naïve Bayes, and Decision Trees to determine likely student dropouts before or in the second year of study for a particular qualification at the Cape Peninsula University of Technology in Cape Town. The developed algorithms were assessed and compared and the most significant variables identified in predicting second-year student dropouts. The best predictive model in terms of model accuracy statistics was deployed in the Higher Education Data Analyzer (HEDA)[3] management information reporting and decision support system in order to generate a means of indicating students with a high probability of dropping out by the second year of study.

This article in a South African context is important for illustrating the use of a practical model within institutional planning in higher education. Being able to predict more accurately which students are likely to drop out will enable institutions to implement focused intervention strategies to assist the students concerned, especially during the first year of study.

REVIEW OF THE STUDENT RETENTION LITERATURE

Many university entrants are not sufficiently prepared to master the academic requirements when entering university. Research studies abroad have highlighted the importance of preparing new entrants for higher education (Woodhead 2002; Herzog 2005; Biswas 2007; Hess 2008; Bottoms and Young 2008; Dekker et al. 2009).

The well-known integration-commitment model of attrition developed by Tinto (1975) and later modified by Pascarella and Terenzini (1983) has been used extensively 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) obligation to earning a degree (goal commitment), and (c) commitment to an institution.

Publications on factors influencing retention rates in South Africa have evolved and include, amongst others, those by Letsaka and Maile (2008), Scott et al. (2007), Lourens and Smit (2003), Van Zyl et al. (2012) and Murray (2014). Several studies considered the relationships between students' background and their retention (Astin and Oseguera 2012; Pike et al. 2014). The effects of the characteristics of entering cohorts (such as race, gender and socio-economic status) on student retention have also been studied extensively. Furthermore,

Pike (2013) reported that institutional academic support was positively related to student retention and graduation rates. Webber and Ehrenberg (2010) indicated in their research that funding student support was positively related to student retention and graduation rates but that expenditure on instruction was not. They also found that entering qualifications (such as grade 12 aggregates) were positively related to student retention and graduation rates. A study by Murray (2014) reported that financial aid and residence-based accommodation helped students who would eventually graduate to do so quicker.

Arulampalam et al. (2004) studied the factors affecting the probability of first-year medical students dropping out in the United Kingdom. They found that the probability of this happening during the first year of study was influenced greatly by the subjects studied and the marks achieved. They also found that the location of a student's accommodation had a significant influence on the probability of progression. Liu (2000) reported that the relationship between integration and satisfaction is important to the success of academic performance and persistence and that student satisfaction is highly related to student retention, whereas dissatisfaction is key to academic withdrawal. Yorke (1999) identified three primary causes of full-time students abandoning their courses: a mismatch between students and their choice 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'.

It is clear that the higher education literature provides a wide range of research studies on the reasons why students drop out from tertiary courses. This article demonstrates a practical contribution to securing student retention, using predictive analytic models to enable an institution to implement student-specific intervention strategies.

INSTITUTIONAL CONTEXT

The interest of CPUT's Department of Institutional Planning in applying predictive analytics to identify 'at risk' students began in earnest in 2014, shortly after the appointment of the university's new vice-chancellor, who initiated a project called 'Know Our Students', with a view to understanding as much as possible about the characteristics of the students at the institution. Descriptive reports regarding the demographic profile of the students, including their high school and its quintile, were among the early outputs of this project.

Also in 2014, the university piloted a 'First-year Experience' project, funded by a Teaching Development Grant from the Department of Higher Education and Training (DHET), with the objective of reducing the number of dropouts between first year and second year. The average second-year dropout rate for three-year national diplomas at CPUT for the period 2005

to 2014 was 23.6 per cent.

The First-year Experience includes a number of related initiatives, including web-based videos on university life and learning tips, the appointment of retention officers, and the provision of tutors in academic programmes, teaching assistants for 'at risk' subjects, and mentors for first-year students. The objective is to reduce the second-year dropout rate and to increase the university's pass rates and throughput rates.

Given that limited resources do not allow the provision of additional support for all first-year students, the question of how to identify those students most in need of intervention is natural. One way of doing this, which CPUT had been aware of for some years but had not yet implemented, was by means of its electronic learning platform, Blackboard.[4] E-learning systems like Blackboard can provide real-time data to lecturers on the engagement of their students with the system and their performance in any system-related tasks. Blackboard can also provide the students with information on how they are performing, relative to others in their cohort. This is considered important information and CPUT intends purchasing the Blackboard Analytics module to facilitate this reporting.

Within the Department of Institutional Planning, the focus was on providing information on students as soon as possible – perhaps even before they began their first year of study. In terms of an early-warning timeframe, this would be the earliest alert, followed by real-time information from Blackboard, followed by results in formal evaluations during the course of the first year. Provided the university could put in place procedures for responding to the resultant information on students at risk in meaningful and sensitive ways, it should be possible to intervene and provide additional support for most students who might become second-year dropouts.

But how to do the predictive analytics? During 2013 and 2014, two members of staff in CPUT's HEMIS Office (which is part of the Department of Institutional Planning) were engaged in studies for the BTech Quality degree. One staff member undertook a data mining project on the National Diploma in Information Technology as part of her BTech.

This study made use of the CHAID (Chi Square Automatic Interaction Detection) tree model to identify variables from historical student biographical data that were significant predictors of students graduating or dropping out. While this study was progressing, the Department of Institutional Planning entered into discussion with IDSC, a private company based in Potchefstroom with the objective of producing an automated tool for predictive analytics, using the capabilities of the Structured Query Language (SQL) Server, already part of the HEDA infrastructure.

To obtain a better idea of the possibilities of this and a greater understanding of the issues involved in predictive analytics generally, the department invited IDSC and a member of the University of Pretoria's Department of Higher Education Research and Innovation to a one-day workshop with departmental staff in October 2014. From discussions at the workshop, the following points became clear to the department:

- There is no single algorithm that provides reliable predictions across a range of qualifications. Significant predictors in one field of study may be insignificant in another.
- CPUT would need to gather additional information about incoming students in order to apply predictive analytics before they begin their studies. CPUT gathers certain information during the application and registration processes, but some data that have proved significant in University of Pretoria studies have not been sought. To take one example: Is the student the first generation in his or her family to study at a university?
- Blackboard could provide real-time descriptive analytics based on student engagement with the E-learning system, but it would struggle to produce predictive analytics of the type required.

In 2015, as described in this article, the predictive analytic component of the Know Our Students project took a somewhat different tack. In the absence of pre-registration data, IDSC undertook to apply predictive analytic procedures using KNIME to identify students enrolled in 2015, for the same National Diploma in Information Technology (IT) qualification, who might be in danger of dropping out.

METHODOLOGY

The six steps in the Cross-Industry Standard Process Model for Data Mining (CRISP-DM), namely, business understanding, data understanding, data preparation, modelling, evaluation, and deployment (Abbott 2014), were applied in this predictive modelling project.

Algorithms for predictive modelling are commonly divided into supervised and unsupervised learning methods. In this case, supervised learning models were used whereby the *supervisor* is the target (dependent) variable, which is a column in the data representing values to predict from the other columns (independent variables) in the data. The algorithms used are those most commonly employed in predicting student success or drop out (Herzog 2006).

The purpose of the study was to predict first-time entering student dropouts before or in the second year of study for a particular qualification at CPUT using various predictor or

independent variables. A first-time entering student is understood as one enrolling for the first time at any higher education institution in a particular year. Second-year student dropouts are those who did not register for the second year of study.

Based on discussions with CPUT, it was decided to focus on the National Diploma in IT for this case study. Institutional operational data for cohorts of first-time entering students enrolled for the qualification at CPUT from 2008 to 2014 were used in the analyses to predict student dropout by the second year of study. KNIME was used to perform the predictive modelling with second-year dropouts as the dependent/target variable. In order to get labels for the supervised learning of predictive models, second-year dropout students were classified in the following way: if the student did not return in the second year of study (i.e. dropped out) he or she was coded as 1, and as 0 if the student returned in the second year of study.

A variety of information on background and demographics (pre-university information) and performance-linked data (first-semester data) from CPUT was extracted, prepared and cleaned in HEDA and used to model the data. As part of the data preparation, a total of 27 variables were used and tested for collinearity in KNIME after which 22 variables were included in the descriptive analysis. From these, only selected variables were used in the analyses since redundant variables were excluded and some variables had to be transformed or combined to prevent overfitting. For example, the modules taken by first-year students were clustered and binned in six categories. The final list of eight prediction variables included in the analyses were: binned first-year module marks[5] in Technical Programming (TP), Information Systems (IS), Development Software (DS), Information Technology Skills (ITS), Systems Software (SS) plus Financial Aid (NSFAS) (Yes or No), Grade 12 Mathematics (Yes or No), and Type of accommodation (resident student or not).

A total of 1 593 student records were used after all missing data had been either removed or binned. In all, 452 (28%) of the 1 593 students were classified as second-year student dropouts. The data were imported from HEDA using SQL directly in KNIME.

The Logistic Regression, Naïve Bayes, and Decision Tree algorithms in KNIME were used in this study. Logistic regression is a linear classification technique that models the relationship between a binary dependent (target) variable and categorical and/or continuous independent (predictor) variables. The logistic regression model uses the predictor variables to predict the probability that the target variable takes on a given value. No two or more predictor variables in a regression model should be highly correlated and a test for collinearity is therefore important.

The Naïve Bayes classifier is based on Bayes' theorem with independence assumptions

between predictors. Naïve Bayes classifiers can handle an arbitrary number of independent variables whether continuous or categorical. Naïve Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the 'naive' assumption of independence between every pair of features (Ng and Jordan 2002).

A decision tree builds classification or regression models in the form of a tree structure. The tree structure is generated by breaking down a dataset into smaller subsets while at the same time an associated decision tree is incrementally developed (Everitt 1998). The classification tree has a built-in variable selection and is non-parametric, that is, no assumptions about distributions for inputs or the target variable are needed. It handles missing data automatically and visually and intuitively supports the results from the logistic regression analysis.

The most important steps before modelling are the test for collinearity and partitioning the data. The typical process for building a model includes constructing it on training data and assessing the model on testing data. Predictive analytics is an iterative process whereby the settings and parameters and/or inputs to the model are changed and the model is re-built on training data and the new model assessed on testing data. This process is repeated until the best 'fit' is created. In order to avoid over-fitting in this process, a third data set (validation data) is used to provide a final estimate of the predictive model's performance or accuracy once it is deployed.

The 2008–2013 data set was randomly subdivided into 70 per cent training and 30 per cent testing data sets; the 2014 data set was kept aside for later use as the validation data set. The Logistic Regression, Naïve Bayes, and Decision Tree training data sets contained 988 records of first-time entering students with which to build the models in order to identify the predictor variables. The remaining 424 records were used as the testing data set to assess the accuracy of the models. As a final step, 181 records from the 2014 data (validation data set) were used to predict the outcome based on the selected model. The Logistic Regression, Naïve Bayes, and Decision Tree nodes were applied to the eight independent variables listed above.

The adequacy of the three models was assessed by making use of the following statistical measures: a confusion matrix indicating the percentage correctly predicted, the sensitivity and the specificity, Cohen's kappa, the area under the receiver operating characteristic curve (AUC), and the validation data set.

RESULTS

The Logistic Regression, Naïve Bayes, and Decision Tree learner and predictor nodes in

KNIME were used to conduct the analyses. The logistic regression scorer retained the following four items as significant predictor variables (at a 5% level of significance, $p < 0.05$) in the model: Financial Aid, the first-year module marks (binned) for Development Software (DS), Information Systems (IS), and Information Technology Skills (ITS). It is evident from Table 1 that the DS Mark, IS Mark, ITS Mark, and Financial Aid (NSFAS_Bursary_Y_N=X) were very strong predictor variables when the p-values (indicated as $p > |z|$ in the last column of Table 1) for the different predictor variables were compared. Therefore, students enrolled for the National Diploma in IT who did not receive any bursary and had low marks for the DS, IS and ITS modules were more likely to drop out by the second year of study.

Table 1: Statistical results for Logistic Regression second-year dropout model

Statistics on Logistic Regression

Logit	Variable	Coeff.	Std. Err.	z-score	P>|z|
0	NSFAS_BURSARY_Y_N=X	-1.4202	0.5309	-2.6749	0.0075
	NSFAS_BURSARY_Y_N=Y	-0.8435	0.5537	-1.5233	0.1277
	Res=Y	0.3783	0.2869	1.3186	0.1873
	DEVELOPMENT_SOFTWARE_Mark_Bin	0.6076	0.0788	7.7129	1.05E-14
	INFORMATION_SYSTEMS_Mark_Bin	0.3928	0.0773	5.0829	3.72E-7
	INFORMATION_TECHNOLOGY_SKILLS_Mark_Bin	0.3313	0.0715	4.6312	3.63E-6
	SYSTEMS_SOFTWARE_Mark_Bin	0.0243	0.0709	0.3429	0.7317
	TECHNICAL_PROGRAMMING_Mark_Bin	0.0802	0.0766	1.047	0.2951
	MATH	0.0742	0.1039	0.7138	0.4753
	Constant	-1.849	0.5617	-3.2919	0.001

The confusion matrix for the Logistic Regression model in Table 2 represents the counts in each quadrant based on a predicted probability threshold of the model of 0.5. The overall percentage correctly classified is 88 per cent. The actual second-year dropout students (dropout = 1) classified correctly are 73.3 per cent (Sensitivity).

Table 2: Confusion matrix for Logistic Regression dropout model

	True positive	False positives	True negatives	False negatives	Sensitivity	Specificity	F-measure	Accuracy	Cohen's Kappa
Dropout = 0	299	28	77	20	0.937	0.733	0.926		
Dropout = 1	77	20	299	28	0.733	0.937	0.762		
Overall								0.887	0.688

A Decision Tree and Naïve Bayes model were included in the study to compare with the Logistic Regression model in terms of best fit. The Decision Tree visually and intuitively

supported the results from the regression model.

The model accuracy in terms of the area under the curve (AUC), percentage correctly classified (PCC), and error rates are displayed in Table 3 for the Logistic Regression, Naïve Bayes and Decision Tree models. The models performed well and the Logistic Regression model had the highest percentage accuracy (88.6%) with an 11.3 per cent error rate and was subsequently used on the validation data set**.**

Table 3: Comparison of the accuracy of the Logistic Regression, Naïve Bayes, and Decision Tree models

Statistic	Logistic Regression	Decision Tree	Naïve Bayes
AUC*	0.9159	0.8457	0.9194
PCC† (%)	88.6	87.5	87.7
Error (%)	11.3	12.5	12.3

*Area under the receiver operating characteristic curve.
†Percentage correctly classified.

The most rigorous test for determining the accuracy of a logistic model is to apply the model to a validation data set. If the model is accurate in its predictions when applied to an independent validation data set (in this case the 2014 data set) that was not used in the initial training of the model parameters, then the model is of true value. Our model performed well in this regard; the logistic model predicted 86.2 per cent of the dropout students correctly in the validation data set. The model could be deployed since the same variables in the training data set were significant in the validation data set and the predictive accuracies were similar.

One of the advantages of predictive modelling is that once a model has been finalised, it can easily be deployed over live data to score the data in real time. Many publications on student retention indicate only the factors influencing the *retention* of students and not the most useful step, namely, *live scoring*. It is easy to gain access to data or to extract data from an institutional database, but to write the results back to the database in order for decision-makers to get results in real time is not always possible. It is also important continuously to improve the predictive models by modifying them periodically by bringing in more factors in order for the models to stay adaptive and remain useful.

Since the main purpose of this project was to provide a list of names of students at risk of dropping out by the second year of study in order to better inform the teaching and learning activities, the individual student dropout probabilities were exported from KNIME into the HEDA management information system used by CPUT. A HEDA report was developed for use

by the Department of Institutional Planning and by the head of department of the specific academic programme, showing the cohort statistics for the programme as well as the list of students identified as possibly being at-risk, based on the Logistic Regression predictive model.

Additional validation of the generated list of at-risk students was provided by a test conducted by Institutional Planning. In checking the names of first-time entering students in the list against the corresponding students on the university's database as at 23 July 2015, we found that six of the students listed had already cancelled their registration. Five of these six students were identified as 'high risk' by the model (with the highest risk percentages), whereas the sixth was identified as 'medium risk'.

The head of the department of the specific programme involved at CPUT will be able to use the list generated from the predictive analytics model to contact students with a high probability of dropping out by the second year of study and provide personalized support where needed.

SUMMARY AND RECOMMENDATIONS

Great attention is paid to understanding the enrolment behaviour of students and many 'first-year experience' projects have been implemented by universities in South Africa in an attempt to improve student retention. A key benefit of predictive analytics in this context is that it can assist higher education institutions to implement targeted intervention strategies in the first year of study in order to reduce the number of students leaving prematurely by their second year.

In this article we presented a predictive analytics case study demonstrating the use of three algorithms on first-year students enrolled for the National Diploma in Information Technology at CPUT. Our results show accuracies of more than 80 per cent in terms of model fit using only pre-university and first-semester data. The test marks for three of the modules in the first semester proved to be very significant predictors of second-year dropouts.

It is important for the university to identify potential at-risk students early in the first year of study in order to implement institutional support and intervention programmes to improve the retention rate in the students' second year. The relevance of this investigation is demonstrated by the practical application of real-time scoring by providing a list of names of students at risk of dropping out by the second year of study. Using a predictive analytics tool such as KNIME and importing/exporting the data/results in a management information system such as HEDA provides a list of students to be targeted for additional academic support.

For each student who can be rescued from the existing 'revolving door' of higher education and guided to graduation, there is immense benefit to the person concerned. But there

is also benefit to his or her family, to the community, and to the economy. For the university, its throughput rate will improve and satisfaction gained in knowing that it is making a greater contribution to society.

There is also an economic incentive for the university. In this instance, if CPUT can reduce its second-year dropout rate for this particular qualification by only 25 per cent, an additional income from tuition fees of at least R256,500 can be realised. This is not taking into account the teaching input and output subsidy income, as well as avoiding penalties that might be incurred by failing to meet enrolment targets as agreed with the Department of Higher Education and Training.

It is recommended that, for future studies, the incorporation of questionnaire data, based on a survey of incoming students at CPUT, in the analyses would result in even more accurate predictions of the second-year dropout characteristics of students.

NOTES

1. More information regarding HEDA can be found at http://www.heda.co.za
2. Konstanz Information Miner (KNIME) is an open source platform for data integration, processing, analysis and exploration and is available at www.knime.org
3. More information regarding HEDA can be found at http://www.heda.co.za
4. www.blackboard.com
5. The first-year module marks were binned in six categories; Bin 0 if module was not taken, Bin 1 if module mark was less than 30 per cent, Bin 2 if module mark was less than 40 per cent, Bin 3 if module mark was less than 50 per cent, Bin 4 if module mark was less than 60 per cent, and Bin 5 if module mark was higher than or equal to 60 per cent.

REFERENCES

Abbott, D. 2014. *Applied predictive analytics: Principles and techniques for the professional data analyst*. John Wiley & Sons Inc. Indiana.

Arulampalam, W., R. Naylor and J. Smith. 2004. Factors affecting the probability of first-year medical student dropout in the UK: A logistic analysis for the entry cohorts of 1980–1992. Warwick Economic Research Papers. No 618. Department of Economics, University of Warwick, Coventry. *Medical Education* 38(5): 492–503.

Astin, A. W. and L. Oseguera. 2012. Pre-college and institutional influences on degree attainment. In *College student retention: Formula for student success,* ed. A. Seidman, 119–146. 2nd edition.

Biswas, R. R. 2007. *Accelerating remedial math education: How institutional innovation and state policy interact*. Boston, MA: Jobs for the Future.

Bottoms, G. and M. Young. 2008. *Lost in transition: Building a better path from school to college and careers*. Atlanta, GA: Southern Regional Education Board.

Campbell, J. P., P. B. DeBlois and D. G. Oblinger. 2007. Academic analytics. *EDUCAUSE Review Online*. https://net.educause.edu/ir/library/pdf/erm0742.pdf (accessed 16 February 2016).

CHE, *see* Council on Higher Education.

Council on Higher Education. 2013. *A proposal for undergraduate curriculum reform in South Africa:*

The case for a flexible curriculum structure. Pretoria: CHE.

Council on Higher Education. 2014. *VitalStats: Public Higher Education 2012*. Pretoria: CHE.

Council on Higher Education. 2010. *Access and throughput in South African higher education: Three case studies*. Pretoria: CHE.

Dekker, G. W., M. Pechenizkiy and J. M. Vleeshouwers. 2009. *Predicting students drop out: A case study*. Educational Data Mining.

Delmater, R. and M. Hancock. 2001. *Data mining explained: A manager's guide to customer-centric business intelligence*. Boston: Digital Press.

Eduventures, 2013. *Predictive analytics in higher education. Data-driven decision-making for the student life cycle*. http://www.eduventures.com/wp-content/uploads/2013/02/Eduventures_ Predictive_Analytics_White_Paper1.pdf (accessed 16 February 2016).

Everitt, B. S. 1998. *The Cambridge Dictionary of Statistics*. Cambridge: Cambridge University Press.

Greller, W. and H. Drachsler. 2012. Translating learning into numbers: A generic framework for learning analytics. *Educational Technology & Society* 15(3): 42–57.

Herzog, S. 2005. Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. *Research in Higher Education* 46(8): 883–928.

Herzog, S. 2006. Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research*: 17–33.

Hess, F. M. 2008. *Still at risk: What students don't know, even now*. Washington, DC: Common Core.

Long, P. D. and G. Siemens. 2011. Penetrating the fog: Analytics in learning and education. *Journal of Interactive Online Learning*. EDUCAUSE Review Online. http://er.educause.edu/articles/ 2011/9/penetrating-the-fog-analytics-in-learning-and-education (accessed 16 February 2016).

Luan. J. 2002. Data mining and its applications in higher education. *New Directions for Institutional Research* Spring: 17–36.

Letsaka, M., S. Maile. 2008. *High university dropout rates: A threat to South Africa's future*. Human Science Research Council.

Liu, R. 2000. Institutional integration: An analysis of Tinto's Theory. Paper presented at the 40th Annual Forum of the Association for Institutional Research Cincinnati, Ohio, May 21–24.

Lourens, A. and I. Smit. 2003. Retention: Predicting first-year success. *South African Journal of Higher Education* 17(2): 169–176.

Murray, V. E. 2008. *The high price of failure in California: How inadequate education costs schools, students, and society*. San Francisco: Pacific Research Institute.

Murray, M. 2014. Factors affecting graduation and student dropout rates at the University of KwaZulu-Natal. *South African Journal of Science* 110(11/12), Art. #2014-0008, 6 pages. http://dx.doi.org/10.1590/sajs. 2014/20140008 (accessed 28 July 2015).

National Planning Commission. 2011. *National Development Plan: Vision for 2030*. November. Pretoria.

NDP, *see* National Planning Commission.

Ng, A. Y. and M. I. Jordan. 2002. *On discriminative vs generative classifiers: A comparison of logistic regression and naïve Bayes*. University of California, Berkeley. http://ai.stanford.edu/~ang/ papers/nips01-discriminativegenerative.pdf (accessed 28 July 2015).

Pascarella, E. and P. Terenzini. 1983. Predicting voluntary freshman year persistence/withdrawal behaviour in a residential university: a path analytic validation of Tinto's model. *Journal of Educational Psychology* 75(2): 215–226.

Pike, G. R. 2013. NSSE benchmarks and institutional outcomes: A note on the importance of considering the intended uses of a measure in validity studies. *Research in Higher Education* 54: 149–170.

Pike, G. R. and S. S. Graunke. 2015. Examining the effects of institutional and cohort characteristics on retention rates. *Research in Higher Education* 56: 146–165.

Pike, G. R., M. J. Hansen and J. E. Childress. 2014. The influences of students' pre-college characteristics, high school experiences, college expectations, and initial enrolment characteristics on degree attainment. *Journal of College Student Retention* 16: 1–23.

Scott, I., N. Yeld, J. Hendry. 2007. A case for improving teaching and learning in South African higher education. *Higher Education Monitor* 6. Pretoria: CHE. http://www.che.ac.za/documents/d000155/index.php (accessed 28 July 2015).

Scott, M., T. Bailey and G. Kienzl. 2006. Relative success? Determinants of college graduation rates in public and private colleges in the U.S. *Research in Higher Education* 47: 249–279.

Tinto, V. 1975. Dropout from Higher Education: A theoretical synthesis of recent research. *Review of Educational Research* 45: 89–125. http://dx.doi.org/10.3102/00346543045001089 (accessed 28 July 2015).

Tinto, V. 1993. *Leaving college: Rethinking the causes and cures of student attrition*. 2nd Edition. Chicago: University of Chicago Press.

Van Zyl, A., S. Gravett and G. P. de Bruin. 2012. To what extent do pre-entry attributes predict first year student academic performance in the South African context? *South African Journal of Higher Education* 18(1): 1095–1111.

Verbert, K., N. Manouselis., H. Drachsler and E. Duval. 2012. Dataset-driven research to support learning and knowledge analytics. *Educational Technology & Society* 15(3): 133–148.

Watters, A. 2012. *Learning analytics: Lot of education data… Now what?* Hack Education. http://hackeducation.com/2012/05/04/learning-analytics-lak12 (accessed 16 February 2016).

Webber, D. A. and R. G. Ehrenberg. 2010. *Do expenditures other than instruction affect graduation and persistence rates in American higher education?* (Working Paper 121). Ithaca, NY: Cornell University Higher Education Research Institute.

Woodhead, C. 2002. *The standards of today and how to raise them to the standards of tomorrow.* London, UK: Adam Smith Institute.

Yorke, M. 1999. *Leaving early. Undergraduate non-completion in higher education.* London and Philadelphia: Falmer Press.