Identification of Early Predictors of Adult Learners' Academic Performance in Higher Education

Sylvia Chong Nguik Yin

Abstract

Universities are inundated with detailed applicant and enrolment data from a variety of sources. However, for these data to be useful there is a need to convert them into strategic knowledge and information for decision-making processes. This study uses predictive modelling to identify at-risk adult learners in their first semester at SIM University, a Singapore University that caters mainly to adult learners. Fourteen variables from the enrolment database were considered as possible factors for the predictive model. To classify the at-risk students, various algorithms were used such as a neural network and classification tree. The performances of the different models were compared for sensitivity, specificity and accuracy indices. The model chosen is a classification tree model that may be used to inform policy. The implications of these results for identification of individuals in need of early intervention are discussed.

Keywords: predictive modelling; adult learners; higher education.

Introduction

The ease of data collection and advances in information technologies, such as storage capability, processing power and access speed, has enabled educational institutions to accumulate vast amounts of data. Universities and their enrolment offices are inundated with detailed applicant and enrolment data from a variety of sources, such as student demographics, professional experience and academic background. However, for these data to be useful there is a need to convert them into strategic knowledge and information for decision-making processes. Over the past decade, data mining has gained increasing attention in academia to generate data driven evidence (Koh and Chong, 2014). Data mining approaches can discover hidden relationships and patterns. These relationships and patterns can, in turn, be developed into models to predict students' performance and behaviour. The predictive models can develop knowledge and insights, on which informed and strategic decisions can be made.

The purpose of this study is to develop predictive models to identify early predictors of academic performance of adult learners who are enrolled in the part-time undergraduate programs at SIM University, Singapore. SIM University is Singapore's only privately-funded university dedicated to working adults. The University has provided pathways for many to pursue lifelong learning and higher education while balancing career, family and social responsibilities (SIM University, 2014).The research scope was developed in the context of the SIM University's enrolment process. The research is timely and significant because of the growing number of adult learners who are returning to higher education (Macfadgen, 2007). his paper focuses on the factors that predict adult learners who may be academically at-risk and proposes incorporating into the enrolment process a predictive model to identify potential at-risk students.

Context of Study

The profile of students in higher education is changing (Chong, Loh and Babu, 2015). There is an increasing number of non-traditional students – these are students who are not in the group of 18-22 year-old full-time undergraduates (Wyatt, 2011; Macfadgen, 2007). There are 13,369 adult students currently enrolled in SIM University (SIM University, 2014). This is significant, as more and more adults who have been out of school for some years are turning to higher education institutions to start, continue or complete undergraduate degrees. In August 2012, the Singapore government declared support for the continuing higher education sector by expanding and diversifying the pathways in higher education (MOE, 2012). The restructuring of higher education pathways and institutions ensures that Singapore develops a more competitive workforce. The Singapore government, in a bid to encourage and support lifelong learning and continuing education, has made available a range of financial support instruments such as government subsidised bursaries and tuition loan schemes for adult learners to take up part-time undergraduate programs in SIM University (MOE, 2012). SIM University must modify and target their enrolment and admission strategies to better serve this growing population of adult learners. It is important for SIM University to identify and profile students who will eventually succeed, as well as applicants who will struggle or are inappropriate for admission.

With growing participation adult learners in higher education, SIM University must sift through an increasing number of applications. Making informed enrolment decisions will require accurate data and analysis for evidenced-based insights and knowledge discovery. Incorporating into the enrolment process predictive models to identify potential at-risk students or student success is highly advantageous. A combination of an explicit knowledge base together with sophisticated analytical approaches and clear domain information can uncover patterns, associations and/or relationships to support enrolment management. By analysing enrolment data, it is possible to develop models that will be able to predict the potential of in-coming students.

Review of Literature

The review of literature is organized in two parts. The first section includes an overview of the data mining process and its use in higher education. The second section provides a review of studies on predictors of academic performance in higher education.

Use of Data Mining in Higher Education

Data mining has emerged in the wake of higher education's ability to capture a rapidly growing amount of data to "develop models for improving learning experiences and improving institutional effectiveness" (Huebner, 2013). The data mining process is often initiated without any preconceived outcomes; it adopts a data analysis methodology (Chong, Mak and Loh, 2016) and is often interchangeable with the term Knowledge Discovery in Databases (KDD) (SPSS, 2009) with the aim of obtaining insightful and useful findings (Giudici, 2013). In its basic form, the data mining process is the extraction of the knowledge within large databases. The data mining process involves several phases among which are: data acquisition, feature selection and extraction from database, model development and pattern recognition using data mining techniques, model interpretation and knowledge generation. Data mining, used in higher education, can strategically combine selected institutional data and statistical analysis to generate information upon which students, educators, administrators and management can improve practices. This highlights the importance of data mining as an approach to build models by transforming raw data into usable knowledge and information (Giudici, 2013).

Chang's (2009) study used data mining techniques to develop a model to predict the academic performance of university applicants. The predictive model was developed based on variables taken from the university's integrated admissions system. The integrated system included databases on application, enrolment and student progress data. The study showed that the neural network and decision tree models developed were able to inform university recruitment strategies, as well as support institutional research. Ramaswami and Bhaskaran (2010) used a CHAID (Chi-square Automatic Interaction Detector) prediction model, based on a classification tree, to identify a set of predictive variables and assess the impact of these variables on the academic performance of university students. A pilot experiment with 224 students from two different universities along with 35 variables was conducted. The model showed a strong correlation between attributes such as location, school type, parents' education, secondary school grades and the students' performance at the universities. Kovacic (2010) developed prediction models of students' success based on enrolment data with statistical techniques such as CART (Classification and Regression Technique) and QUEST (Quick, Unbiased and Efficient Statistical Tree) classification tree methods. He concluded that classifying students based on pre-enrolment data helps to identify students who may be at risk, and recommended orientation, advising and mentoring programs to support these students.

The literature also indicated that algorithmic or data mining approaches to develop predictive models could provide notable results vis-à-vis traditional statistical modelling approaches (Li, Nsofor and Song, 2009; Bogard, James, Helbig & Huff, 2012). Vandamme, Meskens and Superby (2007) used decision trees, neural networks and linear regression for the early identification of three categories of first-year students: low-, medium- and high-risk students. Some of the demographics and academic variables of these students were significantly related to academic performance. Such predictions are useful to identify and support students with appropriate interventions to improve their academic performance.

Predictors of Academic Performance

The antecedents to success in university prior to students' matriculating are well established. Evidence exists to show that pre-university academic performance has a significant impact on subsequent academic performance in university. The relationship between pre-university grades and university performance has been validated in studies (Iam-On and Boongoen, 2015; Adelman, 2006). Adelman's (2006) research on persistence points to the importance of both pre-university (high school) performance and the rigor of the high school curriculum. Iam-On and Boongoen (2015) affirmed the importance of pre-university grade-point average in predicting success in university. However, the predictive ability of pre-university school grades is different for different individuals and groups. Power, Robertson, and Baker (1987) showed that the correlation between pre-university/high school grades and Grade Point Average (GPA) at university is generally about 0.5. They also found that secondary school grades are not as predictive for mature students' performance as they are for school leavers' performance. According to Bhardwaj & Pal, (2011) personal, social, psychological and environmental variables have an impact on students' academic performance. Other variables such as living location, medium of teaching, mother's qualifications, and family annual income also potentially affects student performance (Bharadwaj & Pal, 2011). Demographic variables that have been found to be determinants of academic performance include age, gender, employment responsibilities, and student workload (Palmer, Bexley and James, 2011).

In addition, pre-university factors that are commonly associated with individuals most at risk include: low pre-university school grade-point average (Adelman, 2006), low SAT/ACT scores; minority status (Pascarella and Terenzini, 2005), low family education levels, and low family income (Eagle and Tinto, 2008). Other variables that contribute include non-cognitive factors such as motivation, aspirations (Eagle and Tinto, 2008), and tendencies toward social and academic integration (Braxton and Hirschy, 2005; Eagle & Tinto, 2008). These non-cognitive factors are also seen as predictors of academic performance. Investigating the interaction of more traditional risk factors, such as demographics, with early engagement indicators can lead to a richer understanding of the predictors of success for students.

Research Objectives

The key purpose of this study is to identify early predictors of academic performance during the adult learners' initial semester at university using a data mining approach. Through this, SIM University hopes to identify students or applicants who are academically at risk as early as possible. Decision trees are used to build these models so that appropriate enrolment and intervention strategies can be designed and implemented. Specifically the study aims to achieve the following research objectives:

- Identify characteristics that are available at application and early engagement variables of adult learners who are academically (GPA) at-risk in higher education
- Build models for early prediction of the academically at risk with the identified application characteristics and early engagement variables
- Evaluate these models using cross-validation

Research Methodology

Data Source

The target sample for this study was comprised of first-year students who started their part-time degree program at SIM University in January and July 2013. Data was extracted from an in-house student information management system which collects and catalogues data from numerous sources within the admissions office as well as in other divisions of the university. For the purpose of this study, a range of demographic and academic data was extracted for 2,392 students that fall within the target sample.

Data Understanding

In order to identify potentially useful and credible patterns in the data, several iterative steps were taken in the development of the enrolment model. Students with missing data were removed from the dataset because some data mining algorithms were not able to handle missing data. To assist with data understanding, profiling was also conducted to determine the proportion of at-risk students in relation to the overall student participation rates.

Profile. Students studying in SIM University were relatively equally distributed in terms of gender, with more than half of the sample being between 21 and 25 years of age (M = 26.5, SD = 4.98). 22.7% of the students were identified as at risk students based on their Pre-University Cumulative Grade Point Average (CGPA) (See Table 1 for more details).

Table 1. Profile of Sample Used for Modelling

	N	%
Gender		
Female	1147	48.0
Male	1245	52.0
Age		
21 to 25	1387	58.0
26 to 30	599	25.0
Above 31	406	17.0
At Risk Status		
At Risk	542	22.7
Not At Risk	1850	77.3
SIM University School		
SASS	546	22.8
SBIZ	848	35.5
HDSS	291	12.2
SST	707	29.6

Variables. Predictors for the study can be broadly categorized into the three following groups. The description of the variables is presented in Table 2.

- Demographic variables (gender, age, marital status, race, length of working experience]
- Pre-SIM University academic performance indicators (prior diploma school, diploma CGPA, years since they last studied, field of diploma study, relevance of previous diploma study to current degree, O-Levels English and O-levels Mathematics) and
- University variables (SIM University schools and Credit Units (CUs) registered).

Demographic variables. As a university dedicated to adult learners, SIM University's enrolment is typically characterized by a diverse student profile in terms of their race, marital status, age and working experience. In view of this, demographic variables are of particular interest as students at different life stages handle the demands of a university program differently.

Pre-SIM University academic performance indicators. The concept of students' innate academic aptitude and its impact on their ability to cope with the demands of a university education has been discussed in literature (Pascoe, McClelland and McGaw, 1997). In view of this, proxy indicators like O-levels English and O-levels Mathematics, subjects that most students offer at national examinations were collected to represent the students' academic competence. In the same token, diploma GPAs and the field of their diploma studies may also serve as a good gauge of the student's aptitude in respective programmes.

SIM University variables. SIM University's programmes are offered by four schools that cover a range of disciplines: School of Arts and Social Sciences (SASS), School of Business (SBIZ), School of Human Development and Social Services (HDSS), and School of Science and Technology (SST). As it is likely that programs offered by each school required and emphasised different domain knowledge and skills, it is insightful to identify and study between-school differences on the students' academic performance. Course workload is another predictor of interest in this study. To capture course workload, the number of Credit Units (CUs) that the students registered for at the start of the semester is used as a proxy.

Table 2. Variables used for data exploration and analysis

Variable Role Level	Role	Description/ Type
At Risk Indicator	Target	Student academic risk status (binary: at risk or not at risk)
Demographic Variables:		
Gender	Predictor	Student gender (binary: male or female)
Age	Predictor	Student age at intake (numeric)
Marital Status	Predictor	Student marital status (binary: Single or Married)
Race	Predictor	Student race (nominal: Chinese, Malay, Indian or Others)
Work Experience	Predictor	Student length of working experience in months (numeric)
Name of Diploma Awarding Institution (DIP Institution)	Predictor	The institution that student obtained Diploma from (nominal: A, B, C, D, E)
Diploma CGPA	Predictor	Student diploma final CGPA attained (numeric: 0.0 to 4.0)
Years since study	Predictor	Number of years since student last studied (numeric)
Field of Diploma Study	Predictor	The diploma area of study that student previously graduated from (nominal: Engineering, Business,
Relevance of Diploma Study	Predictor	Whether student diploma field of study is relevant to the degree he/she is pursuing (binary: relevant or not relevant)
Mathematics "O" level Grade	Predictor	Student previous Math grades (ordinal: 1 to 9)
English "O" level Grade	Predictor	Student previous English grades (ordinal: 1 to 9)
SIM University Variables		
SIM University School	Predictor	The school which the student is currently enrolled in (nominal: SASS, SBIZ, HDSS, SST)
CUs Registered	Predictor	Number of credit units student registered for that semester (numeric)

Modelling

A binary target variable 'at risk' was also constructed where students with a CGPA score of 2.3 and below is flagged as at risk while those with a CGPA score of above 2.3 is flagged as not at risk. The threshold CGPA cut-off of 2.3 was used to be consistent with SIM University's practice of offering academic counselling to students with a CGPA score of 2.3 and below.

After data preparation, a data driven approach was used to select statistically significant predictors. Using IBM SPSS Modeler 14.1, a list of significant predictors was identified using Model Feature Selection node, Neural Networks, CHAID, C5.0 and CRT based on their statistical significance (p-value <.05). As each algorithm has its own computation methodology strengths, comparing the list of predictors chosen by different algorithms offered a balanced and insightful approach in short listing variables that are consistently important for subsequent modelling. This controlled for variable selection bias. The list of short listed variables was then evaluated based on inputs from the literature as well as by subject matter experts who have contextual knowledge of the workings of UniSIM and the Singapore education landscape.

In model building, the CHAID decision tree was chosen as the baseline decision tree among the other decision trees that were developed via different algorithms on the full dataset (N = 2,392). The selection of the baseline model was based on an evaluation of a basket of criterion which measured the models' specificity, sensitivity, accuracy, and G-mean (Kubat, Holte & Matwin 1997). Collectively, the different criteria represented the models' ability to correctly classify at risk students, correctly classify not at risk students, and measure the degree of closeness of predicted values to actual values and measure the trade-off between specificity and sensitivity respectively.

Subsequent to this, the team attempted to build a contextualised decision tree for UniSIM which could better the predictive performance of the baseline decision tree. In this phase, greater emphasis was placed on literature and domain knowledge whereby different predictors were used as the first tree splitting criterion. All these predictors were selected based on their statistical significance (p-value < .05) as well as their influence on the students' performance as observed from domain knowledge. After the alternative CHAID trees were grown, a 10-fold cross validation was applied to ascertain their stability and to prevent over-fitting. In instances of significant deviation in performance criterion the outliers were removed, a model was reconstructed and cross-validated using the same process. Lastly, performance criteria of all the alternative CHAID models were compared and evaluated. The final CHAID model which presented an optimal balance in its accuracy, stability in predictive performance and explanatory power was chosen.

Results and Discussion

Findings from Data Understanding

As part of data understanding, a cross-tabulation was done for each variable listed in Table 2 to understand the proportion of at-risk students compared to the student participation rates. The results of selected variables in Table 3 revealed some interesting patterns.

Table 3. Summary results of selected variables of the at-risk model

<table>
<tr><th>1st split criterion: DIP Institution</th><th colspan="2">Institution A</th><th colspan="3">Institution B</th><th colspan="2">Institution C</th><th colspan="2">Institution D</th><th colspan="3">Institution E</th></tr>
<tr><td>Probability of Sem 1_Outcome = at-risk</td><td colspan="2">0.357</td><td colspan="3">0.213</td><td colspan="2">0.243</td><td colspan="2">0.187</td><td colspan="3">0.167</td></tr>
<tr><td>2nd split criterion: DIP CGPA</td><td>≤3.08</td><td>>3.08</td><td>≤2.09</td><td>2.09 to 3.30</td><td>>3.30</td><td>≤2.91</td><td>>2.91</td><td>≤2.09</td><td>>2.09</td><td>≤1.86</td><td>1.86 to 3.08</td><td>>3.08</td></tr>
<tr><td>Probability of Sem 1_Outcome = at-risk</td><td>0.395</td><td>0.186</td><td>0.320</td><td>0.205</td><td>0.078</td><td>0.298</td><td>0.148</td><td>0.377</td><td>0.135</td><td>0.292</td><td>0.179</td><td>0</td></tr>
<tr><td>3rd split criterion: Varies</td><td colspan="2" rowspan="2">No further split</td><td>Yrs since study end</td><td>‘O’ level English</td><td>No further split</td><td>Sch</td><td>Sch</td><td rowspan="2">No further split</td><td>Sch</td><td colspan="3" rowspan="2">No further split</td></tr>
<tr><td>4th split criterion: Varies</td><td colspan="3">No further split</td><td>‘O’ level Maths</td><td>No further split</td><td>‘O’ level Maths</td></tr>
</table>

The pre-UniSIM academic performance indicators such as Diploma CGPA, Mathematics ‘O’ level grades and English ‘O’ level grades, if the students had a lower poly GPA score (< 2.00) or weak ‘O’ level English and Maths grade (C6 or less), a higher percentage of the students were classified as at risk. It seems that the diploma awarding institution may have had some influence on the students’ academic performance as a substantial percentage of graduates from Institution C are classified at risk (21.2%) compared to their participation rate (13.5%).

Evaluation and Validation of Model

Based on the confusion matrices presented in Table 4, the three alternative CHAID models offer a more balanced predictive performance than the baseline reference model given their higher G-Mean scores (defined as a Geometric mean of Specificity and Sensitivity (Kubat, Holte, & Matwin, 1997). Out of the 3 models, the DIP Institution model was selected as it offers comparable specificity and sensitivity indices with no significant trade-off in other evaluation criteria. In this instance, it is important that the model has a good hit rate (evaluated holistically based on specificity and G-Mean indices) since the practical cost of misclassification would mean that actual at-risk students would not be able to benefit from subsequent intervention strategies or support.

Table 4. Comparison of evaluation criteria on CHAID models with different factors as tree splitting criteria

	Reference Model	DIP Institution	Diploma CGPA	UniSIM Schools
Specificity	40.8%	50.9%	50.4%	48.7%
Sensitivity	84.8%	76.8%	75.4%	78.4%
Accuracy	74.8%	70.9%	69.7%	71.7%
G-Mean	58.8%	62.5%	61.6%	61.8%
Error rate	25.2%	29.1%	30.3%	28.3%

The chosen DIP Institution model was then tested for its stability and replicability using the 10-Folds Cross Validation method. The cross validation result that is presented at Table 5 suggests a reasonably stable model and consistent predictive performance.

Table 5. 10-folds Cross Validation Performance of the chosen DIP Institution CHAID model

	Chosen DIP Institution CHAID model	10 Folds Cross Validation
Sensitivity *Ability to correctly identify actual cases (true positives)*	76.8%	77.3%
Specificity *Ability to correctly identify negative cases (true negatives)*	50.9%	48.1%
Accuracy *Closeness of its prediction to the actual values (true positives & negatives)*	70.9%	70.8%
Error Rate *Proportion of incorrect predictions (true negative & positives)*	29.1%	29.2%
G-Mean *Geometric mean of sensitivity and specificity*	62.5%	60.1%

Upon examination of the final CHAID decision tree (see Figure 1), we find that the DIP Institution ($x^2 = 44.66$, p-value < .05) that the students graduated from is significant. It is also observed that the CHAID decision tree divides into three branches with a few DIP Institutions grouped together (for example, DIP Institutions A, B and D are grouped at 1 split, while Dip Institution E remains by itself). This could perhaps be attributed to a lack of comparable grading criteria adopted by different DIP Institutions. This is an indication that the quality of their prior academic preparation is an important influencing factor on the adult learners' ability to cope in the degree program in addition to their innate academic potential.

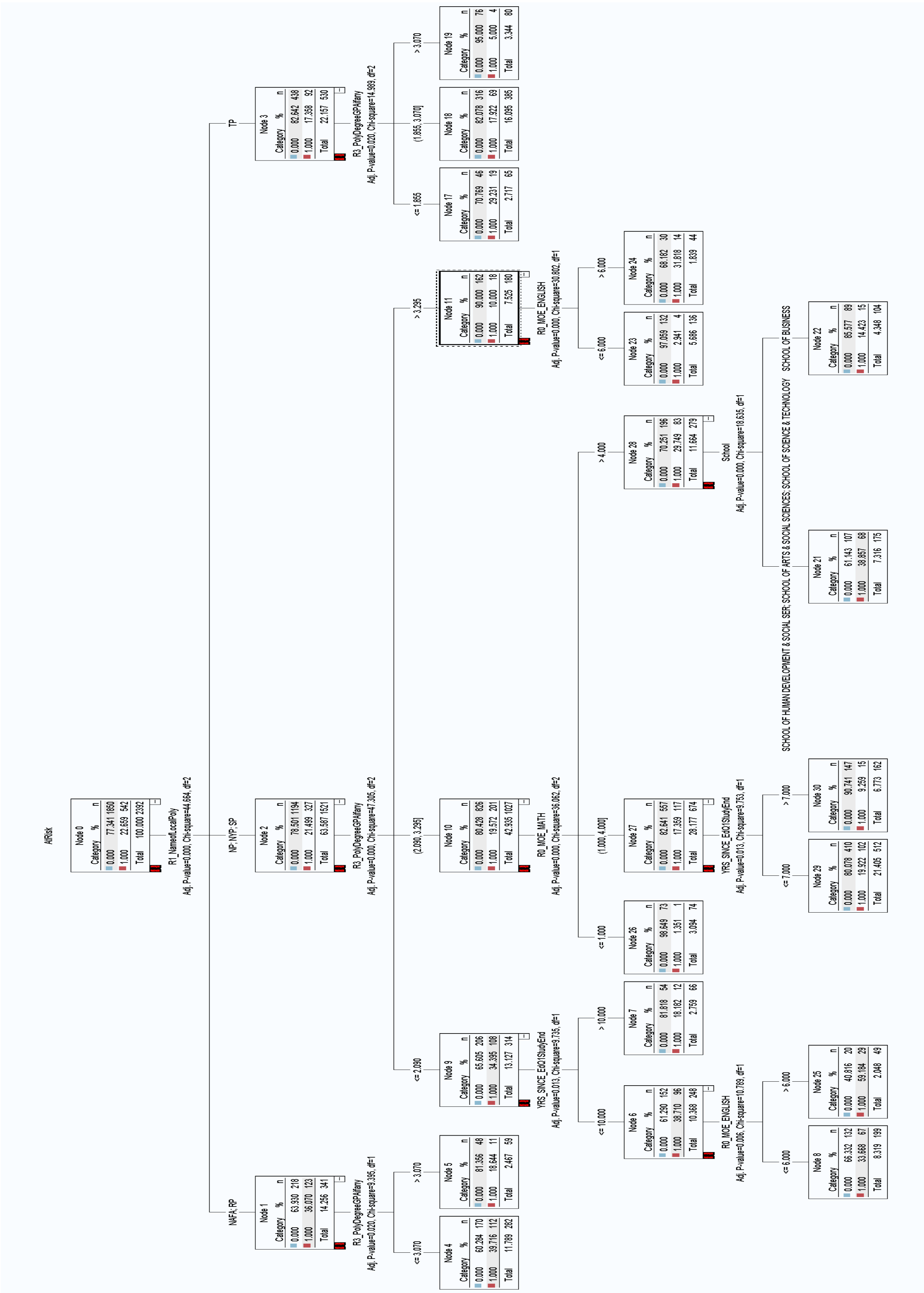

Figure 1. Final CHAID Model

Implications and Application of Findings

In the CHAID model (see Figure 1), pre-SIM University academic performance variables: the Pre-University institution that the students graduated from, students' CGPA score and 'O' level English and Mathematics grades emerged as significant predictors of the adult learner's academic performance for the first semester. The finding that the pre-university institution that the students graduated from is a key predictor indicates that there is a wide variation in the standards of performance among the different diploma institutions. Singapore's education system is essentially centralized and standardized (Lo, 2014). This variation in academic performance standards among feeder institutions is an issue of concern for the University's enrolment office.

The quality and strength of the students' academic foundation prior to entering University impacts how they cope with the demands of a university program. The finding that pre-university diploma CGPA is a significant predictor of academic outcomes is consistent with Geiser and Santelices' (2007) study that concluded that high school GPA is consistently a strong predictor of four-year college academic outcomes.

The importance of a strong pre-university academic foundation is consistent with views that high school English and mathematics proficiencies are critical parts of undergraduate preparation for success. Research also demonstrates that language proficiency is correlated with academic success (Ellis, Chong & Choy, 2013; Gottlieb, 2006). Goldinch and Hughes (2007) investigated the relationships between students' confidence in their generic skills on entry to university, their learning styles and their academic performance in the first year. Their study highlighted a link between students' confidence with language and numeracy proficiencies.

The model developed in this study can be of assistance to university enrolment management in many ways. An awareness of how potential students may perform academically could lead to a more targeted marketing campaign. Promotional materials about academics, mentoring and student support resources can raise applicants' awareness of how these services can aid in adult learner transition to university. With the identification of significant factors that may affect the students' initial academic performance, universities can provide timely interventions through early identification and monitoring of possible at-risk students. A multi-pronged support structure may be more efficacious in assisting these students to remain in their degree program. Concrete steps, such as academic counseling may be offered to targeted students to maximise their learning and overall university experience. In this way, resources can be more effectively and efficiently targeted towards a comprehensive support for these students.

Future Directions

Although this study is limited in that it is based on SIM University's 2013 enrolment dataset the proposed model may serve as a baseline for future research. Another limitation of the present study is that the academically at risk status (operationalized as CGPA) was assessed for only a single academic year. There are three potential future research directions. Firstly, more variables could be included to improve the enrolment models as well as to develop other relevant models with reduced misclassification of students' academic performance. Other decision tree models and ensembles of models may also be explored as educational institutions exploit data mining for effective decision making, efficient operations and to improve teaching and learning (Koh and Chong, 2014). Secondly, there is a need to follow various groups of students, students who are at risk, transferees, withdrawals, as well as students who are high performers. Including a time line in the analysis to follow these groups of students in subsequent semesters and tracking their study outcome would help to model their learning behaviour and patterns. Thirdly, aside from enhancing the accuracy of prediction, one of the directions for future research could be focused on using the data collected to identify and develop the support systems for teaching and learning.

Acknowledgement

This paper is part of an institutional research made possible through a grant (code: RF13LSC01) provided by SIM University.

References

Adelman, C. (2006). *The Tool Box Revisited.* Washington, DC: US Department of Education.

Bhardwaj, K. B., & Pal, S. (2011). Data mining: a prediction for performance improvement using classification. *International Journal of Computer Science and Information Security, 9*(4).

Bogard, M., James, C., Helbig, T., & Huff, G. (2012). Using SAS® Enterprise BI and SAS® Enterprise MinerTM to Reduce Student Attrition. In *SAS Global Forum 2012 Proceedings* (No. 031-2012).

Braxton, J. M., & Hirschy, A. S. (2005). Theoretical developments in the study of college student departure. *College student retention: Formula for student success*, *3*, 61-87.

Chang, T. (2009). Data mining: A magic technology for college recruitment. Retrieved August 20, 2016, from www.ocair.org/files/presentations/paper2008.../tongshan_chang_2009

Chong, S., Loh, W. M., & Babu, M. (2015). The millennial learner: A new generation of adult learners in higher education. *Advances in Scholarship of Teaching and Learning*, *2*(2), 1-14.

Chong, S., Mak, M., & Loh, W. M. (2016). Data-mining applications with the admission data of adult learners in higher education: A pilot study. *International Journal of Management in Education*, *10*(2), 131-144. https://doi.org/10.1504/IJMIE.2016.075555

Eagle, J. & Tinto, V. (2008). *Moving beyond access: College success for low-income, first-generation students*. Washington, DC, The Pell Institute for the Study of Opportunity in Higher Education.

Ellis, M., Chong, S., & Choy, Z. (2013). IELTS as an indicator of written proficiency levels: A study of student teachers at the National Institute of Education, Singapore. *International Journal of Educational Research*, *60*, 11-18. https://doi.org/10.1016/j.ijer.2013.03.003

Geiser, S. and Santelices, M. V. (2007). *Validity of high-school grades in predicting student success beyond the freshman year*. Berkeley Center for Studies in Higher Education: University of California.

Giudici, P. (2013). *Applied data mining: Statistical methods for business and industry*. Hoboken, NJ: John Wiley and Sons.

Goldfinch, J., & Hughes, M. (2007). Skills, learning styles and success of first-year undergraduates. *Active learning in higher education*, *8*(3), 259-273. https://doi.org/10.1177/1469787407081881

Gottlieb, M. (2006). *Assessing English language learners: Bridges from language proficiency to academic achievement*. Corwin Press.

Huebner, R. A. (2013). *A Survey of Educational Data-Mining Research. Research in Higher Education Journal, 19(1) 1-13*.

Iam-On, N., & Boongoen, T. (2015). Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. *International Journal of Machine Learning and Cybernetics*, 1-14.

Koh, H. C., & Chong, S. (2014). *Broadening and deepening of SoTL with learning analytics,* Retrieved August 20, 2016, from http://tlc.unisim.edu.sg/research/AdvSoTL/koh_hian_chye.html

Kovacic, Z. (2010). Early prediction of student success: Mining students' enrolment data. Retrieved August 20, 2016, from http://repository.openpolytechnic.ac.nz/handle/11072/646

Kubat, M., Holte, R., & Matwin, S. (1997). Learning when negative examples abound. *Machine Learning:ECML-97, Lecture Notes in Artificial Intelligence 1224,* 146–153. https://doi.org/10.1007/3-540-62858-4_79

Li, X., Nsofor, G. C., & Song, L. (2009). A comparative analysis of predictive data mining techniques. *International Journal of Rapid Manufacturing*, *1*(2), 150-172. https://doi.org/10.1504/IJRAPIDM.2009.029380

Lo, W. Y. W. (2014). Think global, think local: The changing landscape of higher education and the role of quality assurance in Singapore. *Policy and Society*, *33*(3), 263-273. https://doi.org/10.1016/j.polsoc.2014.09.002

MacFadgen, S. L. (2007). *Mature students in the persistence puzzle: An exploration of the factors that contribute to adult learners' quality of life and retention in post-secondary education* (Doctoral dissertation, Faculty of Education-Simon Fraser University).

MOE (2012). *Report of the Committee on University Education*. Pathways Beyond 2015 (CUEP). Singapore

Palmer, N., Bexley, E., & James, R. (2011). Selection and participation in higher education: University selection in support of student success and diversity of participation. *Centre for the Study of Higher Education*.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research* (Vol. 2). San Francisco: Jossey-Bass.

Pascoe, R., McClelland, A., & McGaw, B. (1997). *Perspectives on selection methods for entry into higher education in Australia.* Commissioned Report No. 58. Retrieved August 20, 2016, from http://www.dest.gov.au/NR/rdonlyres/37F1B0B3-7F8C-4011-A502-E68E94358211/3952/97_18.pdf

Power, C., Robertson, F., & Baker, M. (1987). *Success in higher education* (No. 94). National Institute of Labour Studies.

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. *International Journal of Computer Science Issues*, *7*(1), 10-18.

SIM University (2014). *Facts & figures*. Retrieved August 20, 2016, from http://www.unisim.edu.sg/about-unisim/Overview/Pages/Facts-Figures.aspx

SPSS (2009). *SPSS for windows, Rel. 18.0.0*. Chicago: IBM SPSS Inc.

Vandamme, J. P., Meskens, N., & Superby, J. F. (2007). Predicting academic performance by data mining methods. *Education Economics, 15*(4), 405-419. https://doi.org/10.1080/09645290701409939

Wyatt, L. G. (2011). Nontraditional student engagement: Increasing adult student success and retention. *The Journal of Continuing Higher Education*, *59*(1), 10-20. https://doi.org/10.1080/07377363.2011.544977