Investigating the efficacy of algorithmic student modelling in predicting students at risk of failing in tertiary education.

Geraldine Gray, Colm McGuinness, Philip Owende
Institute of Technology Blanchardstown
Blanchardstown Road North
Dublin 15, Ireland
geraldine.gray@itb.ie

ABSTRACT

The increasing numbers enrolling for college courses, and increased diversity in the classroom, poses a challenge for colleges in enabling all students achieve their potential. This paper reports on a study to model factors, using data mining techniques, that are predictive of college academic performance, and can be measured during first year enrolment. Data was gathered over three years, and focused on a diverse student population of first year students from a range of academic disciplines (n≈1100). Initial models generated on two years of data (n=713) demonstrate high accuracy. Advice is sought on additional analysis approaches to consider.

Keywords

Educational data mining, academic performance, personality, motivation, specific learning difficulties, self-regulation.

1. INTRODUCTION

In tertiary education, learning is typically measured by student performance based on a variety of assessments that are aggregated to generate a single measure of academic performance. Factors impacting on academic performance have been the focus of research for many years [9, 14]. It still remains an active research topic [5, 12], indicating the inherent difficulty in defining robust deterministic models to predict academic performance [16]. Typically, methodologies for quantitative research in this domain focus on statistical analysis of performance metrics and their correlations with, or dependencies on, a wide variety of factors including measures of aptitude, motivation, organisation skills, personality traits, prior academic achievements and demographic data [6, 11, 18]. More recently, Educational Data Mining (EDM) has emerged as an evolving and growing research discipline, covering the application of data mining techniques in educational settings [1, 4, 10, 19]. There have been calls for greater use of data mining by educational institutes to realise the potential of the large amounts of data gathered by institutes each year [7, 20]. While initial studies show promising results, a greater body of work is needed to determine if data mining techniques can offer an improvement over statistical methods [6, 11, 15].

It is increasingly evident that significant numbers of students in Institutes of Technology[1] (IoT) in Ireland do not complete the courses on which they enrolled [13]. Increased numbers enrolling in first year, and increased diversity in the student population, adds to the challenge of both identifying students at risk of failing, and planning appropriate supports to enable students perform optimally [13]. This study aims to investigate the suitability of classification techniques in generating a robust student model at enrolment which could identify students at risk of failing. The study focuses on two areas of research (1) an investigation of additional measures to augment the data currently gathered by college administration which will assist in the identification of students at risk, and (2) an investigation of suitable data mining techniques to accurately model this augmented dataset.

Study Hypothesis:. That educational data mining techniques can generate an accurate, deterministic model of academic performance based on factors measured at enrolment to tertiary education.

Study Objectives:

1. Identify and investigate factors most likely to determine academic performance in tertiary education, with a focus on factors that can be measured at enrolment.
2. Investigate the accuracy and stability of a range of classification techniques in predicting students at risk of failing in first year.
3. Compare the suitability of a data mining approach with a statistical approach for modelling a diverse student population.

2. EXPECTED CONTRIBUTION

This study adds to existing knowledge in the following ways:

[1]The Institute of Technology sector is a major provider of third and fourth level education in Ireland, focusing on the skill needs of the community they serve (www.ioti.ie).

1. Extends existing research in Education Data Mining: EDM has given much attention to datasets generated from students' behaviour on Virtual Learning Environments (VLE) and Intelligent Tutoring Systems (ITS) [4]. Less focus has been given to modelling datasets from outside virtual or online learning environments. This research focuses on models of college students that can be applied early in semester one.

2. Focus on third level students outside the university sector: Students enrolling in IoTs have, on average, a weaker academic history than university students[2], and there are increasing admissions of non-standard students [13]. This is an under-studied group compared to university students. A study of computing students [2] has shown that there is a difference between factors influencing the academic performance of university students compared to students in an IoT. This research extends that work by incorporating a wider range of factors and a diversity of students from across several academic disciplines.

3. The novel inclusion of data on specific learning difficulties: This study is based at the Institute of Technology Blanchardstown (ITB), who in partnership with the National Learning Network Assessment Services[3] (NLN) located on campus, provide assessment and follow up support for all students in four areas of specific learning difficulty: reading and spelling, organisation and co-ordination, social and communication, and attention and concentration. Profiling has shown one in five students at ITB report difficulties in at least one of these areas [8]. There is insufficient research including measures of specific learning difficulties in student modelling.

3. RESULTS SO FAR

3.1 Study criteria

Limited profiling of students in terms of specific learning difficulties and some learning preferences was already underway at ITB. This study extended that initiative, adding measures relating to four additional factors: aptitude, personality, motivation and learning strategies. These were chosen firstly because research highlights these factors as being directly or indirectly related to academic performance [21], and secondly because these factors can be measured early in semester one. An online questionnaire was developed to profile students and give immediate feedback (www.howilearn.ie). Data already available to college administration on prior academic performance was also used[4]. A full list of the factors used is given in Table 1.

[2]The majority of students in the IoT sector will have attained between 200 and 400 points in the Leaving Certificate exam, the state exam at the end of secondary school. The majority of students in the Irish university sector will have attained over 400 points, including some with the maximum score of 600 points [13, Appendix A].

[3]The NLN assessment team includes an educational psychologist, assistant psychologist and occupational therapist (http://www.nln.ie/Learning-and-Assessment-Services.aspx).

[4]Prior academic performance is based on state examinations completed by all students at the end of secondary school in Ireland.

Table 1: Measures included in the study

Prior Academic Performance	
English Grade	Did Honours English
Maths Grade	Did Honours Maths
Highest Mark	Humanities Average
Science Average	Creative/Practical Average
Aggregate Mark (CAO points)	
Personality, Goldbergs IPIP scales (http://ipip.ori.org)	
Conscientousness	Openness
Motivation, based on MSLQ [17]	
Intrinsic Goal Orientation	Self Efficacy
Extrinsic Goal Orientation	
Learning style, based on R-SPQ-2F [3]	
Deep Learner	Shallow Learner
Strategic Learner	
Self-regulated Learning, based on MSLQ [17]	
Self Regulation	Study Effort
Study Time	
Specific Learning Difficulties, Do-IT profiler (www.doitprofiler.info)	
Reading and Spelling	Social and Communication
Organisation and Co-ordination	Attention and Concentration
Preferred learning channel	
Visual, Auditory, Kinaesthetic, or a combination of these	
Other preferences, using Learning Styles Questionnaire from NLN (www.nln.ie)	
Organised or Disorganised	Morning or Evening
Meticulous or Approximate	Group work or solo
Logical or Creative	Like background noise
Other factors:	
Age	Gender

3.2 Study participants

Data was gathered on first year students over three academic years, 2010, 2011 and 2012. All students in the first year of study were invited to complete the online questionnaire as part of first year induction. 1,332 students completed the questionnaire. End of year results are available for two of the three years, giving a current sample size of (n=713). The final sample size is expected to be approximately 1,100 as to date 16% of students either gave an invalid student ID during profiling, or did not give permission for their data to be used in the study. Average CAO[5] points is 257.9 $\pm$ 75. 59% of the students were male. The students are from Computing, Engineering, Business, Social Care, Creative Digital Media, Sports Management and Horticulture.

3.3 Initial results

Modelling was done on the 2010 and 2011 data, using five dimensions, namely: prior academic performance, motivation, learning orientation, personality and age. A binary class label was used based on end of year GPA, range [0-4]. The two classes included poor academic achievers who failed overall (GPA<2, n=296), and strong academic achievers who achieved honours overall (GPA$\geq$2.5, n=340). To focus on patterns that distinguish poor and strong academic achievements, students with a GPA of between 2.0 and 2.5 were excluded from initial models, giving a dataset of (n=636). Six algorithms were used: Support Vector Machine(SVM), Neural Network, k-Nearest Neighbour, Naïve Bayes, Decision tree and Logistic Regression, using RapidMiner V5.2 (rapid-i.com). When modelling all students, model performance was comparable across the six learners, with Naïve

[5]CAO Points are an aggregate measure of prior academic performance in Ireland, range [0,600]. It represents the combined score achieved in six subjects.

Bayes achieving the best accuracy at 75.74%. However when modelling subgroups split by age, model accuracies for algorithms that can learn more complex patterns increased, with SVMs getting the best accuracy (82.62% for students under 21, 93.45% for students over 21). Other subgroups were not consider in the initial analysis.

4. OUTSTANDING QUESTIONS

Feedback and discussion is welcome on all aspects of the study, and specifically in the following areas:

1. The dataset has 44 attributes, primarily generated from a questionnaire using Likert scales, and so have a small range of discrete numeric values. Attributes are based on factors that have widely published correlations and interdependencies, although the reported significance of those dependencies varies. Opinions are sought on modelling approaches to consider for this type of dataset.

2. There have been calls from the EDM community for the use of statistical methodologies in data mining research [15]. Feedback on how this study should adhere to a statistical methodology to validate modelling results would be of value.

3. Also of interest are opinions on the value and limitations of early student modelling, before data on student engagement in course work is available. Is there value in also considering measures of early engagement based on activity on a VLE such as Moodle?

5. REFERENCES

[1] S. R. Barahate and M. Shelake, Vijay. A survey and future vision of data mining in educational field. *Second International Conference on Advanced Computing and Communication Technologies*, 2012.

[2] S. Bergin and R. Reilly. Predicting introductory programming performance: A multi-institutional multivariate study. *Computer Science Education*, 16, No. 4:303–323, 2006.

[3] J. Biggs, D. Kember, and D. Leung. The revised two-factor study process questionnaire: R-spq-2f. *British Journal of Education Psychology*, 71:133–149, 2001.

[4] T. Calders and M. Pechenizkiy. Introduction to the special section on educational data mining. *SIGKDD*, 13(2):3–3, 2011.

[5] S. Cassidy. Exploring individual differences as determining factors in student academic achievement in higher education. *Studies in Higher Education*, pages 1–18, 2011.

[6] G. Dekker, M. Pechenizkiy, and J. Vleeshouwers. Predicting students drop out: a case study. In T. Barnes, M. C. Desmarais, C. Romero, and S. Ventura, editors, *Proceedings of the 2nd International Conference on Educational Data Mining*, pages 41–50, Cordoba, Spain, 2009.

[7] N. Delavari, M. R. A. Shiraze, and M. R. Beikzadeh. A new model of using data mining technology in higher education systems. In *Proceedings of 5th International Conference in Information Technology Based Higher Education and Training*, Istanbul, Turkey, 2004.

[8] D. Duffin and G. Gray. Using ict to enable inclusive teaching practices in higher education. *AAATE, Florence*, Sept 2009.

[9] T. Farsides and R. Woodfield. Individual differences and undergraduate academic success: the roles of personality, intelligence, and application. *Personality and Individual Differences*, 34:1225–1243, 2003.

[10] Y. Gong, D. Rai, J. E. Beck, and N. T. Heffernan. Does self-discipline impact students' knowledge and learning? *Proceedings of the 2nd International Conference on Educational Data Mining*, pages 61–70, 2009.

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

[12] M. Komarraju, A. Ramsey, and V. Rinella. Cognitive and non-cognitive predictors of college readiness and performance. role academic discipline. *Learning and Individual Differences*, 24, 2013.

[13] O. Mooney, V. Patterson, M. O'Connor, and A. Chantler. A study of progression in higher education: A report by the higher education authority. Technical report, Higher Education Authority, Ireland, October 2010.

[14] M. A. Moran and M. J. Crowley. The leaving certificate and first year university performance. *Journal of Statistical and Social Enquiry in Ireland*, XXIV, part 1:231–266, 1979.

[15] B. Nelson, R. Nugent, and A. A. Rupp. On instructional utility, statistical methodology, and the added value of ecd: Lessons learned from the special issue. *Journal of Educational Data Mining*, 4 (1):227–233, 2012.

[16] Z. A. Pardos, R. S. J. D. Baker, S. M. Gowda, and N. T. Heffernan. The sum is greater than the parts: Ensembling models of student knowledge in educational software. *SIGKDD Explorations*, 13(2), 2011.

[17] P. Pintrich, D. Smith, T. Garcia, and W. McKeachie. A manual for the use of the motivated strategies for learning questionnaire. Technical Report 91-B-004, The Regents of the University of Michigan, 1991.

[18] S. B. Robbins, K. Lauver, H. Le, D. Davis, and R. Langley. Do psychosocial and study skill factors predict college outcomes? a meta analysis. *Psychological Bulletin*, 130 (2):261–288, 2004.

[19] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33:135–146, 2007.

[20] C. Romero and S. Ventura. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, 40(6):601–618, 2010.

[21] N. Schmitt, F. L. Oswald, T. Pleskac, R. Sinha, and M. Zorzie. Prediction of four-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology*, 2009.