Data mining: a useful student retention tool or what we knew already?

Chetwynd Frances, Dobbyn Chris, Jefferis Helen, Woodthorpe John, *The Open University, UK*

INTRODUCTION

HE students who enrol at a UK University but do not subsequently complete their degree not only damage their employment prospects, but also waste institutional and tax-payers' resources. In an era of reduced Government funding, rising tuition fees and diminishing revenue, universities must explore every option for reducing their non-completion rates. With the advent of Virtual Learning Environments (VLEs) it has become possible to combine data about students' study progress with their registration background data, to attempt to identify those most at risk of failing. This paper describes some of the findings of an on-going Open University (OU) eSTEeM-funded project set up to explore these issues.

NON-COMPLETION AT THE OPEN UNIVERSITY

According to the National Audit Office (NAO 2007, P20), 'The factor most affecting a student's chance of continuing [from L1 to L2] is whether they are studying full time or part time, with full-time students being much more likely to continue if other factors are held constant...' . With its unique combination of distance learning, part-time study and open entry, the OU has an uphill struggle to achieve acceptable student retention rates. The UK sector average for part-time undergraduates no longer in HE after two years is 35.2%, whereas the OU figure is 44.7% (HESA, 2014).

This project focusses on students studying a Level 1 Computing and IT module – TU100 My Digital Life. According to the NAO (2007, p49), the choice of a strategically important science, technology, engineering or mathematical subject immediately reduces a student's odds of continuing to a second year, compared to any other subject. Earlier HEA-funded work (Authors, 2013 and 2014) investigated the behaviour of students that remained engaged for the duration of this 9 months module, passing the continuous assessment aspects (five tutor marked assignments and seven interactive Computer Marked Assignments (iCMAs)), but failing the final End of Module Assessment (EMA). This work suggested that students who fail the EMA do not follow the same patterns of iCMA completion as the cohort as a whole; they not only miss out on the early learning opportunities that iCMAs afford, but also have a higher workload at the end of the module to pass the various thresholds.

RESEARCH AIMS

Although late completion of the interactive quizzes may be an indicator of impending failure, this factor on its own is unlikely to be sufficient to distinguish, early on, those students capable of passing but in danger of failing, who would benefit from extra study advice from their tutor. The aims of the project we describe are to explore

1. what main factors distinguish an at-risk student;
2. whether an Artificial Neural Network (ANN) can be trained to categorise at risk students;
3. what impact an academic tutor can have on moving students from the at risk category to the pass category.

It is clear that developing a personal relationship with academic staff enhances the student experience (Chickering and Gamson, 1987; Jones, 2006; Lamer, 2009). Given the influence of student motivation on the likelihood of that student persisting (Alarcon and Edwards, 2013; Harterich, 2012; Simpson, 2012), the key role of academic staff in promoting learning motivation cannot be underestimated. However, local academic tutors at the OU have limited part time contracts. It is not possible, therefore, to adopt a scattergun approach when devising a program of extra contacts with their students.

METHODOLOGY

Initially, the project is using data on three student cohorts. TU100 attracts around 4500 registrations annually, with 2500 students starting in October and 2000 starting in February. The three cohorts chosen are

1. Oct 2012 – Jun 2013, (Cohort 12J): here, the pass/fail student outcome is already known, so this data will be used to train and test the ANNs.
2. Oct 2013 – Jun 2014, (Cohort 13J): the final results are not currently available, so this data will be used as a control set.
3. Feb 2014 – Oct 2014, (Cohort 14B): final results are not currently available, so this data will be used as the trial dataset for use with the ANN to classify students for extra study help.

Part of the study will be to identify the most indicative inputs out of a total of 23 possibles culled from the VLE and student personal data. These include timeliness of assignment submissions and grades achieved; sex; stated study motivation; receipt of sponsorship.

The training data from the 12J cohort was classified into four student categories, according to their EMA result:

1. Passed EMA
2. Failed EMA with score of 30-39%
3. Failed EMA with score of 0-29%
4. Did not submit EMA.

RESULTS AND DISCUSSION

Much of the work on ANNs involves preparing the input data so as to make good classification results possible. Early inspection of the data revealed that a neural network would find it hard to produce clear-cut classifications for several reasons.

- The data is very noisy, especially for students not in class (1) above. For example predicting that students achieving assignment marks in the 70s will pass the EMA ought to be straightforward; however, later in the module those students' results may decline for a multitude of reasons unrelated to their study, and unpredictable by any computer system.
- The data at the boundaries of the classes has few distinguishing characteristics. For example one student scoring assignment marks in the low 40s may pass the EMA with a 40% grade. A similar student may fail the EMA (and module) with 39% - the difference being the answer to one small part of one question.
- The classes mainly of interest to us, classes (2) and (3), make up only 4% of the total dataset.

Initial trials suggest that data on students in class (4) – failing through non-submission of the EMA – injects so much noise and uncertainty into the network that discrimination becomes poor. Indeed, non-submission should be treated as a separate problem, and may be best predicted simply by data on timeliness of submission, along with personal factors such as sponsorship and motivation for joining the course. Work continues on the optimal neural architecture for the problem of predicting near fails, and on the best combination of indicators to be presented at the network's input.

REFERENCES

Alarcon, G. M., & Edwards, J. M. (2013). 'Ability and motivation: Assessing individual factors that contribute to university retention', *Journal of Educational Psychology*, vol. 105, no. 1 [Online]. Available at http://libezproxy.open.ac.uk/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=2012-12948-001&site=ehost-live&scope=site (Accessed 18 June 2014)

Chickering, A. W. and Gamson, Z. F. (1987) 'Seven principles for good practice in undergraduate education', *The Wingspread Journal*, vol. 9, no. 2, special insert. Available at http://www.bgsu.edu/downloads/provost/file84390.pdf (Accessed 18 June 2014)

Authors (2013) 'Online assessment: supported learning or "just do it"?, *Annual Conference of the Society for Research in Higher Education Conference*, Newport, Wales, December 2013. Available at http://www.srhe.ac.uk/conference2013/abstracts/0106.pdf (Accessed 18 June 2014)

Authors (2014) 'Online assessment: supported learning or "just do it"?', *HEA third Annual STEM Learning & Teaching Conference*, Edinburgh, Scotland, April 2014. Available at http://www.heacademy.ac.uk/resources/detail/STEM/STEM-conference-proceedings-2014/Computing/Computing-Conf-Proceedings (Accessed 18 June 2014)

Härterich, J., Kiss, C., Rooch, A., Mönnigmann, M., Darup, M.S. and Span, R. (2012) 'MathePraxis - connecting first-year mathematics with engineering applications', *European Journal of Engineering Education*, vol. 37, no. 3 [Online]. Available at http://www.tandfonline.com.libezproxy.open.ac.uk/doi/abs/10.1080/03043797.2012.681295 (Accessed 18 June 2014)

HESA (2014) *'Non-continuation following year of entry (table series T3e)'* https://www.hesa.ac.uk/pis/noncon (Accessed 18 June 2014)

Jones, D. (2006) *Enhancing the learning journey for distance education students in an introductory programming course,* Queensland [Online] Available at http://davidtjones.wordpress.com/publications/enhancing-the-learning-journey-for-distance-education-students-in-an-introductory-programming-course/ (Accessed 18 June 2014)

Lamer, M. (2009) 'Where's Walter? Adjunct Outreach Strategies to Bridge the Virtual Distance and Increase Student Retention', Online *Journal of Distance Learning Administration*, vol. 12, no. 2 [Online]. Available at http://www.westga.edu/~distance/ojdla/summer122/lamer122.html (Accessed 18 June 2014)

NAO (2007) *Staying the Course: The Retention of Students in Higher Education*, House of Commons. Available at http://www.nao.org.uk/report/staying-the-course-the-retention-of-students-in-higher-education/ (Accessed 18 June 2014)

Simpson, O. (2012) *Supporting Students for Success in Online and Distance Education*, Abingdon, Routledge.