Developing Predictive Models of Student Success in Undergraduate Engineering Mathematics Courses

Sarah Dart[a]
[a] *Learning & Teaching Unit and Science & Engineering Faculty, Queensland University of Technology (QUT)*
Corresponding Author's Email: sarah.dart@qut.edu.au

Introduction

Student success and retention are hot topics in higher education, given the strong ties to institutional reputation, funding and government policy. A recent government review found that the attrition rate for Australian universities had been relatively stagnant over the last decade, hovering at around 15% (Australian Government Department of Education and Training, 2018). This review recommended that institutions develop evidence-based risk analytics in order to enhance student support strategies. Given the shifts toward online learning and systematic collection of student data at scale, there is enormous potential to harness this information for predictive models of student success. These models would enable enhanced understanding of key risk factors, methodical identification of at-risk students, and even the ability to directly target these students with tailored support interventions.

The present study focuses on predicting students at risk of failure for two undergraduate engineering mathematics courses by developing models based on data already collected within the institution. Identifying at-risk students as early as possible within the semester is advantageous as it enables early intervention (Lizzio & Wilson, 2013). However, the accuracy of predicting whether a student is at-risk would be expected to increase over time as additional information is generated about student engagement and performance. Thus the key questions to be answered in this research are: what factors are most useful for predicting student success, and how does prediction accuracy vary across a semester as new data emerges.

Background

Measures of Higher Education Student Outcomes

Attrition rates, defined as the percentage of the student cohort who study in one year but then do not graduate or continue study in the following year, are one of the most widely used measures for higher education outcomes. However, outcomes can also be measured in terms of student success and completion (McMillan, 2005). Success measures the rate at which students pass individual courses, while completion rates measure the percentage of students who graduate with degrees, typically within a certain timeframe (Edwards & McMillan, 2015). As students must pass individual courses to advance through multi-year degrees, academic success is not only vital to degree completion but also an important factor underpinning retention (Chen et al., 2008; Crosling et al., 2009). While retention and completion outcome data lags months or years, data on course outcomes is available at the end of each semester. Utilising course data thus enables more responsive modelling. Furthermore, analysis at the course level enables modelling of more customised variables (such as performance in individual assessment items) that could not otherwise be incorporated when considering outcomes at a program or institutional level. As such, the present study argues for the value of course-level data in supporting student retention and completion.

Factors influencing Student Success in Higher Education

There is a large body of literature which seeks to explain the underlying reasons for student success, retention and completion. Much of this work is grounded in the theoretical model of Tinto (1975). Here family background, personal characteristics, and prior educational experiences influence student commitment to their study and to their institution, in turn

influencing student interactions with the academic and social systems of the university. How well a student is able to integrate in both these systems ultimately determines whether they will continue in their studies.

Institutions typically collect a vast amount of data which can be used as measures of academic integration, such as assessment results and grade point averages. Interactions with learning management systems (LMS), such as Blackboard or Moodle, can also be used as a proxy, with previous research showing a relationship between academic success and frequency and type of LMS usage (Campbell, 2007). Institutions also typically collect a range of information at enrolment relating to students' family background, personal characteristics and prior educational experiences. The literature shows a number of these factors can be linked with student outcomes, especially across retention and completion measures. Socio-demographic factors associated with poorer student outcomes include having a lower socio-economic background, being from a regional or remote area, having indigenous heritage, being older (Edwards & McMillan, 2015), and being a first-generation student (Collier & Morgan, 2008). Other attributes include studying in a part-time or distance capacity, entering university from a non-school leaver pathway such as vocational education or following work experiences (Edwards & McMillan, 2015), and having lower tertiary entry scores (Marks, 2007).

Modelling Techniques

Mathematical models have previously been used by institutions to analyse and understand the variation in student outcomes. A range of techniques have been applied including regression analyses, decision trees, support vector machines and neural networks (Kovacic, 2010; Marbouti et al., 2016). The present study focuses on using a logistic regression model as a binary classifier (here passing or failing a course) with multiple predictors that may be continuous, discrete or categorical. This technique is popular within the education field given its simplicity and ease of use. Furthermore, it has been shown to perform comparably to more complex and computationally intensive techniques, including for the prediction of higher education student outcomes (Kotsiantis et al., 2003; Marbouti et al., 2016).

Of the research around student success and retention, a number of studies have attempted to predict individual students' success based on various types of student data. Kovacic (2010) developed a model to predict pass/fail outcome in an information systems course based only on enrolment data. An overall classification accuracy of about 60% was achieved, leading the author to conclude that background information gathered at enrolment was insufficient to make this type of prediction. Kotsiantis et al. (2003) predicted performance in an informatics course delivered via distance, using both demographic data and progressive assignment marks. This found overall classification accuracy could be increased from 61% using only demographic data, to as high as 82% when assignment marks were included. Similar research conducted by Marbouti and Diefes-Dux (2015) considered progressive attendance and assessment results to predict outcomes in a large first-year engineering course. The model was tuned to enhance classification of at-risk students (those who went on to fail), resulting in an overall accuracy of 81%, and 90% accuracy for the at-risk sub-cohort. These findings suggest a combination of background, performance, and engagement data may be most effective for predicting course outcomes. The present study uses background data, assessment results and LMS engagement to investigate the predictors of success in an Australian undergraduate engineering cohort, a population which has yet to be explored in detail.

Methods

Participants and Setting

Participants consisted of students enrolled in two courses delivered on-campus during Semester 2, 2018 at the Queensland University of Technology, a large metropolitan university situated in Brisbane, Australia. The core undergraduate engineering mathematics courses of

Introductory Engineering Mathematics (IEM) and Engineering Computation (EC) were selected due to historically high failure rates, regular assessment items which could be drawn into the modelling, and their identification as an area that could benefit from additional support.

IEM covered fundamental mathematical concepts of differentiation, integration, vectors, matrices and complex numbers. It also included a programming component where students practiced learned concepts in Matlab. The course was typically taken by students in their first semester of study, although those who commenced with a strong background in senior high school mathematics could opt to take an alternative course. IEM had a weekly lecture (2 hours), tutorial (2 hours) and computer lab (1 hour). Assessment consisted of several short Matlab tasks and online quizzes, as well as a longer problem solving task and final exam. The schedule and weightings are shown in Table 1 below.

Table 1: Assessment schedules for IEM and EC; M=Matlab Tasks (2% each), Q=Online Quizzes (4% each), P=Problem Solving Task (20% for IEM and 12.5% each for EC), E=Exam (50%)

Week	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**Exam Period**
IEM			M1	Q1	M2	Q2	M3	Q3	P	M4	Q4	M5	Q5	E
EC				P1			P2			P3			P4	E

IEM was a prerequisite for EC, which was typically taken by students in their second semester. EC covered extended programming concepts (also applied in Matlab), solving ordinary differential equations using analytical and numerical techniques, and foundational probability and statistics. It had a weekly lecture (2 hours) and tutorial (2 hours). Students were assessed in four equally-weighted problem solving tasks, as well as a final exam (see Table 1). In Semester 2, 2018, there were 260 students for Introductory Engineering Mathematics with a pass rate of 68%, and 820 students for Engineering Computation with a pass rate of 85%.

Modelling Approach

Binary logistic regression was applied to predict passing or failing the course of interest. The regression model relates to the log odds by

$$\ln(ODDS) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where p is the probability of passing the course, x_1 through x_n are the independent predictors, and β_0 through β_n are regression coefficients (Kotsiantis et al., 2003). Regression models were generated at three key points in the semester to align with key academic milestones:

- Week 0 – week prior to teaching commencing (orientation week)
- Week 4 – last week to withdraw without financial penalty
- Week 9 – last week to withdraw without academic penalty

The predictors used were background characteristics, as well as the data that had been generated up to the chosen week of semester relating to LMS engagement and academic performance (discussed in detail below). SPSS Statistics 25.0 was used to generate the models using a stepwise forward selection method in line with Marbouti et al. (2016), where the predictors with the most statistical significance were progressively added until further variables did not significantly improve model performance. The cut-off for classifying students as passing or failing the course was set to the overall pass rate of the respective courses as per Marbouti and Diefes-Dux (2015). In order to evaluate the models, cross-validation was performed by randomly splitting the observations into five equally sized groups, controlling for the pass to fail ratio. Model performance was assessed on the hold-out groups by examining the classification table for the percentage of overall cases correctly classified (the overall accuracy), as well as the percentage of failing students correctly classified as failing the course (the recall). The beta coefficients were averaged for the final models.

Predictors

Assessment results were extracted from the LMS. Those available up to the week of semester being considered were incorporated into a single data point representing the percentage of the total marks achieved thus far (Marbouti & Diefes-Dux, 2015). LMS engagement was measured by hits on the course page. This was incorporated into a single data point representing the percentage of weeks thus far with at least one hit, starting at Week 0 (orientation week). Background data was drawn from the Student and Academic Management System. A range of characteristics aligned with those previously shown to correlate to student outcomes were extracted. Due to small numbers of observations in the minor category, the variables of indigenous heritage and part-time study were eliminated. Furthermore, due to substantial and non-randomly distributed missing data, the variables of tertiary entrance score and high school results (only available for school-leavers), as well as socio-economic status and rural background (only available for domestic students) were removed. The background characteristics in the final analysis were age, gender, first-in-family status, school-leaver admission basis, domestic student status, number of times repeating the course of interest, number of courses previously passed, and number of courses previously failed. Incoming grade point average (GPA) was also considered for EC given that unlike IEM, students should have completed at a least one semester of study prior to enrolling.

Results & Discussion

Significant Predictors of Student Success

The final models are shown in Table 2 below. It can be seen that the predictors significant in differentiating successful and unsuccessful students vary by course and week of the semester. Both courses' Week 0 models exclusively relied on variables associated with students' educational history. For IEM, this was the number of times students had previously attempted the IEM course, and the total number of courses previously passed. The latter is an interesting parameter, given that IEM is designed for study in first-year first-semester. Thus students who have passed courses before enrolling in IEM are likely to have studied as part of previous program, be repeating IEM, or be completing an atypical enrolment plan. For EC, the number of courses previously failed and incoming GPA were the significant factors. In both courses, socio-demographic background data was insignificant. This is consistent with Kovacic (2010) who found that characteristics like gender, age, ethnicity, disability and secondary school were insufficient for accurately predicting student success. Interestingly, access to the LMS during orientation week was an insignificant factor in predicting success in both courses. It should be also noted that the constant in EC's Week 0 model was found to be insignificant.

In the Week 4 and 9 models for IEM, the strongest predictor of success was the progressive assessment results. Moreover, the beta coefficient for assessment increased by approximately 1.5 times from Week 4 to 9, indicating that its importance grows through the semester. LMS access was an insignificant predictor of success. This may be because the Matlab and quiz submissions were online and weekly (see Table 1), meaning students who participated in assessment automatically also engaged with the LMS. Thus, it is likely the assessment measure was also capturing the LMS engagement for this course.

The strongest predictor of outcome for EC at both Week 4 and Week 9 was progressive assessment results. As with IEM, its strength as a predictor increased over time and by a similar margin (1.6 times). However unlike IEM, LMS engagement was a significant predictor of student success, with the associated beta coefficient increasing by about 1.5 times between Weeks 4 and 9. This suggests that in courses where contact with the LMS is not forced by an assessment schedule, regularity of student engagement with the LMS becomes an increasingly useful factor in differentiating between successful and unsuccessful students. However, it is not known whether LMS engagement outside of assessment indicates that the LMS is supporting students (LMS-factors), or whether these students are more driven to

Table 2: Final models for predicting student success by course and week of semester

Course	Week of Model	Predictor	Beta Coefficient	Significance
Introductory Engineering Mathematics	Week 0	Courses Passed	0.123	0.027
		Repeating Attempt Number	-1.130	0.000
		Constant	0.878	0.000
	Week 4	Assessment (to Week 4)	0.055	0.000
		Constant	-3.398	0.000
	Week 9	Courses Passed	0.122	0.031
		Assessment (to Week 9)	0.084	0.000
		Constant	-5.577	0.000
Engineering Computation	Week 0	Courses Failed	-0.275	0.000
		Incoming GPA	0.415	0.000
	Week 4	Incoming GPA	0.377	0.000
		LMS Access (to Week 4)	0.021	0.003
		Assessment (to Week 4)	0.044	0.000
		Constant	-4.746	0.000
	Week 9	Incoming GPA	0.333	0.003
		LMS Access (to Week 9)	0.031	0.000
		Assessment (to Week 9)	0.072	0.000
		Constant	-7.523	0.000

engage with their studies (student-factors). Finally, although students' incoming GPA was significant throughout the semester, its importance gradually decreased.

Accuracy of Model Predictions through Semester

Accuracy of each model in terms of classifying the overall student population as well as the at-risk sub-cohort is shown in Figure 1 below. It can be seen that accuracy trends upwards over the semester as expected, but model performance is consistently poorer for the at-risk cohorts compared to the equivalent overall cohorts. For the EC course, little change was observed in the overall classification accuracy throughout the semester. In contrast, a substantial improvement occurred from Week 0 to 4 for IEM. Both courses reached 82-84% overall accuracy from Week 4, which is in line with that achieved by Kotsiantis et al. (2003) and Marbouti and Diefes-Dux (2015).

Inspection of the at-risk sub-cohorts in Figure 1 shows classification accuracy is lower for this group. For IEM, the Week 0 model was only able to identify 51% of at-risk students, but this improved substantially to 76% when early performance data was considered at Week 4. It should be noted that this early assessment only corresponds to 6% of the total across the course, and yet immediately becomes the strongest indicator of outcome. For EC the improvement in classifying the at-risk cohort was smaller, moving from 55% to 67% between Weeks 0 and 4. At Week 4 progressive assessment amounted to 12.5% of the total for the semester, so it is interesting that even though a larger proportion of the final grade was known at this point, it was less indicative for predicting eventual success.

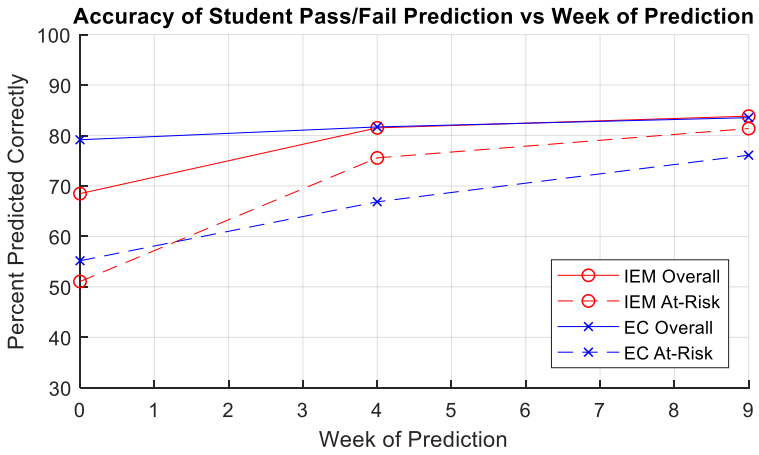

Figure 1: Prediction accuracy by course, cohort and week of semester

Figure 1 demonstrates that identifying students at-risk of unit failure is a trade-off between accuracy and early intervention. It appears waiting for the results of early assessment (around Week 4) is beneficial, as the improvement in accuracy through the start of semester is considerable but the gains gradually diminish. It is important to note that it is only possible to make such improved predictions when course assessment strategies include an early assessment item that can be graded quickly. However, as illustrated by the 6% of assessment completed at Week 4 in IEM, this early assessment does not need to be heavily weighted in order for it to be useful in differentiating successful from unsuccessful students.

Further tuning of the classification cut-off could be employed to reduce the false negative rate (at-risk students being predicted as passing), however this would be at the expense of increasing the false positive rate given that these two quantities trade-off against one another. Thus the optimal cut-off depends on the application of the model and by extension the accepted tolerance to Type II error. If the model was to be used specifically to identify students for direct contact, an additional consideration would be available resourcing, such that those students with the lowest predicted probability of passing were prioritised (Simpson, 2006).

Summary & Conclusions

This study applied logistic regression to predict students' pass/fail outcomes in two undergraduate engineering mathematics courses based on background, performance and LMS engagement data. Moreover, predictions were made at three key points in the semester to understand how prediction accuracy varied over time. Progressive assessment results were found to be the most useful predictor of student success, while LMS engagement was highly useful for EC but not IEM. This was due to the IEM assessment schedule forcing regular contact with the LMS, negating its value in this context. Socio-demographic factors were not significant predictors of student outcomes, even at Week 0. Instead only characteristics related to students' educational background were of significance. Performance of the final models was in line with overall classification accuracies achieved in similar previous studies, reaching 82-84% overall accuracy from Week 4. Reduced accuracy was achieved for the at-risk student sub-cohort, however this improved considerably between Week 0 and Week 4. Based on this finding, identifying at-risk students would be most beneficial around the Week 4 mark to balance the trade-off between prediction accuracy and early intervention.

This study marks a successful step in understanding the important aspects of implementing predictive models of student success in undergraduate engineering mathematics. The value of progressive assessment items is clear, allowing early identification and intervention for struggling students. While it is not known whether the relationship between LMS engagement and student outcomes is due to LMS factors or student factors, these findings provide a clear link between the two, and indicate value in further research into this area. More work is required

to understand the generalisability of the identified predictors across additional courses, given the applicability of the methodology across multiple discipline areas. A worthwhile extension of this study would be to apply models generated in a given semester to future cohorts in order to understand how applicable models are between cohorts, and whether there is a significant reduction in performance as students, assessment tasks and teaching teams evolve. Exploring the usefulness of factors like socio-economic status and high school performance which were only available for a fraction of the student cohort would also be a beneficial investigation. Finally, given student success underpins retention, it would be interesting to compare predictors of student success to predictors of student retention within the selected cohorts.

References

Australian Government Department of Education and Training. (2018). *Higher Education Standards Panel Final Report - Improving Retention, Completion and Success in Higher Education*. Retrieved from https://docs.education.gov.au/system/files/doc/other/final_report_for_publishing.pdf

Campbell, J. P. (2007). *Utilizing student data within the course management system to determine undergraduate student academic success: An exploratory study.* Purdue University,

Chen, H. L., Lattuca, L. R., & Hamilton, E. R. (2008). Conceptualizing engagement: Contributions of faculty to student engagement in engineering. *Journal of Engineering Education, 97*(3), 339-353.

Collier, P. J., & Morgan, D. L. (2008). "Is that paper really due today?": differences in first-generation and traditional college students' understandings of faculty expectations. *Higher Education, 55*(4), 425-446.

Crosling, G., Heagney, M., & Thomas, L. (2009). Improving student retention in higher education: Improving teaching and learning. *Australian Universities' Review, The, 51*(2), 9-18.

Edwards, D., & McMillan, J. (2015). *Completing university in a growing sector: Is equity an issue?* Retrieved from https://research.acer.edu.au/higher_education/43/

Kotsiantis, S. B., Pierrakeas, C., & Pintelas, P. E. (2003, September 3-5, 2010). *Preventing student dropout in distance learning using machine learning techniques.* Paper presented at the International conference on knowledge-based and intelligent information and engineering systems, Oxford, United Kingdom.

Kovacic, Z. (2010, June 19-24, 2010). *Early prediction of student success: Mining students' enrolment data.* Paper presented at the Proceedings of Informing Science & IT Education Conference (InSITE) 2010, Cassino, Italy.

Lizzio, A., & Wilson, K. (2013). Early intervention to support the academic recovery of first-year students at risk of non-continuation. *Innovations in Education and Teaching International, 50*(2), 109-120.

Marbouti, F., & Diefes-Dux, H. A. (2015). *Building course-specific regression-based models to identify at-risk students*. Paper presented at the 122nd ASEE Annual Conference & Exposition, Seattle, Washington.

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education, 103*, 1-15.

Marks, G. (2007). *Completing university: characteristics and outcomes of completing and non completing students*. Retrieved from https://research.acer.edu.au/lsay_research/55/

McMillan, J. (2005). *Course change and attrition from higher education*. Retrieved from https://research.acer.edu.au/lsay_research/43/

Simpson, O. (2006). Predicting student success in open and distance learning. *Open Learning: The Journal of Open, Distance and e-Learning, 21*(2), 125-138.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*(1), 89-125.

Acknowledgements & Ethics Approval

Ben Lewis' contribution through the Research Experience Scheme is gratefully acknowledged. This research was approved by QUT's Human Research Ethics Committee (1800000923).

Copyright statement