PREDICTING UNDERGRADUATE ONSITE STUDENT WITHDRAWALS BASED ON ENROLMENT, PROGRESS, AND ONLINE STUDENT DATA

Sviatlana Burova – Denny Meyer – Wendy Doubé – Pragalathan Apputhurai

Abstract

Student attrition is one of the problems universities target. Predicting student withdrawal can help universities to retain students. The aim of this project is to produce a data mining tool for predicting undergraduate onsite student withdrawals at the Swinburne University of Technology. The tool combines statistical models for predicting course withdrawal with a user friendly software interface.

Initial predictions were made at the beginning of each semester, using logistic regression and classification tree techniques based on enrolment data for commencing students and on both enrolment and progress data for continuing students. Next, evolving weekly hybrid predictions, based on online Blackboard engagement student data and the initial estimated probabilities of withdrawal, were obtained. Usage of Blackboard engagement data improved the accuracy of predictions, with areas under the ROC curves for commencing student models increasing from 0.837 to 0.882, and areas under the ROC increasing from 0.806 to 0.854 for continuing students. Additional validation on the second semester of 2013 data showed even better performance, which confirmed the robustness of the models produced.

Future work on this project includes the development of software for automatically predicting the probability of student withdrawals.

Key words: classification, evolving hybrid predictions, ROC validation, enrolment data, online access data.

JEL Code: C14, C30, C33

Introduction

Student attrition is a problem for many universities. Predicting student withdrawal can help to identify at risk students, and provide them with the necessary assistance in order to successfully complete their degree programs. The more accurate these withdrawal predictions are, the less expense is associated with contacting students, because fewer students who are not at risk are contacted.

Statistical and data mining methods and techniques have been previously applied for researching student attrition. Logistic regression, decision trees, neural networks, support vector machine, K-mean clustering and other methods have been used (Delen, 2010; Luan, 2002; Salazar et al., 2004; Sun, 2010). Multiple types of factors were found to influence student attrition: demographic, psychological, academic, social and institutional (Crosling and Heagney, 2009; Evans, 2000; Franssen & Nijhuis, 2011; Nelson et al., 2009; Park et al., 2011; Scott et al., 2008; Thammasiri et al., 2014).

The aim of this project is to produce a data mining tool for predicting undergraduate on-campus student withdrawals at the Swinburne University of Technology based on student enrolment, progress, and online engagement data. The tool is a combination of statistical models for predicting course withdrawal and a user friendly software interface for non-expert users.

1 Methodology

One of the university's requirements was model transparency, allowing the university to better understand the characteristics of the students most likely to withdraw. This has meant that only relatively simple classification techniques can be applied, namely binary logistic regression and classification trees.

Binary logistic regression is a supervised statistical technique, which is applied for predicting the probability for a binary outcome based on the values of relevant predictor variables, as shown below.

$$\ln\left(\frac{p(x_1,x_2,...,x_n)}{1-p(x_1,x_2,...,x_n)}\right)=a_0+a_1\cdot x_1+...+a_n\cdot x_n$$

where $x_1,x_2,...,x_n$ are predictors, $p(x_1,x_2,...,x_n)$ is the predicted probability, and $a_1,a_2,...,a_n$ are logistic regression coefficients. Logistic regression is preferred to discriminant analysis, because there is no normality assumption (Brace, 2009). Binary logistic regression with stepwise variable selection was used in the research.

Classification trees (Breiman et al., 1984) provide another technique for predicting a categorized outcome which require no normality assumption. A classification tree consists of a hierarchical system of nodes and leaves, where each node is defined using splitting rules based on one of the predictors, and a leaf corresponds to the classification outcome. Various approaches can be used to construct such trees. The classification tree approach used in this research was CART (classification and regression trees) with Gini index as the splitting function. This approach corresponds to Breiman's (1984) implementation and is known to perform well with dummy variables (Shmueli et al., 2007), which is important in this research.

Receiver operating characteristic (ROC) curves show the relationship between the proportion of correctly identified positive cases (sensitivity) and the proportion of correctly identified negative cases (specificity). The area under the ROC curve (AUC) is used as a measure of classifier performance (Weiss, 2004). Variable importance is measured by the significance of the variables in the multivariate models.

All analysis was performed using the R statistical software language (R Development Core Team, 2012). The following R packages were used: caret (Kuhn, 2013), caTools (Tuszynski, 2012), rpart (Therneau et al., 2014).

2 Analysis

Two primary data sources were used for predictions of course withdrawal: (1) enrolment and progress data collected by Swinburne University, and (2) usage of online university facilities extracted from Blackboard, the university's online learning management system. Fewer variables are available for commencing students than for continuing students, mainly because commencing students do not have historical progress data. Six semesters of data from the second semester of 2010 to the first semester of 2013 were used for training and validation of the models for commencing and continuing undergraduate students. The dataset for each cohort was randomly split with 70% for training and 30% for testing. The sizes of datasets for commencing and continuing students were 17,465 and 84,850 cases respectively.

Initial predictions of course withdrawals for undergraduate students were made at the beginning of each semester, based on 53 enrolment variables for commencing students, and 59 enrolment and progress variables for continuing students. Enrolment variables included demographic data (age, gender, socio-economic status, language spoken at home etc.), information about previous education (tertiary entrance score, basis of admission, year left school etc.), and information relevant to the student current education (type of attendance, number of units enrolled, field of education, tuition fee etc.). Progress data contained information about student performance in the previous semester, such as number of distinctions, number of failed subjects, average mark, ratio of passed to enrolled subjects, and so on. For each cohort of students, both logistic regression and classification tree models were developed using training data. The data were unbalanced, with approximately 10% of withdrawals for commencing students and approximately 4% of withdrawals for continuing students. Therefore overweighting of the withdrawal cases was required. Next, evolving weekly hybrid predictions, based on Blackboard data and the initial estimated probabilities of withdrawal, were obtained. These models were validated using the new data for the second semester of 2013 as well as the test data.

3 Results

3.1 Commencing student model

For the initial prediction, based on commencing student enrolment data, the classification tree showed better performance than the logistic regression with an area under the ROC curve of 0.837 (see Fig. 1.a). The findings, based on the most important variables for the initial classifiers are as follows:

- Local students and students with permanent residency, who have relatively low total fees charged for all units in the semester, are more likely to withdraw from their courses;
- Students are more likely to withdraw in first semester[1] than in second semester;
- Withdrawals are more likely for international students with relatively low total tuition fees for the year;
- Withdrawals are more likely when the total sum of credit points for enrolled units in the semester is relatively low.

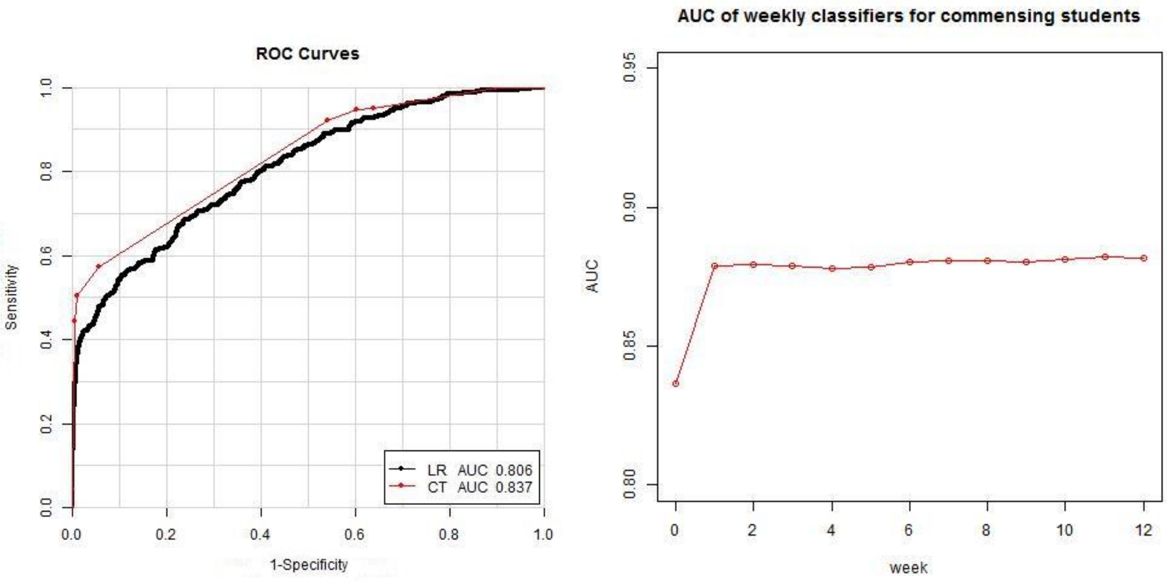
Figure 1.b illustrates how the model fit improves when data for online Blackboard engagement are included in the model. However, after week 2 the inclusion of additional Blackboard engagement data in the model fails to improve the model accuracy further.

The most important Blackboard engagement variable in this model is Blackboard access, with students who fail to log in to Blackboard having a greater risk of course withdrawal. The next most important factor is the number of subjects accessed on Blackboard, with withdrawal less likely when more subjects are accessed. High numbers of Blackboard interactions are also associated with a lower risk of withdrawal, however high numbers of

[1] First semester in Australian tertiary educational institutions starts in February-March and ends in June

Blackboard logins has a slightly positive association with course withdrawal suggesting that these students may have concentration or perseverance problems.

Fig. 1: Performance of classifiers for undergraduate commencing students



a) ROC curves for initial predictions **b) AUC for weekly classifiers**

Validation of the commencing student models using data for the second semester of 2013 showed even better results, The AUC for the initial model was 0.856, but the inclusion of Blackboard engagement data in the model increased this value to 0.923-0.935 depending on the week of the semester. But the strongest validation of all was obtained when it was found that the majority of the commencing semester 1 2014 undergraduate students, with an estimated probability of withdrawal in excess of 0.90, had withdrawn by week 5 of this semester.

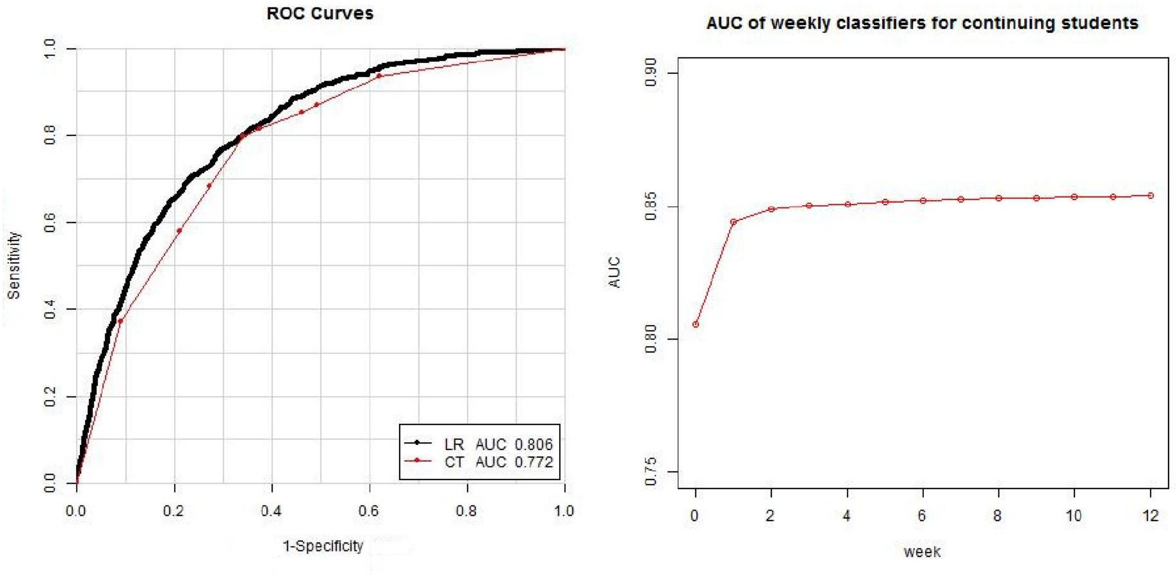
3.2 Continuing student model

Unlike the case for commencing students, for continuing students the initial logistic regression model performed better than a classification tree with an AUC of 0.806 (see Fig.2.a). The predictions of withdrawal for undergraduate continuing student at the start of a semester are based on enrolment and progress data available at the end of previous semester. An interpretation of the important variables for predicting this withdrawal follows.

- The duration of education was the most important variable: the longer students study at the university, the smaller the probability of withdrawal;
- Similarly to the findings for commencing students, students are more likely to withdraw in the first semester than in the second semester;
- If a student does not have any previous academic qualification, the probability of a withdrawal is higher;
- A high ratio of passed units to enrolled units in the previous semester is related to a low probability of withdrawal;
- Withdrawals are less likely when the total sum of credit points for enrolled units in the previous semester is relatively high;
- A high number of failed subjects in the previous semester is associated with a high probability of withdrawal;
- Withdrawals are more likely if one of the subjects in which a student is enrolled in the previous semester was a subject with a high attrition rate (more than 30%).

Figure 2.b shows increasing AUC over a semester suggesting an improving overall accuracy, when more weekly Blackboard engagement data are incorporated in the model.

Fig. 2: Performance of classifiers for undergraduate continuing students



a) ROC curves for initial predictions

b) AUC for weekly classifiers

Similarly to commencing student models, the main Blackboard factors influencing student course withdrawal were failure to log onto the Blackboard system and a relatively small number of units accessed online in any week.

The continuing students models showed even better performance with the second semester of 2013 data, than on test data. The AUC for the initial model was 0.847 but the addition of Blackboard engagement data increased this value to 0.871-0.883 depending on the week of the semester.

Conclusion

Student enrolment and progress data can be successfully used for predicting undergraduate course withdrawals. Classification trees performed better for the commencing student model (area under the ROC curve was 0.837), while a logistic regression showed better performance for continuing students with an area under the ROC curve of 0.806. Usage of Blackboard data improved the accuracy of predictions, with areas under the ROC curves for commencing student models increasing from 0.837 to 0.882, and areas under the ROC increasing from 0.806 to 0.854 for continuing students.

The most important variables in the initial model for commencing students were total fees charged for local and permanent resident students, tuition fee for international students, sum of credit points for all subject enrolments in the semester. All the above listed variables had a negative relationship with the probability of course withdrawal. The variables duration of education, ratio of passed to enrolled units, and sum of credit points had a negative association with course withdrawal in the initial model for continuing students, while number of failed subjects had a positive relationship with course withdrawal. Two factors increased the probability of continuing student withdrawal: absence of previous academic qualification, and studying a subject with a high attrition rate in the previous semester. In both initial models students are more likely to withdraw in the first semester than in the second semester.

For both commencing and continuing students the most important Blackboard variable was Blackboard access, following by number of subjects accessed. Both variables had a negative association with course withdrawal. In the commencing student Blackboard models, number of Blackboard interactions had a negative and number of logins had a positive association with probability of course withdrawal.

Future work on this project includes the development of software to automatically predict the probability of student withdrawals and provide appropriate weekly reports, allowing the university to better support students who are at a high risk of course withdrawal.

Acknowledgment

This project is financially supported by Swinburne University of Technology.

References

Brace, N. (2009). *SPSS for psychologists* (4th ed.). Basingstoke: Palgrave Macmillan.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees.* CRC press.

Crosling, G., Heagney, M., & Thomas, L. (2009). Improving Student Retention in Higher Education: Improving Teaching and Learning. *Australian Universities' Review, 51*(2), 9-18.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*(4), 498-506. doi:10.1016/j.dss.2010.06.003.

Evans, M. (2000). Planning for the transition to tertiary study: A literature review. *Journal of Institutional Research*, *9*(1), 1-13.

Franssen, R., & Nijhuis, J. (2011). Exploring student attrition in problem-based learning: Tutor and student perceptions on student progress. *In Building learning experiences in a changing world* (pp. 139-146). Springer Netherlands.

Kuhn, M. (2013). A Short Introduction to the caret Package.

Luan, J. (2002). Data Mining and Knowledge Management in Higher Education -Potential Applications.

Nelson, K. J., Duncan, M. E., & Clarke, J. A. (2009). Student success: The identification and support of first year university students at risk of attrition. *Studies in Learning, Evaluation, Innovation and Development, 6*(1), 1-15.

Park, C. L., Perry, B., & Edwards, M. (2011). Minimising attrition: strategies for assisting students who are at risk of withdrawal. *Innovations in Education and Teaching International, 48*(1), 37-47.

R Development Core Team. (2012). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/ (ISBN 3-900051-07-0).

Salazar, A., Gosalbez, J., Bosch, I., Miralles, R., & Vergara, L. (2004). A case study of knowledge discovery on academic achievement, student desertion and student retention. In *Information Technology: Research and Education, 2004. ITRE 2004. 2nd International Conference on* (pp. 150-154). IEEE.

Scott, G., Shah, M., Grebennikov, L., & Singh, H. (2008). Improving student retention: A University of Western Sydney case study. *Journal of Institutional Research, 14*(1), 9-23.

Shmueli, G., Patel, N. R., & Bruce, P. C. (2007). *Data mining for business intelligence: Concepts, techniques, and applications in Microsoft Office Excel with XLMiner.* John Wiley & Sons.

Sun, H. (2010). Research on Student Learning Result System based on Data Mining. *IJCSNS, 10*(4), 203.

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications, 41*(2), 321-330.

Therneau, T., Atkinson, B., Ripley, B., & Ripley, M. B. (2014). Package 'rpart'.

Tuszynski, J. (2012). CRAN-package caTools.

Weiss, G. M. (2004). Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter, 6*(1), 7-19.

Contact

Sviatlana Burova
Swinburne University of Technology,
John St, Hawthorn VIC 3122 Australia
sburova@swin.edu.au

Dr Denny Meyer
Swinburne University of Technology,
John St, Hawthorn VIC 3122 Australia
dmeyer@swin.edu.au

Dr Wendy Doubé
Swinburne University of Technology,
John St, Hawthorn VIC 3122 Australia
wdoube@swin.edu.au

Dr Pragalathan Apputhurai
Swinburne University of Technology,
John St, Hawthorn VIC 3122 Australia
papputhurai@swin.edu.au