Chang, M. et al. (Eds.) (2019). Proceedings of the 27th International Conference on Computers in Education. Taiwan: Asia-Pacific Society for Computers in Education

Identifying At-risk Students from Course-specific Predictive Analytics

Chung Lim Christopher KWAN[a*]
[a]*Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong*
*ceclkwan@polyu.edu.hk

Abstract: Identifying at-risk students in a large class of an engineering mathematics course during the delivery of teaching and learning activities is not an easy task to be accomplished by many instructors, particularly in the first few weeks of their studies. In the paper, course-specific predictive analytics, called the multiple linear regression model, the logistic regression model and the classification and regression tree (CART) model are trained, tested and compared with the use of LMS data in the first semester of the academic year 2017-18 such as the level of achievements in online class activities, the mini-project, the mid-term test, assignments, and the final examination for classifying at-risk students as early as possible during the course of study. A feature selection method is used to select statistically significant variables in the development of multiple linear regression and logistic regression models for enhancing the generalizability of both models. It is found that 3 key variables such as the level of achievements in the 6th online class activity, the mid-term test and assignment 2, which may have pedagogically meaningful information, are crucial for classifying at-risk students. Despite the highest accuracy of the CART model, the logistic regression model significantly outperforms the multiple linear regression and the CART models in terms of the recall and f-measure of the testing set. Instead of selecting 3 key variables, the present logistic regression model which only comprises 2 statistically significant variables such as the level of achievements in the 6th online class activity and the mid-term test can be employed to identify at-risk students for early intervention of their studies once the results of the mid-term test and the 6th online class activity are made available at the end of week 7.

Keywords: At-risk Students, Multiple Linear Regression Model, Logistic Regression Model, Classification and Regression Tree, Recall, F-measure

1. Introduction

Monitoring students' learning is one of the important tasks for an instructor to ascertain how well students have learned during the delivery of teaching and learning. Many assessment methods like assignments, a mid-term test, a mini-project, online class activities and a final examination are designed to measure the achievement of the subject intended learning outcomes (Biggs, 2003; Sazhin, 1998). As assessment can serve as feedback to students, students are sometimes informed of their performance in the online class activities immediately and they often get feedbacks on other types of course works like assignments and the mid-term test after one or two weeks. In a large class, more attention should be paid on their progress and attainment to ascertain how well they are on track during learning and how they need further assistance and improvement for students' learning if at-risk students can be identified early in a semester. Through the application of predictive analytics with the use of data extracted from learning management system (LMS), it is possible to identify at-risk students in class and to predict students' success in a course (Lackey, Lackey, Grady, and Davis, 2003; Olani, 2009).

Marbouti, Diefes-Dux, and Strobel (2015) built three logistic regression-based models to identify at-risk students in a large first-year engineering course at weeks 2, 4 and 9 in a semester. The models were optimized for identifying at-risk students with high prediction accuracy, illustrating the value of creating course-specific prediction models rather than generic ones.

2. Course-specific Predictive Analytics

2.1 Dataset

The dataset of an engineering mathematics course offered in the first semester of the academic year 2017-2018 is used for the present study and extracted from Blackboard LMS for the development of course-specific predictive analytics. There are total 240 observations recorded the level of achievement in each assessment task performed by 240 students.

Regarding the dataset, there are 16 input variables comprising 2 assignments, a mini-project, a mid-term test, and 12 online class activities held in each week of a semester. The online class activities are done in face-to-face (F2F) sessions for recording the number of multiple-choice questions correctly attempted as well as students' attendance. The output variable used in the multiple linear regression (MLR) model is the final examination score, but the output variable used in the logistic regression (LR) and CART models is a binary variable (i.e. 0 or 1) which indicates whether the student is at-risk or not. As the result of the final examination in 2017-18 is well known, an integer "1" can be assigned to the binary variable which represents an at-risk student who fails in the final examination. Conversely, an integer "0" is assigned to a not-at-risk student passing the final examination. The input and output variables are summarized in Table 1. The dataset is split into a training set and a testing set with a ratio of 70:30. Initially, all input and output variables are first used for the development of the models.

Table 1

Input and Output Variables used for the Initial Development of Course-specific Predictive Analytics

Input Variable	**Completed by week**	**Type**	**Point**
Assignment 1	5	Numeric	0 - 15
Assignment 2	11	Numeric	0 - 15
Mini-project	8	Numeric	0 - 20
Mid-term test	7	Numeric	0 - 50
1st Online class activity	1	Integer	0 - 3
2nd Online class activity	2	Integer	0 - 8
3rd Online class activity	3	Integer	0 - 4
4th Online class activity	4	Integer	0 - 6
5th Online class activity	5	Integer	0 - 2
6th Online class activity	6	Integer	0 - 3
7th Online class activity	7	Integer	0 - 6
8th Online class activity	8	Integer	0 - 2
9th Online class activity	9	Integer	0 - 3
10th Online class activity	10	Integer	0 - 3
11th Online class activity	11	Integer	0 - 1
12th Online class activity	12	Integer	0 - 1
Output Variable	**Use**	**Type**	**Point**
Final examination	MLR	Numeric	0 - 100
At-risk student	LR & CART	Binary	0 - 1

2.2 Feature Selection Method

Feature selection is a process of selecting a subset of features or variables which are more correlated to the output or predicted variable, yielding a more generalizable model (James, Witten, Hastie, and Tibshirani, 2013). The subset of features or variables may have pedagogically meaningful information to critically identify whether a student is at-risk or not (Macfadyen and Dawson, 2010). It is found that 3 input variables such as the level of achievements in the mid-term test, assignment 2 and the 6th online

class activity are statistically significant with the p-value of being below 0.05 in the development of MLR model with the use of feature selection.

The non-significant variables are then removed if the p-value of the input variable is not below 0.05. Instead of selecting 3 input variables, the MLR model only comprises 2 statistically significant variables such as the level of achievements in the mid-term test and the 6th online class activity for early identification of at-risk students by the end of week 7.

Despite the fact that the model shows a high accuracy of 0.833 on classifying both at-risk student (True Positive) and not-at-risk student (True Negative), the sensitivity and the specificity are 0.467 and 0.930 respectively. A student is classified as an at-risk student if his/her final examination score is predicted to be below a passing score, P. The confusion matrix for the classification of both at-risk and not-at-risk students is tabulated in Table 2.

Table 2

Confusion matrix for the classification of both at-risk student and not-at-risk students in MLR model (Testing set)

	Predicted Negative (Score ≥ P)	Predicted Positive (Score < P)
Observed Negative (Score ≥ P)	53	4
Observed Positive (Score < P)	8	7

The process of selecting statistically significant variables is still adopted for the development of logistic regression (LR) model. It is found that the above-mentioned 3 variables are statistically significant in the LR model with the p-value of being below 0.05. After removing the non-significant variables (i.e. the p-value of the input variable is not below 0.05), the present LR model only comprises 2 statistically significant variables such as the level of achievements in the mid-term test and the 6th online class activity for the early identification of at-risk students. The confusion matrix of the LR model is computed for the testing set using different thresholds of 0.5 and 0.2 respectively. It is found that the sensitivity is increased from 0.667 to 0.867 and the specificity is decreased from 0.930 to 0.824. The accuracy of the present model is decreased from 0.875 to 0.833. The confusion matrix for the classification of both at-risk and not-at-risk students in LR model with the threshold of 0.2 is depicted in Table 3.

Table 3

Confusion matrix for the classification of both at-risk student and not-at-risk students in LR model with the threshold of 0.2 (Testing set)

	Predicted Negative	Predicted Positive
Observed Negative	47	10
Observed Positive	2	13

All input and output variables are considered in the development of the CART model. The following tree which is based on the use of training set and the complexity parameter of 0.02 is built and shown in Figure 1. Whether a student is at-risk or not can apparently be inferred from the CART tree. For example, a student can be classified as a at-risk student if the score of mid-term test is not above 24.75 and the score of a class activity is not above 1.

The confusion matrix of the present CART model is computed for the testing test. It is found that the accuracy, the sensitivity and the specificity of the present CART model are 0.861, 0.667 and 0.912 respectively. A summary of the confusion matrix is depicted in Table 4.

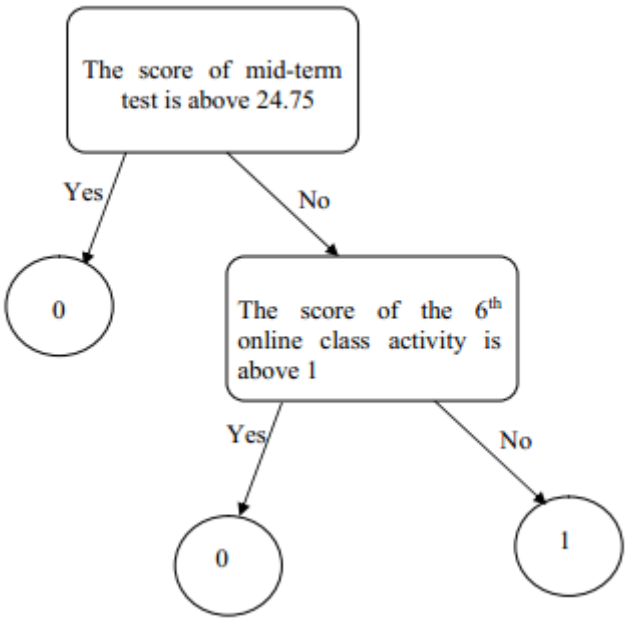

Figure 1. The CART model built with the complexity parameter of 0.02.

Table 4

Confusion matrix for the classification of both at-risk student and not-at-risk students in CART model with the complexity parameter of 0.02 (Testing set)

	Predicted Negative	Predicted Positive
Observed Negative	52	5
Observed Positive	5	10

The performance of the present models is further evaluated in terms of accuracy, precision, recall (i.e. sensitivity), and f-measure as shown in Table 5. Among three present models, it is found that the LR model has both the highest recall of 0.867 and the highest f-measure of 0.684.

Table 5

Accuracy, precision, recall and f-measure of the present models

	MLR	LR	CART
Accuracy	0.833	0.833	0.861
Precision	0.636	0.565	0.667
Recall	0.467	0.867	0.667
F-measure	0.539	0.684	0.667

3. Discussion

The feature selection method is used to select 3 statistically significant variables such as the level of achievements in the mid-term test, assignment 2 and the 6th online class activity in the development of MLR and LR models with the help of the training set. However, the result of assignment 2 can only be made available at the end of week 11. Instead of selecting 3 statistically significant variables, both models merely comprise 2 statistically significant variables such as the level of achievements in the

mid-term test and the 6th online class activity for identifying at-risk students by the end of week 7. In addition, the non-significant variables are removed during the training stage for the purpose of developing the MLR and LR models which can be more generalizable. If the non-significant variables are also included in the models, the models will be less generalizable for prediction.

It is found that 3 key variables such as the level of achievements in the 6th online class activity, the mid-term test and assignment 2, which may have pedagogically meaningful information, are crucial for classifying at-risk students simply because students are required to demonstrate high levels of understanding such as relational and extended abstract for accomplishing these assessment tasks, as distinguished by Structure of Observed Learning Outcomes (SOLO) taxonomy (Biggs and Collis, 1982).

The LR model comprising 2 statistically significant variables such as the level of achievements in the 6th online class activity and the mid-term test can first be employed for the identification of at-risk students and intervention of their studies at the end of week 7. They will be informed immediately by emails, and closely monitored via consultative meetings during the period from week 8 to week 11. The LR model which comprises 3 statistically significant variables such as the level of achievements in the 6th online class activity, the mid-term test and assignment 2 can further be used to decide whether they are still classified as at-risk students or not, once the result of assignment 2 is made available at the end of week 11.

The current course-specific LR model also outperforms the other generic LR model developed by Macfadyen and Dawson (2010) with the use of LMS tracking variables only such as the number of assessments finished, the total number of discussion messages posted and the number of mail messages sent in terms of the accuracy and recall, highlighting the value of creating course-specific predictive analytics as the focus of the research.

4. Conclusion

It is concluded that 3 statistically significant variables such as the level of achievements in the 6th online class activity, the mid-term test and assignment 2, which may have pedagogically meaningful information, are crucial for identifying at-risk students. Despite the highest accuracy of the CART model, the logistic regression model significantly outperforms the multiple linear regression and the CART models in terms of the recall and f-measure of the testing set. Instead of selecting 3 key variables, the present logistic regression model which only comprises 2 statistically significant variables such as the level of achievements in the 6th online class activity and the mid-term test can be employed for early identification of at-risk students and timely intervention of their studies once the results of the mid-term test and the 6th online class activity are made available at the end of week 7.

References

Biggs, J. (2003). Teaching for quality learning at university, 2nd Edition. *Society for Research into Higher education & Open University Press*.

Biggs, J. B., & Collis, K. F. (1982). Evaluating the quality of learning: The SOLO taxonomy. New York: *Academic Press.*

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Linear model selection and regularization an introduction to statistical learning. *Springer*.

Lackey, L. W., Lackey, W. J., Grady, H. M., & Davis, M. T. (2003). Efficacy of using a single, non-technical variable to predict the academic success of freshmen engineering students. *Journal of Engineering Education*, 92(1), 41-48.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, 54(2), 588-599.

Marbouti, F., Diefes-Dux, H. A., & Strobel, J. (2015). Building course-specific regression-based models to identify at-risk students. *In The american society for engineering educators annual conference*. Seattle, WA.

Olani, A. (2009). Predicting first year university students' academic success. *Electronic Journal of Research in Educational Psychology*, 7(3), 1053-1072.

Sazhin, S. (1998). Teaching mathematics to engineering students. *International Journal of Engineering Education* 14, 145-152.

WORKSHOP 11 - New Endeavours of Implementing Computational Thinking in K-12 Education