122nd ASEE Annual
Conference & Exposition
June 14 - 17, 2015
Seattle, WA
Making Value for Society

Paper ID #11229

Building course-specific regression-based models to identify at-risk students

Mr. Farshid Marbouti, Purdue University, West Lafayette

Farshid Marbouti is currently pursuing his Ph.D. in Engineering Education at Purdue University. His research interest is first-year engineering and specifically using learning analytics to improve first-year engineering students' success. He completed his M.A. in the Educational Technology and Learning Design at Simon Fraser University in Canada, and his B.S. and M.S. in computer engineering in Iran.

Prof. Heidi A. Diefes-Dux, Purdue University, West Lafayette

Heidi A. Diefes-Dux is a Professor in the School of Engineering Education at Purdue University. She received her B.S. and M.S. in Food Science from Cornell University and her Ph.D. in Food Process Engineering from the Department of Agricultural and Biological Engineering at Purdue University. She is a member of Purdue's Teaching Academy. Since 1999, she has been a faculty member within the First-Year Engineering Program, teaching and guiding the design of one of the required first-year engineering courses that engages students in open-ended problem solving and design. Her research focuses on the development, implementation, and assessment of modeling and design activities with authentic engineering contexts. She is currently a member of the educational team for the Network for Computational Nanotechnology (NCN).

Dr. Johannes Strobel, Texas A&M University

Dr. Johannes Strobel is Director, Educational Outreach Programs and Associate Professor, Engineering & Education at Texas A&M, College Station. He received his M.Ed. and Ph.D. in Information Science & Learning Technologies from the University of Missouri. His research/teaching focuses on engineering as an innovation in pK-12 education, policy of STEM education, how to support teachers and students' academic achievements through engineering, engineering 'habits of mind' and empathy and care in engineering. He has published more than 140 journal articles and proceedings papers in engineering education and educational technology and is the inaugural editor for the Journal of Pre-College Engineering Education Research.

Building Course-Specific Regression-Based Models to Identify At-Risk Students

Abstract

The first step in helping students who may fail a course is to identify them as early in the semester as possible. Predictive modeling techniques can be used to create an early warning system which predicts students' success in courses and informs both the instructor and the students of their performance. One common problem with existing early warning systems is that they typically employ a general model that cannot address the complexity of all courses. In this study, we built three models to identify at-risk students in a specific large first-year engineering course at three important times of the semester according to the academic calendar. Then the models were optimized for identifying at-risk students. The models were able to identify 79% of at-risk students at week 2, 90% at week 4, and 98% at week 9. This high accuracy illustrates the value of creating course specific prediction models instead of generic ones.

Introduction

Despite numerous efforts to improve students' retention and success over the past 30 years, retention rates remain the same[1]. According to the National Collegiate Retention and Persistence to Degree Rates report in 2012, the first to second year retention rate for undergraduate students was 66.5%[2]. While this rate is higher for four-year institutes (~71%) than for two-year institutes (~55%)[2], low retention rates reveal a serious problem in higher education. Almost one third of students leave college after experiencing their first year. This trend more or less continues through the next years of school - only 45% of students who enter college graduate after 5 years[2]. Graduation rates for students in the fields of Science, Technology, Engineering, and Mathematics (STEM) are even lower than the general rates.

Over the past decade, the use of course or learning management systems (CMS/LMS) in higher education has increased[3]. As a result, higher education institutes have extensive databases that store many aspects of students' performance. Instead of relying only on experience or anecdotal evidence, another way to enhance students' success is to analyze students' performance data[4,5]. The results of such analyses can help instructors understand what leads to a student passing or failing a course.

To help students' who may fail a course, it is crucial to identify them as early in the semester as possible. Prediction of success can help the course instructor to identify at-risk students and help them to be successful in the course. In an attempt to predict students' grades early in the semester, some instructors use course syllabi grading criteria and apply it to the available performance information in order to calculate an early grade for students. This method may only be useful closer to the end of semester when the majority of performance information is available and it may be extremely inaccurate at the beginning of the semester. Using such inaccurate methods at the beginning of semester can result in wrong predictions and lead to mistrust of students in the predictions.

With the use of predictive modeling techniques, it is possible to better predict students' success in a course[6-8]. A predictive model can be used as an early warning system, which predicts students' success in courses and informs both the instructor and the students of their performance. Use of an early warning system in a course, along with guidelines on how to succeed in the course, can increase students' success in the course[9-11].

One common problem with early warning systems, which are currently being used, is that they typically employ a general model that cannot address the complexity of all courses. Course Signals[12] is a pioneer early warning system that is currently being used at Purdue University. Course Signals uses the same model for all courses across campus. The Course Signals' predication model was originally created based on campus-wide data that included more than 600 different courses in 75 departments and nine colleges, without differentiating different courses, departments, or even colleges. The Student Success System (S3) is another early warning system that uses predictive modeling[10]. S3 uses an ensemble approach including four prediction models. However, these four models are the same model for different courses.

While this one-model-fits-all approach is cost effective for higher education institutes, it lowers the accuracy of the model. Over the past decade, more instructors are trying to add more complex learning objectives and using active learning strategies rather than solely lecturing. The new pedagogies are usually implemented along with new assessment methods that do not fit the traditional homework and exam framework. Thus, the assessed components can vary largely from one course to another. Therefore, using one model for different courses can significantly reduce the accuracy of the model. Creating predictive models at the specific course level increases the accuracy of the model[11].

A second shortcoming of current early warning systems is the assumption that students' success during the semester can be predicted with only one model without taking into account time of the semester[11,12]. As the performance information becomes available each week during the semester, students' success should be predicted more accurately. Because the information changes during the semester, different models may need to be built based on the availability of new information to predict students' success each week of the semester. It is expected that closer to the end of the semester, the models converge to the course syllabus grading criteria. In other words, because the final grade is calculated by a (typically linear) combination of course components based on the syllabus, the model at the end of the semester should be very close to the course syllabus grading criteria. However, earlier in the semester, the model may differ significantly from the syllabus grading criteria.

Research Questions

In this study, we built three models to predict students' success in a course at three critical times of the semester using academic factors (i.e., grades during the semester). Then, we evaluated and improved the accuracy of the predictions to identify the students who may fail the course. In addition, we identified course components that are important for students' success in the course. These models are used as a proof of concept to showcase and move toward course-specific prediction models rather than the existing generic ones. The research questions are:

- To what extend does accuracy of prediction of students' success in a course change across the weeks during the semester based on available performance information?

- What are the most important course components (e.g., homework, quiz, project, exam) that link to student success?

Methods

Participants and Settings

Participants are the approximately 3400 first-year engineering (FYE) students at a large mid-western university who enrolled in a required 2-credit FYE course in Spring 2011 and 2012. In this course students learn engineering fundamentals, oral and graphical communication, logical thinking, team work, how to construct engineering solutions to open-ended problems, and how to use engineering tools such as Excel and MATLAB[13].

This FYE course is a good example to showcase how to create and use a predictive model for a course. This course has a large number of students (about 1700 every Spring semester). It also uses several different assessment components (e.g., quizzes, homeworks, project milestones, exams) and most are collected on a weekly basis. Thus, it has a large number of new student performance data each week. As a result, the prediction model and its accuracy will change every week based on newly available information.

Data and Analysis Plan

We used the course performance data (i.e. course grades) from Spring 2011 (about 1700 students) as the training data, to build and refine the models, and then tested the models with Spring 2012 data (about 1700 students). The performance data were selected according to the course syllabus. The data include grades on weekly attendance, homework, and quizzes, as well as exams (Table I). Project milestones, open-ended problem solving activities, and team evaluations were only available at week 9 in one of the semesters. Thus they were not included in this study.
The university academic calendar has three important dates for students who may want to drop a course:

- Beginning of week 3: last day to drop a course without it appearing on record.
- Beginning of week 5: last day to withdraw from a course with a grade of W (withdraw).
- Beginning of week 10: last day to withdraw a course with a grade of W or WF (withdraw fail).

We created three predictive models, one for week 2, week 4, and week 9. For each of these three weeks, we created the model using logistic regression and data from Spring 2011. Then, we refined the system to keep only the most significant factors and minimize the complexity of the system. Finally, we tested the refined model on the Spring 2012 dataset, which had not been used to create the model (test data). The test showed how the model predicts future unseen data. Logistic regression is a reliable predicting method that is commonly used in educational settings[14-18]. In addition, by using logistic regression, it is possible to explain which course components are significant predictors of students' success and to what extent.

Most engineering programs have a minimum requirement of maintaining a C or better GPA to stay in the program. In addition, to accept a course as a transfer course, students should receive a C or better in the course. Thus, in this research, similar to other studies[11], success (passing) is defined as getting at least a C grade.

TABLE I. NUMBER OF GRADES THAT WERE AVAILABLE

Assessment Component	Week 2	Week 4	Week 9
Attendance	4	8	18
Homework	2	4	9
Quiz	4	8	18
Exam	0	0	1
Project	0	0	0*
Open-ended Problems	0	0	0*
Team Evaluations	0	0	0*

[*] At week 9, number of available grades was different for Spring 2011 and Spring 2012. Thus, in thus study they were not included in the predictive model.

For each week, we used the academic information available in that week to create a model. For each syllabus component, we calculated a percentage score based on the average of that component so far. For example at week 4 of the semester, we only used the data points (e.g., attendance, quizzes, homeworks) that were available and created one percentage score for attendance, one for homework, and one for quiz. Then, based on these percentage scores, we created the week 4 model. Finally, we used the week 4 model to predict students' success at the end of the semester and calculated the accuracy and error rate of the predictions. To calculate the accuracy of each model, the number of correct predictions was divided by the total number of predictions (i.e., students).

Because the goal of this study is to identify at-risk students, we also calculated and optimized the false negative (type II) error. To calculate the accuracy of identifying at-risk students, the number of at-risk students that were identified correctly by the model was divided to the total number of at-risk students. By optimizing the models, we minimized the number of students who may fail the course but will not be identified by the model. This slightly increased the false positive (type I) error or the number of students who may pass the course but will be identified as potentially failing. Although in general the accuracy of the model is important, it is crucial to identify as many potentially at-risk students as possible.

Results

Week 2 Model

At the end of week 2, students may decide to drop the course without it appearing on their records. At this week, only homework, quiz grades and attendance records were available. A logistic regression model was created based on Spring 2011 data to predict success (i.e., receiving a grade of A, B, or C) or failure (i.e., receiving a grade of D or F) in the course. The results are shown in Table II. In this model, only quiz and homework grades were significant predictors of success. Thus attendance was removed from the model. The final model for week 2 is reported in Table III and equation (1):

$$Ln(p/1\text{-}p) = (4.06)*homework + (2.73)*quiz - 2.23 \quad (1)$$

where *p* is probability of a student passing the course based on two weeks of *homework* and *quiz* grades.

Homework contributed 20% and quiz 10% in the final grade. The relative importance of quiz to homework in the equation (1) is slightly more than the syllabus percentages for the quiz and homework. Thus quiz has slightly more effect than its syllabus weight.

Based on the final regression model for week 2 (Table III), probabilities of Spring 2012 students' passing the course were calculated using data from only the first two weeks of class. The outcome of the regressions equation (1) is the probability of a student passing the course. The probability threshold for passing the course was originally set to 0.5 and the number of correct predictions was calculated. In other words if the probability *p* was 0.5 or more it was predicted that the student pass the course, and if *p* was less than 0.5 it was predicted that the student do not pass the course. The overall accuracy of the predictions was 93%. Because in this study our primary concern was students who may fail the course, accuracy of predictions for students who failed the course was also calculated. While the model had good performance overall, the accuracy of identifying at-risk students was only 24% (Table IV).

Reviewing the cases in which the model made wrong predictions revealed that the model was too optimistic in its predictions. In other words, most of the error in the system was false negative and the model failed to identify the at-risk students. One solution to reduce the error of identifying at-risk students was to raise the probability threshold for passing the course. In Spring 2011, 8% of students failed the course and 92% passed the course. Thus the new probability threshold was set to 0.92 instead of the original probability of 0.5. This increase in the threshold resulted in improving the accuracy of identifying at-risk students to 79%, however, it decreased the overall accuracy of the system to 77% (Table IV).

TABLE II. PRELIMINARY PREDICTIVE MODEL FOR WEEK 2

	Homework	**Quiz**	**Attendance**
p	< 0.001	< 0.001	0.058
Estimate	3.85	2.42	1.34

TABLE III. FINAL PREDICTIVE MODEL FOR WEEK 2

	Homework	**Quiz**	**Attendance**
p	< 0.001	< 0.001	-----
Estimate	4.06	2.73	-----

TABLE IV. ACCURACY OF WEEK 2 PREDICTIONS

Model	**Threshold**	**Overall**	**At-risk students**
Original	0.5	93%	24%
Modified	0.92	77%	79%

Week 4 Model

At week 4, students may withdraw from the course with a W grade. At this week, similar to week 2, only homework grades, quiz grades, and attendance records were available. A logistic regression model was created based on Spring 2011 data to predict success in the course. In this model, all three available course components were significant predictors of success. The model for week 4 is illustrated in Table V and equation (2):

$$Ln(p/1-p) = (7.46)*homework + (3.37)*quiz + (4.66)*attendance - 9.30 \quad (2)$$

where *p* is probability of a student passing the course based on four weeks of *attendance* records and *homework* and *quiz* grades.

Although the attendance percentage in the syllabus is only 5%, it has more effect in equation (2) than the quiz with 10% share in the final grade. Thus attendance is more important in success of students in the course than its weight in the syllabus.

Based on the final regression model for week 4 (Table V), probabilities of Spring 2012 students' passing the course were calculated using only the first four weeks of data. The probability threshold for passing the course was originally set to 0.5 and the number of correct predictions was calculated. The overall accuracy of the predictions was 94% and for at-risk students was 52% (Table VI).

TABLE V. PREDICTIVE MODEL FOR WEEK 4

	Homework	**Quiz**	**Attendance**
p	< 0.001	< 0.001	<0.001
Estimate	7.46	3.37	4.66

TABLE VI. ACCURACY FOR WEEK 4 PREDICTIONS

Model	**Threshold**	**Overall**	**At-risk students**
Original	0.5	94%	52%
Modified	0.92	81%	90%

Similar to week 2, the model was too optimistic in its predictions. Thus, the threshold was changed from 0.5 to 0.92 to enhance the accuracy of the model for at-risk students. The increase in the threshold resulted in improving the accuracy of identifying at-risk students to 90%, however, it decreased the overall accuracy of the system to 81%.

Week 9 Model

Week 9 is the last chance for students to withdraw from a course with a W or WF grade. At week 9, in addition to homework and quiz grades and attendance records, the first midterm exam grades were also available. In Spring 2011, team evaluations and open-ended problem solving activity and project's grades were also available at week 9. However, since these grades were not available at this week in Spring 2012, we did not include these grades in the model. A logistic regression model was created based on the Spring 2011 data to predict success in the course. The results are shown in Table VII. In this model, exam, homework, and attendance grades were significant predictors of success. Thus, quiz grades was removed from the model. The final model for week 9 is illustrated in Table VIII and equation (3):

$$Ln(p/1-p) = (11.08)*exam + (10.90)*homework + (11.30)*attendance - 22.05 \quad (3)$$

where *p* is probability of a student passing the course based on *exam* grade and nine weeks of *attendance* records and *homework* grades.

In equation (3), exam, homework, and attendance have coefficients very close to each other. However, in the syllabus, exam weight is 30% of the final grade, homework is 20%, and attendance is only 5%.

Based on the final regression model for week 9 (Table VIII), probability of Spring 2012 students' passing the course was calculated using the first nine weeks of data. The probability threshold for passing the course was originally set to 0.5 and the number of correct predictions was calculated. The overall accuracy of the predictions was 95% and for at-risk students was 83% (Table IX). Similar to previous weeks, though to a lesser extent, the model was optimistic in its predictions. Again, the threshold was changed from 0.5 to 0.92 to enhance accuracy of the model for at-risk students. The increase in the threshold resulted in improving the accuracy of identifying at-risk students to 98%. However, it decreased the overall accuracy of the system to 87%.

TABLE VII. PRELIMINARY PREDICTIVE MODEL FOR WEEK 9

	Exam	**Homework**	**Quiz**	**Attendance**
p	<0.001	<0.001	0.039	<0.001
Estimate	10.23	10.76	2.78	10.50

TABLE VIII. FINAL PREDICTIVE MODEL FOR WEEK 9

	Exam	**Homework**	**Quiz**	**Attendance**
p	<0.001	<0.001	-----	<0.001
Estimate	11.08	10.90	-----	11.30

TABLE IX. ACCURACY OF WEEK 9 PREDICTIONS

Model	**Threshold**	**Overall**	**At-risk students**
Original	0.5	95%	83%
Modified	0.92	87%	98%

Discussion

In this study, we created course performance prediction models for weeks 2, 4, and 9 of the semester. The primary focus of the prediction models was to identify at-risk students at critical weeks of the semester. These predictions can help students, advisors, and instructors in decision-making.

The results show that the estimated models (i.e., regression equations), as expected, are not similar to the syllabus weights. Especially attendance is more important in determining whether a student pass the course or not than its syllabus weight. Thus instructors cannot solely rely on their syllabus formula to calculate students' grades during the semester.

Fig. 1 summarizes the accuracy of the predictions for the three weeks. Changing the probability threshold from the default 0.5 to 0.92, which was the passing student percentage for the course in Spring 2011, significantly improved the accuracy of the models for at-risk students. However, it slightly decreased the overall accuracy of the model. This resulted in the identification of more at-risk students but increased false positive error.

The enhancing week 2 model could identify more than two-thirds of at-risk students based on only homework and quiz grades. While this is very promising for course-specific predictions models, it raises concerns about how the instructors can change students' behaviors in a course. Predicting more than two-thirds of failures at week 2 illustrates that these students, despite all instructors' efforts during the semester, never improved their educational behavior in the course. Identifying these students early in the semester and providing personalized feedback to them might be one way to help these students.

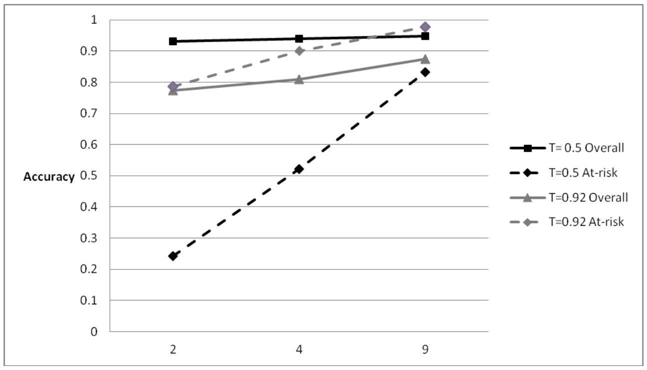

Fig. 1. Accuracy of predictive models

At week 4, the enhanced model could identify 90% of at-risk students, while the overall accuracy was 81%. As it was expected, the accuracy of the model increased at week 9 and 98% of at-risk students were identified. These results show a high accuracy in predicting students who may fail a course with course specific models.

Unlike most prediction models, our model is customized to one specific course. Creating a model for a specific course allows for taking into account the complexity of the course. Therefore, it is more likely to have high predictive power compared to generic prediction models that use the same model for different courses across disciplines.

This study can be a first step in customizing prediction models. Although the model that is being built is specific for one course, the recommendations for how to build the model, how to take into account the complexity of the course, and how to calculate the accuracy for different weeks can be reused by other courses. It is important to remember that to achieve high accuracy and effectiveness in predictions, each course prediction model should be created separately using the course data.

Limitations

The described prediction method gives better results (i.e., accuracy) in large classrooms. For classes with a small number of students, which the instructor is more likely to know the students on a personal level, other methods might be preferred. In addition, the prediction models were created only for three critical weeks in the semester. To enable instructors to identify at-risk students at any time of the semester, a separate prediction model for each week or a model that can predict students' success at any time of the semester should be created. In addition, the models were created based on data for only one semester. Increasing the size of the data set used to create the model by adding more semesters may increase the accuracy of the models.

Future work

The next step is to compare different prediction methods to create a high accuracy model that can identify at-risk students at any given week of the semester. In addition, increasing the datasets to more than one semester may increase the accuracy of the models.

References

[1] J. Gardner and A. Koch, "The First-Year Experience Thirty Years Later: It is Time for an Evidence-Based, Intentional Plan," Purdue SoLar Flare Practitioners' Conference, West Lafayette, IN, 2012.

[2] ACT, National Collegiate Retention and Persistence to Degree Rates, 2012. Available at http://www.act.org/research/policymakers/pdf/retain_2012.pdf

[3] K. Green, "Campus computing, 2009," The 19th national survey of computing and information technology in US higher education, 2009.

[4] S. Huang and N. Fang, "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models," Computers & Education, 61, 2013, pp. 133-143.

[5] R. White, "Predicting likely student performance in a first year Science, Technology, Society course," International Journal of Innovation and Learning, 12(1), 2012, pp. 72 - 84.

[6] L. Lackey, W. Lackey, H. Grady, and M. Davis, "Efficacy of using a single, non-technical variable to predict the academic success of freshmen engineering students," Journal of Engineering Education, 92(1), 2003, pp. 41-48.

[7] Q. Jin, P.K. Imbrie, J. Lin, X. and Chen, "A multi-outcome hybrid model for predicting student success in engineering," 118th ASEE Annual Conference and Exposition, June 2011, Vancouver, BC, Canada.

[8] A. Olani, "Predicting First Year University Students' Academic Success," Electronic Journal of Research in Educational Psychology, 7(3), 2009, pp. 1053-1072.

[9] K. Arnold and M. Pistilli, "Course signals at Purdue: using learning analytics to increase student success," Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 2012, Vancouver, BC.

[10] A. Essa and A. Hanan, "Student success system: Risk analytics and data visualization using ensembles of predictive models," ACM International Conference Proceeding Series, 2012.

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

[12] J. Campbell, Utilizing student data within the course management system to determine undergraduate student academic success: An exploratory study, Purdue University, 2007.

[13] Purdue University, Purdue Course Catalogue, 2013. Available at http://mypurdue.purdue.edu

[14] R. Smith and P. Schumacher, "Academic Attributes of College Freshmen that Lead to Success in Actuarial Studies in a Business College," Journal of Education for Business, 81(5), 2006, pp. 256-260.

[15] B. French, J. Immekus, and W. Oakes, "An examination of indicators of engineering students' success and persistence," Journal of Engineering Education, 94(4), 2005, pp. 419-425.

[16] J. Purdie and V. Rosser, "Examining the Academic Performance and Retention of First-Year Students in Living-Learning Communities and First-Year Experience Courses," College Student Affairs Journal, 29(2), 2011, pp. 95-112.

[17] I. Wait and J. Gressel, "Relationship between TOEFL score and academic success for international engineering students," Journal of Engineering Education, 98(4), 2009, pp. 389-398.

[18] G. Zhang, M. Padilla, T. Anderson, and M. Ohland, “Gender differences in major selection and academic success for students leaving engineering,” 2005 ASEE Annual Conference and Exposition: The Changing Landscape of Engineering and Technology Education in a Global World, June 2005, Portland, OR, United States.