Open Academic Analytics Initiative: Initial Research Findings

Eitel J.M. Lauría, Erik W. Moody , Sandeep M. Jayaprakash, Nagamani Jonnalagadda, Joshua D. Baron

Marist College, Poughkeepsie, NY, USA

ABSTRACT

This paper describes the results on research work performed by the Open Academic Analytics Initiative, an on-going research project aimed at developing an early detection system of college students at academic risk, using data mining models trained using student personal and demographic data, as well as course management data. We report initial findings on the predictive performance of those models, their portability across pilot programs in different institutions and the results of interventions applied on those pilots.

Categories and Subject Descriptors

J.1 [**Administrative Data Processing**] *Education*; K.3.1 [**Computer Uses in Education**] *Collaborative learning, Computer-assisted instruction (CAI), Computer-managed instruction (CMI), Distance learning*

General Terms

Algorithms, Measurement, Design, Experimentation

Keywords

Learning Analytics, Open Source, Data Mining, Course Management Systems, Portability, Intervention

1. INTRODUCTION

During the spring 2012 semester, the Open Academic Analytics Initiative (OAAI), led by Marist College, successfully deployed an open-source Learning Analytics solution, developed by the project during the fall, at two community colleges (Cerritos College and College of the Redwoods) and one Historically Black College and University (HBCU) (Savannah State University) as a means to further research in this emerging field. Our spring pilots involved a total of 1379 students, 67% of whom were considered low-income students, who were enrolled in introductory-level courses with, generally, three sections each being taught by the same instructor (e.g. BIOL 101 Section 1, 2 and 3). Each course section was then assigned to either a control or one of two treatment groups, thereby standardizing the instructional delivery to the extent possible across all three. Students in the two treatment groups who had been identified by our predictive model, which uses student demographic, aptitude and course management system usage data, as being likely to not complete the course received interventions designed to help them succeed.

LAK '13, April 08 - 12 2013, Leuven, Belgium

This paper discusses our results from our spring pilots with regards to our two primary research areas: the "portability" of predictive models from one academic context to another and the effectiveness of different intervention strategies. In summary, we have found the following to date:

- The predictive model developed using student data from Marist College was very similar with regards to the predictive elements (e.g. cumulative GPA) and correlation strengths as the predictive model developed at Purdue University by Dr. John Campbell.
- The accuracy of the predictive model built using Marist student data performed considerably higher than sheer randomness when deployed at both community colleges and HBCUs which exceed our expectations.
- We have found a statistically significant difference between mean course grades when comparing all students in our two treatment groups (awareness and OASE) to our controls.
- In addition, we have found a statistically significant difference with regards to "content mastery" (a C grade or above) as well as withdrawal rates between our combined treatment groups and our controls.

These findings, which are detailed on the following pages, are noteworthy as they indicate that predictive models used in Learning Analytics are more "portable" than initially anticipated as well as the fact that our intervention strategies have been effective in improving student outcomes.

2. PREVIOUS WORK

Academic analytics is still a developing field, which has arisen in higher education as a natural reverberation of the successful application of data mining in the business world. Although it has yet to be implemented broadly across a range of institutional types, student populations and learning technologies, the increasing amount of research shows the level of interest that this domain has attained in both the data mining and the educational communities. A sample of this on-going research is listed below.

As early as 2004, clustering algorithms were used to find affinity patterns of user behavior in course management systems [9]. Data mining of temporal participation indicators was applied to measure student contributions to discussion forums in online courses [6]. In 2005, researchers at the University System of Georgia used discriminant analysis on high school GPA and SAT mathematics scores to predict completion of fully online general education courses[7]. In his dissertation work at Purdue University, Campbell[3] used factor analysis and logistic regression on course management system usage data and student demographics to produce predictive models of student performance. Romero et al [8] applied data mining techniques on Moodle data (Moodle is a popular open source course management system). More recently Bravo Agapito et al [2] used C4.5 decision tree rules to detect symptoms of low performance in e-learning systems. The interest in this domain by the data mining community as a rich applied research field is evidenced by the fact that the KDD Cup 2010's theme was prediction of student performance on mathematical problems from logs of student interaction with Intelligent Tutoring Systems [4].

3. OVERVIEW OF INITIAL PORTABILITY FINDINGS

Many questions exist in the emerging field of Learning Analytics with regards to the degree to which predictive models which are built based on data from one particular institution type and student population can be effectively deployed in academic contexts where the institution type and/or student population differ significantly. The degree to which predictive models may be "portable" has major implications for the scaling of learning analytics across multiple institutions and even higher education itself. The OAAI has researched this issue of portability in two phases, first looking at the degree to which a model built using data from Marist College compared to the original model built at Purdue University (which lead to the development of Course Signals) and then, second, deploying the Marist-developed model at community colleges and HBCUs and investigating the accuracy of the model in these different academic contexts. Results from these two research efforts, which are detailed below, have shown that predictive models appear to be more portable than initially anticipated. See [5] for details regarding model implementation, algorithms and software platforms used by the OAAI project.

3.1 Phase One Portability Research Results: Comparing Marist Predictive Model to Purdue Model

During the summer of 2011 the analytics research team at Marist College developed a predictive model using the same student demographic (e.g. age), aptitude (e.g. SAT scores), and course management system usage (e.g. number of course visits) that were used by Dr. John Campbell at Purdue University in his original dissertation research [3], except our data was from Marist students and based on usage of the Sakai Course Management System (CMS). Although Marist College and Purdue University differ in obvious ways (e.g., institutional type, size, and instructional approaches) they do share a number of similarities which are particularly pertinent to this study. These include percentage of students receiving federal Pell Grants (Marist 11%, Purdue 14%), percentage Asian/Black/African American/Hispanic students (Marist 11%, Purdue 11%), and ACT composite 25th/75th percentile (Marist 23/27, Purdue 23/29).

We compared the model predictors that were correlated with student grades (as was done at Purdue) as means to understand to the degree to which the models differed. In general, we found the same statistically significant elements as Purdue with similar correlation strengths. These initial findings on "portability" were included in a paper presented at the 2012 international Learning Analytics and Knowledge (LAK) conference [5].

3.2 Phase Two Portability Research: Deployment of Marist Model in Different Academic Contexts

After the phase one research efforts, additional time was spent improving the initial predictive model through additional machine learning techniques and the introduction of additional data elements (e.g. gradebook data from the CMS). The enhanced predictive model was then evaluated using a Marist test data set which had been excluded during the development of the model. Table 1 provides a summary of the outcomes from this evaluation (a total of ten trials were run but the table only includes the means from these tests).

Table 1. Prediction analysis using Marist test data

Classification Algorithm	Performance Metrics	Mean	SE
SVN / SMO	Accuracy	85.96%	0.86%
	Recall	81.24%	4.55%
	Specificity	86.31%	1.24%
	Precision	31.81%	2.16%
C5.0 / J48	Accuracy	86.82%	1.14%
	Recall	65.45%	3.85%
	Specificity	88.31%	1.48%
	Precision	31.34%	2.52%
Logistic Reg.	Accuracy	86.07%	0.92%
	Recall	80.26%	5.18%
	Specificity	86.53%	1.35%
	Precision	33.01%	3.44%
Training Dataset: 7344 records			
Test Dataset: 5101 records			

This model was then deployed as part of our course pilots during representing vastly different educational contexts as compared to Marist. For example, Savannah State has a 94% Black non-Hispanic student population with 67% of the student receiving Pell Grants, College of the Redwoods' student population has 22% minorities and 60% are receiving Pell Grants and Cerritos College's student body is 41% Hispanic and 45% of the students are receiving Pell Grants.

At the conclusion of the spring semester, a similar evaluation was completed to determine how well the model performed when deployed in these different academic contexts. This evaluation included assessing the model's performance at three points during the semester (25%, 50% and 75% of the semester completed), which correspond to when Academic Alert Reports were provided to instructors, to evaluate how the model's performance improved as more CMS and gradebook data was available.

Table 2. Prediction analysis from Spring 2012 pilots

College	% of Semester Completed	# of students	Accuracy	Recall	Specificity	Precision
Savannah	25%	504	66.96%	70.76%	64.64%	55.00%
	50%	504	71.52%	78.22%	67.56%	59.41%
	75%	504	77.75%	72.53%	80.94%	69.84%
Cerritos	25%	502	59.13%	69.23%	56.31%	30.73%
	50%	601	70.92%	66.14%	72.51%	44.44%
	75%	649	74.77%	74.42%	74.88%	47.76%
Redwoods	25%	195	70.50%	86.27%	61.36%	56.41%
	50%	195	79.86%	72.55%	84.09%	72.55%
	75%	195	79.14%	72.55%	82.95%	71.15%

Table 2 provides a summary of the results of this evaluation. Looking at the accuracy of the model (percentage of students that were correctly identified), just as one indicator, it is clear that the results from the predictive analysis were considerably higher than sheer randomness. For example, when one compares the accuracy at the "75% of the semester completed" point, which ranged from 75% to 79%, to the accuracy of the model when tested with

Marist data (which was historical and thus represented 100% of the semester completed), which was 86-87%, we find only a 6-10% difference. Given that we expected a much larger difference between how the model performed when tested with Marist data and when deployed at community colleges and HBCUs, this was an encouraging finding.

3.3 Portability Research Discussion

Our findings from Phase One and Two of our portability research seem to indicate that predictive models that are developed based on data from one institution may be scalable to other institutions, even those that are different with regards to institutional type, student population and instructional practices. We believe this unexpected finding may be the result of the specific elements of the predictive models which have shown to be the most powerful predictors. The elements that are most predictive of student outcomes, which were also identified by Purdue in their research, are cumulative GPA and data from the CMS' gradebook. Given that these two elements are such fundamental aspects of academic success it is not all that surprising that the predictive model has fared so well across these different institutions.

If this explanation is correct, it does point to the importance of instructors using the gradebook within their CMS if they wish to take advantage of learning analytics. It also indicates that models may not "port" well to institutions where cumulative GPA is not available (e.g. non-credit training programs) or if the student population is entering an institution and thus GPA is not yet available. Thus, although our initial findings are encouraging with regards to portability, important questions remain with regards scaling up models across more diverse academic settings.

4. OVERVIEW OF INITIAL INTERVENTION RESEARCH

OAAI has also conducted research on two different intervention strategies which were deployed to "at risk" students as means to help them succeed in the course. One of these strategies is "awareness messaging", which is closely aligned with the approach taken by Purdue's Course Signals project, which entails the instructor sending a message to the "at risk" student noting their concern over the student's academic performance and then suggesting specific steps the student should take to improve (e.g. meet with a tutor, attend a study group session, etc.). The text of these messages, outside of what the instructor suggests as means to improve, was standardized across all of the treatment groups. The other intervention strategy is referred to as the "Online Academic Support Environment" or OASE and it parallels the "awareness messaging" strategy in that students receive the same initial message noting concern but are then invited to join an Sakai-based online support site in which they are given access to Open Educational Resources (OER) instructional materials (e.g. Khan Academy videos, Flat World Knowledge textbooks, etc.). In addition to these materials, they are provided with a range of mentoring from peers and professional support staff. As with the "awareness messaging", the text of the messages is standardized across instructors and the text becomes increasingly serious in tone as students receive their second and third message.

4.1 Intervention Research: Impact on General Student Academic Success

We began our analysis by comparing, through a one-way ANOVA, the overall academic success, as measured by average course grades, between the two treatment groups and the controls, The results indicate there was no difference between the Awareness and the OASE treatment groups. However, we did find a statistically significant difference between both treatment groups and the control group (Awareness: M =77.47, SD = 13.34; OASE: M =77.5, SD = 13.44; control group: M =75.17, SD = 13.8; $F(2,299) = 3.065$, $p = .047$*, see Fig.1) which we feel is impressive as the groups include all students in the classes assigned to the three groups regardless of whether they were identified by the model as being "at risk" or not.

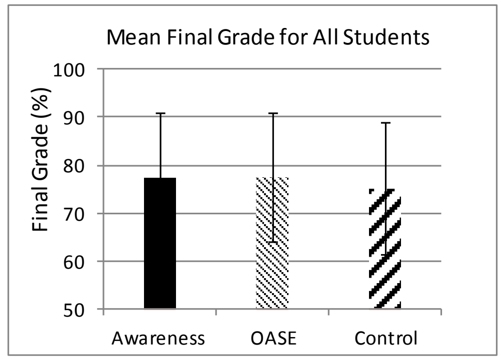

Figure 1. Average Course Grade Analysis

We then refined this analysis by comparing, also through a one-way ANOVA, what we believe represents those students who were the most "at risk". In the treatment groups ("awareness messaging" and OASE), we consider only students who had received at least one intervention based on any of the three Academic Alert Reports that were posted during the semester.

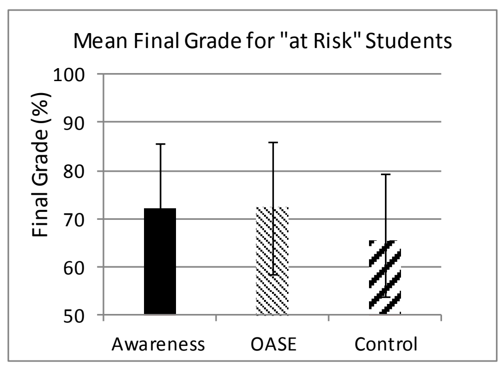

Figure 2. Refined Course Grade Analysis

For controls, which by definition did not receive interventions, we selected those students who had been identified as having an average "risk level" of three or higher across all three Academic Alert Reports. Students were categorized, based on scoring by the predictive model, as having: "no risk" (1), low risk (2), medium risk (3) or high risk (4) with regards to their likelihood of not completing the course successfully. Once again we did not find a difference between the Awareness and OASE treatment groups. However, in this analysis we found even greater statistical significance between both treatment groups and the control group (Awareness: M=72.05, SD=13.5; OASE: M=72.43, SD = 14.06; control group: M =65.38, SD = 11.8; $F(2,448) = 8.484$, $p = .000$*, see Fig. 2) which may indicate that the interventions were of particular benefit to those "at risk" to not succeed academically.

Finally, we also analyzed the impact the interventions had on "low income" students with regards to their average course grades. This comparison, like the first ANOVA, includes all students in every class regardless of risk level. This pilot study did not provide adequate *N* to all for an analysis of the “at risk” students with low income[1] status.

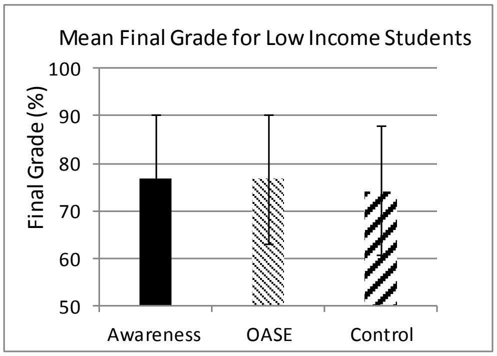

Figure 3. Low Income Student Analysis

In these case there were no statistically significant differences between any of the groups. There is however, a trend consistent with the two previous ANOVAs where the Awareness and OASE groups are approaching significance difference from the control group (Awareness: M =76.85, SD = 13.2; OASE: M =76.61, SD = 13.7; control group: M =74.14, SD = 13.7; $F(2,691) = 2.601$, p = .075, see Fig.3). The trend seems to indicate that “low income” students in the treatment groups performed better than those in the control groups. Since the number is close to being significant we are particularly interested in performing analysis on a larger sample containing Pell Grant students which we hope to collect during the Fall pilots.

4.2 Intervention Research: Impact on Content Mastery

We also examined, through a Chi-square analysis, the impact our interventions had on student “content mastery” (i.e. received a C or better in the course). We performed this analysis by comparing the controls to both treatment groups combined to increase our sample size.

Grading Scale

Letter Grade	Percentage Grade
F	0 - 55%
D	56% - 65%
C	66% - 75%
B	76% - 85%
A	86% - 100%

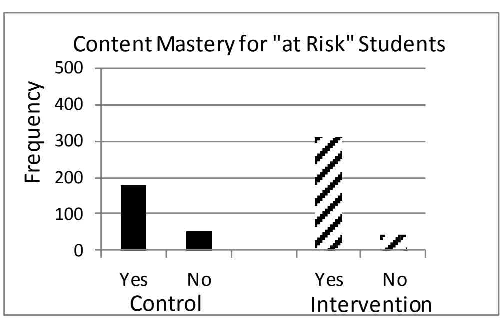

Figure 4. Students who Mastered Content (C or above)

We found a statistically significant difference ($\chi^2(1) = 8.913$, p = .003*, see Fig. 4) indicating that our treatment groups achieved better learning outcomes than those in our control groups.

4.3 Intervention Research: Impact on Withdrawal Rates

Finally, we examined the impact our interventions had on withdrawals by comparing withdrawal rates among those students identified as being “at risk” on the first Academic Alert Report in our control groups to the withdrawal rate of the combined treatment group (both “awareness messaging” and OASE). We chose to focus on those students identified during the first Academic Alert Report as we have additional evidence that interventions early in the semester have the largest impact on student outcomes and come at a time when withdrawal is possible without incurring the maximum penalties (e.g. no tuition refund).

Based on the Chi-square analysis we have identified a statistically significant difference ($\chi^2(1) = 7.097$, p = .008*, see Fig. 5) between the two aforementioned groups which indicates that students in our combined treatment group were more likely to withdraw than those in our control groups. Although this finding may at first seem to be a negative outcome, it is consistent with findings at Purdue and may be an indication that students who receive interventions are withdrawing earlier in the course (as opposed to remaining enrolled and failing) than those who do not. While we would much prefer that students complete their course successfully, withdrawing, particularly before incurring major penalties, is preferable over failing and thus we see this as a potentially positive result. We are working to explore at what point in the semester students did withdraw as means to investigate this further.

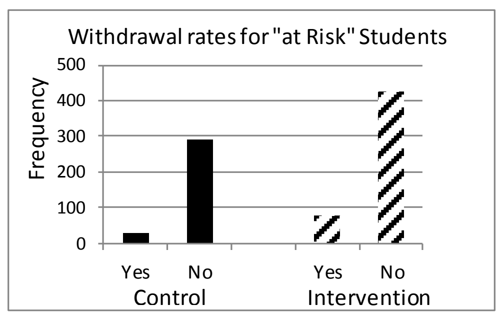

Figure 5. Analysis of Withdrawal Rates

5. CONCLUSION AND FURTHER RESEARCH AGENDA

This paper reports on initial research findings of the Open Academic Analytics Initiative[2]. Predictive models were trained and tested using Marist College data, and those models were then applied on pilot runs using data from several partner institutions. The research tested the portability of those models, and the success of intervention strategies in improving “at Risk” student outcomes. The results are promising as they seem to point at a higher portability of those models than initially anticipated. These

[1] Low-income students were defined as students who were receiving Pell Grants.

[2] We apologize if we’ve left out any material desired by prospective readers, due to lack of space. However, we will provide more complete coverage of these topics at the conference.

results had a subsequent positive impact of the effectiveness of interventions on students at academic risk. We hope that these results will encourage researchers from other institutions to develop similar strategies of early detection and intervention of academic risk.

Additional pilots were completed during the Fall 2012. We hope to use the Fall 2012 data to: (a) confirm our findings from the Spring 2012 data set; (b) increase our overall sample sizes (n) which we believe will allow us to identify additional correlations which are statistically significant; (c) look at our impact on persistence rates which require two semesters worth of data; (d) explore how differences between course subject matter or size might affect the impact of our interventions and/or be a factor in our predictions; (e) look for relationships between student responses' to questions provided by the National Survey of Student Engagement (NSSE) (which were incorporated into our OAAI Student Survey) and the receipt of an intervention. We are also planning on conducting interviews with select instructors and students to identify strengths and weakness of the interventions. These insights will help us identify ways to refine our intervention approaches.

We have also begun to discuss and identify areas of research that we believe will be important to the field of Learning Analytics as it begins to be deployed more widely. These research questions, which are outlined below, could form the basis for a national research agenda in this new and emerging field.

Does Learning Analytics allow us to identify students "at risk" to not complete a course that the typical instructor would miss?

. Although many of the instructors in our pilots have noted with us that they have found the identification of "at risk" students very helpful, it remains unclear to us if they would have identified the same students as our model if they attempted to do so on their own. Thus, we believe it will be important to conduct research studies in which "instructor predictions" are compared to "model predictions"

What are the characteristics of students who seem to have "immunity" to the treatment (those who got interventions but never improved) and those who were effectively treated after just one intervention?

From our initial research we have found that students seem to fall into one of two broad categories, those who improved after receiving just one "treatment" or intervention and those who did not improve regardless of the number of treatments received. Very few students who did not improve after the first intervention went on to improve after the second or third. Our theory is that some students respond very well to the "treatment" and thus improve after just one intervention while other seem "immune" to the "treatment" and do not improve regardless of how many treatment their receive. Understanding why this is the case and what characteristics are associated with these two categories of students would help us better understand how to most effectively deploy interventions.

How portable are predictive models that are designed for one type of course delivery (e.g. face-to-face) when they are deployed in another delivery format (e.g. fully online)?

We are particularly interested in exploring the issue of portability with regards to face-to-face and fully online programs given how much more CMS usage takes place in the later mode of instruction.

6. ACKNOWLEDGEMENTS

This research is supported by EDUCAUSE's Next Generation Learning Challenges, funded through the Bill & Melinda Gates Foundation and The William and Flora Hewlett Foundation. It is also partially supported by funding from the National Science Foundation, award numbers 1125520 and 0963365.

7. REFERENCES

[1] Baepler, P., & Murdoch, C. J. (2010). Academic Analytics and Data Mining in Higher Education. International Journal for the Scholarship of Teaching and Learning 4(2).

[2] Bravo Agapito, J., Sosnovsky , S., & Ortigosa, A. (2009). Detecting Symptoms of Low Performance Using Production Rules. Paper presented at the 2nd International Conference on Education Data Mining, Cordoba, Spain.

[3] Campbell, J. P. (2007). Utilizing Student Data within the Course Management System to Determine Undergraduate Student Academic Success: An Exploratory Study (Doctoral dissertation, Purdue University, 2007). (UMI No. 3287222).

[4] KDD Cup 2010. This Year's Challenge. Available: http://www.sigkdd.org/kddcup

[5] Lauría E., Baron J., Devireddy M., Sundararaju V., Jayaprakash S. (2012), "Mining academic data to improve college student retention: An open source perspective", Proceedings of LAK 2012, Vancouver, BC, Canada, April 29 - May 2, 2012

[6] Laurie, P. D., & Timothy, E. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. Comput. Educ., 45(1), 141-160.

[7] Morris, L. V., Wu, S., & Finnegan, C. (2005). Predicting retention in online general education courses. The American Journal of Distance Education, 19(1), 23-36.

[8] Romero, C., Ventura, S., & Garcia, E. (2008). Data mining in course management systems: Moodle case study and tutorial. Comput. Educ., 51(1), 368-384.

[9] Talavera, L., & Gaudioso, E. (2004). Mining student data to characterize similar behavior groups in unstructured collaboration spaces. Paper presented at the Workshop on AI in CSCL.

8. AUTHORS' ADDRESSES

Eitel J.M. Lauría, Nagamani Jonnalagadda, School of Computer Science and Mathematics, Marist College, Poughkeepsie, NY 12601. USA

Erik W. Moody, School of Social and Behavioral Sciences, Marist College, Poughkeepsie, NY 12601. USA

Sandeep M. Jayaprakash, Learning Analytics Specialist, Marist College, Poughkeepsie, NY 12601. USA

Joshua D. Baron, Senior Academic Technology Officer, Marist College, Poughkeepsie, NY 12601. USA

Email: {Eitel.Lauria, Nagamani.Jonnalagadda1, Erik.Moody , Sandeep.Jayaprakash1, Josh.Baron}@marist.edu