Internet and Higher Education 28 (2016) 68–84

Contents lists available at ScienceDirect

Internet and Higher Education

Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success

Dragan Gašević [a,b,*], Shane Dawson [b], Tim Rogers [b], Danijela Gasevic [c]

[a] *Moray House School of Education and School of Informatics, University of Edinburgh, United Kingdom*
[b] *Teaching Innovation Unit, University of South Australia, Australia*
[c] *Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, United Kingdom*

ARTICLE INFO

Article history:
Received 19 May 2014
Received in revised form 16 September 2015
Accepted 16 October 2015
Available online 17 October 2015

Keywords:
Learning analytics
Learning success
Student retention
Instructional conditions
Self-regulated learning

ABSTRACT

This study examined the extent to which instructional conditions influence the prediction of academic success in nine undergraduate courses offered in a blended learning model (n = 4134). The study illustrates the differences in predictive power and significant predictors between course-specific models and generalized predictive models. The results suggest that it is imperative for learning analytics research to account for the diverse ways technology is adopted and applied in course-specific contexts. The differences in technology use, especially those related to whether and how learners use the learning management system, require consideration before the log-data can be merged to create a generalized model for predicting academic success. A lack of attention to instructional conditions can lead to an over or under estimation of the effects of LMS features on students' academic success. These findings have broader implications for institutions seeking generalized and portable models for identifying students at risk of academic failure.

1. Introduction

The field of learning analytics has received much attention as a means for addressing institutional teaching and learning problems linked to the early identification of students at-risk of attrition or academic failure (Dawson, Gašević, Siemens, & Joksimovic, 2014). Despite the broad interest and implementation of learning analytics there remain numerous questions regarding the portability of any developed predictive models across student sub-populations and pedagogical contexts within an institution (Jayaprakash, Moody, Lauría, Regan, & Baron, 2014). This paper responds to this issue by investigating the impact of instructional conditions on a predictive model of learner success. In so doing, the paper aims to empirically demonstrate the importance for understanding the course and disciplinary context as an essential step when developing and interpreting predictive models of academic success and attrition (Lockyer, Heathcote, & Dawson, 2013).

1.1. Learning analytics and predictive modelling

The analysis of data collected by institutional student information systems (SIS), and from student interactions with their learning management system (LMS) (e.g. Moodle, Sakai, or Desire2Learn) has attracted much attention among researchers, teachers and managers for its potential to address some of the major challenges confronting the education sector (Baer & Campbell, 2012; Macfadyen & Dawson, 2010; Siemens & Long, 2011). Learning analytics approaches typically rely on data emanating from a user's interactions with information and communication technologies (ICTs), such as LMS, SIS and social media. For example, the trace data (also known as log data) recorded by LMS contains time-stamped events about views of specific resources, attempts and completion of quizzes, or discussion messages viewed or posted. Data mining techniques are commonly applied to identify patterns in these trace data (Baker & Yacef, 2009). The interpretation of these patterns can be used to improve our understanding of learning and teaching processes, predict the achievement of learning outcomes, inform support interventions and aid decisions on resource allocation. This process has been described as *learning analytics* (Siemens & Gašević, 2012).

Research in learning analytics and its closely related field of educational data mining, has demonstrated much potential for understanding and optimizing the learning process (Baker & Siemens, 2014). To date, much of this research has focused on developing predictive models of academic success and retention (Siemens, Dawson, & Lynch, 2014). Specifically, the prediction of students at risk of failing a course (i.e., the dependent variable is binary with two categories — fail and pass) and the prediction of students' grades (i.e., the dependent variable is continuous representing a final percent mark) have been

* Corresponding author at: School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh, Midlothian EH8 9LE, United Kingdom.
E-mail address: dgasevic@acm.org (D. Gašević).

http://dx.doi.org/10.1016/j.iheduc.2015.10.002
1096-7516/

two commonly reported tasks in the learning analytics and educational data mining literature (Dawson et al., 2014). These two types of success predictions have been based on the following sources of data:

I. data stored in institutional student information systems, e.g., high school grades, socio-economic status, citizenship and immigration status, parents' education, and language skills (Araque, Roldán, & Salguero, 2009; Kovacic, 2012);
II. trace data recorded by LMSs and other online learning environments (Agudo-Peregrina, Iglesias-Pradas, Conde-González, & Hernández-García, 2014; Romero, López, Luna, & Ventura, 2013; Romero, Ventura, & García, 2008; Zafra, Romero, & Ventura, 2011); and
III. combinations of data sources described under i) and ii) (Alstete & Beutell, 2004; Barber & Sharkey, 2012; Jayaprakash et al., 2014).

Regardless of the data source, the prediction of student grades is generally determined by applying logistic regression (Barber & Sharkey, 2012; Campbell, DeBlois, & Oblinger, 2007; Lauría, Baron, Devireddy, Sundararaju, & Jayaprakash, 2012; Palmer, 2013). However, many authors, especially those from educational data mining backgrounds, have also reported highly accurate predictions using different classification algorithms such as C4.5, EM, Naïve Bayes, and support vector machines (SVM).

The underlying rationale of these studies is to uncover variables that are common in the undergraduate environment that will either individually or in concert inform a generalized model of predictive risk that acts independently of contextual factors such as institution, discipline, or learning design. These omissions of contextual variables are also occasionally expressed as an overt objective. For example, the large scale Open Academic Analytics Initiative (OAAI) (Jayaprakash et al., 2014) had the explicit aim of testing an open source risk identification solution that was applicable to most forms of US tertiary education—from community colleges to private liberal arts universities—was impervious to institutional variances, and thereby could prove suitable for "scaling… across all of higher education" (Jayaprakash et al., 2014, p. 7).

While this rationale suggests that pooling data across contexts to increase the sample size and predictive utility is ideal, studies that employ this approach are the exception rather than the rule. Most of the reported studies investigating the prediction of academic success have been based on trace data extracted from a single, or small number, of courses within a particular discipline (Macfadyen & Dawson, 2010; Romero et al., 2013). The small sample sizes and disciplinary homogeneity adds further complexity in interpreting the research findings, leaving open the possibility that disciplinary context and course specific effects may be contributing factors.

Research in predictive analytics has an obvious and significant level of importance for contemporary higher education institutions. The capacity for early identification of students at-risk of academic failure or attrition allows for a proactive approach to implementing learning interventions and strategies that target teaching quality and student retention. Thus, it is not surprising that the insights gleaned from research on student academic risk are being so readily and eagerly adopted across the sector (Siemens et al., 2014). Despite the encouraging progress in this research, a significant challenge remains. That is, how best to interpret such findings in order to ascertain generalizability.

1.2. Need for educational theory underpinning in learning analytics

Despite the titular reference to 'learning', learning analytics has only recently begun to draw on learning theory, and there remains a significant absence of theory in the research literature that focuses on LMS variables as key indicators of interaction and success (Lust, Juarez Collazo, Elen, & Clarebout, 2012).

Expectations of academic risk drawn from the learning theory literature are largely antithetical to the universalist assumptions underpinning the practice of identifying student risk from pooled LMS data. Most post-behaviorist learning theories would suggest the importance of elements of the specific learning situation and student and teacher intentions. For example, contemporary process theories would emphasize the dialectic between instruction and learning (Engeström, 2014), while motivational approaches focus (in part) on the beliefs that students hold regarding their capabilities with respect to specific content (Zimmerman & Schunk, 2011), and constructivist theories investigate the interplay of instructional design and student internal conditions (Winne, 2006; Winne & Hadwin, 1998). All therefore acknowledge the contextual conditions that shape student learning, and so posit that learning is fundamentally situated (Lave & Wenger, 1991), suggesting that there are potentially important differences between disciplines and courses. Furthermore, there is a long history of research on the particular characteristics of students within disciplines and courses suggesting that, for example, self-regulation of learning may be course specific (Black & Deci, 2000), and that self-efficacy (Chung, Schwager, & Turner, 2002) and information seeking behavior (Whitmire, 2002) can vary by courses and discipline. Altogether the preponderance of evidence indicates that disciplines and courses are not cut from the same cloth and that their respective student constituents may not be of one kind.

Yet, to our knowledge, only a study by Finnegan, Morris, and Lee (2009) examined the possibility of a mediating role for contextual variables. Finnegan et al. (2009) found disciplinary differences in the effects of trace data to predict grades on 22 courses from English and Communication; Social Sciences; and Mathematics, Science, and Technology. Not only did the authors report on the differences in the explained variability of the final grades by multiple regression models (from 26% to 36%), but they also noted there was no single significant predictor shared across all three disciplines. Although some variables (e.g., time spent on content pages) were identified as significant predictors of academic success in regression models for individual disciplines, the same effect was not apparent when data from all three disciplines were combined. Similarly, the multiple linear regression model of the three disciplines combined showed no significant effects and/or the overestimated/underestimated importance of some variables (e.g., time spent on follow-up posts and time spent on reading discussions) in comparison to the regression models performed for individual disciplines.

The under-explored role of contextual variables may help explain the mixed findings in the field, with even large scale studies reporting differences in their results in relation to the overall predictive power of the same individual LMS variables. For example, where Macfadyen and Dawson (2012) identified a strong correlation between student discussion forum activity and overall academic grades at a large research intensive Canadian university ($N = 52{,}917$), Lauría et al. (2012) found only weak correlations (ranging from 0.098 to 0.233) between students' grades and LMS activity, including discussions read and posted, at a private liberal arts college in the USA ($N = 18{,}968$). Although the approaches adopted were similar, the observed results markedly differed. Plainly, if the hypothesis that LMS tool use is predictive of student risk is valid, then there are contextual differences at work here, and plausibly these are located in the distinctive elements of the courses that comprised the studies.

There are several advantages in leveraging existing learning theory to investigate the nature of these contextual factors, discussed at length elsewhere in the literature (Gašević, Dawson, & Siemens, 2015; Rogers, Gašević, & Dawson, in press). Briefly, studies designed with clear theoretical frameworks will a) connect learning analytics research with decades of previous research in education and b) make clear what is contended by research designs, and so make explicit what the research