Implementing Learning Analytic Tools in Predicting Students' Performance in a Business School

R. Sujatha and B. Uma Maheswari

1 Introduction

Developments in the field of information technology with respect to big data have resulted in disruptive implications across all sectors (Baradwaj and Pal 2011). Data is now available in abundance, and therefore, there is a need to employ tools and techniques to mine such data. Data mining tools and techniques have found applications across various disciplines including customer profiling, fraud detection, DNA sequencing, etc. (Lauria and Baron 2011). Educational data mining (EDM) and learning analytics (LA) are two communities evincing a keen interest in how big data could be exploited for the larger benefit of the education sector (Baker and Inventado 2014). EDM deals with "developing methods that discover knowledge from the data originating from educational environments" (Han and Kamber 2006). Such mining techniques result in pattern recognition which forms the basis for decision making and support interventions (Siemens et al. 2011) ultimately leading to optimizing the learning process (Baker and Siemens 2014).

LA on the other hand emphasizes more on the data visualization and human intervention and has evolved into a critical domain in the education space (Gasevic et al. 2016). LA is defined as "the measurement, collection, analysis and reporting of data about learners and their context, for purposes of understanding and optimizing learning and the environment which it occurs" (Ferguson 2012; Elbadrawy et al. 2016). Research in the field of EDM and LA has clearly demonstrated the need for understanding the teaching–learning process and using this information for improving the same (Baker and Inventado 2014; Gasevic et al. 2016).

R. Sujatha (✉) · B. Uma Maheswari
PSG Institute of Management, Coimbatore, India
e-mail: sujatha@psgim.ac.in

B. Uma Maheswari
e-mail: uma@psgim.ac.in

A. K. Laha (ed.), *Applied Advanced Analytics*, Springer Proceedings in Business
and Economics, https://doi.org/10.1007/978-981-33-6656-5_12

2 Purpose of Research

A solid theoretical framework in this domain has not evolved yet. The vast differences in the kind of learning tools and educational systems in different institutions across developed versus developing nations clearly delineate the impracticality of a model which could be generalized across geographies as well as across disciplines (Agudo-Peregrina et al. 2014). Studies clearly proved that self-regulation of learning (Black and Deci 2000), self-efficacy (Chung et al. 2002) and information-seeking behavior (Whitmire 2002) vary with course and discipline. Early identification of students who are likely to fail in a particular subject gives scope for early interventions and therefore helps faculty to provide more attention to a specific set of students.

The review of literature showed that existing studies in this domain has been based on data extracted from learning management system (LMS) (Romero et al. 2013). One such LA model was developed at Purdue University called Course Signals (Arnold and Pistilli 2012) which successfully transformed a research agenda into a practical application. LMS data was used in pattern recognition of user behavior (Talavera and Gaudioso 2004). Studies have also been undertaken with data obtained from Moodle (an open-source course management system) (Romero et al. 2013). Demographic characteristics and course management system usage data was used to develop predictive machine learning models (Campbell 2007). Unstructured data obtained from online discussion forums was used to perform sentiment analysis in a study conducted by Laurie and Timothy (2005). Lykourentzou et al. (2009) used the scores of a quiz activity to cluster the students using neural networks.

Applying analytics in education is the need of the hour, especially in the context of a developing economy like India. The inferences drawn from prior studies have been eagerly accepted by the academic community (Gasevic et al. 2016). Hence, it is time for educational institutions to use machine learning tools to enhance teaching–learning experience. This study deploys learning analytics technique using the data of students undergoing a post-graduate management program and attempts to create a system of preventive feedback mechanism for faculty and students.

3 Research Objectives

The objective of the study was to create an early intervention mechanism to enhance students' performance. This study aims to develop a predictive model using machine learning algorithms to predict the academic risk of a student passing or failing a course (binary response) and the marks of the student (continuous data) in a course based on past data. The research objectives of the study are

o RO1: To develop a model to predict the academic status of a student in a course?
o RO2: To develop a model to predict the grade of a student in the course?

4 Methodology

The learning analytics model adopted in this study is based on supervised learning algorithm. The training data set had both predictive features including demographic characteristics of the students, as well as the response feature which is the students' academic status. The methodological framework developed by Lauria and Baron (2011) has been adopted in this study. The framework includes five steps such as data collection, data preparation, data partition, building the models and evaluating the models.

4.1 Data Collection

The study was conducted among post-graduate management students undergoing the Master of Business Administration (MBA) program. The data was collected during the admission and during the progression of the course. Demographic data of the students was collected from the admissions portal and the academic performance of the students from the examination portal. The academic performance of the students relating to six foundation courses undertaken by the student in the first semester and one capstone course strategic management was considered for the purpose of this study. Past three years data was considered ($n = 522$).

4.2 Data Preparation

The data was pre-processed for missing values, outliers and incomplete data. Few students' data who left the college was identified and removed from the database. Next step, the identity of the students represented by names and roll numbers was removed to ensure anonymity. A few features had to be derived for the purpose of this study. The data relating to the 'date of birth' of the students was extracted from the portal, and the age of the student at the time of joining the course was derived. The other derived feature was 'break in study.' The data relating to the undergraduate degree completion year was extracted. This data showed that most of the students did not have a break in study; therefore, this feature was dichotomized into 'break in study,' Yes/No. Data transformation was done in the case of four features (community, tenth board, higher secondary board, undergraduate degree). The feature 'community' had seven categories such as OC, BC, MBC, DNC, SC, ST and Others. A summary of the feature showed a small percentage of the students belonged to MBC, DNC, SC and ST. Therefore, these categories were combined with 'Others' category resulting in only three categories of 'community,' namely OC, BC and Others. The tenth and the higher secondary board also went through a similar process to result in four categories, namely Tamil Nadu Board, Kerala Board, CBSE and Others. As the MBA program did not have any restrictions in terms of undergraduate degree, this program attracts students from versatile disciplines. This feature also had to be transformed into three major groups. namely Arts, Engineering and Science. The features used for model building is shown in Table 1.

Table 1 Features in input dataset

S. No.	Features	Continuous/categorical	Type of features
1	Age	Continuous	Numeric
2	Gender	Categorical	Male/Female
3	Community	Categorical	OC/BC/Others
4	Tenth Percentage	Continuous	Numeric
5	Tenth Board	Categorical	Tamil Nadu/Kerala/CBSE/Others
6	Higher Secondary Percentage	Continuous	Numeric
7	Higher Secondary Board	Categorical	Tamil Nadu/Kerala/CBSE/Others
8	Undergraduate Course Discipline	Categorical	Arts/Engineering/Science
9	Undergraduate Percentage	Continuous	Numeric
10	Break in Study	Categorical	Yes/No
11	Work Experience	Continuous	Numeric
12	Entrance exam score	Continuous	Numeric
13	Marks obtained in Organizational Behavior course	Continuous	Numeric
14	Marks obtained in Business Environment course	Continuous	Numeric
15	Marks obtained in Managerial Economics course	Continuous	Numeric
16	Marks obtained in Accounting for managers course	Continuous	Numeric
17	Marks obtained in Business Communication course	Continuous	Numeric
18	Marks obtained in Quantitative Techniques course	Continuous	Numeric
19	Marks obtained in Strategic Management course	Continuous	Numeric

4.3 Partition the Data

The data was partitioned into two sets, the training dataset and testing dataset. Eighty percentage of the dataset was used for training the model, and the model was tested using the remaining twenty percentage of the data set.

4.4 Build Models

Logistic regression was used to predict the response variable. Logistic regression is a generalized linear model, where the response variable is a function of the linear combination of all the predictor variables (Lauria and Baron 2011). The categorical predictor variables used in this study include age, gender, community, tenth board, higher secondary board, undergraduate degree and break in study. The continuous predictor variables include tenth percentage, higher secondary percentage, undergraduate percentage, work experience and entrance exam score. The response variable in this study is the academic status which is denoted by pass (50% or more marks) or fail (less than 50% marks) (Palmer 2013; Barker and Sharkey 2012). The academic status of all the six foundation courses was predicted using this model. The prediction of students' marks in capstone course is done using stepwise multiple linear regression. The coefficients derived for each of the predictor variables helped in identifying which of these variables influences the response variable.

4.5 Evaluate the Models

The accuracy of the logistic regression model can be evaluated based on three metrics such as accuracy, specificity and sensitivity which are derived from the confusion matrix. In this study the overall accuracy [(TP + TN)/(TP + TN + FP + FN)] is not considered as a metric for model evaluation. Instead the focus is on sensitivity [TP/(TP + FN)] and specificity [(TN/(TN + FP)], where TP stands for True Positive, TN for True Negative, FP for False Positive and FN for False Negative. The multiple linear regression model developed using stepwise regression was validated through tenfold cross-validation technique. This technique uses ten rounds of cross-validation using multiple cross-validation training and testing sets. The result of this technique estimates the validity of the machine learning model.

Table 2 Frequency analysis

Features	Categories	Frequency	Percentage
Gender	Male Female	286 236	55 45
Community	OC BC Others	72 272 178	14 52 34
Tenth board	Tamil Nadu Kerala CBSE Others	367 19 93 43	70 4 18 8
Higher Secondary Board	Tamil Nadu Kerala CBSE Others	411 22 65 24	79 4 12 5
Undergraduate Degree	Arts Engineering Science	161 325 36	31 62 7
Break in Study	Yes No	201 321	39 61

5 Results and Discussion

5.1 *Descriptive Analytics*

The results of frequency analysis depicted in Table 2 showed almost equal representation of male students (55%) and female students (45%). Descriptive analysis related to the socio-economic status of the students showed that 52% of the students belonged to BC category. Students who have undergone tenth and higher secondary from Tamil Nadu Board represented 70% and 79%, respectively. Students with Engineering as their undergraduate study was 62% which is the highest compared to Arts (31%) and Science (7%). Thirty-nine percentage of the students have a break in their study indicating that they would have taken up a job after their undergraduation.

5.2 *Scatter Plot*

Before deciding on the modeling strategy, it was essential to understand the correlation between response and predictor variables. A scatter plot was used to identify the correlation. The scatter plot between the response variable (marks obtained in strategic management course) and other predictor variables like tenth percentage, higher secondary percentage, undergraduate percentage, work experience, entrance exam scores is shown in Fig. 1. Figure 1 shows a positive relationship between marks obtained in strategic management course and the other predictor variables.

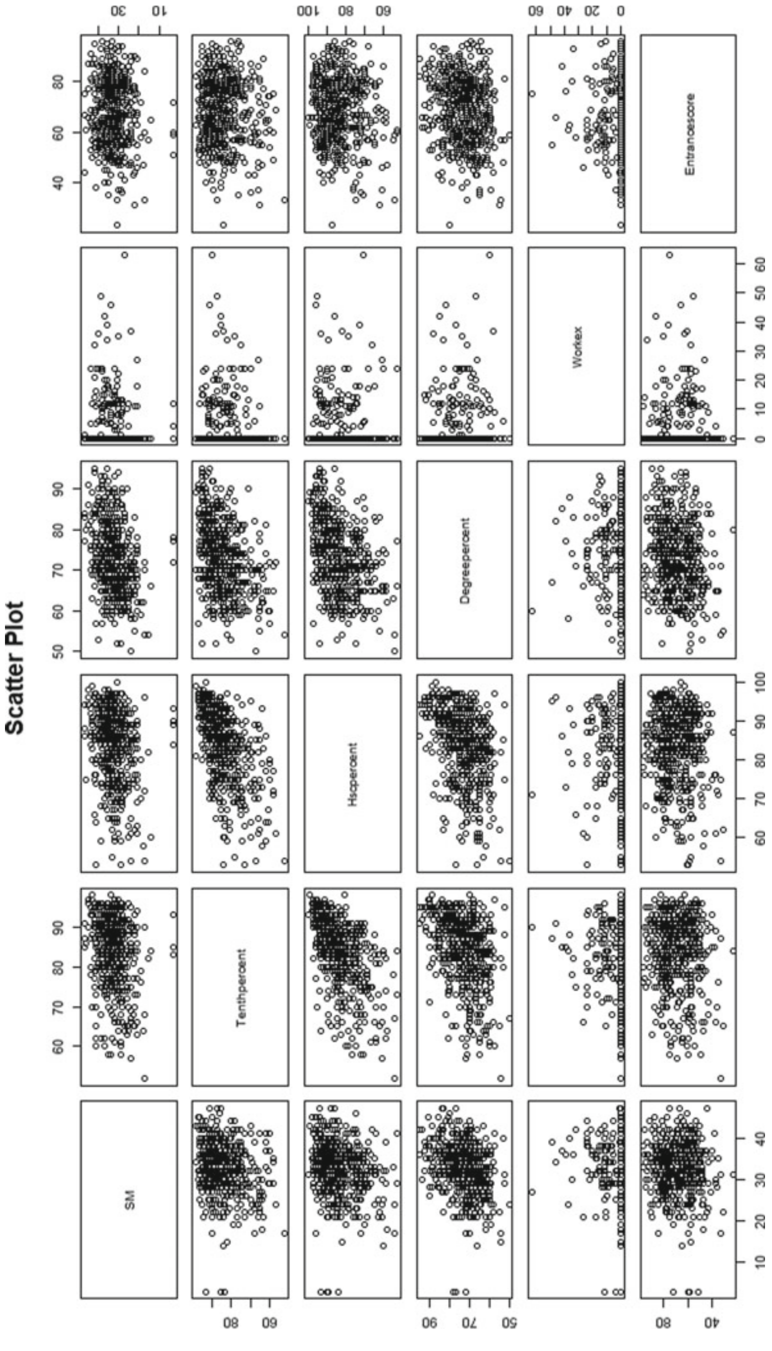

Fig. 1 Scatter plot of variables derived from admission portal

The scatter plot between the response variable (marks obtained in strategic management course) and predictor variables (marks obtained in foundation courses) is shown in Fig. 2. Figure 2 shows a positive relationship between marks obtained in strategic management courses and marks of foundation courses like organizational behavior, business environment, managerial economics, accounting for managers, business communication and quantitative techniques.

5.3 Model to Predict the Academic Status in a Course

Logistic regression was used to build the model. Sensitivity and specificity scores of this model was used for predicting the academic status of the foundation courses. In this study, it is important to note that the model so developed should be capable of predicting a student who has failed in the course as “FAIL,” and more important that the model *does not* predict a student who actually failed as “PASS.” Therefore, specificity as a metric gains more prominence than sensitivity. The specificity of the logistic regression models developed for the foundation courses was 82.54% for organizational behavior, 59.57% for business environment, 92.75% for managerial economics, 60.42% for accounting for managers, 92.47% for business communication and 87.01% for quantitative techniques.

5.4 Model Building to Predict the Grade of a Student in the Capstone Course

Multiple linear regression model was deployed to capture the unique contribution of the predictor variables in explaining the variation in the response variable. Since this study had six categorical variables which could not be included directly into the model, the categorical variables were re-coded using dummy variables. For example, since the variable undergraduate degree had three categories, i.e., Arts, Engineering and Science, two $(n - 1)$ dummy variables were included. The same process was applied for all the other categorical variables.

In stepwise regression, the entering criterion for a new variable to enter the model is based on the smallest p value of the partial F test and the removal criterion for a variable is based on the β value. In this study, $\alpha = 0.05$ was considered and if the p value $< \alpha$ then the variable was entered in to the model, and if the p value $> \beta = 0.10$, the variable will be excluded from the model. At each stage, the variable was either entered into the model or removed from the model.

The stepwise regression model excluded the variables such as community, tenth board, higher secondary board, break in study, work experience, entrance exam scores. The final model had retained the variables gender, tenth percentage, higher secondary percentage, undergraduate degree percentage, and the marks of all the six

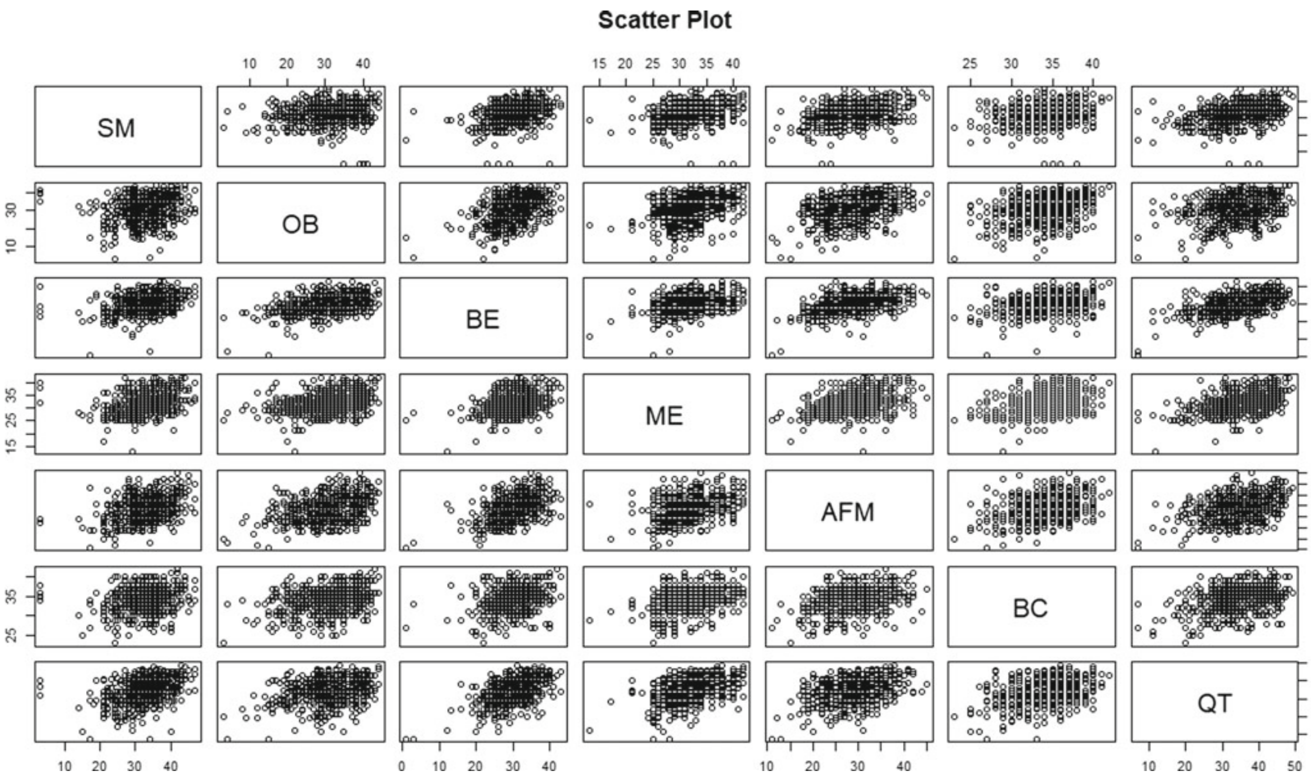

Fig. 2 Scatter plot of variables derived from examination portal

Table 3 Results of stepwise regression model

Variables	Unstandardized coefficients		Standardized coefficients	*t*	Sig
	B	Std. Error	Beta		
(Constant)	11.439	3.148		3.633	0.000
Gender	0.454	0.535	0.036	1.949	0.007
Tenth percent	0.012	0.035	0.017	2.343	0.001
Hsc percent	0.015	0.033	0.023	2.470	0.008
Degree percent	0.070	0.036	0.096	1.938	0.043
Business environment	0.103	0.061	0.087	2.701	0.035
Managerial economics	0.110	0.069	0.076	2.597	0.001
Accounting for Managers	0.229	0.053	0.211	4.327	0.000
Business communication	0.172	0.044	0.198	3.897	0.000
Organizational behavior	0.084	0.039	0.096	2.142	0.033
Quantitative techniques	0.122	0.090	0.062	2.357	0.005

foundation courses. The *p* values of these variables are less than 0.05 indicating the significance of these variables in the model. The results of final regression model is given in Table 3. The model was further validated using the tenfold cross-validation technique. The root mean square error is 5.6.

6 Implications and Conclusion

The results of this study indicate that learning analytics could be effectively implemented in enhancing the quality of teaching–learning experience (Macfadyen and Dawson 2012). In this paper, two predictive models were used to predict academic risk of students who were not performing well in the course as an early intervention mechanism. In the first part, logistic regression was used to identify the academic status of foundation courses in the first semester. Data obtained during the admission process is used as input for model building. Since an MBA program is open for all streams of undergraduate studies, it is essential to have an early intervention in order to ensure a smooth progression of the students into the second semester where they are introduced to advanced management courses.

In the second part of the study, the stepwise regression model was used to predict the marks of the students in capstone course. The results showed that as the students' progress into second semester courses, the tenth and higher secondary board become irrelevant. Performance in the first semester courses greatly influences the results of the second semester. The student who scored well in the first semester also scored well in the second semester. Therefore, this early intervention would help enhance student performance, thereby preparing him to face forthcoming semesters more confidently. This understanding further helps students to select courses in which they can perform better.

Model deployment would help build a transparent system by which both the stakeholders, faculty and student would get insights about the students' progress. This study could be further extended to all courses in the forthcoming semesters. This would gradually evolve into a learning analytics system which can be inbuilt in to the curriculum. Further, this model could be extended to predict the probability of the students succeeding in placement. Deployment of the models developed in this study would go a long way in not only enhancing students' performance but also more fruitful faculty engagement. Embedding analytics in the education system would transform the education landscape to greater heights.

References

Agudo-Peregrina, Á. F., Iglesias-Pradas, S., Conde-González, M. Á., & Hernández-García, Á. (2014). Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior, 31,* 542–550.

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267–270). ACM.

Baker, R., & Siemens, G. (2014). Educational data mining and learning analytics. In R. K. Sawyer (Ed.), *Cambridge handbook pf the learning sciences*. Cambridge, UK: Cambridge University Press.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics* (pp. 61–75). Springer, New York, NY.

Baradwaj, B. K., & Pal, S. (2011). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications, 2*(6), 63–69.

Barber, R., & Sharkey, M. (2012), Course correction: Using analytics to predict course success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 259–262). ACM.

Black, A. E., & Deci, E. L. (2000). The effects of instructors' autonomy support and students' autonomous motivation on learning organic chemistry: A self-determination theory perspective. *Science Education, 84*(6), 740–756.

Campbell, J. P. (2007), *Utilizing student data within the course management system to determine undergraduate student academic success: An exploratory study*. Purdue University.

Chung, S. H., Schwager, P. H., & Turner, D. E. (2002). An empirical study of students' computer self-efficacy: Differences among four academic disciplines at a large university. *Journal of Computer Information Systems, 42*(4), 1–6.

Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G., & Rangwala, H. (2016). Predicting student performance using personalized analytics. *Computer, 49*(4), 61–69.

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(5/6), 304–317.

Gašević, D., Dawson, S., & Rogers, T. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education, 28,* 68–84.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). Boston, MA: Elsevier.

Lauría, E. J., & Baron, J. (2011). Mining Sakai to measure student performance: Opportunities and challenges in academic analytics. *Download at:* https://eccmarist.edu/conf2011/materials.

Laurie, P. D., & Timothy, E. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computer Education, 45*(1), 141–160.

Lykourentzou, I., Giannoukos, I., Mpardis, G., Nikolopoulos, V., & Loumos, V. (2009). Early and dynamic student achievement prediction in e-learning courses using neural networks. *Journal of the American Society for Information Science and Technology, 60*(2), 372–380.

Macfadyen, L. P., & Dawson, S. (2012), Numbers are not enough. Why e-learning analytics failed to inform an institutional strategic plan. *Journal of Educational Technology & Society*, *15*(3).

Palmer, S. (2013). Modelling engineering student academic performance using academic analytics. *International Journal of Engineering Education, 29*(1), 132–138.

Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers and Education, 68,* 458–472.

Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S. P., Shum, S., Ferguson, R, & Baker, R. (2011). Open learning analytics: An integrated & modularized platform.

Talavera, L., & Gaudioso, E. (2004). Mining student data to characterize similar behavior groups in unstructured collaboration spaces. In *Workshop on artificial intelligence in CSCL. 16th European conference on artificial intelligence* (pp. 17–23).

Whitmire, E. (2002). Disciplinary differences and undergraduates' information-seeking behavior. *Journal of the American Society for Information Science and Technology, 53*(8), 631–638.