2024

EARLY IDENTIFICATION OF STUDENTS AT ACADEMIC RISK BASED ON LEARNING MANAGEMENT SYSTEM LOG DATA

Roger Sheng So

EARLY IDENTIFICATION OF STUDENTS AT ACADEMIC RISK BASED ON LEARNING MANAGEMENT SYSTEM LOG DATA

A dissertation submitted in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

to the faculty of the

DEPARTMENT OF ADMINISTRATIVE AND INSTRUCTIONAL LEADERSHIP

of

THE SCHOOL OF EDUCATION

at

ST. JOHN'S UNIVERSITY
New York
by
Roger Sheng So

Date Submitted October 25, 2023 Date Approved January 31, 2024

Roger Sheng So	Dr. James R. Campbell

ABSTRACT

EARLY IDENTIFICATION OF STUDENTS AT ACADEMIC RISK BASED ON LEARNING MANAGEMENT SYSTEM LOG DATA

Roger Sheng So

Understanding student engagement with the institution from the first day of classes to the end of the semester would help inform the institution of the potential risk that a student will drop out of a class or of the school. Learning Management Systems (LMS) record student interactions with the system and might be able to be used to identify students who are at academic risk. The scope of this study is to retrospectively analyze first-year student activity for the Spring 2022 semester for early warning signs worthy of intervention. A student risk assessment will be determined by reviewing student LMS activity, compared with peers, during the semester.

DEDICATION

This work is dedicated to my parents Pearl and P.T. who instilled in me a desire for knowledge and to my wife Alice who supported my unrealistic dream.

ACKNOWLEDGMENTS

The professors at the Department of Administrative and Instructional Leadership (DAIL) School of Education Research deserve my gratitude for helping me to appreciate quality research. In particular, I thank Dr. James Campbell, my dissertation advisor; Dr. Catherine C. DiMartino, my first doctorial professor; Dr. Seokhee Cho; and Dr. Erin M. Fahle, my academic advisor.

I want to acknowledge Joe Tufano for bringing me to St. John's University.

The support and encouragement of my St. John's co-workers, particularly Eric Alvarado and Lauren Drakakis.

Gratitude is also deserved by my fellow students who challenged me to study harder, especially my dear friend Van Havercome.

Appreciation is also expressed to Andrew Mungai, my patient editor.

I thank my Metlife friends, NYU undergraduate family and extended family for their patience and willingness to schedule around my classwork.

Finally, I thank my family for their love and support during this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

According to the NCES The Condition of Education 2020 report (Hussar, Zhang, Hein, Wang, Roberts, Cui, Smith, Bullock Mann, Barmer & Dilig, 2020), only sixty-seven percent of bachelor degree students at four-year private higher education institutions obtained their degree from the same institution within six years. For fall 2017, the first-year overall retention rate for four-year institutions was eighty-one percent. This means that close to one in five students leave after only one year, potentially with college debts to pay and colleges with empty seats. Understanding student engagement with the institution from the first day of classes to the end of the semester would help inform the institution of the potential risk that a student will drop out of a class or drop out of the school thus providing an opportunity to retain these students. Learning Management Systems record student interactions with their courses and potentially could be used to identify students who are at risk.

Purpose of the Study

The purpose of this study is to retrospectively analyze archived Learning Management System student activity for first-year students during the Spring 2022 semester for early warning signs that the students would withdraw from courses or the University. Understanding student engagement with the institution from the first day of classes to the end of the semester would help inform the institution of the potential risk that a student will drop out of a class or of the school. Learning Management Systems record student interactions with the system and might be able to be used to identify students who are at academic risk.

Theoretical/Conceptual Framework

The theoretical framework is that students with higher levels of activity are more engaged with the institution and are therefore less likely to leave (Tinto & Pusser, 2006). The premise of this study is that LMS activity is a surrogate for academic engagement and therefore could inform institutions if student engagement was lacking based on low or no LMS activity.

Review of Related Research

The literature review builds on the theoretical framework. Student academic engagement is a critical component of reducing student dropouts by examining past research on early identification of learner dropouts. Within this context, it reviews prior work on prediction models based on LMS data. Then, it explores related technology areas: Among them, Education Data Mining which surveys different analytical approaches to surface meaning from LMS log data.

Significance of the Study

The size of the outstanding student loan debt in the United States has been growing over the last decade. Among the causes are: high tuition costs, inadequate employment compensation to repay these student loans, and low college graduation and retention levels. For academic year 2015-2016, the average cumulative loan amount for undergraduate degree students attending private nonprofit institutions was $33,200 (McFarland, Hussar, Zhang, Wang, Wang, Hein, Diliberti, Forrest Cataldi, Bullock Mann, & Barmer, 2019).

For the cohort entry year 2010 and 2011, the six-year graduation rate from first institution attended for full-time four-year bachelor degrees from all private nonprofit

institutions was 66% (McFarland et al., 2019). The current low levels of graduation and retention result in lower income for higher educational institutions, lower employment potential for students who do not complete studies and, in effect, increased unpaid student loans.

Students are less likely to leave if they are engaged with the institution (Tinto, 1975; Tinto & Pusser, 2006). The optimum time to establish this alignment is during students' initial period with the institution and the principal interaction lies with instructors.

This study will investigate whether LMS data can be used to improve retention.

LMS data currently tracks course activity, and the premise is that this course activity can be used to measure academic student engagement. Higher student engagement will lead to increased retention and conversely lower student engagement may be a signal of lower engagement and therefore higher risk of withdrawal.

Connection with Social Justice in Education

A large portion of St. John's University's student population is economically disadvantaged. Significant efforts have taken expended to get this class of students accepted into the University. It is incumbent on the University to do as much it can to to help all of its students complete their education in a timely manner. While some resources are available to help students, it would be beneficial to have a reliable ability to identify students at risk of withdrawing while there is time to intervene.

Research Question

The research question that guided this study was:

Research Question: To what extent can Learning Management System (LMS) data inform large private higher education institutional student retention actions? Specifically, can Learning Management System (LMS) data for individual course activity predict student retention?

Definition of Terms

Sessions – Count of unique student logon sessions recorded by the Learning Management System (LMS) in table REQUESTS, column SESSION_ID.

Interactions – Count of LMS records for each student in the REQUESTS table. A record is generated whenever a student interacts with the LMS.

Interactions-Read – Count of LMS records for each student in the REQUESTS table which do not change LMS contents. A record is generated whenever a student interacts with the LMS.

Interactions-Write – Count of LMS records for each student in the REQUESTS table which change LMS contents. A record is generated whenever a student interacts with the LMS.

Interaction Days – Count of the number of days that any interactions took places.

Interaction Weeks – Count of the number of Weeks that any interactions took places. Weeks are defined as Sunday through Saturday.

Not-Retained – Departure from the University (Coded as 1) versus retained (Coded as 0). The data for this dependent variable is from the Canvas enrollment table (ENROLLMENT_DIM) at several periods (Fall 2021 start, Spring 2022 start, and Fall

2022 start). Students registered in Spring 2022 courses but not in Fall 2022 courses are considered not-retained (Coded as 1).

CHAPTER 2 REVIEW OF RELATED RESEARCH

Theoretical Framework

Theory: Student academic engagement is a critical component of reducing student dropouts

Tinto and Pusser (2006) theorized that institutions can take effective action to increase student persistence and in turn affect student success. Their perspective is that the reasons that students leave is different from why they might stay. They also believe that it is important to focus on concepts that can translate into courses of action. And while there are many issues external to the institution, they are less relevant because they can't be affected by any actions of the institution. Tinto and Pusser also point out that persistence can be viewed at an individual course or for student enrollment perspective in other words, dropping out of one course or dropping out of school. But the larger issue concept is that all of these factors need to be considered together by institutions, because comprehensive approaches might be needed to effect better outcomes.

Institutions have a lot of levers within their control to influence the outcomes. Among them are academic, financial, and social. The academic structures start with pedagogy and curriculum, but include academic support, monitoring, assessment and early warning. One additional part of this theory is that outcomes can be influenced from an institutional perspective (Campbell, DeBlois, & Oblinger, 2007).

Building on this perspective, persistence can be positively affected by using Learning Management System (LMS) to provide academic engagement through assignments, discussions and assessments as well as a means to measure and monitor these activities.

Review of Related Literature

Over the last few years, there have been several research studies about the use of LMS and related student data to predict student outcomes in online courses. At the same time advances in data mining techniques and speed and reduction in processing costs have made real time analysis of LMS data more feasible.

Early Identification of Learner Dropouts

To understand whether online course material could assist in providing early identification of learner dropouts, Cohen (2017) conducted a quantitative analysis of computer log data related to 362 students in three mathematics and statistics courses. The study's literature review began with an explanation of the Moodle Learning Management system which provided the source data for the study. Several studies (Black et al., 2008; Brandl, 2005; Graf and List, 2005) collected Moodle data about student website activity and student performance. Cohen (2017) then discussed the areas of learning analytics and educational data mining which permit non-intrusive accumulation of information without faculty or staff intervention. Learning analytics applies to the capture an analysis of fields of learning to create a new educational knowledge base. This has the potential to improve teaching and learning, improve learning system and the operations (Ai and Laffey, 2007; Lu et al., 2003; Romero and Ventura, 2007).

Cohen's study explained the importance of studying student dropout and touched on some of the theoretical models which describe or predict it. (Astin, 1999; Bean, 1985; Cabrera et al., 1992; Xenos, 2004; and Tinto, 1975, 1993). In addition, several similar higher education studies which used LMS log data to identify at risk students were identified (Campbell et al., 2007; Diet-Uhler and Hurn, 2013; Romero et,al., 2013). The

focus of some of these studies differed in the way they analyzed the data. MacFadyen and Dawson (2010) identified thirteen variables (number of messages in discussion, number of items uploaded and submitted assessments) that correlated with the students' final grade. Lykourentzou, et al (2009) used quiz scores and considered assignment submission on time were relevant. Nistor and Neubaurer (2010) looked for patterns of behavior which lead to course dropout. While earlier studies were predominately focused on fully online course (Santana et al, 2015; You, 2016), more recently attention has been paid to web-supported courses.

Educational Data Mining

The purpose of the Tang, Zing, and Pei (2019) study was to use educational data mining (EDM) techniques to uncover the time dimension of online participation and to identify key moments for intervention. This was based on an event-based view of learning, this study investigated longitudinal patterns in online participation based on the premise that online learning is a cumulative process with participation at critical moments more significant than at other points. Participation has been identified as a key predictor of online performance (Cheng & Chau, 2016; Davies & Groff, 2005) but the other research also shows that learners in asynchronous courses are dependent on peer participation (Dringus & Ellis, 2010). Learners who actively participate are likely to do well in asynchronous courses (Hew & Cheung, 2008). Therefore, there have been efforts to boost participation, but the timing of intervention is critical (Xing et al., 2016). Reimann (2009) proposed an event-based view of learning that portrays the learning process as a sequence of developments, as opposed to a change in process variables.

Think of learners as actors and their individual participation during a specific period as events.

Tang, et. al.'s study used a dataset from the JuxtalLearn Project at a Spanish University. It was a three-month project from September 2013 to December 2013 involving 111 participants, 82 media and communications and 29 computer engineering across two campuses. The researchers divided the project into four three-week segments and four activity categories (create, annotate, delete and update).

An unsupervised clustering algorithm (Longitudinal k-means clustering) was used to determine the number of clusters to use, then used the longitudinal KmL algorithm to cluster the participants into groups using their longitudinal patterns of participation. Then, statistical analysis was performed using these group assignments.

The two clusters (A and B) were statistically different from each other. Cluster A, which constituted 66.4% of the participants, demonstrated lower participation at the beginning, and a slight increase in the second segment, but still lower in the third. However, these participants were very active in the 4th segment. In contrast, Cluster B, which comprised 33.6% of the participants were extremely active during the first segment, declined a little during the second and third segment, but increased in the fourth segment, but this was still not as much as the first. Overall, Cluster B's participation was greater than Cluster A. Cluster B's peer assessment scores were statistically better than Cluster A. The conclusion was that the optimum time to encourage active participation was at the beginning during the first segment. In other words, learner engagement developed early on had a downstream effect on subsequent participation.

The study's results supported the conclusion that EDM was a promising technique to understand the nature of learning because it could provide more granular results, handle large volumes of data and detect low-level features, such as the number of times a student visits of forum. This study was of particular interest due to use of educational data mining tools as part of the analysis. Its segmentation of the semester into four three-week periods was an interesting approach. This is more granular than looking at a semester as a whole, but a step up from looking at the semester week by week. In the end, this four-segment approach is attractive because it is consistent with the semester start, mid-term, after mid-term, and then final segmentation typical of most classes.

Tang, et. al. noted several limitations, including the learning environment design, the lack of a test for individual improvements. They recommended that future studies should use larger data sets and investigate knowledge produced using online environments.

Prediction Models based on LMS data

Gašević, Dawson, Rogers, and Gasevic (2016) analyzed the influence of instructional conditions on the prediction of academic success in nine undergraduate blended learning courses to understand the predictive power and significant predictors for course-specific and generalized predictor models.

Several past studies explored the analysis of data in institutional student information systems (SIS) and learning management systems (LMS) to address educational challenges (Baer & Campbell, 2012; Macfadyen & Dawson, 2010; Siemens & Long, 2011). For LMS data, trace data, otherwise known as log data, contains time stamped events of views of resources, assignments, discussions, and similar activities.

Educational Data Mining (EDM) techniques are commonly used to identify patterns in this data (Baker & Yacef, 2009).

The research questions were: 1) What is the level of similarity in student characteristics and LMS usage across different courses in a blended mode of study?; 2) What is the portability of a general model for predicting academic success across courses?; and 3) To what extent does the predictive power of individual variables derived from trace data differ in the prediction of academic success across courses? (p. 70).

Gašević, et al. (2016) used Learning Management System trace data for course activity and student information system data for student characteristic data and completion status so therefore this was a correlational (non-experimental) design. The data used was from a public research-intensive Australian university consisting of four divisions, involving 4,134 students in 2012. Data was collected from nine first year courses that were part of the institution's retention initiative. These courses each had > 150 students and had historically displayed a consistent pattern of low success. The LMS system was Moodle.

The characteristics data included age, gender, international student, language, home, term access, and previous enrollment. The LMS data included tools/feature use (formums, course logins, resources, Turnitin, assignments, book, quizzes, feedback, map, virtual classroom, lessons, and chat). The LMS data was originally collected as continuous data and aggregated. However, due to low utilization, some features such as quizzes, feedback, map, virtual classroom, lessons and chat were dichotomized into accessed and not-accessed. Similarly, Turnitin activity was arbitrarily categorized into did not log; logged 1-2; looked 3 or more times to facilitate data analysis. ANOVA

statistical tests were applied to the continuous data and Chi-square used for the categorical data. Multiple linear regression models were performed on the total sample and for each course. Two logistic regression models were also performed on the total sample and for each course separately to explore the associate between use of LMS features and students' outcomes. The measures used was a percent mark – a continuous variable from 0% to 100% and an academic status which had three outcomes (pass, fail, and withdraw). In the analysis, since only 88 students withdrew, the analysis used pass or fail.

Differences in student characteristics were found across the courses, possibly due to the subject matter. For example, students in the biology class were on average older and had a higher representation of females. Regarding student performance, there were significant differences across courses. Also feature use varied between courses and between subjects. For example, discussion forums were used in Biology 1, but less frequently or not at all by Biology 2. For the total population, approximately 5% of the variability in the student percentage mark was explained by student characteristics and an additional 16% of the variability was explained by online interactions.

When looking at the data at a course level, significant differences in the association between student characteristics and online interactions with student percent marks was found. Course level student characteristics ranged from 2.9% to 14.8% and online interactions ranged from 2.0% to 70.3%.

Course logins was a significant factor when looking at the total population, however, it was not significant at a course level for several courses. In other words,

analysis of the total sample had a tendency to produce underestimates or overestimates of the effect of certain variables.

For almost all courses (exception: Graphics Design), overall prediction accuracy was high using a pass/fail performance status. And after adjusting for student characteristics, a total population analysis indicates that each additional course login decreased the odds of failing the course. However, course login activity is not significant with a course level analysis.

The results revealed significant differences in student characteristics and use of LMS features across the nine courses. Particularly in the extent and frequency LMS features were used. The implication is that there is a need to create prediction models for individual courses incorporating instructional conditions and institutions need to be careful when making academic success prediction if they do include instructional conditions.

The authors also concluded that the use of generalized prediction models for academic success pose a threat to learning analytics because they are inaccurate. More granular course-specific models can produce better insight to help student success and improve course design, however, they "may be unwieldy to implement, despite being more accurate" (p. 83).

Conclusions

Learning Analytics research is over ten years old and has spanned several educational systems in many countries. Much of the earlier work has been focused on determining whether outcomes of individual courses could be improved or predicted. Learning analytics could potentially have increased relevance due to the growing use of

LMS in face-to-face courses (Rhode, Richter, Gowen, Miller, & Wills, 2017) and improved analytical tools, techniques, speed, and affordability.

Many of the studies contained information about earlier work that can be leveraged in other studies. For example, Conijn, et al. (2017) identified and employed predictor variables which were used in earlier studies. Romero, Ventura, and Garcia (2007) explained several computational approaches and tools. While some of these might be dated, the majority are still relevant. This literature research has also suggested the possibility that LMS data alone may be insufficient to reach statistically relevant conclusions (Conijn, et al, 2017). This conclusion suggests an opportunity for additional research.

It wasn't clear why there are relatively few studies about Learning Analytics experiences in the United States aside from Purdue University (Arnold, & Pistill, 2013) and the University of Maryland, Baltimore County (Fritz, 2017). International studies are very useful, however, cultural differences may affect the outcomes.

While this review is oriented towards use of LMS data to address retention, a report authored by Colvin, Wade, Dawson, Gasevic, Buckingham Shum, Nelson, and Fisher (2015) on student retention and learning analytics in Australia note that there is also interest in using LMS to understand learning and teaching practices.

Despite all of the promise of Academic and Learning Analytics, the realization of its potential is not guaranteed. Ifenthaler (2017) believes that one of the obstacles is lack of staff and technology available for learning analytics projects. Kellen (2019) suggests that political silos prevent consolidation of systems and that IT organizations and end-users have not kept pace with advances in analytical approaches and tools. And parts of

the academic community may push back due to perceived and real threats that their privacy and academic independence is threatened. One hope is that a balance is found between information access and benefits – just as people accept data sharing apps such as Waze despite the exposure.

Relationship Between Prior Research and Present Study

The majority of prior research focused on situations where there was high LMS activity because the courses were online. However, in reality, there are many more face-to-face courses than online and many face-to-face courses are taking advantage of LMS features and functions. In other words, LMS activity is increasing in all teaching modes. Furthermore, the percentage of students who are teaching at least one course which uses LMS functions is increasing. Within this context, the ability to identify students whose activity levels are outliers within and across classes may have significant value and has not been studied.

CHAPTER 3 METHOD

Introduction

The objective of this study is to predict new students who are retention risk during the first few weeks of the semester. Early identification would permit intervention while there might be time for the student to change. Indications of potential risk factors would also provide student advisors with additional data to help them prioritize their workload. The desire was to identify an approach which could easily be applied at a large scale.

This study first created a working set of data by selecting relevant records from an archived data set of the Canvas Learning Management System. From this data, a list of new Fall 2021/Spring 2022 students was created based on the date of their first enrollment semester term. Finally, a list of students who left the University was determined by matching the list of new Fall 2021 students with Fall 2022 enrollment. Students in the Fall 2021 list, but were not in the Fall 2022 list, were considered to be not retained.

Second, the Canvas data was transformed into a form which is more conducive for analysis. For example, REQUESTS table entries contain the record number of an entry in a Quiz table. This is represented in the analysis data as a quiz interaction count. Additional data transformations were made to convert activity counts to percentile values to facilitate predictions.

The third step was to evaluate different approaches to analyze the data. Canvas use in courses varies greatly across the University. Also, the meeting mode of course varies from fully online to hybrid to fact-to-face. Some courses meet once a week and

others meet several times of week. In addition, there are differences in course sizes, and consistency of use during the semester. And obviously, student use within courses varies.

This study used binomial logistic regression with different sets of transformed data to determine whether statistically valid conclusions could be reached. The study looked at all activity together at the early weeks of the semester and subsequently looked at course activity during the same period.

Specific Research Questions and Hypotheses

$H1_0$ Learning Management System data for individual course activity at week 3 and week 5 cannot predict student retention. There will be no relationship between LMS Indicators and Not-Retained.

$H1_1$ Learning Management System data for individual course activity at week 3 and week 5 can predict student retention. There will be a relationship between LMS Indicators and Not-Retained.

The alpha level of .05 was chosen to test for significance.

Research Design and Data Analysis

A non-experiment study was conducted due to the absence of an Active Independent Variable. Logistic Regression analysis was conducted due to the dichotomous dependent variable, Not-Retained. The Dependent Variable, Not-Retained was dummy coded as a dichotomous variable with 1 indicating that the student did not continue enrollment and 0 indicating that the student was retained. There were multiple Independent Variables representing Sessions, Requests, Active-Days, Discussion, Content, Design, Quiz, Submission, Other and Assignment.

Reliability and Validity of Research Design

The Canvas Learning Management System provided the raw source of the data for this analysis in the form of the Requests table data. Aside from potentially any defects in the Canvas application, it is accurate and reliable. However, the validity of this data as it relates to providing academic student interaction indicators is dependent on the course design and instructional teaching method of the specific course. If the instructor does not use LMS, then there is no data. Furthermore, instructors independently determine the LMS features (i.e., discussions, quizzes, assignments) to incorporate in the courses and their employment may vary throughout the semester. Student activity varies according to many factors including their schedule and study habits.

Furthermore, LMS data in its raw form is not generally conducive to research analysis and needs to be transformed into a different format (Gašević et al, 2016). For example, if a REQUEST table entry contains the identification number of the Quiz, a Quiz count entry is defined with a value of 1. Also, to facilitate analysis, independent variable values were additionally recoded as percentiles in 5-unit values instead of the raw numeric value.

The processing techniques and underlying theories can introduce questions about the overall validity of this study. Additional discovery of related research is needed to surface academically sound methods and procedures.

Sample and Population

The data sample of this study is from the Canvas LMS for the Spring 2022 semester at St. John's University. To create a subset of the Canvas data pertaining to new students, the Canvas Enrollment table (ENROLLMENT_DIM) was consulted to identify

new students as of the Fall 2021 semester. Enrollment in any academic course prior to the Fall 2021 semester indicated a continuing student. The resulting list of new students was then used to identify any courses which any of them were registered to attend.

This same student enrollment data was used to determine whether a student was not retained after the Spring 2022 semester. This was accomplished by comparing Spring 2022 enrollment with Fall 2022 enrollment. Students registered in the Spring but not in the Fall were categorized as not retained. This satisfies the requirement for new student retention data while limiting data sources to the Canvas LMS. Note that some transfer students are included in the new student group.

See Table 1 below for a summary of the data sample:

Table 1

Spring 2022 Canvas LMS Activity

Variable	Number
New Students	3,088
Courses with New Student Enrollment	1,939
Interactions	1,134,229,992
Sessions	50,057,153

After new students were identified, a list of courses attended by any of these new students was created. The enrollment records were then consulted again to identify other (continuing) students enrolled in each of these courses. As a result, for every course with new student enrollment, there was a list of students with new student and not-retained indicators.

Instruments

The primary data source for this study is Canvas Learning Management System (LMS) log data, otherwise referred to as trace data. This data is produced by the LMS applications as a byproduct of its operation and is not critical to their operation. These timestamped data sets have entries for significant LMS interactions such as logon, logoff, course access, and course activity. Note that it only provides some indication that an action occurred, for example, the submission of an assignment. However, details associated with the assignment such as the name, actual data submitted, and grades are not recorded in the log data. It is also important to note that LMS log data is not the same as server web logs. Web logs record technical interactions, such as mouse clicks, with the system, but what actually occurs by the LMS application behind the scenes is generally difficult to ascertain. LMS Logs provide more insight at a functional level. For example, a logon attempt would be represented by several entries in a web log, but only one LMS log file entry. Conversely, some mobile phone applications periodically take actions to maintain connectivity and appear as multiple logon attempts, even in the absence of human interaction. While other LMS applications, namely Moodle, Desire2Learn, and Canvas produce log data, the format and content is specific to that product.

Canvas log data is automatically by the system as a byproduct of normal operation. This study was conducted on a subset of this archival data pertaining to the Spring 2022 academic semester.

Proxy Variables

Due to the raw, unstructured nature of LMS log data, most studies translate these files into proxy variables and use the proxy variables for analysis (Jo, Kim & Yoon, 2014). While the log data might have several course access entries in a given period, the proxy variable might have just one entry per course per period which contained a count of the number of course accesses. Similarly, data during a time period could be aggregated to a more management grouping. Daily, weekly, or perhaps longer periods. The risk is that these data transformations may affect the validity of the data.

Instructional Conditions

Many sections of the same course might be taught differently and, more importantly, use LMS features differently or not at all. This may materially affect the LMS activity as reflected in the log data. Gašević, Dawson, Rogers, and Gasevic (2016) cautioned about making predictions based on generalized models.

Validity of the Instruments

The focus of this study is to determine whether LMS data can inform higher education retention actions. As noted earlier, raw LMS log data is not in a format that is conducive to statistical analysis without some aggregation or data coding actions. The specific approach taken can materially affect the outcomes of the analysis. Ideally, this analysis would use commonly used data transformation approaches in the absence of official, standards. The research accumulated to date, has not identified one.

The LMS intensity and use patterns differed greatly between each course. This finding is consistent with other studies (Gašević et al., 2016; Casany et al., 2012; Conijn

et al., 2017). From the surface, while developing a model for each course is desirable, it may not be practical if course enrollment is small or LMS interactions are low.

In addition, further analysis is needed to classify when a student drops out of a class to distinguish between 1) a student was absent because the student never intended to attend, 2) a student dropped out of the class due to a scheduling change 3) a student dropped out of all classes 4) a student dropped out of this specific section and 5) whether the student achieved a passing grade in the course. There may be some exogenous reasons for the student to drop the course that may not be visible in their use of the LMS. For example, students may change their enrollment because another course may be a better fit for their schedule or there was an opening in a more desirable course.

There may be other factors that are more relevant than LMS behavior when considering student success such as SAT or GPA. Having understood this, this study attempts to determine whether LMS activity alone is significant.

Reliability of the Instruments

In addition to the aforementioned issues regarding the data transformation of the original log file data, in many courses, there simply might not be sufficient LMS activity associated with the course to be statistically reliable. For example, a course instructor might not have sufficiently used the LMS to make any conclusions regarding student success. Perhaps courses with low LMS activity should be excluded because there is insufficient data to reach any conclusions. Still, the expectation is that there will be adequate data from other courses to determine if the student is exhibiting traits which should be investigated.

Research Ethics

This study was conducted with IRB approval. No data was created for this study because it used existing archival data. Furthermore, no personally identifiable information was used for the analysis.

Procedures for Collecting Data

Learning Management Systems, such as Canvas, collect telemetry about user interactions. Canvas records this information in a relational database table named REQUESTS (n =4,676,654,319) as of February 2, 2023. To facilitate processing for this Spring 2022 study, a subset copy was created with timestamps between January 1, 2022 and May 31, 2022 (n=1,134,229,992). The relevant fields are in Table 2.

Table 2

Selected Canvas LMS REQUESTS Table Fields

Field	Description	Record Count
Timestamp	Date and time of the activity	1,134,229,992
USER_ID	User if applicable	20,409
COURSE_ID	Course if applicable	19,451
HTTP_METHOD	Web update method	1,134,229,992
WEB_APPLICATION_CONTROLLER	Canvas application information	1,134,229,992
SESSION_ID	Logon session	50,057,153

Table 3

Selected Canvas LMS USER_DIM Table Fields

Field	Description
ID	Unique identifier
CANVAS_ID	Unique identifier (short)
WORKFLOW_STATE	Current state of user in Canvas

Table 4

Selected Canvas LMS ENROLLMENT_DIM Table Fields

Field	Description
ID	Unique identifier
TYPE	Indicates student, instructor or other type
WORKFLOW_STATE	Indicates whether the enrollment is accurate
COURSE_ID	Enrolled course
USER_ID	Enrolled user

Table 5

Selected Canvas LMS ENROLLMENT_FACT Table Fields

Field	Description
ENROLLMENT_ID	Corresponding ENROLLMENT_DIM record identifier
USER_ID	User identifier
COURSE_ID	Course identifier
ENROLLMENT_TERM_ID	Enrollment Term

Table 6

Selected Canvas LMS COURSES_DIM Table Fields

Field	Description
ID	Unique course identifier
CANVAS_ID	Unique course identifier (short)
WORKFLOW_STATE	Status of course

The Canvas LMS data were preprocessed for statistical analysis. The official Canvas LMS data files reside in a cloud service provided by Snowflake, Inc. The Snowflake company also provides web-based tools to store and manipulate the data in database and table structures using a structured query language (SQL).

First, a database table to house the REQUESTS data for the study was created using the original table name and structure. Each main REQUESTS record contains the timestamp of the original event. A SQL command was run to copy the REQUESTS

records from the main table into the study table for records within the study's Spring 2022 timeframe resulting in the aforementioned subset copy.

A set of reference tables was then created to identify students in scope and whether these students were not retained. The ENROLLMENT_DIM table at the start of the Spring 2022 semester provided a list of students and their enrolled courses for Spring 2022. Then, students in this list were eliminated from scope if they were registered for courses before the Fall 2021 semester. The remaining students were considered new students. Note that some transfer students may be included in this list of new students. The ENROLLMENT_DIM table at the middle of the Fall 2022 semester was then used to identify the students who were not retained. Specifically, the absence of Fall 2022 enrollment evidence in the Fall 2022 semester indicated that the student was no longer enrolled. To facilitate data analysis, similar derivations of the reference tables were created. Users alone, courses alone and a combined user course table. As mentioned earlier, this latter table contained all students, both new and continuing.

Figure 1 shows enrollment by course.

Figure 1

Enrollment by Course

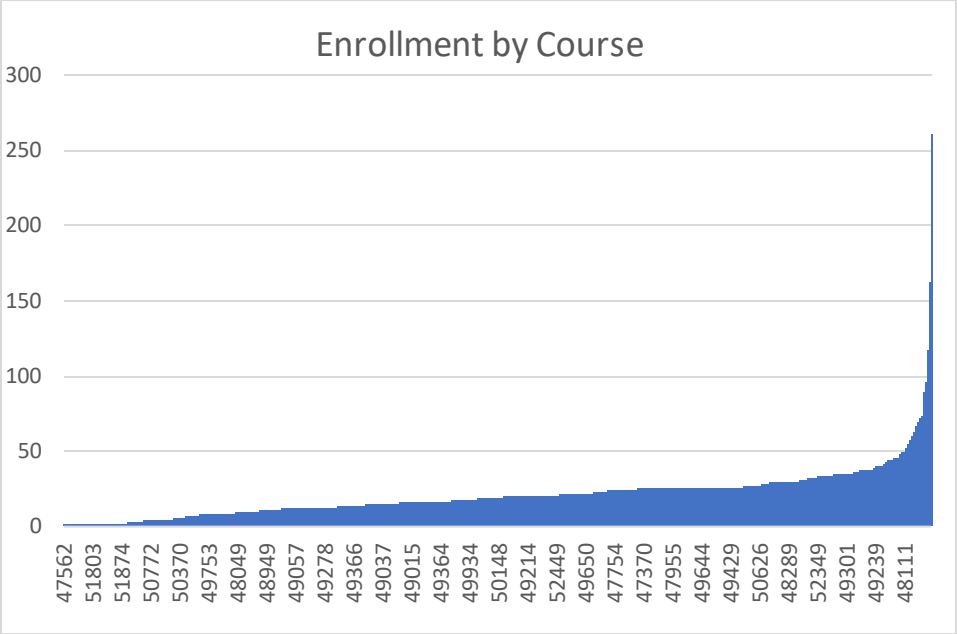


Figure 2

Course Distribution by Enrollment Size

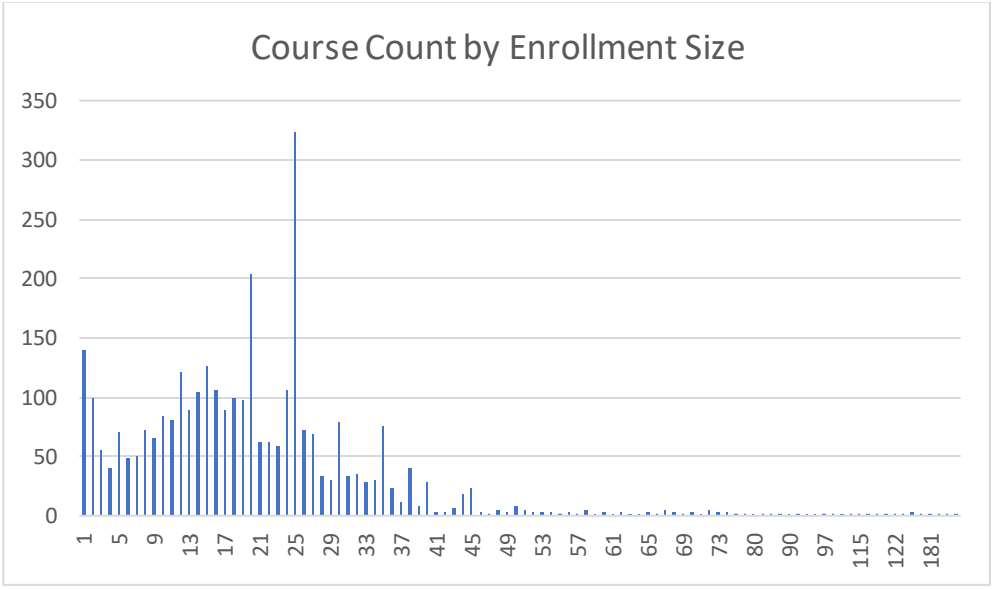
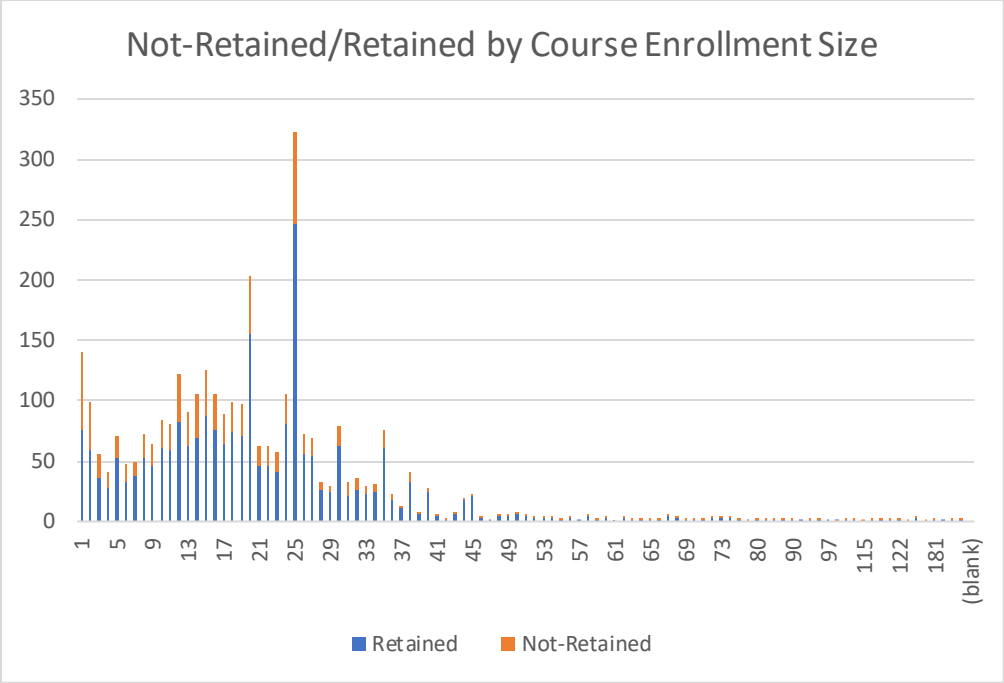


Figure 3

Not-Retained/Retained by Course Enrollment Size



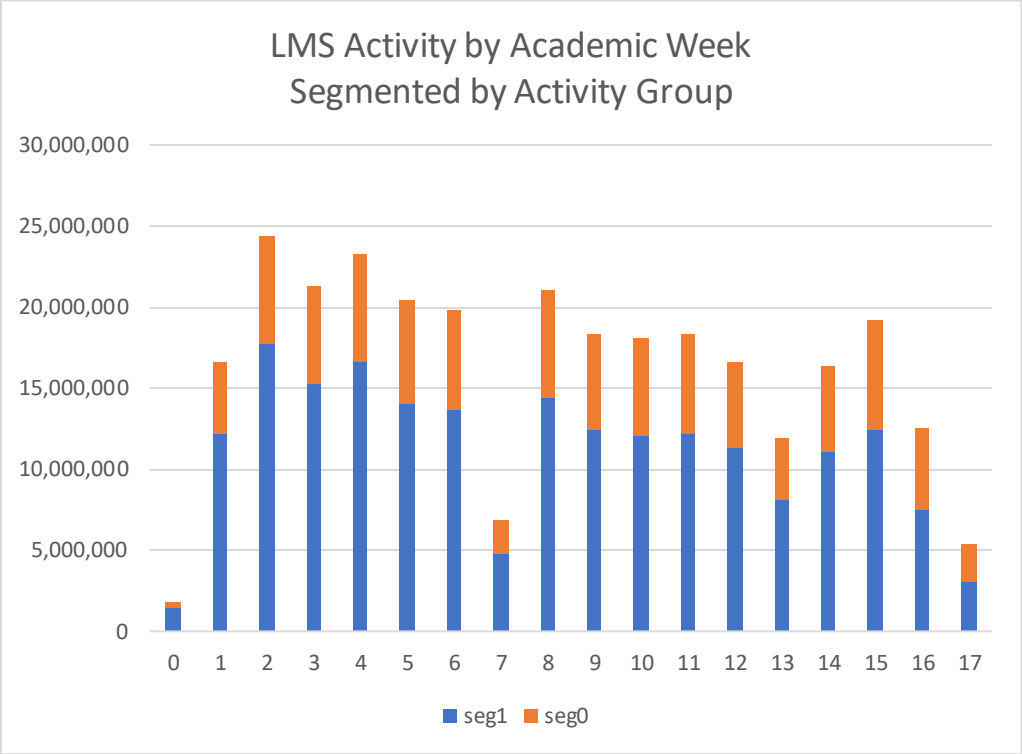
As Figure 3 illustrates, there are many courses with few students. Small sample sizes are not conducive to statistical analysis. Also, courses with low LMS activity would not render any useful statistical insight using binomial logistic regression. Therefore, courses with fewer than thirty students and fewer than 750 average requests by the third week of the semester were placed into a different pool. The low activity pool had 2,725 courses (1,802 included new students) with 50,828 enrolled students (15,047 new student, 35,781 continuing student). The active pool had 205 courses, 1,787 new students, 3,779 continuing students, (total 5,566) with 8,901 enrollments. While this might seem large, however, only 2,427 (74.9%) courses with 60,372 enrollment out of 3,241 courses with 78,904 enrollment were available to students as of that date.

While aware that previous research had recommended course specific analysis, this study first attempted to understand whether there was a statistical relationship

between aggregate student LMS activity at the end of the third week for all courses and retention. Figure 4 depicts LMS activity by week. Student activity begins to increase after the first class session and begins to stabilize by the third week. The desire is to gauge activity as early as possible during the semester in order to have time to intervene.

Figure 4

LMS Activity by Academic Week



Similarly, measurements were taken at the fifth week to see if the additional time improved the prediction. Datasets were created for each of the hypotheses to complete binomial regression analysis.

Data Analysis

IBM SPSS statistical software version 28 was used to analyze the data. Logistic Regression analysis was performed to identify relationships between student activity and non-retention starting for the third and fifth week of the semester.

Data extracts were created using Snowflake Structured Query Language (SQL) statements through a web based Snowsight console. These statements transformed and aggregated the raw REQUESTS table data into one record for each student for each course number for the third and fifth week of the semester. Included in each record were new student and not-retained indicators. For statistical analysis, new students were coded as 1 and continuing students as 0. The not-retained indicator was coded as 1 if the student was not retained and 0 if the student was retained. This base file contained actual activity counts. In an attempt to normalize the results, a Python program was used to add additional columns to the file containing the percentile rank and z score for the student activity within the course. Specifically, Pandas dataframe operations were run for each of the course numbers in the input file to calculate the percentile rank and z score values for each of the independent variables. These additional columns were appended to the end of the original data extract. These operations facilitated analysis using the original counts, rank counts, or z scores.

An additional query was used to create a list of course numbers where the course had at least 30 students and at least one student with a not-retained indicator. The SPSS program's graphical user interface was used to perform the SPSS actions to conduct the initial set of binomial logistic regression and Receiver Operating Characteristic (ROC) Curve operations. The SPSS commands in the output file were copied into a Microsoft Excel spreadsheet with a list of 159 course numbers to analyze. As mentioned earlier, these courses had at least 30 students and at least one student who was not-retained. Excel formulas inserted the course number into SPSS selection statement and the prediction and studentized residual column names. An SPSS Syntax window was started

and populated with SPSS commands to import the data extract in comma separated value (.csv) format into the SPSS Editor. The SPSS statements created from the Excel spreadsheet were pasted into the Syntax window after double quotes characters were removed. The running of the Syntax window resulted binomial logistic regression and ROC Curve analysis of the majority of the 159 selected courses. Note that SPSS calculates and saves each PREDiction for every row of the Data Editor. The Data Editor and Output files for each iteration was saved.

CHAPTER 4 PRESENTATION OF RESULTS

Introduction

The purpose of this study is to determine whether Learning Management System (LMS) data can be used to identify students who are at risk of not completing their studies and to make this assessment while there is still time to intervene. The underlying premise is that facets of the LMS data is statistically related to non-completion. This study considered daily sampling frequency and interaction activity.

Findings

The binomial logistic regression is appropriate to consider for this study because there is one dichotomous dependent (retained or not retained) with a variety of continuous independent variables to consider. The independent variables considered included:

- Sessions – a continuous variable which indicates a collection of interactions by a user during a login period.
- Requests – a continuous variable reflecting the number of Canvas log entries. In this analysis, the requests are directly associated with a course as opposed to interactions such as login which cannot be directly associated with a course.
- Active Day – a continuous variable which indicates the number of calendar days any activity was observed. This usually relates to the number of sessions.
- Discussion – a continuous variable which indications the reading, writing, or editing of Canvas discussion posts.
- Content – a continuous variable which reflects actions associated with reading or writing content.

- Design – a continuous variable indicating changes to the design of a course.
- Quiz – a continuous variable which reflects actions associated with taking or reviewing a Canvas quiz.
- Submissions – a continuous variable triggered by an submission into the system.
- Other – A continuous variable for other activities.
- Assignment – a continuous variable associated with submitting Canvas assignments.

SPSS Version 28.0 was used to analyze these variables in several forms:

- Observation counts – the accumulated number of occurrences of each activity (i.e., Requests) for specific measurement periods. This study used the end of the 3rd and 5th academic weeks.
- Rank – the ranked order of the accumulated observation counts of each student for each course for the measurement period grouped by 5th percentile
- Z Score – the Z score of the accumulated observation counts of each student for each course for the measurement periods
- Frequency Bin – the grouping of accumulated observation counts of each student for each course for the measurement periods within 15 buckets

Binomial Logistic Regression Assumptions

Seven assumptions need to be satisfied for Binomial Logistic Regression analysis (Lund & Lund, n.d.; Laerd Statistics, 2017). The first assumption that there is one dependent variable that is dichotomous is satisfied by the Not-Retained variable, coded as 1 for not retained and 2 for retained. The second assumption that there are one or more independent variables that are measured on either a continuous or nominal scale is

satisfied by all of the defined independent variables. The third assumption that there is independence of observations and categories of the dichotomous depended variable and nominal independent variables is met. For the fourth assumption related to the minimum number of test cases, as mentioned in Chapter 3, the courses were segmented by enrollment size of 30 following the guidelines of ten samples for each variable (Agresti, 2007). Lund & Lund (n.d.) recommend fifteen per independent variable. Two hundred five courses with enrollment equal to or greater than 30 were analyzed with SPSS. One hundred eighty two of the two hundered five courses meeting the minimum size constraint included Not-retained students. The fifth assumption requires that there needs to be a linear relationship between the continuous independent variables and the logic transformation of the dependent variables. To accomplish this, the SPSS compute function was used to calculate the natural log value of each variable and interaction terms for each independent variable and respective natural log transformed variable was created. The Box-Tidwell (1962) procedure included in the SPSS procedure test for linearity. The sixth assumption is that two or more continuous independent variables there should not be any show any multicollinearity. This can be detected through linear regression tests. The seventh assumption is that there should be no significant outliers. This will be disclosed in SPSS casewise diagnostics.

LMS activity varies from course to course. First, it is dependent on whether the instructor chooses to use the LMS to teach the course. Second, it is dependent on the which LMS features and functions used in the class and third, the schedule of these features or functions during the semester. In addition, the level of activity recorded in the LMS logs depend on how course material is stored in the LMS. For example, course

material stored in several LMS pages will invoke more activity than course material stored as an uploaded file. To isolate the activity, several studies recommend analyzing student activity at an individual course level. As mentioned earlier, this study analyzed courses which had at least 30 students to satisfy binomial logistic regression sample sizes. In addition, these courses needed to have at least one student in the Not-retained dependent variable. To facilitate analysis across all courses and to stay within size guidelines, tests were conducted using the REQUESTS variable. SPSS runs were conducted using combinations of the other tests, however, due to course design differences, only REQUESTS demonstrated value among the courses. In addition, several transformations of the REQUEST activity counts were made to determine if any of them would be more appropriate for prediction. These variations were Rank and Frequency Group.

Table 7 sumarizes the 77 individual binomial logistic regression runs against REQUESTS activity for week 3 that did not have casewise outliers or other errors. Course 47291 will be reported in detail as a model for the rest of the samples.

Table 7

Summary of Selected SPSS Results for Week 3

Course	X^2	Nagelkerke R2	Classification	*p*	ROC	Constant B	Variable B	Variable Odds Ratio
47237	0.099	1.2%	98%	0.717	0.563	-3.874	0.005	1.005
47240	0.000	0.0%	94%	0.997	0.437	-2.768	0.000	1.000
47281	0.147	0.6%	86%	0.710	0.437	-1.595	-0.006	0.994
47291	6.607	26.7%	60%	0.056	0.563	-1.842	0.038	1.039
47298	0.000	0.0%	93%	0.991	0.437	-2.632	0.000	1.000
47301	0.384	4.8%	97%	0.505	0.563	-4.128	0.019	1.020
47343	2.156	9.4%	72%	0.280	0.437	-0.400	-0.023	0.977
47391	2.003	12.7%	91%	0.165	0.563	-3.884	0.059	1.061

Course	X^2	Nagelkerke R2	Classification	*p*	ROC	Constant B	Variable B	Variable Odds Ratio
47447	0.418	1.7%	55%	0.527	0.563	-0.108	0.011	1.011
47453	0.123	1.3%	98%	0.746	0.437	-3.625	-0.015	0.985
47463	1.173	13.9%	98%	0.293	0.563	-5.990	0.058	1.059
47472	0.000	0.0%	95%	0.989	0.437	-2.914	0.000	1.000
47694	2.563	31.4%	97%	0.185	0.563	-6.548	0.062	1.064
47729	0.779	3.7%	77%	0.373	0.563	-1.352	0.009	1.009
47744	0.077	0.3%	55%	0.782	0.563	-0.186	0.002	1.002
47795	0.639	7.6%	98%	0.515	0.437	-2.354	-0.061	0.940
47907	2.238	8.5%	62%	0.163	0.563	-0.818	0.019	1.020
47927	0.104	0.9%	93%	0.757	0.437	-2.327	-0.015	0.985
47938	0.000	0.0%	38%	0.996	0.563	-0.002	0.000	1.000
47954	2.906	21.8%	95%	0.204	0.437	-0.317	-0.171	0.843
47971	0.043	0.2%	83%	0.837	0.437	-1.458	-0.007	0.993
47977	2.405	5.3%	56%	0.159	0.563	-0.606	0.015	1.015
47998	3.563	15.9%	77%	0.140	0.437	0.475	-0.079	0.924
48000	0.418	2.5%	98%	0.557	0.437	-2.819	-0.022	0.978
48117	3.350	13.0%	58%	0.092	0.563	-1.386	0.025	1.025
48144	0.214	1.8%	94%	0.630	0.563	-3.225	0.013	1.013
48147	0.005	0.1%	97%	0.947	0.437	-3.371	-0.001	0.999
48186	2.449	9.7%	58%	0.216	0.563	-1.035	0.018	1.018
48188	1.284	6.9%	94%	0.328	0.563	1.794	0.036	1.037
48201	0.836	10.4%	97%	0.465	0.437	-2.022	-0.080	0.923
48300	0.050	0.2%	62%	0.822	0.437	0.606	-0.004	0.996
48329	1.226	14.6%	97%	0.430	0.437	-1.792	-0.093	0.911
48332	2.199	9.0%	70%	0.149	0.563	-1.575	0.036	1.036
48338	4.828	38.4%	97%	0.064	0.563	-5.136	0.042	1.043
48385	0.246	1.1%	77%	0.614	0.437	1.427	-0.007	0.993
48407	1.162	4.4%	79%	0.275	0.563	-1.856	0.011	1.011
48451	0.004	0.0%	78%	0.948	0.437	-1.205	-0.001	0.999
48510	0.003	0.0%	59%	0.955	0.563	0.354	0.000	1.000
48793	0.003	0.0%	97%	0.956	0.437	-3.389	-0.002	0.998
48857	0.062	0.7%	98%	0.786	0.563	-3.920	0.005	1.005
48935	0.553	3.3%	93%	0.427	0.437	3.070	-0.013	0.987

Course	X^2	Nagelkerke R2	Classification	*p*	ROC	Constant B	Variable B	Variable Odds Ratio
48952	1.444	5.5%	77%	0.226	0.563	-1.813	0.015	1.016
48976	0.872	5.0%	88%	0.339	0.437	2.668	-0.011	0.989
48995	0.122	0.4%	92%	0.713	0.437	2.571	-0.004	0.996
49006	0.076	0.2%	87%	0.779	0.437	2.079	-0.005	0.995
49020	0.472	1.7%	79%	0.515	0.563	0.910	0.011	1.011
49059	0.708	4.7%	91%	0.476	0.563	1.587	0.025	1.025
49063	1.432	6.5%	92%	0.212	0.437	3.365	-0.016	0.984
49077	1.201	5.3%	89%	0.402	0.437	-1.439	-0.033	0.968
49134	0.568	3.0%	84%	0.505	0.437	-0.991	-0.025	0.976
49198	0.020	0.0%	86%	0.888	0.437	1.893	-0.002	0.998
49240	0.525	1.8%	58%	0.473	0.563	-0.544	0.006	1.006
49288	0.006	0.0%	87%	0.937	0.563	-1.933	0.002	1.002
49312	1.357	5.5%	82%	0.245	0.437	1.911	-0.016	0.984
49527	2.066	16.3%	94%	0.155	0.563	-4.979	0.063	1.065
49565	0.026	0.1%	88%	0.874	0.563	1.857	0.002	1.002
49567	4.279	27.8%	90%	0.148	0.563	-0.018	0.112	1.118
49624	0.006	0.1%	98%	0.940	0.563	-3.857	0.006	1.006
49712	0.416	2.9%	96%	0.556	0.563	2.483	0.020	1.020
49718	0.040	0.1%	91%	0.851	0.563	2.292	0.001	1.001
49897	0.017	0.1%	77%	0.897	0.437	-1.079	-0.004	0.996
50004	0.799	3.8%	96%	0.439	0.563	2.263	0.032	1.032
50032	0.038	0.2%	82%	0.850	0.563	1.425	0.002	1.002
50045	1.048	5.3%	90%	0.272	0.437	2.712	-0.011	0.989
50050	0.163	1.4%	94%	0.703	0.563	2.123	0.015	1.016
50265	1.458	8.5%	93%	0.304	0.563	1.461	0.052	1.053
50389	0.516	2.2%	66%	0.500	0.563	0.360	0.010	1.010
50392	4.009	24.7%	91%	0.145	0.563	-0.115	0.155	1.167
50696	0.029	0.1%	80%	0.865	0.437	1.471	-0.001	0.999
51688	1.061	2.5%	85%	0.284	0.437	2.111	-0.007	0.993
51981	0.332	1.9%	88%	0.551	0.437	2.502	-0.013	0.987
52259	0.081	0.7%	94%	0.785	0.563	2.347	0.011	1.011
52349	0.159	0.7%	75%	0.697	0.437	-0.917	-0.003	0.997
52413	1.002	2.7%	65%	0.357	0.437	-0.294	-0.010	0.990

Course	X^2	Nagelkerke R2	Classification	*p*	ROC	Constant B	Variable B	Variable Odds Ratio
52617	4.204	27.4%	90%	0.167	0.563	-0.745	0.122	1.129
52669	1.615	9.7%	92%	0.322	0.563	1.345	0.040	1.041
52670	0.867	1.4%	95%	0.323	0.437	3.232	-0.011	0.989

Course 47291 REQUEST Activity Counts – Week 3

The first three assumptions were met. Course 47291 had a total of 30 students. 12 were coded as 1 meaning Not-Retained and 18 coded as 0, meaning retained. The natural log of REQUESTS was calculated. A binomial logistic regression was performed to ascertain the effects of REQUESTS activity on the likelihood that students will not be retained. Linearity of the continuous variables with respect to the logit of the dependent variable was assessed via the Box-Tidwell (1962) procedure. Based on this assessment, the continuous independent variable was found to be linearly related to the logit of the dependent variable. No outliers were observed. The logistic regression model was statistically significant, $\chi2(1) = 6.607$, $p = .010$. The model explained 26.7% (Nagelkerke R2) of the variance in REQUESTS and correctly classified 60.0% of cases. Sensitivity was 25.0%, specificity was 83.3%, positive predictive value was 50.0% and negative predictive value was 62.5%. REQUESTS was not found to be statistically significant. The area under the ROC curve was .563, 95% CI [.551, .576] which is a poor discrimination according to Hosmer et al. (2013).

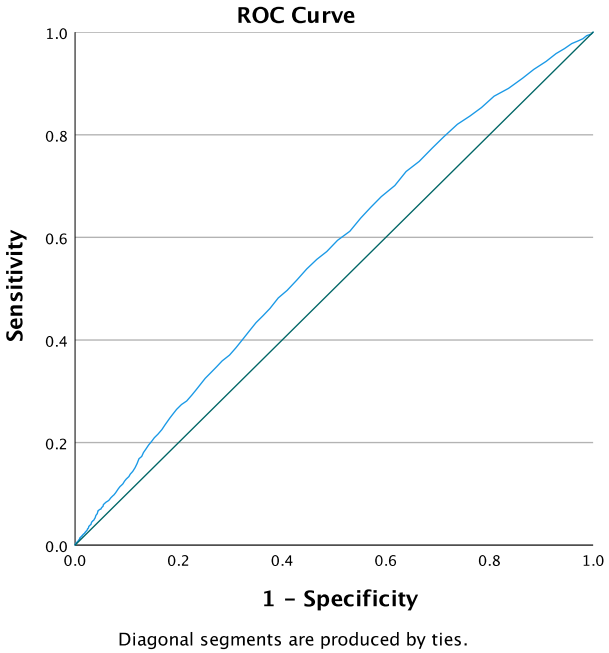
Table 8

Logistic Regression Predicting Not-Retained based on REQUESTS Activity at Week 3

	B	*SE*	Wald	*df*	*p*	Odds Ratio	95% for Odds ratio	
							Lower	Upper
Requests	.038	.020	3.654	1	.056	1.039	.999	1.080
Constant	-1.842	.820	5.043	1	.025	.158		

Figure 5

ROC Curve for Not-Retained based on REQUESTS Activity at Week 3



Course 47291 REQUEST Activity Counts – Week 5

The first three assumptions were met. Course 47291 had a total of 30 students. 12 were coded as 1 meaning Not-Retained and 18 coded as 0, meaning retained. The natural log of REQUESTS was calculated. A binomial logistic regression was performed to ascertain the effects of REQUESTS activity on the likelihood that students will not be retained. Linearity of the continuous variables with respect to the logit of the dependent variable was assessed via the Box-Tidwell (1962) procedure. Based on this assessment, the continuous independent variable was found to be linearly related to the logit of the dependent variable. No outliers were observed. The logistic regression model was statistically significant, $\chi 2(1) = 6.955$, $p = .008$. The model explained 28.0% (Nagelkerke R2) of the variance in REQUESTS and correctly classified 70.0% of cases. Sensitivity was 50.0%, specificity was 83.3%, positive predictive value was 66.7% and negative

predictive value was 71.4%. REQUESTS was found to be statistically significant. The area under the ROC curve was .554, 95% CI [.542, .567] which is a poor discrimination according to Hosmer et al. (2013).

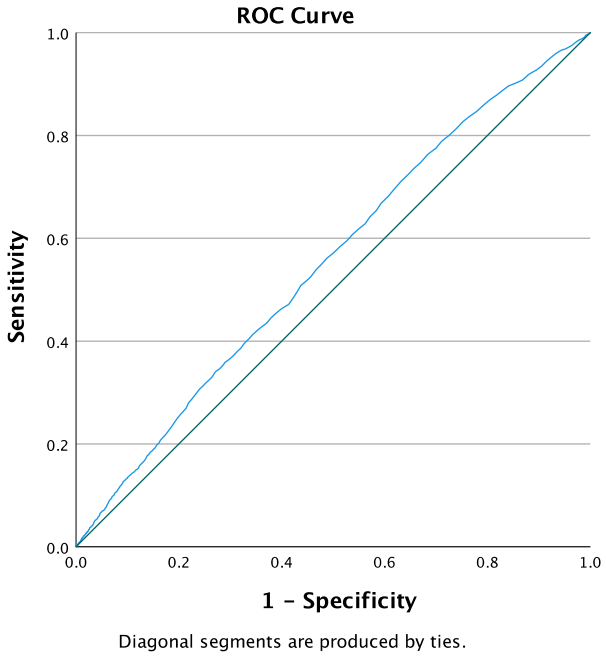
Table 9

Logistic Regression Predicting Not-Retained based on REQUESTS Activity at Week 5

	B	*SE*	Wald	*df*	*p*	Odds Ratio	95% for Odds ratio	
							Lower	Upper
Requests	.021	.010	4.173	1	.041	1.021	1.001	1.042
Constant	-1.807	.777	5.410	1	.020	.164		

Figure 6

ROC Curve for Not-Retained based on REQUESTS Activity at Week 5



Course 47291 REQUEST Rank at Week 3

The first three assumptions were met. Course 47291 had a total of 30 students. 12 were coded as 1 meaning Not-Retained and 18 coded as 0, meaning retained. The natural log of REQUESTS_RANK was calculated. A binomial logistic regression was

performed to ascertain the effects of REQUESTS_Rank activity on the likelihood that students will not be retained. Linearity of the continuous variables with respect to the logit of the dependent variable was assessed via the Box-Tidwell (1962) procedure. Based on this assessment, the continuous independent variable was found to be linearly related to the logit of the dependent variable. There was one standardized residual with a value of 2,022 standard deviations, which was kept in the analysis.The logistic regression model was statistically significant, $\chi 2(1) = 6.619$, $p = .010$. The model explained 26.8% (Nagelkerke R2) of the variance in REQUESTS_Rank and correctly classified 66.7% of cases. Sensitivity was 50.0%, specificity was 77.8%, positive predictive value was 60.0% and negative predictive value was 30.0%. REQUESTS_rank was found to be statistically significant (as shown in Table 2.) The area under the ROC curve was .517, 95% CI [.504, .529] which is a poor discrimination according to Hosmer et al. (2013).

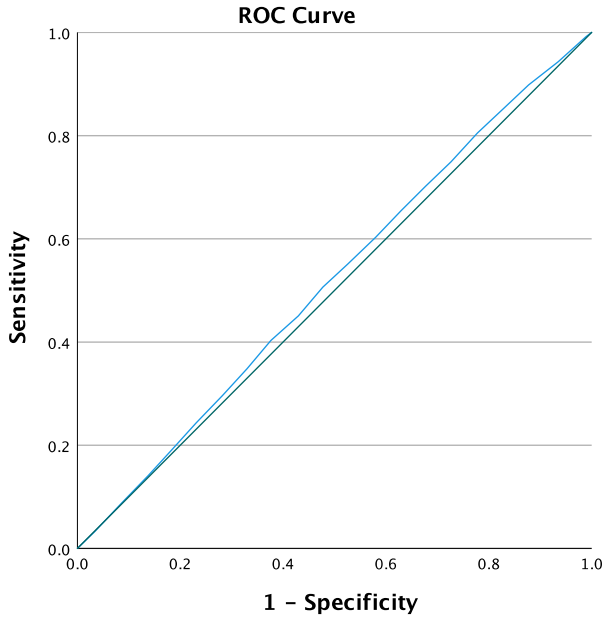
Table 10

Logistic Regression Predicting Not-Retained based on REQUESTS_Rank Activity at Week 3

	B	*SE*	Wald	*df*	*p*	Odds Ratio	95% for Odds ratio	
							Lower	Upper
Requests_Rank	.038	.017	5.261	1	.022	1.039	1.006	1.073
Constant	-2.493	1.041	5.740	1	.017	.158		

Figure 7

ROC Curve for Not-Retained based on REQUESTS_Rank Activity at Week 3



Course 47291 REQUEST Rank at Week 5

The first three assumptions were met. Course 47291 had a total of 30 students. 12 were coded as 1 meaning Not-Retained and 18 coded as 0, meaning retained. The natural log of REQUESTS_RANK was calculated. A binomial logistic regression was performed to ascertain the effects of REQUESTS_Rank activity on the likelihood that students will not be retained. Linearity of the continuous variables with respect to the logit of the dependent variable was assessed via the Box-Tidwell (1962) procedure. Based on this assessment, the continuous independent variable was found to be linearly related to the logit of the dependent variable. There was one standardized residual with a value of 2.018 standard deviations, which was kept in the analysis. The logistic regression model was statistically significant, $\chi 2(1) = 6.857$, $p = .009$. The model explained 27.6% (Nagelkerke R2) of the variance in REQUESTS_Rank and correctly

classified 73.3% of cases. Sensitivity was 58.3%, specificity was 83.3%, positive predictive value was 70.0% and negative predictive value was 75.0%. REQUESTS_rank was found to be statistically significant (as shown in Table 2.) The area under the ROC curve was .517, 95% CI [.504, .529] which is a poor discrimination according to Hosmer et al. (2013).

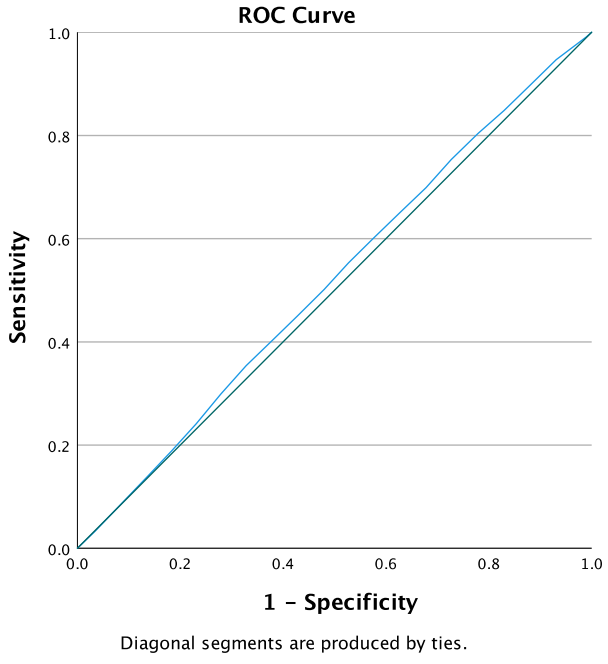
Table 11

Logistic Regression Predicting Not-Retained based on REQUESTS_Rank Activity at Week 5

	B	*SE*	Wald	*df*	*p*	Odds Ratio	95% for Odds ratio	
							Lower	Upper
Requests_Rank	.038	.016	5.408	1	.020	1.039	1.006	1.073
Constant	-2.493	1.031	5.847	1	.016	.083		

Figure 8

ROC Curve for Not-Retained based on REQUESTS_Rank at Week 5



Course 47291 REQUEST Frequency Group at Week 3

The first three assumptions were met. Course 47291 had a total of 30 students. 12 were coded as 1 meaning Not-Retained and 18 coded as 0, meaning retained. The natural log of REQUESTS_BIN was calculated. A binomial logistic regression was performed to ascertain the effects of REQUESTS_Bin activity on the likelihood that students will not be retained. Linearity of the continuous variables with respect to the logit of the dependent variable was assessed via the Box-Tidwell (1962) procedure. Based on this assessment, the continuous independent variable was found to be linearly related to the logit of the dependent variable. No outliers were observed. The logistic regression model was statistically significant, $\chi 2(1) = 6.329$, $p = .012$. The model explained 25.7% (Nagelkerke R2) of the variance in REQUESTS_Bin and correctly classified 66.3% of cases. Sensitivity was 25.0%, specificity was 88.9%, positive predictive value was 60.0% and negative predictive value was 36.0%. REQUESTS_Bin was not found to be statistically significant (as shown in Table 3.) The area under the ROC curve was .505, 95% CI [.493, .518] which is a poor discrimination according to Hosmer et al. (2013).

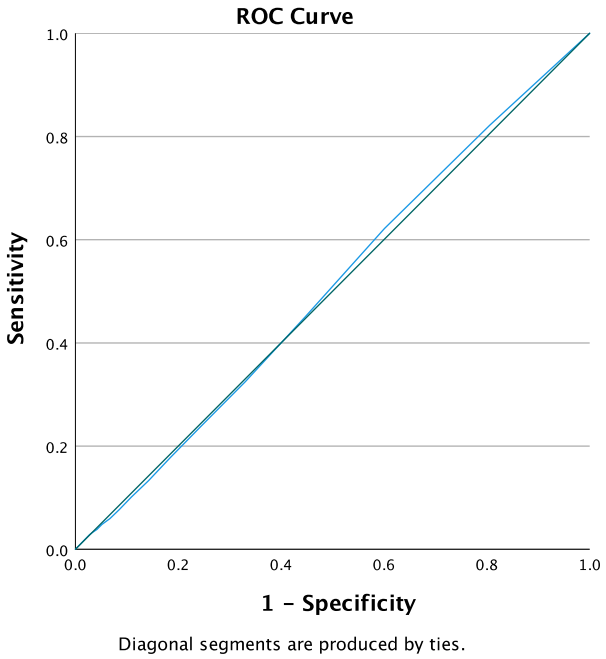
Table 12

Logistic Regression Predicting Not-Retained based on REQUESTS_Bin Activity at Week 3

	B	*SE*	Wald	*df*	*p*	Odds Ratio	95% for Odds ratio	
							Lower	Upper
Requests_Bin	1.025	.547	3.513	1	.061	2.786	.954	8.132
Constant	-2.330	1.060	4.827	1	.028	.097		

Figure 9

ROC Curve for Not-Retained based on REQUESTS_Bin at Week 3



Course 47291 REQUEST Frequency Group at Week 5

The first three assumptions were met. Course 47291 had a total of 30 students. 12 were coded as 1 meaning Not-Retained and 18 coded as 0, meaning retained. The natural log of REQUESTS_BIN was calculated. A binomial logistic regression was performed to ascertain the effects of REQUESTS_Bin activity on the likelihood that students will

not be retained. Linearity of the continuous variables with respect to the logit of the dependent variable was assessed via the Box-Tidwell (1962) procedure. Based on this assessment, the continuous independent variable was found to be linearly related to the logit of the dependent variable. No outliers were observed. The logistic regression model was statistically significant, $\chi 2(1) = 6.735$, $p = .009$. The model explained 27.2% (Nagelkerke R2) of the variance in REQUESTS_Bin and correctly classified 70.0% of cases. Sensitivity was 58.3.0%, specificity was 77.8%, positive predictive value was 63.6% and negative predictive value was 73.7%. REQUESTS_Bin was found to be statistically significant (as shown in Table x.) The area under the ROC curve was .508, 95% CI [.496, .521] which is a poor discrimination according to Hosmer et al. (2013).

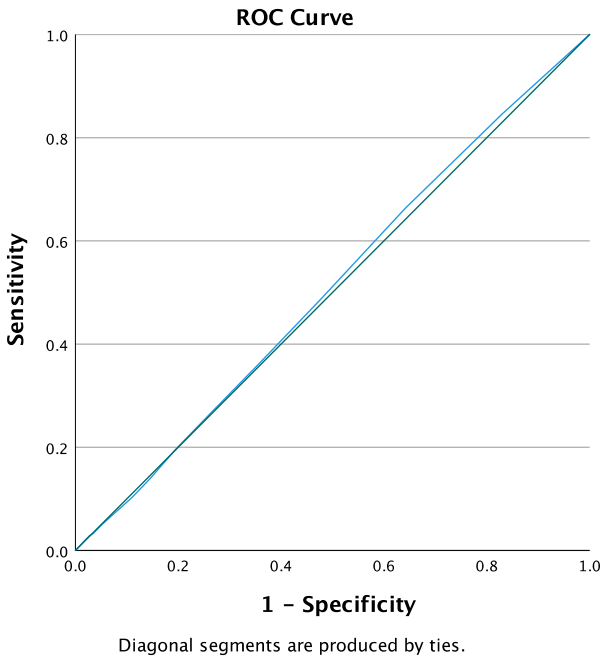
Table 13

Logistic Regression Predicting Not-Retained based on REQUESTS_Bin Activity at Week 5

	B	*SE*	Wald	*df*	*p*	Odds Ratio	95% for Odds ratio	
							Lower	Upper
Requests_Bin	.767	.385	3.975	1	.046	2.153	1.013	4.575
Constant	-2.237	.976	5.258	1	.022	.107		

Figure 10

ROC Curve for Not-Retained based on REQUESTS_Bin at Week 5



This study attempted to identify a statistically verifiable method to identify students at risk of not being retained by the university based on the level of their activity reflected LMS activity data logs during the early weeks of the semester while intervention was still viable. Various techniques were used to transform the raw activity data into a form that could be normalized to a larger sample. Then, a set of courses was identified which satisfied all of the assumptions required for binomial logistic regression analysis. After conducting statistical analysis of student activity at a course level during the third and fifth weeks of the spring 2022 semester this study failed to identify any statistically significant relationship between the level of REQUESTS activity and retention.

The results from a representative Course 42791 were used to illustrate the analysis applied to 182 courses. Course 42791 satisfied all binomial logistic regression assumption tests, however, the results were either marginally statistically significant or not statistically significant. Furthermore, ROC Curve analysis disclosed poor discrimination in all variations. As expected, the statistical significance of the binomial logistic regressions was greater in the fifth week than the third week.

Table 14

Comparison of Course 42791 Results

	χ^2	Nagelkerke R2	Classification	*p*	ROC
Requests Activity Week 3	6.607	26.7%	60.0%	.056	.563
Requests Activity Week 5	6.955	28.0%	70.0%	.041	.554
Requests Rank Week 3	6.619	26.8%	66.7%	.022	.517
Requests Rank Week 5	6.857	27.6%	73.3%	.020	.517
Requests Bin Week 3	6.329	25.7%	66.3%	.061	.505
Requests Bin Week 5	6.735	27.2%	70.0%	.046	.508

Table 15

Comparison of Course 42791 Predictions

	Constant *B*	Variable *B*	Variable Odds Ratio
Requests Activity Week 3	-1.842	.038	1.039
Requests Activity Week 5	-1.807	.021	1.021
Requests Rank Week 3	-2.493	.038	1.039
Requests Rank Week 5	-2.493	.038	1.039
Requests Bin Week 3	-2.330	1.025	2.786
Requests Bin Week 5	-2.237	.767	2.153

Research Question 1

- Learning Management System data for individual course activity at week 3 and week 5 cannot predict student retention.

This study failed to identify a statistically significant relationship between LMS Indicators and Not-Retained. The null hypothesis is accepted. The REQUESTS data did not consistently demonstrate statistically significance and failed to show strong discrimination. Sample size constraints due to class sizes limited analysis and the diversity of LMS employment in courses complicated analysis. For example, discussions were only used in a portion of the course sample, so inclusion of this data variable could not be used in a model for all courses.

Conclusions

Binomial Logistic Regression analysis was conducted against 182 courses using several variations of cumulative LMS activity counts at the end of week 3 and week 5. This study failed to identify a statistically significant method to identify students at retention risk based on their Canvas activity as recorded in LMS data logs.

CHAPTER 5 DISCUSSION

Introduction

This study endeavored to identify students at risk of not completing their studies based on their academic activity as represented in Canvas Learning Management System log data. With the assumption that individual instructors have the means to monitor student course activity, the results were intended to be used to create a list of students at higher retention risk, based on all of a students' course activity, for consideration by academic advisors during the early weeks of the semester while intervention would be more likely to have positive effects. Within this context, this study sought to identify a method to assess student retention risk based on student interactions with all of their courses without any understanding of individual course design.

The Canvas Learning Management System data set contains timestamped activity records of the Canvas system. For example, reading a course discussion post would be represented by one record. Re-displaying the same page would be represented in a different record. This study transformed the dataset into an aggregated form for analysis. Each record of the transformed data contained the student number, course number, academic week number and counts of several selected fields, including records, discussions, and quizzes.

The study focused on Week 3 and Week 5 of the semester. By this time, the course enrollment has normally stabilized and there is still time to intervene. Data extracts for Week 3 and Week 5 were created based on aggregation of the log data. In addition, the ranking, grouped by units of 5, of students within each course was calculated, and student activity within each course was placed into one of fifteen groups

based on the highest level of activity. The intent of these additional transformations was to normalize the student activity. Raw activity counts of student activity varies by course, by week, and by semester. Attempts were also made to develop an alternative measure which could be applied to courses in other semesters.

Binomial logistic regression was selected because there is one dichotomous dependent (retained or not retained) with a variety of continuous independent variables to consider. Preliminary application of this analysis against the full sample resulted in several hundred casewise exceptions. Because most of the exceptions were Not-retained students, simply deleting these exceptions from the sample was undesirable. The alternative of expanding the acceptance criteria was also problematic because it would skew the model. As a result, the study was focused on course-level analysis as recommended by Gašević, et al. (2016).

Actual course enrollment reduced the number of courses which could be included in course-level analysis due to the sample size minimums needed for binominal logistic regression. Still, student participation in this smaller subset by new students was broad enough to be worthwhile. SPSS binomial logistic regression was run for 162 courses. Seventy-seven courses remained after removing courses due to casewise exceptions and other statistical issues. The statistical results were not consistently significant and those which were significant did not provide adequate significant discrimination to be meaningful.

Implications of Findings

This study was based on a large sample set based on courses (n=1,939), students (n=3,088) and course enrollments (n=19,451). It attempted to determine if there were any relationships between LMS log data activity solely and retention solely on student log activity without delving into individual course design or use of LMS features. This study's failure to identify a statistical relationship could be due to several reasons. Due to the technical issues encountered with applying binomial logistic regression to the data, the primary reason could be that the statistical approach is inappropriate to directly apply to the data used in the study. The absence of observable activity typically presents as an outlier, and outliers are typically excluded from consideration because they skew the results. However, in this situation, the lack of any or low observable activity is important. Another reason to avoid use of binomial logistic regression is due to class enrollment sizes found in the sample. Many course sizes were smaller than ideal for this approach.

Relationship to Prior Research

Online courses were the focus of many earlier research studies of LMS data because the majority of LMS courses were asynchronous online courses. These studies attempted to determine the relationship of LMS activity and completion of the course or achieving a specific grade. Student activity of course specific activities such as submission assignments was a factor in their models. Several studies (Cohen, 2017; Gašević, Dawson, Rogers, & Gasevic, 2014; Macfadyen & Dawson, 2010) concluded that student course dropout could be predicted for select courses where there was an opportunity to calibrate the LMS data analysis to the design of the course. However,

Gašević et. al (2014) found that student activity differed across the nine courses in their study with enrollment varying from 192 to 746. The large enrollment permitted Gašević et. al. to consider as many as 12 LMS features.

Significant differences in instructor and student use of LMS features across courses, particularly in the extent and frequency LMS features were used was found, and this experience is consistent with prior research (Gašević et al, 2014; Gašević, 2016; Conijn, et al, 2017).

Recent research (López-Zambrano, Lara & Romero, 2021; López-Zambrano, Lara & Romero, 2020) has recommended using high-level attributes with more semantic meaning, called ontologies instead of the low-level attributes used in this study. This research indicates increased model portability with improved predictive accuracy.

Limitations of the Study

The original LMS data was processed to create the data used for this study. This work created attributes related to Canvas functions such as discussions, however, the analysis did not utilize them because the sample size limited the number of variables, but more importantly because many of these features was not across all courses. Attempts to include them in regression models resulted in missing data situations. Also, in many cases, a combination of course design and student behavior resulted in many data outliers causing statistical processing issues.

Recommendation for Future Practice

This study considered all courses taken by new students and multiple sections of the same course were considered separate courses. New students at this University are all required to take a set of core courses. Focusing on this subset of courses would possibly

provide a sample more appropriate for statistical analysis. Combining multiple sections of a course using the same design would also facilitate the identification of students whose participation is less than their peers.

SPSS was difficult to use for multiple samples. This study ran several regression runs against 162 course samples. For each of the three primary data forms: counts, rank and groupings, an assumption test and regression run was completed for week 3 and for week 4. In other words, up to 12 runs were made for each of the 162 course samples. While it was relatively easy to set up, SPSS did not provide any easy way to extract the output results into a spreadsheet for comparison across courses. Other tools such as Python, Stata, or Mathematica should be considered.

Recommendations for Future Research

Binomial Logistic Regression analysis of the raw or directly processed data appears to have significant limitations. Several other approaches including Python Chefboost decision tree algorithms C4.5 and CART were used but rejected due to technical complexity and learning curve. Tang, Zing, and Pei (2019) used educational data mining. This and other technologies may provide solutions if configured and trained properly.

Another opportunity relates to treatment of outliers, particularly those associated with very low or no activity. Still, no or low activity from student inattentiveness needs to be treated differently from no or low activity due to course design. Transactional distance analysis might be appropriate.

Translating LMS activity into a score representing LMS activity is an additional approach which would avoid issues with outliers and courses not using Canvas. Some

calibration might be needed to develop a scoring methodology which provides significant value. A score could be developed for each student-course combination and then combined into a student score.

One of the challenges of this study was that there was little or no use by several courses. A quantitative study could be undertaken to understand LMS use in face-to-face courses. A qualitative study would help to understand the instructors' attitudes towards use of their LMS in face-to-face courses and perhaps suggest some ways to increase instructor use.

This study assumes a relationship between student LMS use and retention. It would be interesting to learn via a qualitative study whether there is any difference in attitudes towards LMS activity between students who left and those who stayed.

Conclusions

The objective of this study was to effect improvements in student retention by identifying students at higher academic risk based on activity recorded in Canvas Learning Management System (LMS) logs. Based in part on Tinto and Pusser (2006) theories, this study's underlying premise that academic engagement is related to retention and that academic engagement can be measured from Canvas LMS logs.

Canvas logs are not naturally in a format conducive to analysis. Log data needs to be transformed and aggregated by student, course, date, and other attributes. This study primarily used log record counts through selected early weeks of the semester to represent the independent variable academic engagement and used binomial logistic regression to determine if there was a statistical relationship to retention.

Instructors have wide latitude in determining the extent to which they use Canvas in their course. Some instructors use many features and some do not use Canvas at all. Also, use of Canvas features is determined by individual instructor course design (Gašević, Dawson, Rogers, & Gasevic, 2014). Ideally, instructors in all courses should actively engage with their students during the students' initial period with the institution (Tinto, 1975; Tinto & Prusser, 2006) and Learning Management Systems features can be used to easily track these interactions.

Currently, many instructors do not significantly incorporate LMS interactions in their courses, limiting the ability to discern levels of student engagement. In addition, there might not be any direct relationship between the number of log records and academic engagement by the student. As a result of both factors, it is necessary to review student activity at an individual course level and accept activity counts as an indication of activity despite acknowledgement of flaws.

Binomial Logistic Regression sample size constraints reduced the number of courses at week 3 that could be analyzed to 161. Of those, only 77 courses produced statistical results without any statistical warnings or casewise exceptions. Unfortunately, the 77 cases reported low or no statistical significance and ROC curve analysis disclosed poor discrimination.

Logistic regression issues with data outliers could possibly have been avoided if the log data was transformed differently. Still, the loose relationship between activity counts and academic engagement might still produce inconclusive results. Measuring actual deliverables such as assignments, discussion posts or quizzes would be an improvement over activity counts. Currently, these are not consistently or predictably

used by new students. However, if the design of core courses, which are required to be taken by all students, could be enhanced with meaningful deliverable activities, then quality data might be available to understand how new students are doing in specific courses (MacFadyen & Dawson, 2010; Lykourentzou, et al, 2009; Santana et al, 2015; You, 2016).

Tinto (2016) advises us that there are many reasons why students drop out. De Silva, Chounta, Rodriquez-Triana, Roa, Gramberg & Valk (2022) identified students' personal information; financial and professional status; academic background; and course engagement and motivation as indicators that retention prediction models use in addition to student engagement with LMSs and virtual learning environments. So there is a possibility that academic activity in isolation might not be a predictor of retention (Conijn, et al, 2017).

Final Thoughts

This has been a journey. In retrospect, I really wasn't prepared when I attended my first class in January 2019. The assignments, starting with an autoethnography and reflections on my kindergarten through high school education helped me appreciate how these experiences helped me in college. I appreciate more now how Brooklyn Technical High School prepared me for New York University by its size, variety of subjects, exposure to new technology and the confidence I gained. And many of the friends I made in college, I now consider family.

But it has taken me up until now to realize that research is my friend. Research provides a way to share information at a very detailed level as to what one did and learned. It communicates information and insight that can't be transmitted in a youTube

video. It provides knowledge that one can build on. While my study strives to improve retention, which everyone understands, I only discussed the details with my colleague, Eric Alvarado. In retrospect, I should have used the opportunity provided by online conferences to exchange more information, ask questions and seek help. My excuse is that I was intimidated because I didn't think I knew enough.

When I was teaching, it seemed obvious that some of my students were going to have problems because they were not checking my Canvas course and I wondered how these students were doing in their other courses. I realized that this inattention was probably due to other reasons: work, health, or family issues. But without a change with or without intervention, these students were not going to learn.

So I am frustrated that I could not identify similarly disengaged students from the LMS data. However, I am hopeful that I will eventually find a solution.

APPENDIX A IRB APPROVAL

Federal Wide Assurance: FWA00009066

Sep 12, 2023 9:48:56 AM EDT

PI: Roger So
CO-PI: James Campbell
Dept: Ed Admin & Instruc Leadership

Re: Initial - IRB-FY2023-64 Academic Risk Assessment Based on LMS Data

Dear Roger So:

The St John's University Institutional Review Board Institutional Review Board has approved your initial submission for Academic Risk Assessment Based on LMS Data. The approval is effective from September 12, 2023 through September 10, 2024.

The committee voted to approve your IRB application by a vote of 7-0-0. We regret that you downloaded the data and did get the analysis before you actually for IRB approval. Given our discussions, we know that you know this was a mistake and agree to follow the IRB regulation in all future research. Because you shared your data file with Dr. DiGiuseep, and it was clear from his review of your files that the data was de-identified, that no one could ascertain the identity of any subjects, and there, was no harm to any subjects, we approve the application.

Please make sure all future research is approved before proceeding with downloading data files, collecting data, or any other activities.

Sincerely,

Raymond DiGiuseppe, PhD, ABPP
Chair, Institutional Review Board
Professor of Psychology

APPENDIX B PERMISSION FOR USE OF INSTITUTIONAL DATA

Anne Pacione, Chief Information Officer
Phone: 718-990-2007
Email: roccoa@stjohns.edu
8000 Utopia Parkway
Queens, NY 11439

October 5, 2022

St. John's University Institutional Review Board
St. John's University
8000 Utopia Parkway
Queens, New York 11439

As St. John's University's Chief Information Officer, I grant Roger S So permission to use spring 2022 semester log data generated by the Canvas Learning Management system for his doctoral dissertation. This authorization is for data that has already been collected and does not include any personal identifiable information and will be in effect from this date until the end of June 2023.

Sincerely,

Anne Pacione

Anne Pacione

Chief Information Officer (CIO)

REFERENCES

Agresti, A. (2007). An introduction to categorical data analysis, second edition. John Wiley & Sons, Inc.

Ai, J., & Laffey, J. (2007). Web mining as a tool for understanding online learning. Merlot Journal of Online Learning and Teaching, 3(2), 160–169.

Alier, M., Casany, P., & Casado, P. (2007). A mobile extension of a web based moodle virtual classroom. In Proceedings of the E-challenges conference (pp. 11–26).

Arnold, K., & Pistill, M., (2013). CASE STUDY A: Traffic lights and interventions: Signals at Purdue University. Learning Analytics in Higher Education. Retrieved from https://analytics.jiscinvolve.org/wp/files/2016/04/CASE-STUDY-A-Purdue-University.pdf

Astin, A. (1999). Student involvement: A developmental theory for higher education. Journal of College Student Development, 40(5), 518–529.

Bean, J. (1985). Interaction effects based on class level in an explanatory model of college student dropout syndrome. American Educational Research Journal, 22(1), 35–64. doi:10.3102/00028312022001035.

Baer, L., & Campbell, J. P. (2012). From metrics to analytics, reporting to action: Analytics' role in changing the learning environment. In D. G. Oblinger (Ed.), Game Changers - Education and Information Technologies (pp. 53–65). EDUCAUSE (Retrieved from http://net.educause.edu/ir/library/pdf/pub7203.pdf).

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: Review and future visions. Journal of Educational Data Mining, 1(1), 3–17.

Barbera, E., Gros, B., & Kirschner, P. A. (2015). Paradox of time in research on educational technology. Time & Society, 24(1), 96–108.

Black, E. W., Dawson, K., & Priem, J. (2008). Data for free: Using LMS activity logs to measure community in online courses. Internet and Higher Education, 11, 65–70. doi:10.1016/j.iheduc.2008.03.002.

Brandl, K. (2005). Are you ready to "MOODLE"? Language Learning & Technology, 9(2), 16–23. Breier, M. (2010). From 'financial considerations' to 'poverty': Towards a reconceptualization of the role of finances in higher education student drop out. Higher Education, 60(6), 657–670. doi:10.1007/s10734- 010-9343-5.

Buckley, K., Fairman, K., Pogge, E., & Raney, E. (2022). Use of Learning Management System Data to Predict Student Success in a Pharmacy Capstone Course. 86(4).

Cabrera, A., Nora, A., & Casaneda, M. (1992). The role of finances in the persistence process: A structural model. Research in Higher Education, 33(5), 571–594. doi:10.1007/BF00973759.

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. EDUCAUSE Review, 42(4), 42–57.

Casany, M. J., Alier, M., Galanis, N., Mayol, E., & Piguillem, J. (2012). Analyzing Moodle/LMS logs to measure mobile access. ˆ, (c), 35–40

Chen, R. (2012). Institutional characteristics and college student dropout risks: A multilevel event history analysis. Research in Higher Education, 53(5), 487–505. doi:10.1007/s11162-011-9241-4.

Cheng, G., & Chau, J. (2016). Exploring the relationships between learning styles, online participation, learning achievement and course satisfaction: An empirical study of

a blended learning course. British Journal of Educational Technology, 47(2), 257–278.

Cheng, J., Kulkarni, C., & Klemmer, S. (2013). Tools for predicting drop-off in large online classes. In Proceedings of the computer supported cooperative work companion (pp. 121–124). doi:10.1145/ 2441955.2441987

Cohen, A., & Shimony, U. (2016). Dropout prediction in a massive open online course using learning analytics. E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, 2016(1), 616–625.

Cohen, A. (2017). Analysis of student activity in web-supported courses as a tool for predicting dropout. Education Technology Research Development, 65, 1285-1304.

Colvin, C., Wade, A., Dawson, S., Gasevic, D., Buckingham Shum, S., Nelson, K., … Fisher, J. (2015). Student retention and learning analytics : A snapshot of Australian practices and a framework for advancement. Draft final report. In Research.Ed.Ac.Uk. Retrieved from http://www.research.ed.ac.uk/portal/files/21121591/Final_Report_190615.pdf%5Cnhttp://he-analytics.com/wp-content/uploads/SP13-3249_-Master17Aug2015-web.pdf

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS. IEEE Transactions on Learning Technologies, 10(1), 17–29. doi:10.1109/TLT.2016.2616312

Davies, J., & Graff, M. (2005). Performance in e-learning: Online participation and student grades. British Journal of Educational Technology, 36(4), 657–663.

De Silva, L. M. H., Chounta, I.-A., Rodríguez-Triana, M. J., Roa, E. R., Gramberg, A., & Valk, A. (2022). Toward an Institutional Analytics Agenda for Addressing Student Dropout in Higher Education. Journal of Learning Analytics, 9(2), 179–201. https://doi.org/10.18608/jla.2022.7507

Dringus, L. P., & Ellis, T. (2010). Temporal transitions in participation flow in an asynchronous discussion forum. Computers & Education, 54(2), 340–349.

Essa, A., & Ayad, H. (2012). Student success system: Risk analytics and data visualization using ensembles of predictive models. ACM International Conference Proceeding Series, 158–161. https://doi.org/10.1145/2330601.2330641

Fritz, J. (2017). Using Analytics to Nudge Student Responsibility for Learning. New Directions for Higher Education, (161), 65–75. doi: 10.1002/he

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. Internet and Higher Education, 28, 68–84. https://doi.org/10.1016/j.iheduc.2015.10.002

Hew, K. F., & Cheung,W. S. (2008). Attracting student participation in asynchronous online discussions: A case study of peer facilitation. Computers & Education, 51(3), 1111–1124.

Hosmer, D. W. Jr., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression, 3rd Edition. John Wiley & Sons, Inc., Hoboken, New Jersey.

Hu, Y. H., Lo, C. L., & Shih, S. P. (2014). Developing early warning systems to predict students' online learning performance. Computers in Human Behavior, 36, 469–478. https://doi.org/10.1016/j.chb.2014.04.002

Hussar, B., Zhang, J., Hein, S., Wang, K., Roberts, A., Cui, J., Smith, M., Bullock Mann, F., Barmer, A., and Dilig, R. (2020). The Condition of Education 2020 (NCES 2020-144). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved [November 24, 2020] from ttps://nces.ed.gov/pubsearch/pubsinfo. asp?pubid=2020144.

Hwang Wu-Yuin & Wang Chin-Yu. (2004). A study of learning time patterns in asynchronous learning environments. Journal of Computer Assisted Learning, 20(4), 292–304. doi:10.1111/j.1365-2729.2004. 00088.x.

Ifenthaler, D. (2017). Are Higher Education Institutions Prepared for Learning Analytics? TechTrends, 61(4). doi:10.1007/s11528-016-0154-0

Jo, I. H., Kim, D., & Yoon, M. (2014). Analyzing the log patterns of adult learners in LMS using learning analytics. ACM International Conference Proceeding Series, 183–187. https://doi.org/10.1145/2567574.2567616

Kellen, V. (2019). 21st-Century Analytics: New Technologies and New Rules. Educause Review.

Knight, S., Wise, A. F., Chen, B., & Cheng, B. H. (2015, March). It's about time: 4th international workshop on temporal analyses of learning data. In Proceedings of the Fifth International Conference on Learning Analytics and Knowledge (pp. 388–389). New York, NY: ACM.

Laerd Statistics (2017). Binomial logistic regression using SPSS Statistics. Statistical tutorials and software guides. Retrieved from https://statistics.laerd.com/

Lee, Y., & Choi, J. (2011). A review of online course dropout research: Implications for practice and future research. Educational Technology Research and Development, 59(5), 593–618. doi:10.1007/s11423- 010-9177-y.

Levi-Gamlieli, H., Cohen, A., & Nachmias, R. (2015). Detection of overly intensive learning by using weblog of course website. Technology, Instruction, Cognition and Learning (TICL), 10(2), 151–171.

Levin, J., Barak, A., & Yaar, E. (1979). College dropout and some of its correlates. Megamot, 24(4), 564– 573 [Hebrew].

Levy, Y. (2007). Comparing dropouts and persistence in e-learning courses. Computers & Education, 48, 185–204. doi:10.1016/j.compedu.2004.12.004.

López-Zambrano, J., Lara, J. A., & Romero, C. (2020). Towards portability of models for predicting students' final performance in university courses starting from moodle logs. Applied Sciences (Switzerland), 10(1). https://doi.org/10.3390/app10010354

López-Zambrano, J., Lara, J. A., & Romero, C. (2021). Improving the portability of predicting students' performance models by using ontologies. Journal of Computing in Higher Education, 15. https://doi.org/10.1007/s12528-021-09273-3

Lu, J., Yu, C. S., & Liu, C. (2003). Learning style, learning patterns, and learning performance in a WebCT- based MIS course. Information & Management, 40(6), 497–507. doi:10.1016/S0378-7206(02)00064-2.

Lund, A. & Lund, M. (n.d.). Laerd Statistics. https://statistics.laerd.com.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. Computers & Education, 53, 950–965. doi:10.1016/j.compedu.2009.05.010.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. Computers & Education, 54(2), 588–599. http://dx.doi.org/10.1016/j.compedu.2009.09.008.

McFarland, J., Hussar, B., Zhang, J., Wang, X., Wang, K., Hein, S., Diliberti, M., Forrest Cataldi, E., Bullock Mann, F., and Barmer, A. (2019). The Condition of Education 2019 (NCES 2019-144). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved November 11, 2019 from https://nces.ed.gov/ pubsearch/pubsinfo.asp?pubid=2019144.

Me´ndez, G., Ochoa, X., & Chiluiza, K. (2014). Techniques for data-driven curriculum analysis. In Proceedings of the fourth international conference on learning analytics and knowledge (pp. 148–157). doi:10.1145/2567574.2567591.

Mercer, N. (2008). The seeds of time: Why classroom dialogue needs a temporal analysis. The Journal of the Learning Sciences, 17(1), 33–59.

Na, K. S., & Tasir, Z. (2018). Identifying at-risk students in online learning by analysing learning behaviour: A systematic review. 2017 IEEE Conference on Big Data and Analytics, ICBDA 2017, 2018-January. https://doi.org/10.1109/ICBDAA.2017.8284117

Nistor, N., & Neubauer, K. (2010). From participation to dropout: Quantitative participation patterns in online university courses. Computers & Education, 55, 663–672. doi:10.1016/j.compedu.2010.02.026.

Park, J.-H., & Choi, H. J. (2009). Factors influencing adult learners' decision to drop out or persist in online learning. Educational Technology & Society, 12(4), 207–217.

Parker, A. (2003). Identifying predictors of academic persistence in distance education. USDLA Journal, 17 (1), 55–62.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Reimann, P. (2009). Time is precious: Variable- and event-centred approaches to process analysis in CSCL research. International Journal of Computer-Supported Collaborative Learning, 4(3), 239–257.

Rhode, J., Richter, S., Gowen, P., Miller, T., & Wills, C. (2017). Understanding faculty use of the learning management system. Online Learning Journal, 21(3), 68,Äì86. doi:10.24059/olj.v%vi%i.1217

Rice, W. H. (2006). Moodle e-learning course development (3rd ed.). Birmingham, UK: Packt Publishing.

Romero, C., Lo´pez, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. Computers & Education, 68, 458–472. doi:10.1016/j. compedu.2013.06.009..

Romero, C., & Ventura, S. (2006). Data mining in E-learning. Southampton, UK: WIT Press. Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33(1), 135–146. doi:10.1016/j.eswa.2006.04.005.

Romero, C., Ventura, S., & Garcı´a, E. (2008). Data mining in course management systems: Moodle case study and tutorial. Computers & Education, 51(1), 368–384. doi:10.1016/j.compedu.2007.05.016.

Saqr, M., Fors, U., & Tedre, M. (2017). How learning analytics can early predict under-achieving students in a blended medical education course. Medical Teacher, 39(7), 757–767. https://doi.org/10.1080/0142159X.2017.1309376

Santana, M. A., Costa, E. B., Neto, B. F. D. S., Silva, I. C. L., & Rego, J. B. (2015). A predictive model for identifying students with dropout profiles in online courses. In Proceeding of the 8th international conference on educational data mining, EDM workshops.

Siemens, G., & Long, P. D. (2011). Penetrating the fog: Analytics in learning and education. Educause Review, 46(5), 31–40.

Stewart, B., Briton, D., Gismondi, M., Heller, B., Kennepohl, D., McGreal, R., et al. (2007). Choosing moodle: An evaluation of learning management systems at Athabasca University. Journal of Distance Education Technologies, 5(3), 1–7.

Tait, H., & Entwistle, N. (1996). Identifying students at risk through ineffective study strategies. Higher Education, 31(1), 97–116. doi:10.1007/BF00129109.

Tang, H., Xing, W., & Pei, B. (2019). Time really matters: understanding the temporal dimension of online learning using educational data mining. Journal of

Educational Computing Research, 57(5), 1326-1347. Doi: 10.1177/07356331 18784705

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. Review of Educational Research, 45(1), 89–125. doi:10.3102/00346543045001089.

Tinto, V. (1993). Leaving college: Rethinking the causes and cures of student attrition (2nd ed.). Chicago: University of Chicago Press.

Tinto, V. & Pusser, B. (2006). Moving from theory to action: building a model of institutional action for student success. National Postsecondary Education Cooperative, 57.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. London, England: Harvard University Press.

Wang, J., Doll, W. J., Deng, X., Park, K., & Yang, M. G. M. (2013). The impact of faculty perceived reconfigurability of learning management systems on effective teaching practices. Computers & Education, 61, 146–157. doi:10.1016/j.compedu.2012.09.005.

Wang, A. Y., & Newlin, M. H. (2002). Predictors of web-student performance: The role of self-efficacy and reasons for taking an on-line class. Computers in Human Behavior, 18(2), 151–163. doi:10.1016/ S0747-5632(01)00042-5.

Winne, P. H. (2006). How software technologies can improve research on learning and bolster school reform. Educational Psychologist, 41(1), 5–17. http://dx.doi.org/10. 1207/s15326985ep4101_3.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), Metacognition in educational theory and practice (pp. 277–304). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Wise, A. F., Zhao, Y., Hausknecht, S. N., & Chui, M. M. (2013). Temporal considerations in analyzing and designing for online discussions in education: Examining duration, sequence, pace and salience. In E. Barbera & P. Reimann (Eds.), Assessment and evaluation of time factors in online teaching and learning (pp. 198–231). Hershey, PA: IGI Global.

Xenos, M. (2004). Prediction and assessment of student behavior in open and distance education in computers using Bayesian networks. Computers & Education, 43(4), 345–359. doi:10.1016/j.compedu. 2003.09.005.

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. Computers in Human Behavior, 58, 119–129.

Yildirim, D., & Gülbahar, Y. (2022). Implementation of Learning Analytics Indicators for Increasing Learners' Final Performance. Technology, Knowledge and Learning, 27(2), 479–504. https://doi.org/10.1007/s10758-021-09583-6

Vita

Name	*Roger Sheng So*
Baccalaureate Degree	*Bachelor of Science, New York University, New York, Major: Computer Science, Accounting*
Date Graduated	*May, 1973*
Other Degrees and Certificates	*Master of Business Administration, New York University, New York, Major: Business*
Date Graduated	*May, 1994*