A Generalized Classifier to Identify Online Learning Tool Disengagement at Scale

Jacqueline Feild
McGraw-Hill Education
Boston, Massachusetts
jacqueline.feild@mheducation.com

Nicholas Lewkow
McGraw-Hill Education
Boston, Massachusetts
nicholas.lewkow@mheducation.com

Sean Burns
Colorado State University
Fort Collins, Colorado
sean.burns@colostate.edu

Karen Gebhardt
Colorado State University
Fort Collins, Colorado
karen.gebhardt@colostate.edu

ABSTRACT

Student success, a major focus in higher education, in part, requires students to remain actively engaged in the required coursework. Identifying student disengagement, when a student stops completing coursework, at scale has been a continuing challenge for higher education due to the heterogeneity of traditional college courses. This research uses data from Connect by McGraw-Hill Education, a widely used online learning tool, to build a classifier to identify learning tool disengagement at scale. This classifier was trained and tested on four years of historical data, representing 4.5 million students in 175,000 courses, across 256 disciplines. Results show that the classifier is effective in identifying disengagement within the online learning tool against baselines, across time, and within and across disciplines. The classifier was also effective in identifying students at risk of disengaging from Connect and then earning unsuccessful grades in a pilot course for which the assignments in Connect were worth a relatively small portion of the overall course grade. Because Connect is widely used, this classifier is positioned to be a good tool for instructors and institutions to identify students at risk for disengagement from coursework. Instructors and institutions can use this information to design and implement interventions to improve engagement and improve student success at the institution in key courses.

CCS CONCEPTS

• **Applied computing** → **E-learning**; • **Computing methodologies** → *Supervised learning by regression*; • **Information systems** → *Data mining*;

KEYWORDS

learning tool disengagement, higher education, economic education, logistic regression, student success, Apache Spark

LAK'18, March 7–9, 2018, Sydney, NSW, Australia

ACM ISBN 978-1-4503-6400-3/18/03…$15.00
https://doi.org/10.1145/3170358.3170370

ACM Reference Format:
Jacqueline Feild, Nicholas Lewkow, Sean Burns, and Karen Gebhardt. 2018. A Generalized Classifier to Identify Online Learning Tool Disengagement at Scale. In *LAK'18: International Conference on Learning Analytics and Knowledge, March 7–9, 2018, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3170358.3170370

1 INTRODUCTION

Retention and degree completion is a major focus of many higher education institutions. The Association of Public and Land Grant Universities (APLU) and the American Association of State Colleges and Universities (AASCU) along with nearly 500 public colleges and universities have pledged to increase graduation rates and reduce time to graduation in order to achieve the goal of having 60% of the working age population receive a college degree by 2025 [15]. Retention and degree completion is closely correlated with student course success [7]. One important component of course success requires students to remain actively engaged in the coursework throughout the duration of the course. If students become disengaged, they are less likely to pass the course. This can contribute to increasing of time to graduation, especially if the course is a core or gateway course [1].

Research on innovations for increasing student engagement typically focuses on redesigning course assignments, enhancing teaching methods, and improving institutional practices[13, 19, 22]. Other studies purport to evaluate the effects of these innovations on student success [12]. However, much existing research focuses on teaching practices or impacts, not the quantitative methodology to measure engagement and its relationship with course success. The goal of this research is to fill this gap. The measurement of engagement is essential for the timely identification of students who may not successfully complete a course.

From an institutional perspective, large class sizes make it difficult for faculty to identify students who are at risk of disengaging from coursework. One can look to massive open online courses (MOOCs) as a potential guide to measuring engagement at scale. Researchers have developed related methods to identify course disengagement in MOOCs based on student assignment submission and online course behavior [5, 9, 14, 16, 18, 20, 21, 23]. One of the strengths of MOOC research is that it is performed at scale, meaning there are thousands of students to analyze in a single course (i.e., the *course* is widely used). Although this research is robust, its

conclusions may not be generalizable to the *traditional* college class [2]. MOOCs have enrollments that can reach into the tens of thousands, few students intend to complete the course (e.g., only 22% of students who start a MOOC at HarvardX intend to complete [17]), these courses are not typically part of a degree program, and there is a short-term relationship between the faculty/institution and the student. This contrasts with the traditional college class where enrollments tend to be 150-300 students in a *large* class, students enroll with the intention to complete the course, the majority of classes taken are part of a degree program, and there is a long-term relationship between the faculty/institution and the student.

Since research conclusions from studies on MOOCs may not be generalizable to a traditional college course, a reliable and robust measurement of engagement in coursework is needed that can be used in a variety of college institutional settings, teaching modes (e.g., on campus, online, blended), class sizes, and disciplines. Developing such a tool is essential for reaching the stated goal of increasing graduation rates for the 20.4 million students who were expected to attend American colleges and universities in fall 2017 (a 33% increase in enrollment since fall 2000) [8]. Some research has studied disengagement behavior in traditional courses [3, 4, 10, 11], but these studies use data from a limited number of student enrollments or courses. In order to develop a measurement of engagement in traditional college courses at scale, methodologies developed from MOOC engagement research must be combined with a larger quantity of student behavior data from the traditional college class.

Fortunately, there are online learning tools used in traditional college classes that collect student behavior data and are used at scale. This research used data from Connect by McGraw-Hill Education, a widely used online learning tool, to draw data from 4.5 million students in traditional college classes across 256 disciplines. Instructors use Connect in a variety of ways: to provide access to the ebook, assign practice or graded low-stakes assignments, or function as the primary course website where students complete all assignments and readings, including high-stakes assignments. The large pool of student data has as much scale as MOOC analyses and because of these characteristics (the *tool* is widely used), the data available from Connect is ideal for building a disengagement classifier.

Using Connect's data and data from a traditional university, this research will show that (1) a generalized disengagement classifier was designed, trained, and tested on historical data from multiple disciplines that can successfully identify student disengagement from an online learning tool used in non-MOOC courses, and (2) this classifier produces useful predictions for a discipline-specific course. This analysis was accomplished at scale with four years of historical data from multiple disciplines using parallel computation with the Apache Spark framework. For the discipline-specific analysis, additional data was analyzed from two large sections of Principles of Microeconomics at Colorado State University.

In Section 2, the definition of online learning tool disengagement is provided as well as an analysis of general identification of disengagement in Connect. Section 3 describes the details of the data set, data cleaning methods and how the generalized disengagement classifier was built and optimized. Section 4 provides the experimental validation of the classifier with pilot discipline-specific Principles of Microeconomics sections. Finally, Section 5 concludes and discusses future work.

2 IDENTIFYING ONLINE LEARNING TOOL DISENGAGEMENT

Research related to student engagement in MOOCs shows that disengagement can be observed in these widely used courses. In this section, it is confirmed that disengagement is observable in Connect, which is a widely used online learning tool. The main purpose of Connect is delivering content and assessing students; it can be used for assigning homework and practice assignments, as well as quizzes and tests. Typically, these assignments are a portion of a student's total grade and that portion varies by course, instructor, discipline, and institution. Student behavior data is collected as assignments are attempted. For each assignment, Connect records the due date and type of assignment (e.g., homework, quiz), and for each assignment, the student start and submission date and time, time spent, attempt number, and grade. These data can be used to better understand disengagement in Connect.

This analysis uses the following rule to define disengagement:

A student who submits assignments through the online learning tool up until a time t and never submits another assignment has disengaged at point t.

In this general analysis to identify student disengagement in Connect, students who do not submit assignments during the last two weeks of the course are excluded. There may be a number of different reasons, positive or negative, to stop submitting work so close to the end of the semester (e.g., students have already achieved the grade they want).

To identify general disengagement patterns, four years of historical data from fall 2013 to spring 2017 were used. This four year period included data from 4.5 million students enrolled in 175,000 sections representing 256 academic disciplines (see section 3.1 for additional data set details). Overall, it is observed that 17.2% of students disengage from Connect before the end of the semester. These results show that the percent of students disengaging from Connect in these traditional college courses is much lower than the rates reported for MOOCs, but is still much higher than desired in higher education.

Disengagement varies by discipline. Figure 1 shows the top ten disciplines ranked by total number of student enrollments and the corresponding disengagement percent. For example, in economics courses, 14.6% of students disengage before the end of the course, but in chemistry, this rises to 18.7%.

The distribution of disengagement varies by the week of the semester. Figure 2 shows that while a sizable portion of the disengagement occurs in the first week of the semester, disengagement also occurs steadily throughout the semester. This spike in disengagement at the start of the semester could signify students finalizing their semester course schedule.

3 MODELING ONLINE LEARNING TOOL DISENGAGEMENT

The previous section confirmed that disengagement is observable in Connect. The high degree of disengagement suggested that a classifier could be developed to identify this disengagement. This section

Discipline	Students (#)	Disengagement (%)
Accounting	927,241	16.5
Anatomy & Physiology	335,181	17.6
Biology	326,462	16.8
Chemistry	316,155	18.7
Economics	296,060	14.6
Finance	249,678	11.2
Management	211,807	9.4
Psychology (Intro)	195,582	12.4
Marketing	192,547	9.2
Spanish (Intro)	160,777	21.1
All Disciplines	**4,691,768**	**17.2**

Figure 1: Percent disengagement from Connect for the top ten disciplines ranked by student enrollments and for all disciplines, fall 2013 - spring 2017.

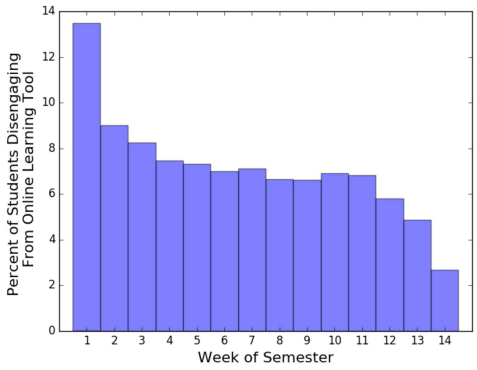


Figure 2: Histogram of student disengagement from Connect by week, fall 2013 - spring 2017.

describes the development of a generalized classifier for Connect disengagement. The data set, classifier features, and the classifier training and testing at scale with Apache Spark are described.

3.1 Historical Data Set

To train and test the disengagement classifier, four years of historical data from fall 2013 to spring 2017 was used. To avoid potential data issues related to small class size, low online learning tool usage, incorrect or unreliable assignment dates, or extremely compressed courses, several filtering steps were necessary to clean this historical data set. Data was filtered to eliminate sections with less than ten students or less than ten assignments to ensure data was used from sections with at least moderate usage. Next, assignments with unreliable date values in the due date fields were filtered out. These included assignments with no due date or incorrecrt due date values (e.g., a due date that is nonsensical, such as due in 2600). Last, the sections in the data set were limited to those with a duration between 8 and 16 weeks, which is a typical length for a higher education course during the fall, spring, or summer semesters. After cleaning, the data set included data from 4,691,768 students enrolled in 175,850 sections representing 256 academic disciplines.

3.2 Generalized Disengagement Classifier

The research goal was to build a time-invariant classifier so that one classifier could be used to predict student disengagement from Connect at any time during the semester. Logistic regression was used because it is an easy algorithm to interpret, provides a floating point estimate between 0 and 1 instead of a hard classification label, and allows for the interpretation of student behavior or feature importance.

A variety of student features can be derived from the data captured by Connect which necessitated identification and testing of the most important features. Previous work on identifying disengagement in MOOCs used time invariant features (e.g. percent of assignments completed instead of number of assignments completed) to build a classifier that could be used at any point in the semester [21]. For this classifier, all features were also time-invariant for the same reason.

Features were hand selected by the researchers to identify learner attributes that were different between engaged students and disengaged students. Features were used in the model if the distribution of values were significantly different between disengaged and engaged students. Ten features were identified and all permutations of this set were tested to ensure that inclusion of each feature improved the performance of the classifier. These ten features are:

(1) Average grade
(2) Average time spent on assignments relative to class average
(3) Percent assignments submitted
(4) Percent assignments submitted late
(5) Percent assignments submitted on time
(6) Percent assignments not submitted
(7) Percent assignments submitted late or not at all
(8) Days since last submission
(9) Submission time relative to due date
(10) Maximum number of consecutive late assignments

Although these features capture essential student behavior in Connect, feature values may fall into a different range depending on the section, course, instructor, discipline, or institution. For example an average grade of 0.9 may be 'good' in one section, while an average grade of 0.7 may be 'good' for another. Without specific information related to grading and assignments for each section, it is not possible to evaluate grade outcomes across sections. To alleviate this problem, the raw student feature values were transformed into z-scores providing values that are relative to each section. Z-scores are calculated by subtracting the section mean from each feature value and then dividing by the section standard deviation. The result is that each transformed feature value represents the number of standard deviations a student is from the mean for that feature within that section. These two sets of features were concatenated for twenty total features.

The classifier generates a score representing the likelihood of future disengagement from Connect. For this classifier, the closer the classifier score is to 1, the more likely the student is to disengage from Connect. For example, if Student A has missed three assignments in a row and Student B has completed all assignments, the classification score for Student A will be greater than for Student B.

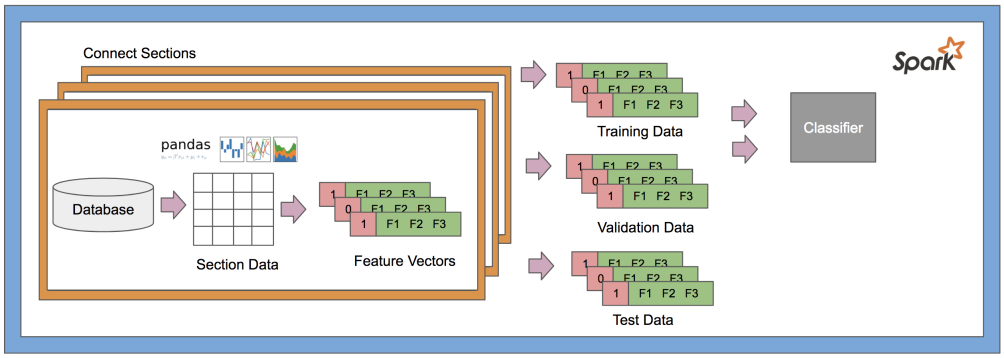


Figure 3: Feature vector and label generation pipeline using Apache Spark. Student data were queried from Connect database in parallel by section. Feature vectors were then generated and then labeled. Labeled feature vectors by student were assigned into training, validation, and test sets. The training and validation sets were used to build and optimize the classifier.

3.3 Building the Training, Validation, and Testing Data Sets

Since logistic regression is a supervised learning algorithm, the model is trained with a series of labelled feature vectors. The historical data set was used to build training, validation and test data sets spans the entire fall, spring, and summer semesters. For each student in every section, a feature vector and disengagement label was created for each week of the semester. Each feature vector was labelled according to whether the student has disengaged from Connect at the current time step as measured by weeks. If the student disengages at that time step or had disengaged at any previous time step, that time step label is 1, otherwise it is 0. To accurately simulate how the classifier would be used in practice, the feature vectors use data from the start of the semester up to the time of labeling. For example, the first week's feature values and disengagement label were calculated using the data from week 1, the second week's feature value and label were calculated using the data from weeks 1 and 2, and so on. Features and labels were generated for all students in the historical data set at the section level to create a weekly student feature vector with label. These student feature vectors with labels were then recombined to pool all students in all sections and then split randomly by student such that 60% of the students were assigned the training set, 20% were assigned the validation set, and the 20% were assigned the testing set. This typical process ensures that all feature vectors for a given student reside in a single data set, eliminating the possibility of training and testing on data from the same student.

3.4 Optimizing the Classifier

Most machine learning models have external model parameters, called hyperparameters, that require tuning or optimization to obtain the best possible classifier. The logistic regression model used for the classifier was optimized for two hyperparameters, an elastic net blending parameter and a regularization parameter. The regularization parameter determines the strength of a penalty applied to high classifier weights and is a commonly used logistic regression hyperparameter that helps prevent overfitting. Regularization comes in two forms: L1 which is robust but does not guarantee a single solution and L2 which is less robust but guarantees a single solution. The elastic net blending parameter, which ranges between 0 and 1, determines how much L1 vs L2 regularization to use. With a value of 0, the elastic net blending parameter uses L1 regularization exclusively while a value of 1 uses L2 exclusively. A linear combination of both L1 and L2 occurs for all other values of the elastic net blending parameter.

To determine the best hyperparameter to use, a grid search was run over both hyperparameters using the validation data set and optimizing on the area under the precision-recall curve (AUC-PR). The AUC-PR was used for an optimization parameter since it is more robust to a skewed set of labels than other metrics [6]. Optimizing on the f1 parameter and area under the receiver operating characteristic curve (AUC-ROC) were tested, both of which resulted in models not significantly different from the one found using the AUC-PR for optimization. The elastic net blending parameter was tested on a grid of 0, 0.25, 0.5, 0.75 and one with the optimal parameter being 0.75. The regularization parameter was tested on a grid from 1e-6 to 1e-2 by factors of ten with the optimal parameter being 1e-6. Together, these hyperparameters optimize the model, allowing for higher accuracy.

3.5 Parallelization with Apache Spark

The Apache Spark parallelization framework, which is open source and designed for large scale data analytics, was used to compute feature vectors and disengagement labels for the training, validation, and test sets. Apache Spark uses the basic concept of a master computer, or node, distributing work in parallel to worker nodes and then collecting the results from the workers nodes back to the master node. This parallel computing was needed due to the massive size of the data set. For example, generating features sets and labels for the historical data set would require approximately 20 days running on a single CPU. Figure 3 shows the pipeline used to parallelize the process. For each course section, the Connect database was queried and individual student behavior data was brought into a Python Pandas DataFrame. A feature vector and disengagement label was then computed for each student by time step and these vectors were randomly assigned, splitting by student, into training, validation, and test sets as described in Section 3.3. Apache Spark was used to do this for each section in parallel on

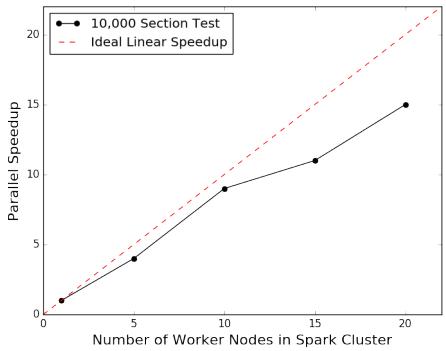


Figure 4: Parallel speedup achieved using Apache Spark to generate student feature vectors. Each worker node consisted of 4 CPUs and 30.5GB of memory.

different worker nodes since the feature and label computations are organized by individual sections.

Figure 4 shows the speedup resulting from use of a singular worker node up to 20 worker nodes. Parallel speedup is defined as

$$S_n = \frac{t_1}{t_n}, \tag{1}$$

where S_n is the parallel speedup, t_1 is the time for the computation on 1 worker node and t_n is the time for computation on n worker nodes. This calculation can be thought of as the time saved on a computation by using n workers in parallel.

Figure 4 shows this parallelization process created a speedup of 15x when using 20 worker nodes. This measured speedup can be compared to linear speedup which is the theoretical best speedup that you can achieve in which doubling the number of worker nodes reduces the computation time by half. The reason linear speedup of 20x was not achieved was due to the computation being limited by the uneven number of section weeks and number of students between sections. If every section had the same number of semester weeks and students, exact linear speedup would be expected. Despite the limitations, the speedup was essential for this analysis.

4 EXPERIMENTAL VALIDATION AND ANALYSIS

4.1 Historical Data Set Analysis

The classifier was evaluated with the historical data set described in Section 3 using several standard evaluation methods described next. These analyses were to test whether the classifier correctly scored students who already disengaged or were about to disengage.

4.1.1 Metrics Against Baselines. To evaluate the performance of the disengagement classifier, the area under the curve of the receiving operating characteristic curve (AUC-ROC) was calculated. This curve plots the true positive rate against the false positive rate and the area under this curve is a value with a standard interpretation in machine learning [6]. Additionally, the area under the curve of the precision-recall curve (AUC-PR) was calculated. This curve is a plot of precision against recall.

	AUC-ROC	AUC-PR
Baseline: Most Frequent	0.50	0.54
Baseline: Stratified	0.50	0.12
Baseline: 75 Percent Submitted	0.80	0.55
Generalized Disengagement Classifier	**0.94**	**0.71**

Figure 5: Baseline models compared to the trained classifier model.

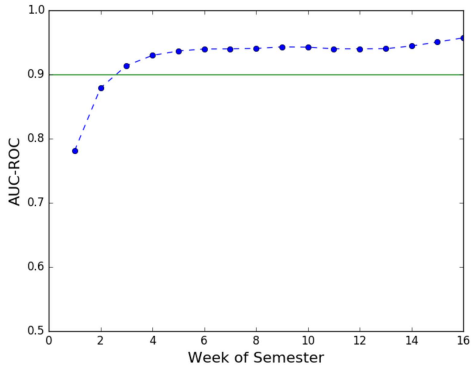


Figure 6: AUC-ROC values for the classifier by week. The horizontal line at 0.9 signifies the AUC-ROC threshold considered 'excellent'.

To determine if the classifier is a superior model, it was compared to three different baseline models and the AUC-ROC and AUC-PR were calculated for each of those models. The baseline models consisted of a 'most frequent' model, a 'stratified' model and 'percent submitted' model. The most frequent model simply assigns the most common label in the data as a prediction for all feature vectors. In the historical data set, 91% feature vectors are 0 labels and 9% are 1 labels. This means the most frequent model assigns every feature vector a 0 label. The stratified model randomly assigns a label to a feature vector based on the labels in the historical data set. In this model 91% of the feature vectors are assigned a 0 label, and the remaining 9% of the feature vectors are assigned a 1 label. The percent submitted baseline assigns a feature vector a 0 label if the student has submitted at least 75% of their assignments, otherwise it assigns a 1 label. All baseline models and the trained classifier are shown in Figure 5 along with their respective AUC-ROC and AUC-PR values. The results show that the trained classifier outperformed all three baseline models with AUC-ROC of 0.94 and AUC-PR of 0.71.

4.1.2 Metrics Over Time. Student behavior changes throughout a semester; therefore, it is important to evaluate the classifier as a semester progresses. To do so, the AUC-ROC was calculated separately for each week. Figure 6 shows the AUC-ROC for the classifier as a function of week in the semester. Across all sections in the data set, at the start of the semester the AUC-ROC value is 0.78. This value increases as additional student behavior becomes available, rising above 0.9, or into the 'excellent' category, by week three. The classifier continues to improve until week six, holds steady through

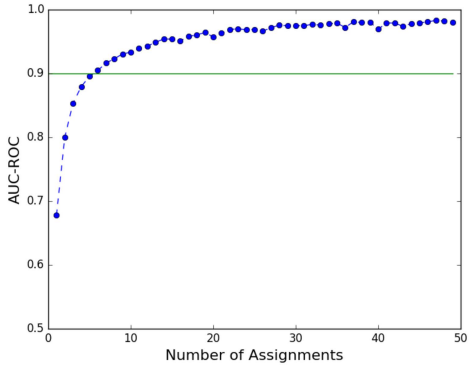


Figure 7: AUC-ROC values for the classifier by number of assignments submitted. The horizontal line at 0.9 signifies the AUC-ROC threshold considered 'excellent.'

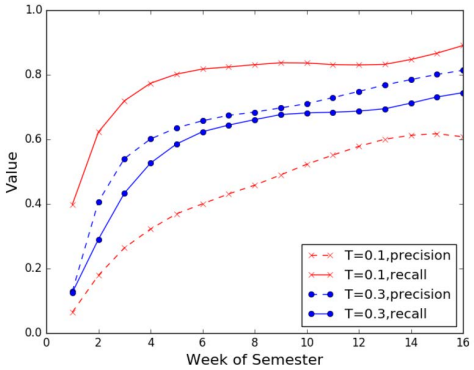


Figure 8: Precision and recall values for the classifier for a threshold of 0.1 and 0.3 by week.

week thirteen, and then improves slightly at the end of the semester. Figure 7 shows a slightly different view of this analysis where the AUC-ROC value is plotted against the cumulative number of assignments. This shows that the disengagement classifier has an AUC-ROC value above 0.9 by the fifth assignment, and it continues to increase as students complete more assignments and additional behavior data is collected.

Precision and recall metrics were also analyzed by week during the semester, because they evaluate performance of the classifier differently than the AUC-ROC in this case. Precision and recall are calculated at a given threshold, and Figure 8 shows two different thresholds of 0.1 and 0.3. There is a trade-off between precision and recall that becomes evident when comparing the threshold values. The values at a threshold of 0.1 show fairly high recall by the fifth week (0.8) but low precision (0.3). The values at a threshold of 0.3 show less of a trade-off, with both precision and recall around 0.6 by the fifth week. These results show through the evaluation by several metrics that the disengagement classifier reaches high levels of confidence relatively early in the semester.

4.1.3 Metrics Across Disciplines. Student behavior varies across courses and disciplines; therefore, it is important to evaluate how well the classifier generalizes within and across disciplines. To do this, two different types of experiments were conducted and the outcomes compared. The first experiment evaluated how well a classifier performed with training and test sets from a single discipline (i.e., within a discipline), and the second experiment evaluated how well a classifier performed when trained on one set of disciplines and tested on a different set of disciplines (i.e., across disciplines).

		Exp 1	Exp 2
Accounting	AUC-ROC	0.95	0.95
	AUC-PR	0.76	0.75
A & P	AUC-ROC	0.93	0.93
	AUC-PR	0.70	0.69
Biology	AUC-ROC	0.94	0.94
	AUC-PR	0.73	0.73
Chemistry	AUC-ROC	0.94	0.94
	AUC-PR	0.72	0.72
Economics	AUC-ROC	0.94	0.94
	AUC-PR	0.71	0.71

Figure 9: Exp. 1 trained and tested the classifier on labeled feature vectors within the same discipline. Exp. 2 trained the classifier on labeled feature vectors from 251 other disciplines, and tested on data exclusively from each discipline in Exp. 1.

For experiment 1, all student feature vectors from within a single discipline were split by student into training and test sets and a model was trained, optimized, and evaluated. This experiment was conducted for the top five disciplines as measured by student enrollment (see Figure 1). The results of this experiment are shown in Figure 9 in the column labelled 'Exp 1.' These results show that the classifier performed well within disciplines with an AUC-ROC between 0.93-0.95, in line with the results when the classifier was trained and tested on the entire historical data set (see Figure 5).

To evaluate if the classifier can generalize across disciplines, a second experiment was conducted in which the classifier was trained using the 251 lower enrollment disciplines and then the classifier was tested on each of the top five disciplines used in the first experiment. This second experiment is important because no feature vectors from the disciplines in this training set were included in the test set. The results from this experiment are shown in the column labelled 'Exp 2' in Figure 9. A classifier that can generalize across disciplines should perform similarly in experiment 1 and experiment 2. Results show that the AUC-ROC and AUC-PR values are nearly identical between these two experiments. These results show that the disengagement classifier generalizes well across disciplines, correctly identifying students who disengage from Connect.

4.2 Enhanced Data Set Analysis

The results from Section 4.1 show that the classifier is successful in identifying disengagement from Connect on the historical data set using a variety of metrics. But many courses that use this online learning tool also use a variety of additional assessments administered through other online learning tools or in the classroom. To determine if the disengagement classifier is generalizable to a

traditional college course, demographic and course performance data from pilot sections of Principles of Microeconomics taught at Colorado State University were combined with the historical data set.

Principles of Microeconomics at Colorado State University is a freshman or sophomore level introductory economics course. It is a university core curriculum course and is included as a required class in more than 40 campus majors. The course is taught in a traditional on-campus (i.e., residential) setting in a lecture-recitation format where students attend two class sessions in a large lecture (180-270 students) and one class session in a smaller recitation section (30 students). Approximately 20% of the final course grade in these pilot sections were associated with assignments completed once or twice per week in Connect for most weeks of the semester. The remaining assignments were writing assignments, in-class exams, and participation. The fall 2016 and spring 2017 pilot sections enrolled 814 students, of which 810 completed assignments in Connect. Student demographic data such as GPA at the start of term and end of term, high school GPA, Principles of Microeconomics course grade, and course and university withdrawal status were gathered after the end of the semester, linked with the historical data, and then de-identified to build an enhanced data set.

For this pilot section analysis, two thresholds were identified which divided the classification score into three risk categories; low risk, moderate risk, and high risk. The thresholds were chosen based on reviewing the precision recall curves and determining the acceptable values for the different risk categories. The low-moderate classification threshold was identified at 0.1 and the moderate-high threshold was identified at 0.3. These thresholds were experimental and can be altered in the future. The research goal in this section is to evaluate how well the classifier scored each student who disengaged from Connect during the semester and how this disengagement classification related to course outcomes and other student characteristics.

4.2.1 Comparison of Metrics between Pilot Sections and Historical Data. Using the same standard evaluation metrics used in the historical data analysis in section 4.1.1, the classifier identified disengagement for these pilot sections well. The evaluation metrics for the pilot sections resulted in an AUC-ROC value of 0.94 and an AUC-PR value of 0.70, both of which are nearly identical to the trained classifier. This suggests that these pilot sections are representative of the sections in the historical data set; the classifier identified student disengagement from Connect for these sections as well as it did for the historical sections.

4.2.2 Analysis of Students Who are Most 'At-Risk' for Disengagement in Pilot Sections. There are subpopulations of students who are most at-risk of course disengagement. These 'at-risk' students ultimately earned a low final grade (D or F) or withdrew from either the course or university.

Figure 10 shows the risk category (i.e., low, moderate, or high) for each of these groups of students during each week of the semester. White squares represent low risk, light grey squares represent moderate risk and black squares represent high risk. The first week at which each student is disengaged from Connect is marked with an 'x.'

Figure 10a shows the students who disengaged from Connect and then earned an D in the course. Of the 13 students in this group (Student 1-13), the disengagement classifier identified 12 as having at least moderate risk for weeks leading up to or on the week of actual disengagement from Connect. Of those 12, the classifier identified nine as high risk at some point before disengagement. The student that was not identified by the classifier prior to disengagement was identified as moderate risk two weeks after disengaging.

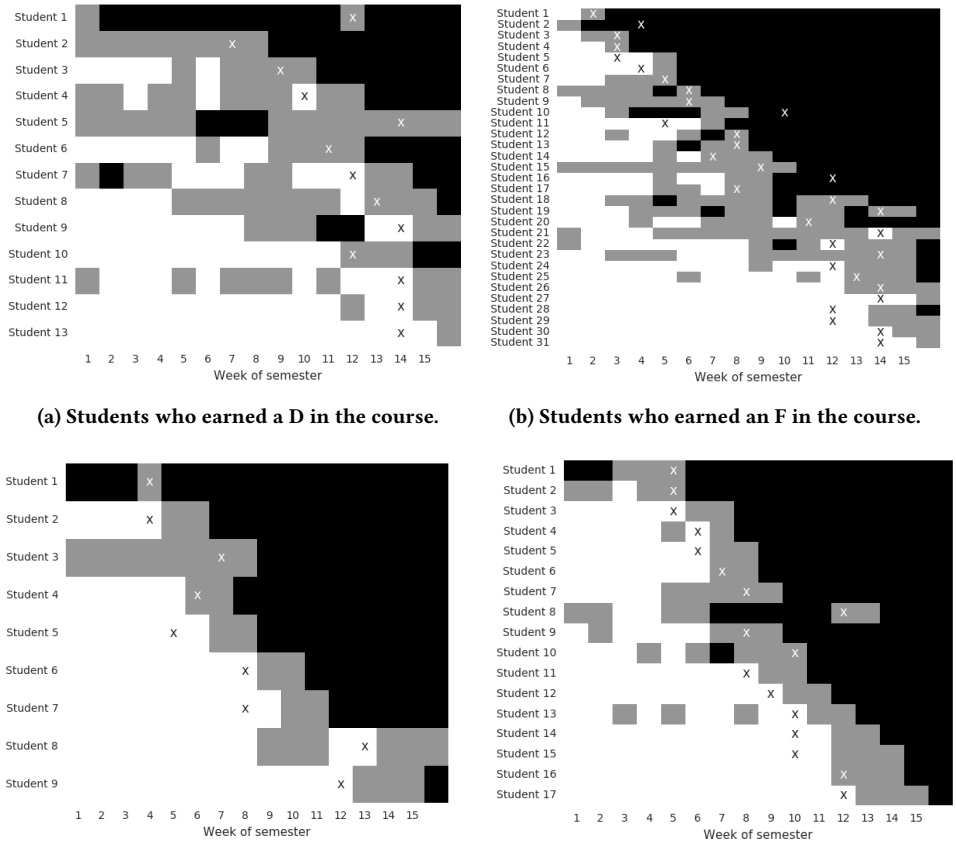
Figure 10b shows the students who disengaged from Connect and then earned an F in the course. Of the 31 students in this group (Student 1-31), the disengagement classifier identified 23 as having at least moderate risk for weeks leading up to or on the week of actual disengagement from Connect. Of those 23, the classifier identified nine as high risk at some point before disengagement. Of the eight students that were not identified by the classifier prior to disengagement, two were identified as moderate risk one week after disengaging, and six were identified as moderate risk two weeks after.

Figure 10c shows the students who disengaged from Connect and then dropped the course sometime between week 2 and week 15. These students did not earn a course grade, but instead, a W was transcribed on their transcript for the course. In this case, the classifier identified four of the nine students before they disengaged from Connect. Of the five students that were not identified by the classifier prior to disengagement, three were labelled as moderate risk one week after disengaging and two were labelled as moderate risk two weeks after disengaging.

Figure 10d shows the students who disengaged from Connect and then withdrew from the university. Students who withdraw from the university often have a traumatic event (e.g., death of a parent or sudden and difficult illness) that precipitates the withdrawal. The classifier successfully identified ten of the seventeen students before they disengaged from the course. Of the seven students that were not identified prior to disengagement, five were identified as having moderate risk within one week, and two were identified within two weeks.

Over all four groups, this classifier identified 70% (49/70) of students before or on the week they disengaged from Connect and identified all students within two weeks of disengaging.

4.2.3 Pilot Sections Course Grade Analysis. Students' average classification scores as a function of week in the semester were analyzed. Students were grouped by their final course grade as strong pass (i.e., A or B), weak pass (i.e., C or D), or students who fail or withdraw from the course or university (i.e., F/drop/withdraw). This was an important analysis for the pilot sections because of the institutional setting. At Colorado State University, typically half of the students taking the course are planning to apply to the business school. To be accepted into the business school, students must earn a B- or better. Additionally, for many other students at the university, earning a C or even a D means they can continue with their majors. Therefore, the distinction between strong and weak pass is an important for these pilot sections. Figure 11 shows that the strong pass students are on average in the low risk category for the entire semester. Weak pass students are also in the low risk category for the entire semester but have slightly higher classification scores on average as compared to the strong pass students. The classification



(a) Students who earned a D in the course.

(b) Students who earned an F in the course.

(c) Students who withdrew from the course.

(d) Students who withdrew from the university.

Figure 10: Heatmaps identifying students in the pilot who disengage from Connect and then earn a D or F, or withdrew from course or university. White squares represent weeks with low risk predictions, gray squares represent moderate risk, and black squares represent high risk. The week a student disengages is marked with an 'x.'

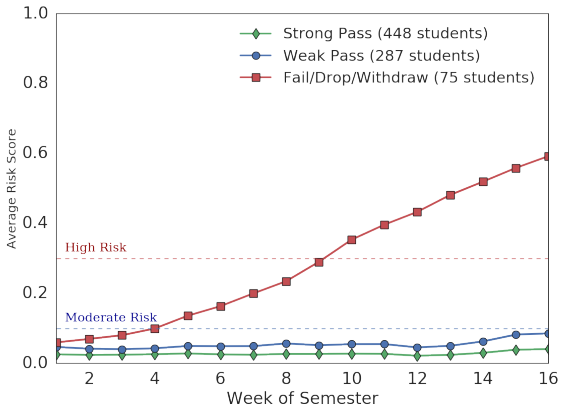


Figure 11: Average risk score by week by student group. Dashed lines show the thresholds between low-moderate risk and moderate-high risk.

score for the weak pass students approached the moderate risk category towards the end of the semester. In stark contrast, the F/drop/withdraw students' average classification score crosses into the moderate risk category on week four and into the high risk category on week ten of the semester.

Another way to analyze these students it to use box plots showing median classification scores, with outliers, as a function of week in the semester (see Figure 12). The students in this analysis are again grouped by their final grades as strong pass, weak pass and F/drop/withdraw. Results show that while the median classification score for strong and weak pass students lie within the low risk category throughout the semester, both groups of students have outliers which reside in the moderate and high risk categories for all weeks in the semester. The F/drop/withdraw students in Figure 12 look much different than the strong and weak pass students starting around week four and throughout the rest of the semester.

Finally, the number of weeks that the students spent in either moderate or high risk categories was analyzed. Figure 13 shows how many weeks all of the pilot students spent in either the moderate or high risk categories. The students in Figure 13 are again grouped by final course grades as strong pass, weak pass, and F/drop/withdraw. This shows that strong pass and weak pass students spend much less time in moderate or high risk categories. The differences between

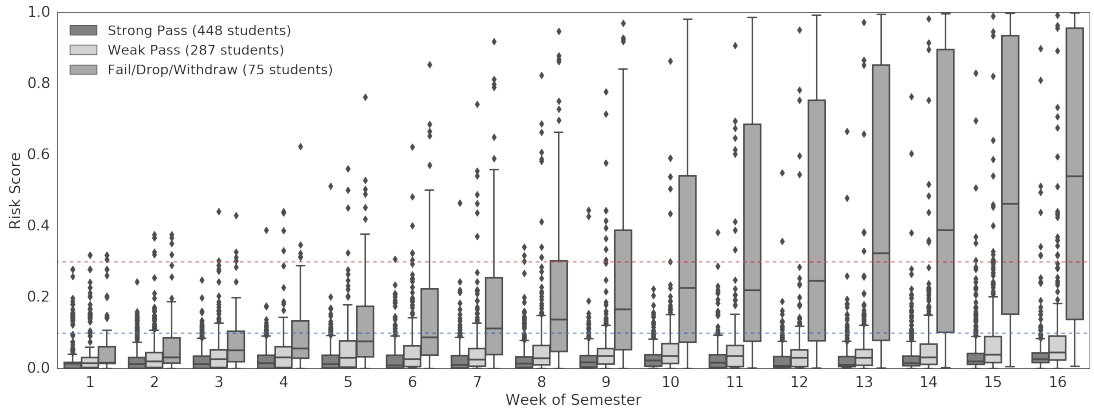


Figure 12: Box plot for risk scores by week for student groups. Dashed lines show the thresholds between low-moderate risk, and moderate-high risk.

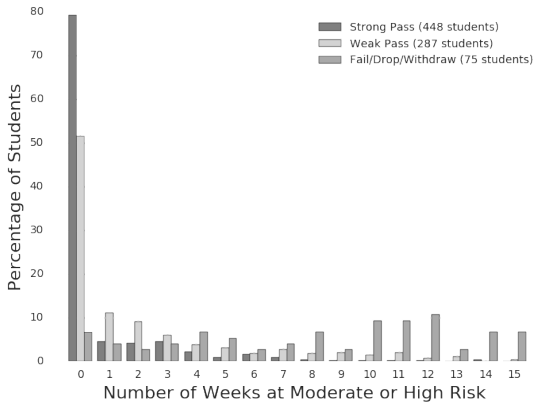


Figure 13: Histogram which identifies the number of weeks that students were labelled as moderate or high risk by student group.

the groups is most obvious when looking at how many students spend zero weeks in either moderate or high risk categories. A total of 79% of strong pass students, 51% of weak pass students, and 7% of F/drop/withdraw students spent the entire semester in the low risk category. The remaining 93% of the F/drop/withdraw students spent at least 1 week in the moderate or high risk categories with about 40% of these students spending ten or more weeks in moderate or high risk categories.

The course grade analyses show the classifier consistently identifies distinct patterns for students most at-risk for disengaging from Connect. Additionally, results show that this disengagement classification score, when interpreted as a risk category, can better differentiate F/drop/withdraw students from the passing students much more than differentiating between strong and weak pass students.

4.2.4 Pilot Sections GPA Analysis. Last, we investigated how the disengagement classifications correlated with GPA data for the students in the pilot sections. A singular disengagement score was created for each student by averaging the weekly disengagement scores over all weeks in the semester. Then, the Pearson's *r* correlations for the averaged disengagement scores were compared with the students' GPAs from high school, GPA from the previous semester, GPA at the end of the semester, as well as the grade earned in Principles of Microeconomics. Figure 14 shows the Pearson's *r* and P values for each of the correlations. Results show that all correlations are negative and statistically significant as shown by a P value less than 0.05, meaning lower disengagement classification scores are correlated with higher GPAs and course grade. When analyzing the correlations between specific variables, results show that high school GPA is the least correlated with Connect disengagement, followed by start of semester GPA, and last by course grade. The current term GPA had the highest correlation. Results indicate that disengagement classifications correlated with GPA data, which is encouraging since the classifier was trained and tested purely on learning tool behavior data for a single course.

Correlation	Pearson's *r*	P value
High School GPA	-0.10	0.01
Start of Semester GPA	-0.28	0.00
End of Semester GPA	-0.57	0.00
Course Grade	-0.47	0.00

Figure 14: Pearson's *r* correlations between student risk scores and GPAs for high school, start and end of semester, and course grade.

5 CONCLUSIONS AND FUTURE WORK

Identifying student disengagement at scale has been a persistent challenge for higher education. The heterogeneity observed in traditional college courses between sections, instructors, disciplines, and institutional structure, means that a robust and generalizable classifier to identify disengagement was difficult to design. This research shows that a classifier was developed to identify disengagement at scale. This classifier was trained and tested on historical

data from Connect, a widely used online learning tool, and was shown to be effective in identifying disengagement within Connect across various metrics (against baselines, across time, and across disciplines). Additionally, the classifier risk score was shown to be related to course disengagement in a pilot course where the assignments in Connect were worth a relatively small portion of the overall course grade. These results are robust despite lack of knowledge of the weight of Connect assignments within the course or specifically how Connect is used throughout the semester.

This research has shown that the disengagement classifier performs better as more data is analyzed. In the first week of the semester, the AUC-ROC value is below 0.8, by the third week in the semester the AUC-ROC value is above 0.9 and by the fifth week it is 0.94. A limitation of this research is that this classifier is not as accurate early in the semester because of the relative lack of student behavior data in Connect during the first several weeks of class. Incorporating institutional data (e.g., student start of semester GPA) and behavior and other assignment grade data captured by the course learning management system into the training of the disengagement classifier could improve early identification of disengagement and enhance the classifier overall.

This research showed that the disengagement classifier identified all students who disengaged from Connect and then earned an F or withdrew from the course or university before or within two weeks of actual disengagement. This is important because the classifier was trained and tested only on data observed in Connect. A potential limitation of this research is that a student may be classified as low risk because the student remained engaged in Connect but then are F/drop/withdraw because of poor performance on the other required course assignments. For example, 16 students earned a course grade of F but remained engaged in Connect. For these students, the classifier labeled 11 as moderate or high risk at some point in the semester. Overall, the classifier labeled 89% of students failing the course and 93% of all F/drop/withdraw students as moderate or high risk at some point during the semester. These outliers can be observed in the pilot course in Figure 12. This is a limitation, but there may be a solution. The inclusion of assignment grades completed outside of Connect and behavior data captured by the learning management system into the training of the disengagement classifier could provide a means to compare Connect disengagement with overall course disengagement.

Despite these limitations, due to the wide use of Connect, this classifier is uniquely positioned to be a good tool for instructors and institutions to use to help identify students at risk for disengagement. In particular, the pilot course analysis showed that the disengagement classifier provides useful information to the instructor. Instructors and institutions can then use this information to design and implement interventions to improve engagement and student success.

ACKNOWLEDGMENTS

This research would not have been possible without the support of CSU Online Research & Analytics, CSU's VP for IT Patrick Burns, Gwen Gorzelsky, Director of The Institute for Learning and Teaching, for work on the APLU Grant, and MHE VP of Research & Analytics Alfred Essa for support of these collaborative projects.

REFERENCES

[1] Clifford Adelman. 2006. The toolbox revisited: Paths to degree completion from high school through college. *US Department of Education* (2006).
[2] Truong-Sinh An, Christopher Krauss, and Agathe Merceron. 2017. Can Typical Behaviors Identified in MOOCs be Discovered in Other Courses?. In *Proceedings of The Tenth International Conference on Educational data Mining (EDM 2017)*.
[3] Jay Bainbridge, James Melitski, Anne Zahradnik, Eitel JM Lauría, Sandeep Jayaprakash, and Josh Baron. 2015. Using learning analytics to predict at-risk students in online graduate public affairs and administration education. *Journal of Public Affairs Education* (2015), 247–262.
[4] Ryan S Baker, David Lindrum, Mary Jane Lindrum, and David Perkowski. 2015. Analyzing Early At-Risk Factors in Higher Education E-Learning Courses. *International Educational Data Mining Society* (2015).
[5] Girish Balakrishnan and Derrick Coetzee. 2013. Predicting student retention in massive open online courses using hidden markov models. *Electrical Engineering and Computer Sciences University of California at Berkeley* (2013).
[6] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 233–240.
[7] Stephen L DesJardins, Dennis A Ahlburg, and Brian P McCall. 2002. A temporal investigation of factors related to timely degree completion. *The Journal of Higher Education* 73, 5 (2002), 555–581.
[8] National Center for Education Statistics. 2017. Enrollment in elementary, secondary, and degree-granting postsecondary institutions, by level and control of institution, enrollment level, and attendance status and sex of student: Selected years, fall 1990 through fall 2026. (Sept. 2017). Retrieved September 8, 2017 from https://nces.ed.gov/programs/digest/d16/tables/dt16_105.20.asp?current=yes
[9] Sherif Halawa, Daniel Greene, and John Mitchell. 2014. Dropout prediction in MOOCs using learner activity features. *Experiences and Best Practices in and around MOOCs* 7 (2014), 3–12.
[10] Shaobo Huang and Ning Fang. 2013. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education* 61 (2013), 133–145.
[11] RR Kabra and RS Bichkar. 2011. Performance prediction of engineering students using decision trees. *International Journal of Computer Applications* (2011).
[12] George D Kuh, Ty M Cruce, Rick Shoup, Jillian Kinzie, and Robert M Gonyea. 2008. Unmasking the effects of student engagement on first-year college grades and persistence. *The Journal of Higher Education* 79, 5 (2008), 540–563.
[13] George D Kuh, Jillian Kinzie, Ty Cruce, Rick Shoup, and Robert M Gonyea. 2006. Connecting the dots: Multi-faceted analyses of the relationships between student engagement results from the NSSE, and the institutional practices and conditions that foster student success. *Indiana University, Bloomington* 547556 (2006).
[14] Bote Lorenzo, L Miguel, and Eduardo Gómez Sánchez. 2017. Predicting the decrease of engagement indicators in a MOOC. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM.
[15] The Association of Public and Land grant Universities. 2017. Project Degree Completion. (Sept. 2017). Retrieved September 8, 2017 from http://www.aplu.org/projects-and-initiatives/project-degree-completion/project-degree-completion-in-depth/index.html
[16] Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daumé III, and Lise Getoor. 2017. Modeling learner engagement in MOOCs using probabilistic soft logic. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM.
[17] Justin Reich. 2014. MOOC completion and retention in the context of student intent. *EDUCAUSE Review Online* (2014).
[18] Colin Taylor, Kalyan Veeramachaneni, and Una-May O'Reilly. 2014. Likely to stop? Predicting stopout in massive open online courses. *arXiv preprint arXiv:1408.3382* (2014).
[19] M Lee Upcraft, John N Gardner, and Betsy O Barefoot. 2004. *Challenging and Supporting the First-Year Student: A Handbook for Improving the First Year of College*. ERIC.
[20] Han Wan, Jun Ding, Xiaopeng Gao, and David Pritchard. 2017. Dropout Prediction in MOOCs using Learners' Study Habits Features. In *Proceedings of The Tenth International Conference on Educational data Mining (EDM 2017)*.
[21] Jacob Whitehill, Joseph Jay Williams, Glenn Lopez, Cody Austun Coleman, and Justin Reich. 2015. Beyond prediction: First steps toward automatic intervention in MOOC student stopout. *Educational Data Mining* (2015).
[22] Linda G Wyatt. 2011. Nontraditional student engagement: Increasing adult student success and retention. *The Journal of Continuing Higher Education* 59, 1 (2011), 10–20.
[23] Diyi Yang, Tanmay Sinha, David Adamson, and Carolyn Penstein Rosé. 2013. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings: 2013 NIPS Data-driven Education Workshop*, Vol. 11. 14.