ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-647

Is Big Data Better? LMS Data and Predictive Analytic Performance in Postsecondary Education

Kelli A. Bird
University of Virginia

Benjamin L. Castleman
University of Virginia

Yifeng Song
University of Virginia

Renzhe Yu
Teachers College
Columbia University

Data science applications are increasingly entwined in students' educational experiences. One prominent application of data science in education is to predict students' risk of failing a course in or dropping out from college. There is growing interest among higher education researchers and administrators in whether learning management system (LMS) data, which capture very detailed information on students' engagement in and performance on course activities, can improve model performance. We systematically evaluate whether incorporating LMS data into course performance prediction models improves model performance. We conduct this analysis within an entire state community college system. Among students with prior academic history in college, administrative data-only models substantially outperform LMS data-only models and are quite accurate at predicting whether students will struggle in a course. Among first-time students, LMS data-only models outperform administrative data-only models. We achieve the highest performance for first-time students with models that include data from both sources. We also show that models achieve similar performance with a small and judiciously selected set of predictors; models trained on system-wide data achieve similar performance as models trained on individual courses.

VERSION: September 2022

Suggested citation: Bird, Kelli A., Benjamin L. Castleman, Yifeng Song, and Renzhe Yu. (2022). Is Big Data Better? LMS Data and Predictive Analytic Performance in Postsecondary Education. (EdWorkingPaper: 22-647). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/8xys-ym74

Is Big Data Better?

LMS Data and Predictive Analytic Performance in Postsecondary Education

Kelli A. Bird, University of Virginia

Benjamin L. Castleman, University of Virginia

Yifeng Song, University of Virginia

Renzhe Yu, Teachers College Columbia University

Abstract

Data science applications are increasingly entwined in students' educational experiences. One prominent application of data science in education is to predict students' risk of failing a course in or dropping out from college. There is growing interest among higher education researchers and administrators in whether learning management system (LMS) data, which capture very detailed information on students' engagement in and performance on course activities, can improve model performance. We systematically evaluate whether incorporating LMS data into course performance prediction models improves model performance. We conduct this analysis within an entire state community college system. Among students with prior academic history in college, administrative data-only models substantially outperform LMS data-only models and are quite accurate at predicting whether students will struggle in a course. Among first-time students, LMS data-only models outperform administrative data-only models. We achieve the highest performance for first-time students with models that include data from both sources. We also show that models achieve similar performance with a small and judiciously selected set of predictors; models trained on system-wide data achieve similar performance as models trained on individual courses.

Acknowledgements

We are very grateful to Dr. Cat Finnegan and the team at the Virginia Community College System for their collaboration on this project. We are grateful to our colleagues at VCCS and University of California-Irvine for their assistance obtaining the LMS data.

Introduction

Data science applications are increasingly entwined in students' educational experiences (Fischer et al., 2020). These applications are both widespread and varied, ranging from adaptive learning algorithms that tailor instruction to students based on their performance on initial tasks (e.g. Murphy et al., 2020), to natural language processing tools that automate writing guidance and assessment (e.g. McNamara et al., 2013; Gayed et al., 2022), and chatbots that ingest textual or verbal input from students to provide guidance through the college application process (Page and Gehlbach, 2017).

One of the most prominent applications of data science in education has been to predict students' risk of failing a course in or dropping out from college. A third of all higher education institutions have invested in predictive analytics and collectively spend hundreds of millions of dollars to generate these predictions (Barshay and Aslanian, 2019). Most institutions use the "early alerts" generated by predictive models to initiate outreach from academic advisors, or to encourage faculty to reach out to students in their classes who are struggling to succeed (Ekowo and Palmer, 2016; Klempin, Grant, and Ramos, 2018).

There is growing interest among higher education administrators and researchers in whether the underlying predictive models accurately predict student success, especially as early in the academic term or the program of study as possible (Arnold and Pistilli, 2012; Treaster, 2017). A related area of interest is what combination of data sources can be leveraged to create meaningful predictors and in turn the most accurate predictions. The most common data source is institutions' administrative data, which include students' sociodemographic characteristics and academic history that have been found to be strongly correlated with student success in education research (Kuh et al., 2007).

More recently, the steady rise in digital learning (most prominently during the COVID-19 pandemic, but also in the years preceding) has generated unprecedentedly rich data about students' day-to-day (and in some cases, moment-to-moment) academic engagement captured by their behavior in learning management software (LMS). Prior studies have used various data mining techniques to demonstrate how fine-grained behavioral traces in LMS can depict students' learning processes and predict students' academic performance (e.g., Li, Baker, and Warschauer, 2020; Lim, 2016; Park et al., 2018). However, the richness of the LMS behavioral trace data (referred to as "LMS data" in the remainder of this paper) requires substantial analytic time and computing capacity, since raw LMS data often include records for each single action a student performs when interacting with the system (Baker et al., 2020).

In this paper we systematically evaluate whether incorporating LMS data into course performance prediction models substantially improves model performance when administrative data (henceforth "admin-only data") are already used. Our analysis builds on prior studies that have conducted exploratory analyses, at the level of a small number of courses, of the comparative utility of LMS data to other data sources for predictive analytics in education. For instance, in a study using data from ten introductory STEM courses at a public research university, Yu et al. (2020) find that predictive models trained on small sets of predictors derived from admin-only data or LMS-only data both have reasonably strong performance, and that models trained on admin and LMS data together have the highest levels of performance. Aquilar et al. (2014) demonstrate that, among first-semester engineering students at Notre Dame, students' ePortfolio entries, which include required student project updates and reflections, enhance predictive accuracy for whether students will persist into the next course in the engineering sequence. Crossley et al. (2016) use data from several hundred participants in a MOOC course to demonstrate that students' clicks in

the MOOC interface and discussion content (as analyzed through natural language processing) accurately predict whether students will complete the course.

While the predictive models investigated in these prior papers included several courses and hundreds to a couple thousand students, our paper includes 2,646 courses across 23 institutions and 226,784 students across an entire state community college system, thus greatly increasing the generalizability of our results. We also build on prior studies by conducting our analysis among community colleges, which account for approximately 40 percent of all postsecondary enrollments and in which course failure and dropout rates are much higher. Insights on whether LMS data improves course performance predictions at the community college level could thus inform outreach and support efforts that have the potential to benefit a much larger and more at-risk population of students. Relative to prior papers, we also make the novel contribution of investigating whether the predictive utility of LMS data varies based on whether students have prior enrollment terms or are new to college.

We conduct our investigation using data from the Virginia Community College System (VCCS), which consists of 23 community colleges in the Commonwealth of Virginia. We have access to detailed student records for all students who attended a VCCS college from 2000 to the present. Because the VCCS recently navigated to a new LMS (Canvas), we only use data from after all colleges switched to the new system in Summer 2019. Across course modalities (i.e. in-person, online, or hybrid), instructors can use Canvas to organize and manage a variety of teaching and learning activities, such as submitting and grading assignments, sharing course materials, creating discussion forums, proctoring quizzes and exams, and in the case of synchronous online courses, hosting virtual meetings.

Our analytic sample consists of observations from course sections that used Canvas during the Summer 2019 through Spring 2021 terms (omitting Spring 2020, for reasons we detail below). We classify student data into two broad categories: admin-only data and LMS-only data. Admin data includes measures such as student's cumulative GPA, prior credit accumulation, current enrollment intensity, and demographics, as well as course-level information like average historic grades and modality. LMS data includes measures such as total time spent logged into the LMS and the number of on-time assignment submissions. We construct LMS measures during the first quarter of the current term, for both the target course and concurrently taken courses, to balance the need to incorporate some information from the current term as predictors while leaving ample time in the remainder of the term to intervene with at-risk students based on predictions generated using LMS data. For returning students, we also construct the LMS predictors for previously taken courses (aggregated by averaging over all course sections) -- measured in both the first quarter of the term and for the full term.

We use a Random Forest prediction model to predict student performance, using a binary measure of success based on the student's final grade (A/B/C versus D/F/W). We test several primary variations of the course performance prediction model. Our primary variation of interest is whether models incorporate admin-only data, LMS-only data, or both admin and LMS data. This comparison allows us to assess whether incorporating LMS data into course performance prediction models improves their accuracy and performance. Our second variation of interest is whether our analytic sample includes students with prior course enrollments at VCCS, or focuses solely on first-time VCCS students. This comparison allows us to assess whether LMS data on early-term engagement and performance adds more to course-performance prediction accuracy for first-time students who lack a record of prior course performance at the postsecondary level.

Finally, we vary whether models include the full set of admin and LMS measures we can leverage or instead incorporate a small set of predictors with high feature importance. This variation enables us to test whether we can achieve similar course performance predictions with a less analytically- and computationally-intensive set of measures.

Our paper yields several primary conclusions and corresponding contributions to the existing research literature. First, among students with prior course history at VCCS, course performance prediction models trained on admin-only data substantially outperform models trained on LMS-only data, and are quite accurate at predicting whether students will struggle in a course or not. The admin-only model has a c-statistic (a general measure of model performance) of 0.855 while the LMS-only model has a c-statistic of 0.779.[1] Adding LMS data to the models trained on admin data results in only a slight marginal improvement in model performance (2 percent increase in c-statistic). This suggests that, for students with course history in college, detailed measures of students' early-term engagement and academic performance do not meaningfully improve our ability to predict their success in the course beyond the predictions we could generate just relying on measures of their prior academic performance. By contrast, among first-time students without course histories at VCCS, course performance prediction models trained on LMS-only data outperform models trained on admin-only data (c-statistics of 0.775 and 0.728, respectively), and we achieve the highest model performance for first-time students with models trained on a combination of admin and LMS data (c-statistic of 0.825). That being said, the relative value of the LMS data in increasing model accuracy is highly variable across courses, with LMS data having the lowest value for predicting performance in math courses. For example, the c-statistic for the LMS-only model ranges from 0.81 for College Composition I to 0.70 for Pre-

[1] As we describe in the Methods section below, a model performance is generally considered strong if the c-statistic is 0.8 or higher.

Calculus I. This pattern of results indicates that early-term LMS measures on engagement and performance can be useful in predicting whether students will struggle in a course, particularly among new students, but not across all courses.

Second, we show that course performance prediction models can achieve effectively the same high level of performance with a small set of high-feature importance predictors (~30 total) as models trained on hundreds of admin and LMS predictors. Interestingly, the ~30 high-feature importance predictors that result in the largest gains in model performance include both admin and LMS measures, despite our earlier finding that, among students with prior course histories, incorporating LMS measures into course-performance prediction models trained on admin data results in only negligible gains in model performance. This finding indicates a high degree of correlation between the most predictive admin and LMS measures.

Third, we show that models trained on system-wide data achieve similar performance as models trained on individual courses with larger enrollments. This suggests course performance prediction models trained at system level can be flexibly used to generate course-level prediction models for a broad range of courses.[2]

Finally, we release all of our code to build and evaluate the predictive models so that they are accessible to the public. Colleges can leverage this code base to predict course-level success in their own contexts. This contribution is useful since prediction analytics platforms in use are typically proprietary and expensive for colleges to operate.

Data

[2] While building one system-wide prediction model may be time saving compared to building many course-specific models, it is also true that course-specific models have significantly fewer observations compared to system-wide observations and therefore require less computing power.

The data for this study come from two sources within the VCCS: (1) administrative records; and (2) behavioral trace data from Canvas LMS. The administrative data include detailed information from each term in which a student enrolls (beginning in Summer 2000), including their program of study, courses taken, grades earned, credits accumulated, financial aid received, and degrees or certificates awarded. We observe unique instructor identifiers beginning in 2008 that allow us to track instructors across courses and terms; we also observe whether instructors are full-time or adjunct faculty. The LMS data include records of any activities students perform in the system, such as visits to any learning content pages, discussion posts and direct messages, and assignment and quiz submissions. Each record comes with rich metadata such as timestamp of the activity, the URL the activity triggers, and the Canvas object ID (e.g., assignment ID) the activity is associated with.

Outcome

Our outcome of interest is a binary measure for successful course completion, and is equal to one if the student earned a grade of A, B, or C, and equal to zero for grades of D, F, or W. While a grade of D earns the student credit for the course and is considered a passing grade, within VCCS, students cannot satisfy some program requirements with a D, and other colleges and universities typically do not accept transfer credit for D grades.

Sample

Our analytic sample includes students taking VCCS courses that use Canvas from Summer 2019 through Spring 2021. Seventy-five percent of all VCCS course sections use Canvas, and our analytic sample consists of 81 percent of all VCCS student-by-course section observations during this time frame.[3] We exclude Spring 2020 from the sample due to the extreme disruptions of

[3] Appendix Table A1 shows the summary statistics for the full VCCS population. Comparing this to Table 1, our analytic sample is quite similar to the full population.

COVID-19 on higher education, which included the VCCS shifting to an emergency grading policy that changed the standard grading scale such that the possible grades were P+, P-, Incomplete, or Withdraw. We further restrict the sample to focus on college-level coursework for regularly-enrolled students. Specifically, we exclude observations corresponding to dual-enrollment (i.e. high school students taking college-level coursework). We also exclude all observations outside the traditional A-F grading scale. The vast majority of these observations correspond to developmental courses, which are graded as pass / fail.

As shown in Panel A of Table 1, the final sample includes 1,173,878 student-by-course-by-section observations from Summer 2019 through Spring 2021. This translates to 226,784 unique students; 2,646 unique courses; and 63,994 unique course sections pairs. We split the analytic sample into a training set and a validation set. The training set includes observations from the Summer 2019, Fall 2019, Summer 2020, and Fall 2020 terms; the validation set includes observations from the Spring 2021. We use Spring 2021 as the validation sample with the intention of building a more generalizable model; specifically, if the observation window of the validation sample occurs after the observation window of the training sample, then the evaluation results provide a more accurate estimation of the performance of the model when applied to a practical setting (i.e. predicting current student success using a model trained on historical cohorts). We further split the analytic sample based on whether the student was enrolled at VCCS prior to the target term. As we detail below, if a student is in their first term and therefore has no prior academic history at VCCS, then we have far less information to include as predictors. Therefore, we build separate models for the observations in the analytic sample with no prior VCCS enrollment ("1st term" sample) versus observations in the analytic sample with at least one term of VCCS enrollment history ("2+ terms" sample).

Panel B of Table 1 shows basic student characteristics for the full analytic sample and separately for the training and validation sets of the 2+ terms and 1st term samples. Within the 2+ terms sample, students in the training and validation sets are similar on average. However, for the 1st term sample, there are some differences in student-level characteristics. Compared to the training set, the validation set contains a significantly lower share of Hispanic students (13.1 percent versus 5.9 percent), a larger share of Female students (55.2 percent and 58.4 percent) and significantly older students (22 versus 27). These differences are likely due in large part to changes in the composition of the new student population, with community colleges experiencing a 20.8 percent drop in new enrollments between Fall 2019 and Fall 2021 (National Student Clearinghouse, 2021). Panel C of Table 1 shows basic course characteristics across the relevant samples. As expected, given that observations in the 2+ terms sample are students who are further along in the academic careers, courses represented in the 2+ terms sample are more likely to be for 200-level, Medical Science, or Applied Technologies courses, less likely to be Social Sciences or Humanities, and have smaller enrollments.

Administrative predictors

We construct 279 predictors from the administrative data. We describe them at high-level here, and include a full list in Appendix Table A2:

- Demographic information (age)[4];
- Non-course-specific academic records:
    - General academic information about the student (e.g. cumulative GPA, total credits accumulated);

[4] While the admin data includes other basic demographic characteristics (race and gender), we omit these variables so that otherwise identical students from different racial or gender categories will not have different predicted scores. Recent work finds that the inclusion of demographic characteristics has negligible impact on the performance of a similar model predicting degree completion (Bird et al, 2021).

    - General information about the term in which the student is enrolled (e.g. program of study the student is currently pursuing, the student's enrollment intensity);
- Course-specific information:
    - Characteristics of the target course (e.g. course enrollment, average grade from the most recent five years);
    - Student academic history related to the target course (e.g. whether the student attempted the target course before; the student's GPA in all the target course's prerequisites);
- Course-subject-specific information:
    - Student academic history related to academic clusters. We categorize VCCS courses into ten high-level clusters (Mathematics, Natural Sciences, Social Sciences, etc; a full list is shown in Table 1) and include predictors indicating the student has taken any course within each cluster, and their cluster-specific GPA. We construct different predictors for each cluster of the target course, i.e. there is a separate Mathematics GPA predictor for Natural Sciences target courses versus Social Sciences target courses;
- Instructor-related information (e.g. instructor tenure, full-time versus adjunct, average grade assigned in the target course in past terms);

For 1st term observations, we include 59 predictors to include basic demographic information, general term-level information, characteristics of the target course, and instructor characteristics.

LMS predictors

We construct 50 predictors from the LMS data. We describe them at high-level here, and include a full list in Appendix Table A1:

- Early-term target course: measures of engagement in the target course during the first quarter of the course period (e.g., total number of click actions, total time spent online, percent of on-time assignment submissions)[5];
- Early-term concurrent courses: the same early-term measures of engagement in all other courses taken in the same term as the target course, averaged across these courses;
- Prior early-term: the same early-term measures of engagement in all courses taken in prior terms, averaged across these courses;
- Prior full-term: the same measures of engagement metrics in all prior courses, computed across the entire term instead of the first quarter and averaged across these courses.

If a student is in their first term, then we only include the 21 predictors measuring engagement in current courses. All of the LMS predictors are normalized within each term x course x section cell.[6] This standardization accounts for differences in engagement due to differences in the use of the online LMS across courses, instructors, and modalities.[7]

Handling Missing Values

There is expected missingness in the data -- for example, if a student has not taken prior courses in an academic cluster, then the average grade of prior courses in this cluster are missing. Similarly, if a particular course section does not use discussion forums, then all LMS predictors

[5] Within an academic term, different courses may vary in start date, end date, and length, so the measures are computed in relation to the specific period of each course.

[6] Z-score normalization is applied to the predictors within each term x course x section cell, such that normalized values of each predictor within the cell has mean 0 and variance 1.

[7] Because all course sections, regardless of modality, may use the LMS for a variety of course aspects, we include observations from online, in-person and hybrid course sections. We explore differences in model performance for modality-specific models in the Appendix.

related to forum posts are missing. We handle this missingness by setting missing values equal to zero and include indicators for whether a given predictor is missing.

Methods

We use a Random Forest model to predict successful course completion. Random Forest is a tree-based ensemble model commonly used in data science research for predictive analytics. Recent work that investigates similar degree completion prediction models find nearly identical levels of accuracy for Random Forest and other commonly used models (Bird et al, 2021). For this paper, we initially tested other models to predict course success, and Random Forest slightly outperformed the others.

Our primary objective is to compare the prediction accuracy of the admin data versus LMS data. Therefore, we first estimate models using (1) admin-only predictors; (2) LMS-only predictors; and (3) full set of predictors. For each of these three settings, we build separate models on the 1st term and the 2+ terms samples. To compare these six main models, we report the following evaluation metrics:

- C-statistic: a "goodness of fit" measure that is equal to the probability that a randomly selected positive observation (i.e. a student who passed a particular course) has a higher predicted score than a randomly selected negative observation. A c-statistic of 0.5 corresponds to a model being no better than choosing at random, while a c-statistic of 1 corresponds to a model perfectly predicted the outcome. A c-statistic of 0.8 or higher is considered strong performance; and a c-statistic of 0.9 or higher is considered outstanding (Hosmer, Lemeshow, and Sturdivant, 2013).

- Precision: share of observations that the model predicts will succeed that actually succeed (i.e. "true positives).
- Recall: share of true positives that the model correctly predicts as succeeding.
- Feature Importance (FI) score: we compute the FI score for each predictor in the model based on the mean decrease in impurity, which roughly speaking is a measures how often a predictor is used to split decision tree "branches", and provides a metric of each predictor's contribution to the overall model's accuracy.[8]

We next estimate course-specific models for five of the largest courses offered by VCCS: General Biology I (BIO101); College Composition I (ENG111); College Composition II (ENG112); Quantitative Reasoning (MTH154); and Pre-Calculus I (MTH161). We estimate admin-only, LMS-only, and full predictor models for each of these courses. With the exception of ENG111, the vast majority of observations (particularly in the validation sets) for these courses are non-first-term; therefore, we combine the 1st term and 2+ terms samples for the course-specific models.[9]

Because our outcome is binary, the immediate output of the model is a predicted score for each observation ranging from zero to one, with a value closer to one indicating a higher predicted probability of course success. Therefore, we set a threshold in predicted score to delineate observations into two categories: those predicted to successfully complete the course (i.e., those with a predicted score at or above the set threshold), and those predicted to not. We set the

[8] In the decision tree growth procedure, at each node splitting, the difference in Gini impurity (a quantitative measure regarding how close the observations in the node are to having the same outcome) between the child nodes and the parent node measures to what extent the predictor used to split the node contributes to the model's ability of differentiating outcome one from outcome zero. The mean decrease in impurity of each predictor is computed by taking the weighted average over all node splittings within all decision trees in the ensemble where the predictor is used to make the splitting, with the weights determined by the corresponding node sizes (Breiman, 2002).

[9] We include an indicator variable for whether each observation corresponds to a course taken during the 1st term. If a predictor (e.g. cumulative GPA prior to taking the course) is not available for the 1st term observations, the value of that predictor is set to 0 for all 1st term observations.

threshold equal to the course completion rate within the training sample used for each model (77.8% for the 2+ terms training sample, and 70.9% for the 1st term training sample).[10]

Results

In Figure 1 we present several measures that evaluate the performance of the course performance prediction model using admin-only data, LMS-only data, or both admin and LMS data. Panel A presents c-statistics while Panels B and C present precision and recall rates, respectively. Within each panel, we present performance measures for the 2+ terms sample (students with two or more terms enrolled at VCCS, inclusive of the current term) on the left and for the 1st term sample (students in their first term at VCCS) on the right. As we show in Panel A, the course performance prediction model trained on the 2+ terms sample with admin-only data achieves a high degree of accuracy, with a c-statistic of 0.855. The model trained on LMS-only data and the 2+ terms sample of students has substantially lower accuracy, with a c-statistic of 0.779. Combining both admin and LMS data with the 2+ terms sample leads to modestly higher accuracy than we obtain with the admin-only data, with a c-statistic of 0.872.

Among the 1st term sample, on the other hand, we find comparatively greater predictive value from the LMS-only data: Whereas the model trained on admin-only data has a c-statistic of 0.728, the model trained on LMS-only data has a c-statistic of 0.775. Combining both admin and LMS data leads to a proportionally greater gain in accuracy (c-statistic of 0.825) than we observed in the 2+ terms sample.

This pattern of relationships makes intuitive sense. First, the course performance prediction models are monotonically more accurate across data sources for the 2+ terms sample, which we

[10] While c-statistics are independent of the threshold, precision and recall can be highly sensitive to the threshold chosen. Other common methods used to set the threshold, such as maximizing the F1-score, can result in significant differences in thresholds set from model to model. Our approach allows for better comparison of precision and recall across models.

would expect given that we have more data--and in particular more observed academic performance--on which to train the model. Second, and by association, the comparative value of LMS measures of engagement is higher for 1st term students, for whom baseline data on academic performance is much more limited.

This same basic pattern of relationships holds when we consider model precision or recall as our accuracy measures instead of c-statistics. All models achieve high rates of both precision and recall, with the highest precision and recall rates among the 2+ terms sample and with the model trained on both admin and LMS data. Specifically, 89.3 percent of students the model predicts to complete the course actually earn an A, B, or C (precision); 89.9 percent of students who actually earn an A, B, or C are predicted by the model to complete the course (recall). Appendix Table A3 displays confusion matrices for the six models, from which precision and recall are derived, and provide a more fine-grained comparison of the models' predictions with students' actual outcomes.[11] The percent of observations with accurate predictions (those with actual grades of A, B, C predicted to complete + those with actual grades of D, F, W predicted to not) reported in each confusion matrix follows a very similar pattern to the other evaluation metrics.

In Figure 2 we investigate the relationship between the number of predictors we include in the course-performance prediction models and the corresponding gain in model performance (as measured by the c-statistic). Plots A, B, and C display this relationship for the model trained on the 2+ terms sample; Plots D, E, and F display this relationship for the model trained on the 1st term sample. Within each sample, the top plot presents the relationship for the model trained on

[11] Specifically, a confusion matrix shows the number of observations for each combination of predicted outcome and actual outcome. We create a confusion matrix where the predicted outcome is binary (A/B/C versus D/F/W), and the actual outcome is the actual grade received (A, B, C, D, F, or W).

admin-only data, while the middle and bottom plots present the relationship for the models trained on LMS-only and admin + LMS data, respectively. In order to create these plots, we begin with the five most important predictors (as determined by feature importance score, defined above) and add predictors in groups of five in descending order of feature importance.

Across plots, we show that the vast majority of gain in model performance is achieved from a small but influential set of measures. This is most apparent in the model trained on admin-only data and the 2+ terms sample. In this model, including the 5 predictors with the highest feature importance scores achieves a c-statistic of 0.82.[12] With 25 high-influence predictors the model achieves a c-statistic of 0.85 and beyond 50 predictors, there is only slight gain in performance up to the full set of over 279 predictors.

We observe the same general relationship with the other five models, though in the case of the 1st term sample and particularly the LMS-only models, the total number of predictors is fewer and the maximum performance achieved is lower. For instance, much of the gain in performance in the LMS-only data trained on the 2+ terms sample is achieved after the first 15 out of 50 predictors (c-statistic of ~0.76); after 30 predictors there is negligible gain in performance from adding additional LMS measures.

This pattern of results demonstrates that course performance prediction models can achieve quite high levels of accuracy with a relatively small set of predictors.[13] In Table 2 we examine the specific substantive groups of predictors that contribute most to model accuracy. The rows in the table correspond to substantive groupings of predictors, e.g. predictors that measure overall

[12] These predictors are: share of previously attempted credits that were withdrawn, cumulative GPA at the beginning of the target term, total credits attempted in the target term, indicator for whether term GPA is available in the term right prior to the target term, term GPA of the term right prior to the target term

[13] While we find that most of the models' performance is achieved from a small subset of high-feature importance predictors (~10 percent of predictors in the full data model), it is important to note that these top predictors are only revealed after building the model using full data.

academic performance, and for each row we report the c-statistic associated with a separate model trained just on that set of predictors. Panel A presents the c-statistics for the sample of students with prior VCCS experience while Panel B presents the c-statistics for the sample of students in their first term. The first two rows in Panel A confirm what we have shown earlier: we obtain the highest performance level from the model that leverages all admin and LMS predictors (N=329, c-statistic=0.877), but the model trained on admin-only predictors (N=279) achieves high accuracy as well (c-statistic=0.855). Within this total set of admin predictors, a subset of 41 predictors that measure a combination of overall (i.e. not course-specific) academic performance and demographic characteristics achieves similar performance (c-statistic=0.841).[14] By comparison, a model trained on 238 course-specific (or course-by-instructor specific) measures, such as the mean grade of students in the target course in prior terms, has notably lower performance (c-statistic=0.778).

This pattern of results reaffirms both that a small and judiciously-selected set of predictors can achieve nearly the same performance as a model with several times the number of predictors, and that measures of students' prior academic performance most strongly predict performance in the target course. The remaining rows in Panel A present c-statistics for models trained on different combinations of LMS measures. The subset of LMS predictors that measure students' engagement in the target course and concurrent courses (N=21 predictors) contribute substantially more to model performance (c-statistic=0.751) than LMS measures of students' engagement in prior courses. Among the 1st term sample (Panel B), we observe a generally similar pattern of results, though as we show earlier, overall model performance among this sample of students is lower.

[14] We combine these two sets of predictors because we only have one demographic predictor: age.

In Table 3, we display the predictors--both admin and LMS--with the highest feature importance score. Predictors with the highest feature importance score contribute the most to overall model performance. For instance, for the model constructed on the 2+ terms sample with admin + LMS predictors, the feature importance scores of the "percent of prior attempted credits withdrawn" and "cumulative GPA" are 0.075 and 0.04, respectively, meaning those two predictors are substantially more important than "credits attempted in last term" and "age at time of target course enrollment," whose feature importance scores are both 0.009.[15] In Panel A we present the thirty predictors with the highest feature importance in the model trained on the 2+ terms sample; in Panel B we present the twenty predictors with the highest feature importance in the model trained on the 1st term sample. Among students with prior academic performance at VCCS, eight of the ten predictors with highest feature importance in this sample are admin measures. Consistent with what we show in Table 2, most of the highest feature importance predictors capture some aspect of students' prior credit accumulation and GPA. For instance, the predictor with the highest feature importance measures the percentage of prior attempted credits from which the student withdrew, while the predictor with the second highest feature importance measures students' cumulative GPA in prior terms. Another important set of admin predictors are two measures of historic performance in the target course (i.e. the average course completion rate in that course in the five prior years). The two LMS predictors in the top ten list measure students' overall engagement in the first quarter of the term: the total number of click actions students perform , and the total time (in minutes) students stay online.

[15] Due to the complex nature of how feature importance scores are calculated (see footnote 2), we cannot make precise comparative statements about the relative importance of predictors using this metric. For example, it is NOT necessarily the case that a predictor with an FI score of 0.08 is precisely twice as important as a predictor with an FI score of 0.04.

Five of the top ten predictors in terms of feature importance are common between the two samples of students: the number of total credits attempted in the target term, the two measures of historic performance in the course, and the two LMS measures of student engagement. Additional LMS measures of student engagement, such as the number of discussion forum posts and the number of assignments submitted, are among the highest feature importance predictors for the 1st term sample.

We have focused up until now on the accuracy of the course performance prediction model across all courses included in our training and validation samples. In practice, however, the model is most applicable at the level of a specific course, where a course instructor could use the predictions to proactively reach out to students in their course who are predicted to get a D, F, or W. In the next set of tables and figures, we therefore investigate the overall accuracy of models trained on course-specific samples (e.g., all students who enroll in English 111, the College Composition course offered across the VCCS), as well as whether we observe generally similar patterns in the contribution of admin versus LMS data to model performance. We focus this analysis on five large-enrollment courses in core subjects that typically function as "gateways" for students to take higher-level courses within each core subject and thus to fulfill degree requirements across most VCCS programs of study. Specifically, we test course-specific performance prediction models for the two-course sequence on College Composition (ENG 111 and ENG 112); General Biology (BIO 101); and two introductory, college-level math courses, Quantitative Reasoning (MATH 154) and Pre-Calculus I (MATH 161).

In Table 4 we present summary statistics for these courses. Each course is offered in hundreds of sections each term across the 23 VCCS institutions, and each enrolls thousands or even tens of thousands of students per term. Performance across these courses tends to be relatively

low, with mean GPAs ranging from 2.22 in MTH 161 to 2.73 in ENG 112. All five courses also have a high rate of students earning a D, F, or W. These rates of academic struggle range from 26.7 percent in BIO 101 to 41.5 percent in MTH 161. A sizable share of enrollments in four of the five courses (all except ENG 112) are students in their first term at VCCS. For instance, 25 percent of students in MTH 154 and 52.1 percent of students in ENG 111 are in their first term.

In Figure 3 we present c-statistics for course-performance models trained separately on the sample of students enrolled in each of the five courses. Across all five courses, models that combine admin and LMS data achieve the highest levels of performance, and performance levels are generally high for the course-specific models. Specifically, the course-specific performance prediction models for ENG 111, ENG 112, and BIO 101 all achieve c-statistics of 0.85 or higher. The MTH 154 and MTH 161 performance prediction models have somewhat lower performance (c-statistics of 0.82 and 0.79, respectively). Across four of the five courses (all except ENG 111), we find that models trained on admin-only measures meaningfully outperform models trained on LMS-only data. In the case of ENG 111, the model trained on LMS-only data does outperform the model trained on admin-only data (c-statistic of 0.81 compared to 0.78); this makes intuitive sense as a sizeable share of ENG 111 students (56.2% of the training sample) are in their first term at VCCS. Appendix Figure A1 shows very similar patterns for precision and recall across the 15 course-specific models represented in Figure 3.

In Figure 4 we show which groups of predictors contribute most to overall model performance, within the course-specific performance prediction models. We again observe a very similar pattern to what we found with the prediction model trained on all courses. Across the five courses, models trained on admin predictors measuring overall academic performance and demographic characteristics achieve nearly as high performance as models trained on the full set

of admin measures (Panel A). By comparison, models trained on the subset of course- (or course-by-instructor-) level measures of academic performance do not achieve as high accuracy. Among the LMS predictors (Panel B), we see across the five courses that models trained on measures of students' early-term engagement in the target course and/or concurrent courses achieve nearly as high performance as models trained on the full set of LMS measures. Models trained on LMS measures of prior-term engagement achieve substantially lower performance.

The differences in LMS-only models across courses we observe in Figures 4 and 5 are notable, with c-statistics ranging from 0.81 for ENG 111 to 0.70 for MTH 161 for models using all LMS predictors. As other researchers have noted (e.g. Baker et al, 2020), the value of LMS predictors is driven in some part due to course-specific context. For instance, English instructors may structure their courses on the LMS significantly differently than Math instructors. We explore this explicitly by comparing the mean values of the top three LMS predictors across the five courses in Figure 5. Overall, we see that the courses for which LMS predictors add the greatest value are those with the highest averages of LMS predictors. We see that while total time online is more similar across courses (ranging from 507 minutes for MTH 161 to 641 minutes for ENG 111), the two math courses have approximately one-third fewer clicks than the English and Biology courses. The starkest difference is average word count in discussion posts, which is three to six times higher in English courses compared to the Math and Biology courses.

Discussion

As LMS software is becoming increasingly more prevalent in higher education -- particularly in a post-COVID era characterized by flexibility of instruction modality -- researchers and higher education institutions are increasingly interested in harnessing the LMS-generated data

for various instructional and analytic purposes. However, making use of LMS data can be very costly in terms of personnel time, data storage, and computing power. For example, the VCCS LMS data for a single term is roughly one to two terabytes. Converting the raw data (which includes a row for each navigation or "click" a student makes within the LMS) into usable predictors requires expertise and a significant time investment. Particularly given limited resources at institutions like community colleges, it is important to understand the potential value of LMS data in predictive analytics.

In this paper, we evaluate the value of including LMS data in prediction models of course performance, relative to administrative data. We find that the added performance gain from LMS data varies significantly across contexts, even within a community college system that uses the same LMS software across all courses and institutions. Specifically, LMS data adds little value in predicting course performance for returning students. However, in the case of new students, LMS-only data outperforms admin-only data, and the combination of LMS and admin data has significantly higher performance compared with using only one data source.

We also find significant variation across five major courses in the added performance from LMS data, with the lowest value-add for math courses -- these are also the courses with the least interaction between students and the LMS system. These results suggest that LMS data adds substantial predictive value and may be worth the investment for courses that (1) enroll many first-time students, and (2) actively use LMS for instructional design. The relatively poor performance of the admin-only data for first-time students that we find (c-statistic of 0.728) suggests that if LMS data is not available for first-time students, then other data collection efforts (e.g. incorporating high school transcripts) could substantially benefit predictive analytics in that setting.

For researchers or administrators interested in learning more specifically about how we work with the LMS and admin data to construct predictors, and how we build the predictive models described in this paper, we have made our codebase public at https://github.com/nudge4/admin_vs_lms_data_public.

More broadly, our results demonstrate that, despite the steady onset of big data in education, along with the corresponding methods and expertise to analyze these data, researchers and educators should continue to critically investigate whether making use of these data result in meaningfully better models or performance than can be achieved with more traditional data sources and methods.

References

Arnold, Kimberly & Pistilli, Matthew. (2012). Course signals at Purdue: Using learning analytics to increase student success. ACM International Conference Proceeding Series. 10.1145/2330601.2330666.

Baker, R., Xu, D., Park, J., Yu, R., Li, Q., Cung, B., Fischer, C., Rodriguez, F., Warschauer, M., & Smyth, P. (2020). The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: opening the black box of learning processes. *International Journal of Educational Technology in Higher Education*, 17(1). https://doi.org/10.1186/s41239-020-00187-1

Barshay, J., Aslanian, S. (2019). Under a watchful eye: Colleges are using big data to track students in an effort to boost graduation rates, but it comes at a cost (APM Reports). https://www.apmreports.org/story/2019/08/06/college-data-tracking-students-graduation

Bird, Kelli A., Benjamin L. Castleman, Zachary Mabel, and Yifeng Song (2021). Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education. *AERA Open*, January 2021. doi:10.1177/23328584211037630

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3. 1. Statistics Department University of California Berkeley, CA, USA 1 58.

Crossley, Scott, Luc Paquette, Mihai Dascalu, Danielle McNamara, & Ryan Baker (2016). Combining click-stream data with NLP tools to better understand MOOC completion. 6-14. 10.1145/2883851.2883931.

Ekowo, M., Palmer, I. (2016, October). The promise and peril of predictive analytics in higher education: A landscape analysis (New America Policy Paper). https://www.newamerica.org/education-policy/policy-papers/promise-and-peril-predictive-analytics-higher-education/

Fischer C, Pardos ZA, Baker RS, Williams JJ, Smyth P, Yu R, Slater S, Baker R, Warschauer M (2020). Mining Big Data in Education: Affordances and Challenges. *Review of Research in Education*, 44(1):130-160. doi:10.3102/0091732X20903304

Gayed, John Maurice, May Kristine Jonson Carlon, Angelu Mari Oriola, Jeffrey S. Cross (2022). Exploring an AI-based writing Assistant's impact on English language learners, *Computers and Education: Artificial Intelligence*, Volume 3, 2022, 100055, ISSN 2666-920X, https://doi.org/10.1016/j.caeai.2022.100055.

Hosmer, David W. Jr., Stanley Lemeshow, and Rodney X. Sturdivant (2013). Applied Logistic Regression, Third Edition. John Wiley & Sons, Inc., Hoboken New Jersey. ISBN 978-0-470-58247-3.

Klempin, S. C., Grant, M., Ramos, M. (2018). Practitioner perspectives on the use of predictive analytics in targeted advising for college students (CCRC Working Paper No. 103). https://ccrc.tc.columbia.edu/publications/practitioner-perspectives-predictive-analytics-targeted-advising.html

Kuh, G. D., Kinzie, J., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2007). Piecing Together the Student Success Puzzle: Research, Propositions, and Recommendations. ASHE Higher Education Report, 32(5), 1–182. https://doi.org/10.1002/aehe.3205

Li, Qiujie, Rachel Baker, Mark Warschauer (2020). Using clickstream data to measure, understand, and support self-regulated learning in online courses. *The Internet and Higher Education*, Volume 45, 100727, ISSN 1096-7516, https://doi.org/10.1016/j.iheduc.2020.100727.

Lim, Janine M. (2016). Predicting successful completion using student delay indicators in undergraduate self-paced online courses. *Distance Education*, 37:3, 317-332, DOI: 10.1080/01587919.2016.1233050

McNamara, D.S., Crossley, S.A. & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behav Res* 45, 499–515. https://doi.org/10.3758/s13428-012-0258-1

Murphy, Robert, Jeremy Roschelle, Mingyu Feng & Craig A. Mason (2020) Investigating Efficacy, Moderators and Mediators for an Online Mathematics Homework Intervention, *Journal of Research on Educational Effectiveness*, 13:2, 235-270, DOI: 10.1080/19345747.2019.1710885

National Student Clearinghouse (2021). COVID-19: Stay Informed with the Latest Enrollment Information, November 18, 2021. Retrieved on June 13th, 2022: https://nscresearchcenter.org/stay-informed/

Page LC, Gehlbach H. How an Artificially Intelligent Virtual Assistant Helps Students Navigate the Road to College. *AERA Open*. October 2017. doi:10.1177/2332858417749220

Park, J., Yu, R., Rodriguez, F., Baker, R., Smyth, P., & Warschauer, M. (2018). Understanding Student Procrastination via Mixture Models. Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018), 187–197.

Treaster, J. B. (2017, February 2). Will you graduate? Ask big data. The New York Times. https://www.nytimes.com/2017/02/02/education/edlife/will-you-graduate-ask-big-data.html?_r=1

Yu, Renzhe, Qiujie Li, Christian Fischer, Shayan Doroudi and Di Xu (2020) "Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data" In: Proceedings of The 13th International Conference on Educational Data Mining., Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 292 - 301

Table 1: Summary statistics of analytic sample

		2+ terms		1st term	
	Full Analytic Sample	Training Set	Validation Set	Training Set	Validation Set
	(1)	(2)	(3)	(4)	(5)
Panel A: Sample sizes					
Student x course x section observations	1,173,878	698,361	270,664	181,673	23,180
Unique students	226,784	164,245	87,022	63,603	8,196
Unique courses	2,646	2,246	1,989	1,399	966
Unique course sections	63,994	47,145	16,645	33,942	8,284
Panel B: Student characteristics					
White	51.6%	52.3%	50.7%	48.9%	50.0%
Black	19.5%	19.0%	18.9%	19.1%	22.7%
Hispanic	13.0%	13.9%	14.4%	13.1%	5.9%
Asian	8.0%	7.9%	8.1%	8.6%	8.0%
Other	5.4%	5.2%	5.4%	6.1%	6.1%
Female	58.9%	59.9%	60.2%	55.2%	58.4%
Age	24.8	25.2	25.4	22	27
Cumulative GPA (at start of the target term)	2.91	2.91	2.88		N/A
Credits accumulated prior to target term	32.6	32.6	32.5		N/A
Panel C: Course characteristics					
200-level	50.1%	48.5%	50.8%	39.3%	39.2%
Average course-level enrollment	153.9	156.4	147.7	257	276.2
Average section-level enrollment	18.3	18.6	17.7	20.6	20.4
Applied Technologies	18.0%	16.7%	16.8%	14.2%	13.9%
Arts	9.7%	9.9%	10.1%	10.3%	7.8%

(Table 1, continued)

Business/Finance	7.3%	7.4%	7.9%	8.6%	8.8%
Engineering	21.8%	21.0%	22.1%	20.4%	24.0%
Foreign Languages	2.6%	2.8%	2.8%	3.6%	3.6%
Humanities	6.9%	7.2%	7.6%	9.3%	9.4%
Mathematics	1.0%	1.2%	1.1%	1.7%	2.1%
Medical Sciences	19.9%	20.2%	17.5%	14.4%	10.8%
Natural Sciences	3.3%	3.4%	3.6%	4.3%	5.4%
Social Sciences	9.6%	10.2%	10.4%	13.3%	14.3%

Notes: student race and sex are averaged across unique students, while student age and prior academic history are averaged across unique student x term cells. Course characteristics are averaged at the course-level (with the exception of section-level enrollment, which is averaged at the course x section level). The unit of observation in the prediction model is student x term x course x section. For both the 1st term and 2+ terms samples, the Training set consists of data from the Summer 2019, Fall 2019, Summer 2020, and Fall 2020 terms; the validation set contains observations from the Spring 2021 term.

Table 2: C-statistics of models using different predictor subcategory combinations

Panel A: Model with 2+ terms observations

Predictor categories	# predictors	C-statistic
All	329	0.872
All Admin	279	0.855
Non-course-specific academic records + demographic	41	0.843
Course-specific + course-subject-specific + instructor-related	238	0.778
All LMS	50	0.778
Early-term target course + early-term concurrent	21	0.751
Early-term target course	12	0.733
Early-term concurrent	9	0.604
Prior early-term + prior full-term	29	0.713
Prior early-term	13	0.665
Prior full-term	16	0.709

Panel B: Model with 1st term observations

Predictor categories	# predictors	C-statistic
All	80	0.825
All Admin	59	0.728
Course-specific + instructor-related	34	0.602
Non-course-specific academic records + demographic	25	0.664
All LMS (early-term target course + concurrent)	21	0.775
Early-term target course	12	0.754
Early-term concurrent	9	0.595

Notes: each row corresponds to a separate random forest prediction model using the set of predictors indicated in the first column. All prior LMS predictors and course-subject-specific predictors are not available for 1st term observations; some course-specific and non-course-specific academic records are unavailable for 1st term observations.

Table 3: Feature importance from models using Full set of predictors (Admin + LMS)

Panel A: Model with 2+ terms observations (329 predictors)

Ranking	Predictor	Category	Subcategory	FI Score
1	% prior attempted credits "Withdrawn"	Admin	Non-course-specific academic records	0.075
2	Cumulative GPA	Admin	Non-course-specific academic records	0.04
3	Total # clicks in the 1st quarter	LMS	Early-term	0.036
4	Term GPA of the last term prior to the target term	Admin	Non-course-specific academic records	0.036
5	# credits attempted in the target term	Admin	Non-course-specific academic records	0.035
6	Total minutes spent in 1st quarter	LMS	Early-term	0.028
7	Term GPA of second-to-last term prior to the target term	Admin	Non-course-specific academic records	0.024
8	% prior attempted credits completed	Admin	Non-course-specific academic records	0.023
9	Average grade assigned by the instructor in the target course	Admin	Instructor-related	0.022
10	Average historical grade in the target course	Admin	Course-specific	0.022
11	# original discussion forum posts created in 1st quarter	LMS	Early-term target course	0.022
12	# assignment submissions in the 1st quarter	LMS	Early-term target course	0.022
13	Stddev of term-level credit completion rate	Admin	Non-course-specific academic records	0.02
14	# discussion forum replies in 1st quarter	LMS	Early-term target course	0.018
15	Slope of term-level GPA in prior terms	Admin	Non-course-specific academic records	0.015
16	Average position of posts in forum thread (original post = 1) in 1st quarter	LMS	Early-term target course	0.014
17	Average # of words per discussion forum thread in 1st quarter	LMS	Early-term target course	0.013
18	Total # of clicks in the 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.013
19	Total credits accumulated prior to target term	Admin	Non-course-specific academic records	0.013
20	Stddev of session lengths in 1st quarter	LMS	Early-term target course	0.013
21	Average historical grade in the concurrent courses	Admin	Course-specific	0.013
22	% on-time assignment submissions in the 1st quarter	LMS	Early-term target course	0.012
23	Total minutes spent in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.012
24	Average session length in 1st quarter	LMS	Early-term target course	0.012
25	Stddev of session lengths in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.011
26	Average session length in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.011
27	Slope of credits attempted in prior terms	Admin	Non-course-specific academic records	0.01
28	Enrollment in target course section	Admin	Course-specific	0.01
29	Credits attempted in last term (prior to target term)	Admin	Non-course-specific academic records	0.009
30	Age at time of target course enrollment	Admin	Demographic	0.009

Panel B: Model with 1st term observations (80 predictors)

Ranking	Predictor	Category	Subcategory	FI Score
1	Total # clicks in the 1st quarter	LMS	Early-term	0.109
2	# credits attempted in the target term	Admin	Non-course-specific academic records	0.079
3	Total minutes spent in 1st quarter	LMS	Early-term	0.073
4	# discussion forum replies in 1st quarter	LMS	Early-term target course	0.049
5	Average historical grade in the target course	Admin	Course-specific	0.048
6	# original discussion forum posts created in 1st quarter	LMS	Early-term target course	0.047
7	# assignment submissions in the 1st quarter	LMS	Early-term target course	0.044
8	Average grade assigned by the instructor in the target course	Admin	Instructor-related	0.04
9	Average position of posts in forum thread (original post = 1) in 1st quarter	LMS	Early-term target course	0.037
10	Total # of clicks in the 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.035
11	% on-time assignment submissions in the 1st quarter	LMS	Early-term target course	0.035
12	Average # of words per discussion forum thread in 1st quarter	LMS	Early-term target course	0.03
13	Total minutes spend in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.028
14	Stddev of session lengths in 1st quarter	LMS	Early-term target course	0.028
15	Target course is in a Summer term	Admin	Course-specific	0.028
16	Average historical grade in the concurrent courses	Admin	Course-specific	0.026
17	Average session length in 1st quarter	LMS	Early-term target course	0.024
18	Stddev of session lengths in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.022
19	Age at time of target course enrollment	Admin	Demographic	0.022
20	Average session length in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.021
21	Enrollment in target course section	Admin	Course-specific	0.019
22	Average # assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.017
23	% attempted credits during target term that are the 200-level	Admin	Non-course-specific academic records	0.014
24	% on-time assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.011
25	% attempted credits during target term that are online	Admin	Non-course-specific academic records	0.011
26	Enrolled in any development courses in the target term	Admin	Non-course-specific academic records	0.009
27	Enrolled in a transfer-oriented associate degree program	Admin	Non-course-specific academic records	0.008
28	% attempted credits during target term that are evening	Admin	Non-course-specific academic records	0.008
29	Student is taking concurrent courses with historic grades available	Admin	Course-specific	0.004
30	Target course section is online	Admin	Course-specific	0.004

Notes: we calculate the FI (Feature Importance) Score using mean decrease in importance. The predictors that rank in the top 10 in both models are highlighted in orange.

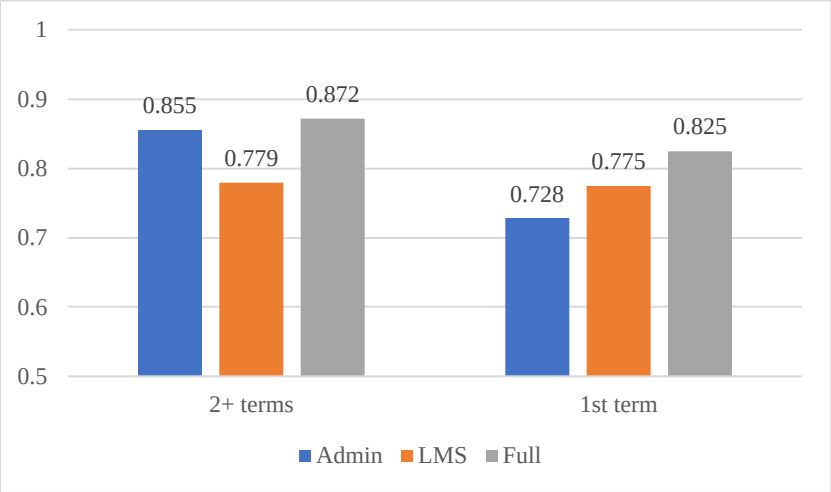
Table 4: Summary statistics for course-specific samples

	ENG 111		ENG 112		BIO 101		MTH 154		MTH 161	
	Training (1)	Validation (2)	Training (3)	Validation (4)	Training (5)	Validation (6)	Training (7)	Validation (8)	Training (9)	Validation (10)
Average number of course sections per term	584	482	253	602	338	391	218	254	185	213
Average total course enrollment per term	11,310	8,973	4,991	11,466	4,816	5,237	4,857	5,736	3,969	4,075
Average section-level enrollment per term	18.4	18.6	19.6	19.1	21.4	22.7	21.3	22.6	20.5	19.2
During target term…										
Average grade (GPA points)	2.45	2.21	2.72	2.74	2.54	2.51	2.34	2.26	2.21	2.27
Share D/F/W	31.3%	40.7%	27.8%	25.6%	26.1%	28.6%	35.3%	38.9%	41.4%	42.0%
% of analytic sample that are 1st term	56.2%	31.3%	7.0%	1.6%	20.0%	6.8%	28.9%	11.7%	34.7%	11.4%
Number of student x section observations	45,232	8,979	19,986	11,471	29,925	8,881	19,437	5,738	15,901	4,080

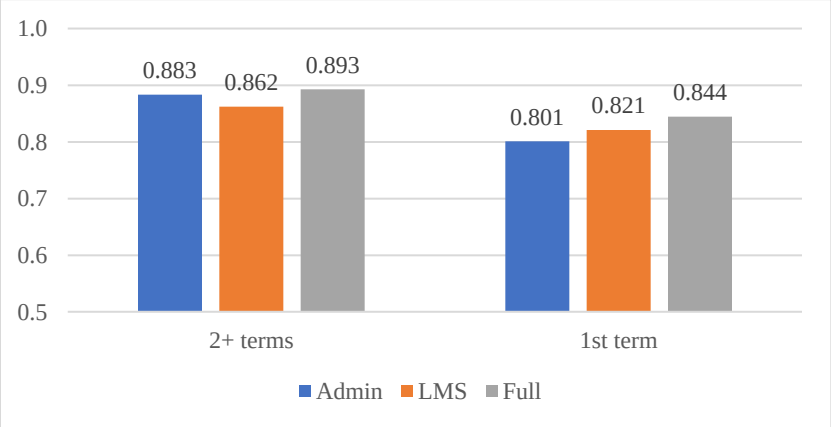
Notes: the first three rows are averaged at the course x term level, or the section x term level, as indicated. The four rows under the "*During target term…"* heading are averaged across student x section observations. Only course sections that are in the analytic sample are included in these calculations.

Figure 1: Performance of prediction model, by category of predictors and sample of students

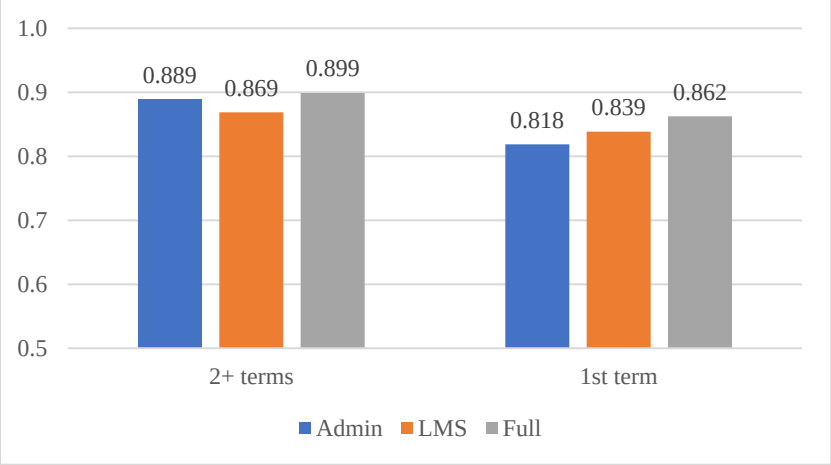
Plot A: C-statistic



Plot B: Precision



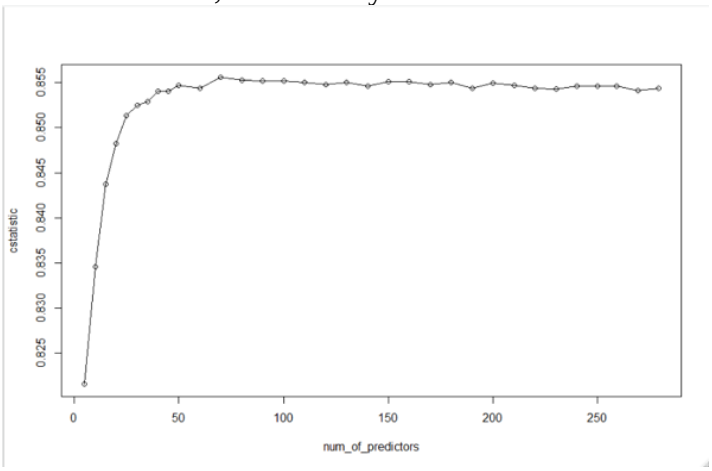
Plot C: Recall



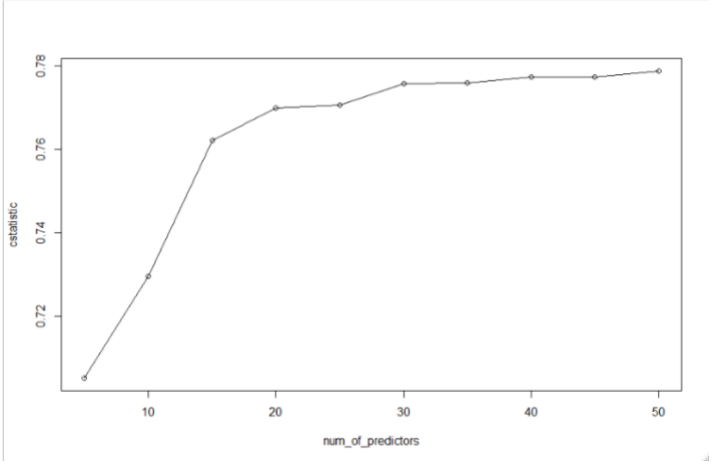
Notes: each bar corresponds to a separate random forest prediction model using the set of predictors indicated by the color of the bar, and observations from the sample of students based on academic history indicated by the x-axis label.

Figure 2: Relationship between c-statistic and number of predictors

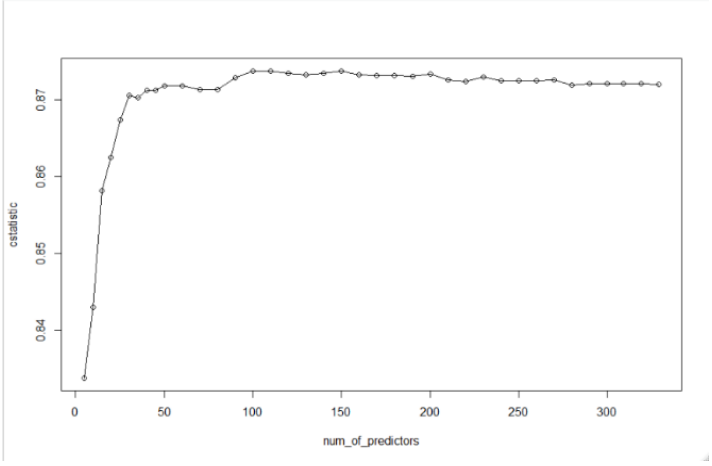
Plot A: 2+ terms, Admin only



Plot B: 2+ terms, LMS only

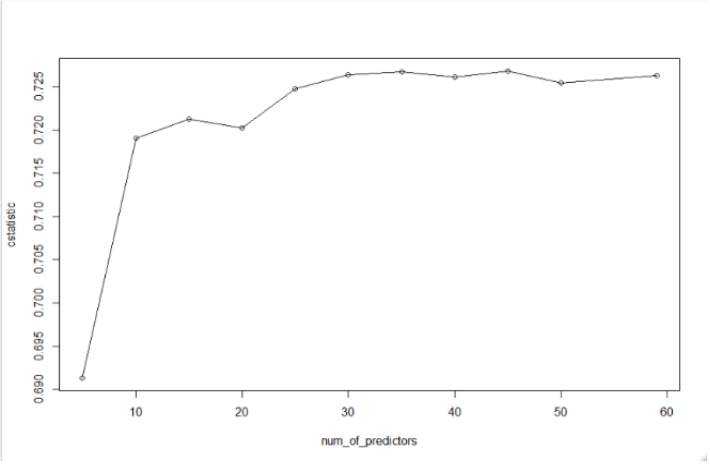


Plot C: 2+ terms, Admin + LMS

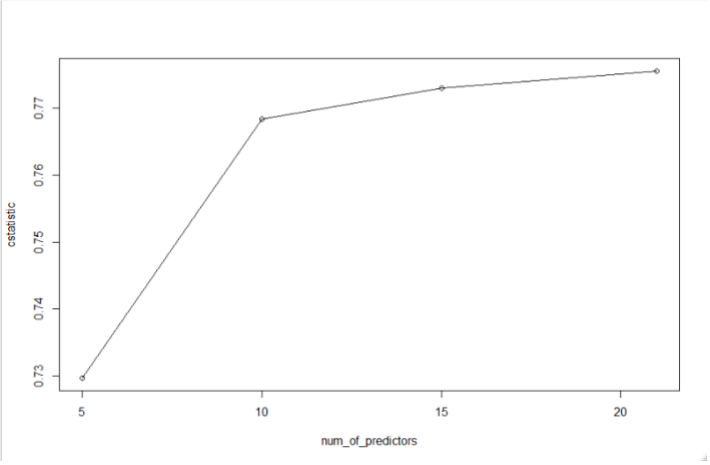


(Figure 2, continued)

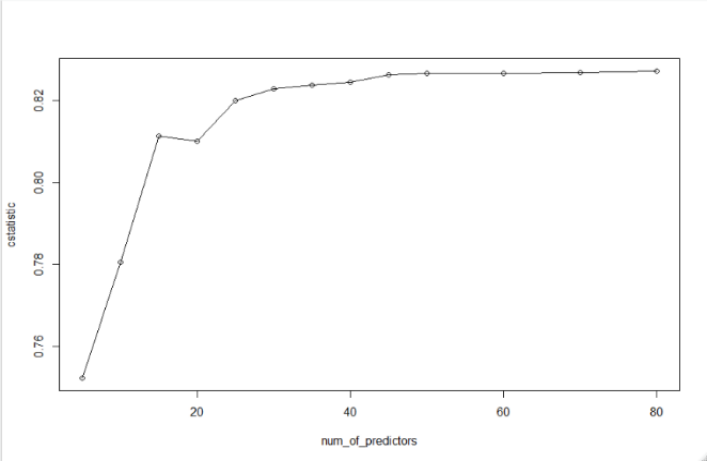
Plot D: 1st term, Admin only



Plot E: 1st term, LMS only

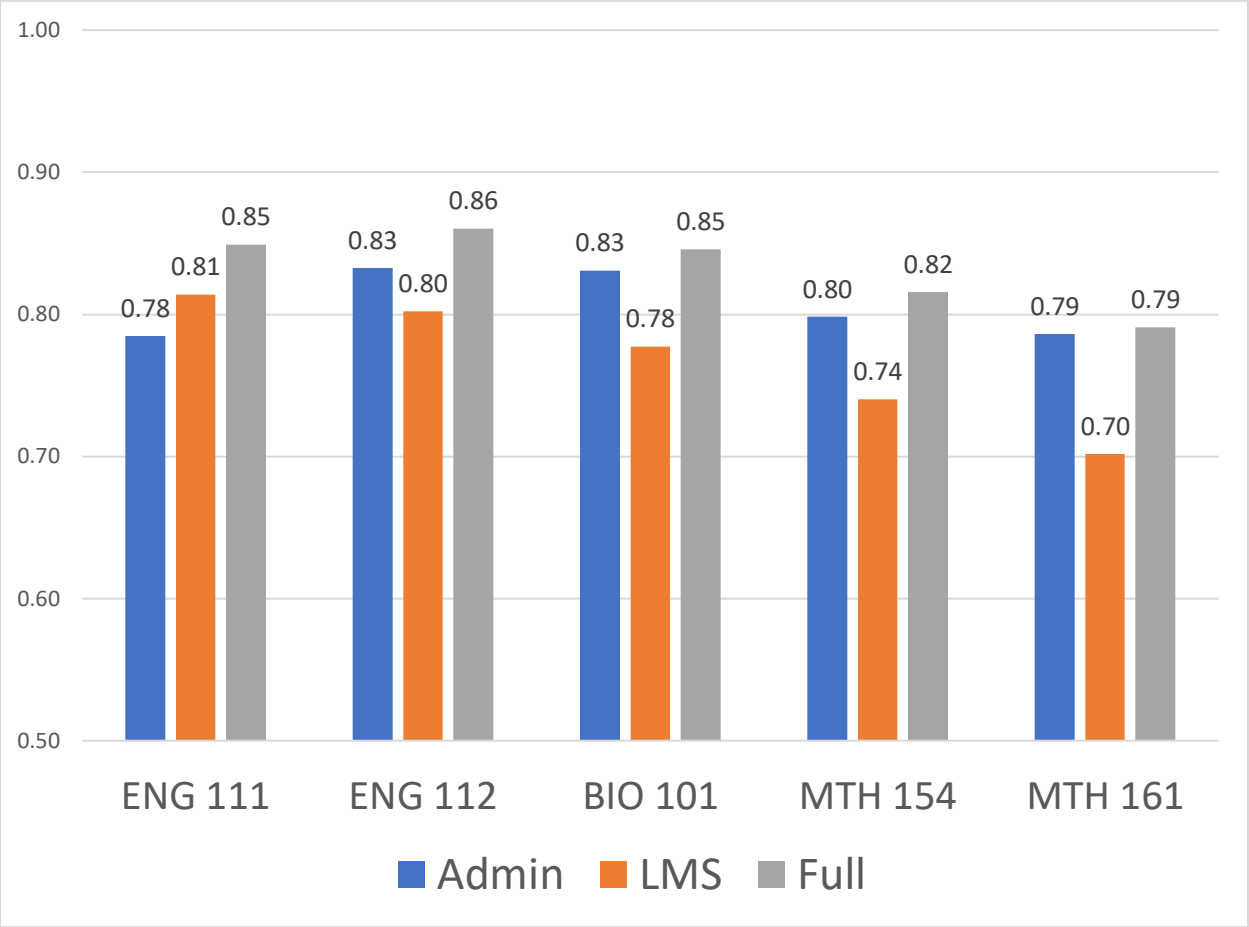


Plot F: 1st term, Admin + LMS



Notes: to construct each of these plots, we use calculate the c-statistic for the five most important predictors (as determined by feature importance score) and add predictors in multiples of five.

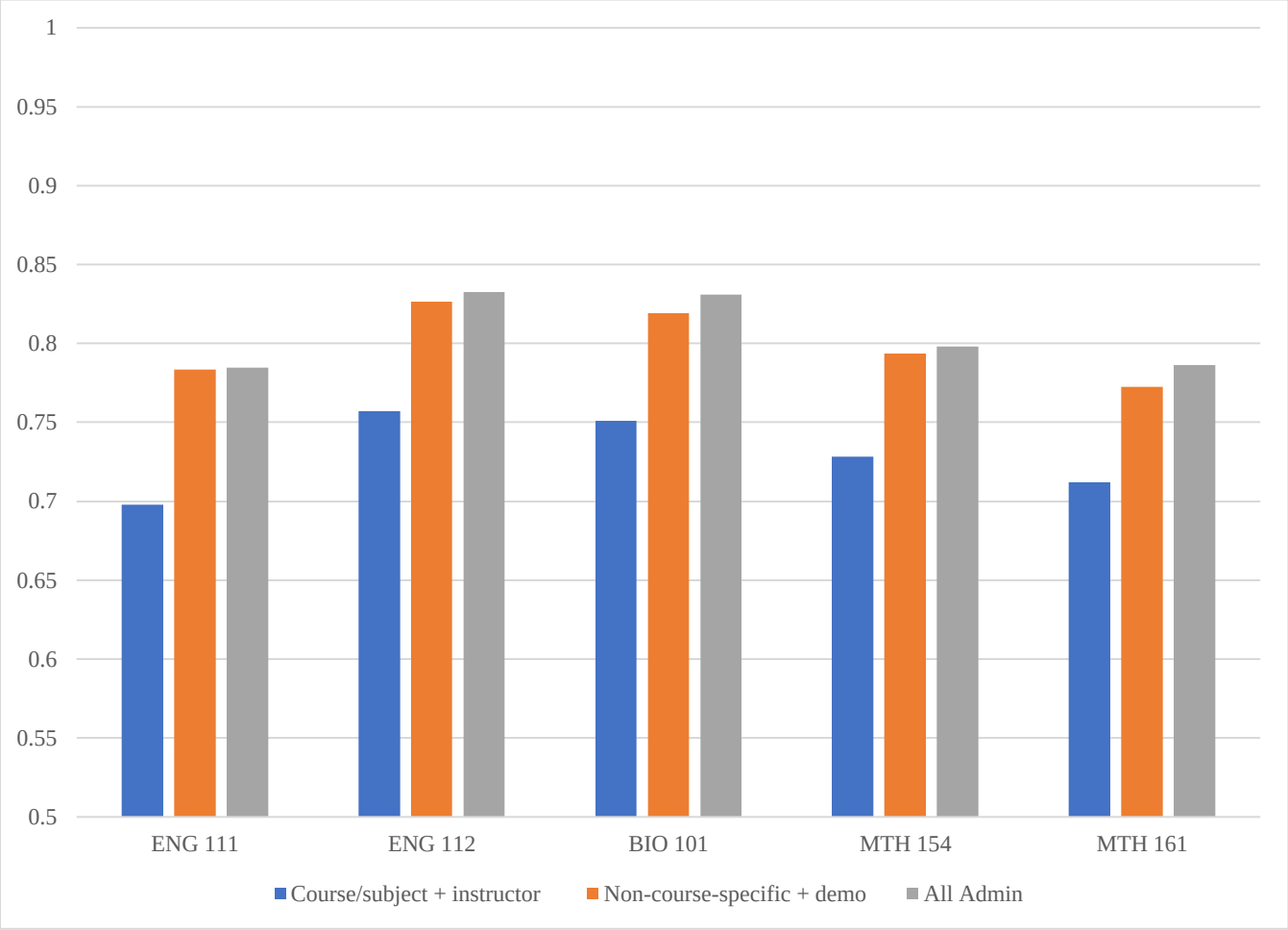
Figure 3: C-statistics for course-specific models, by predictor category



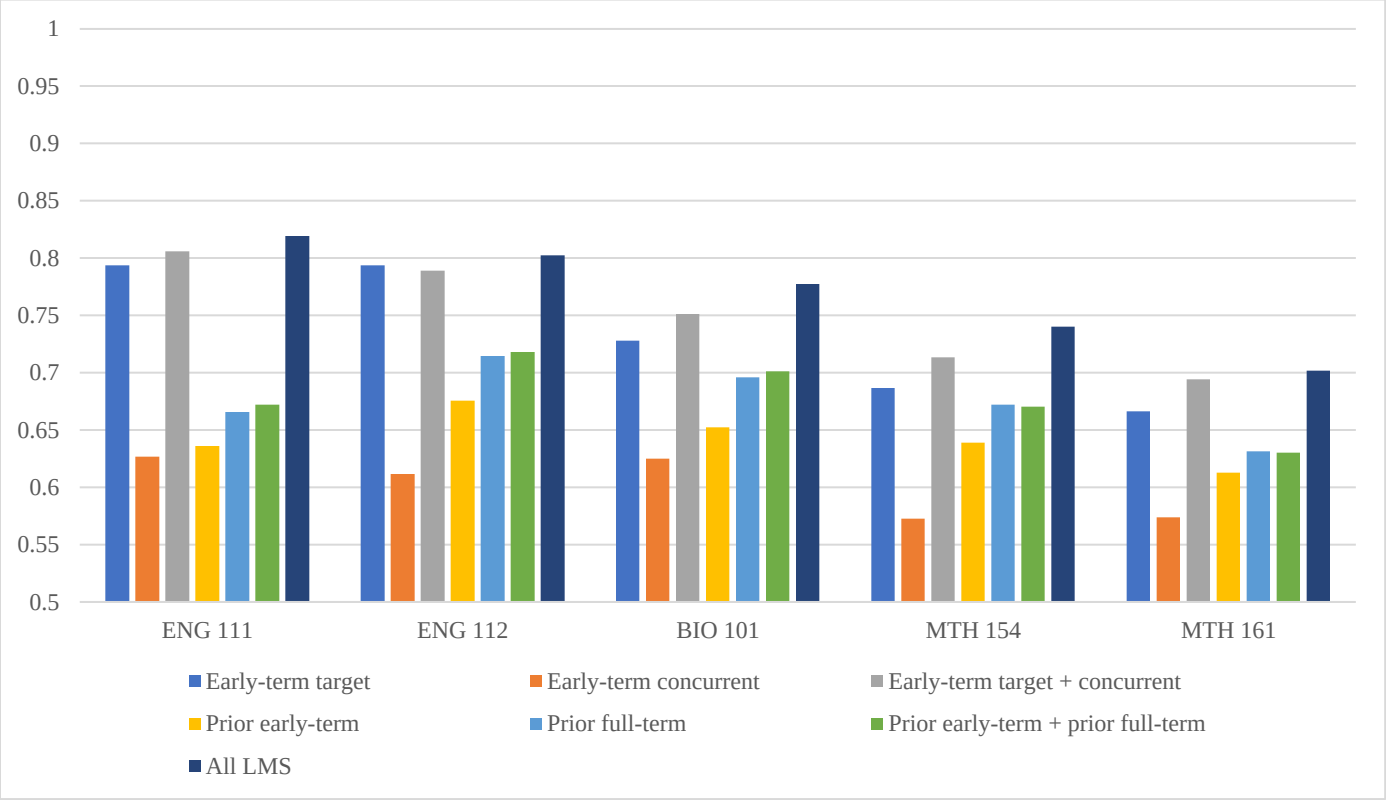
Notes: each bar corresponds to a separate random forest prediction model using the set of predictors indicated by the color of the bar, and observations from the course indicated by the x-axis label.

Figure 4: C-statistics of course-specific models using different predictor subcategory combinations

Panel A: Admin predictor categories

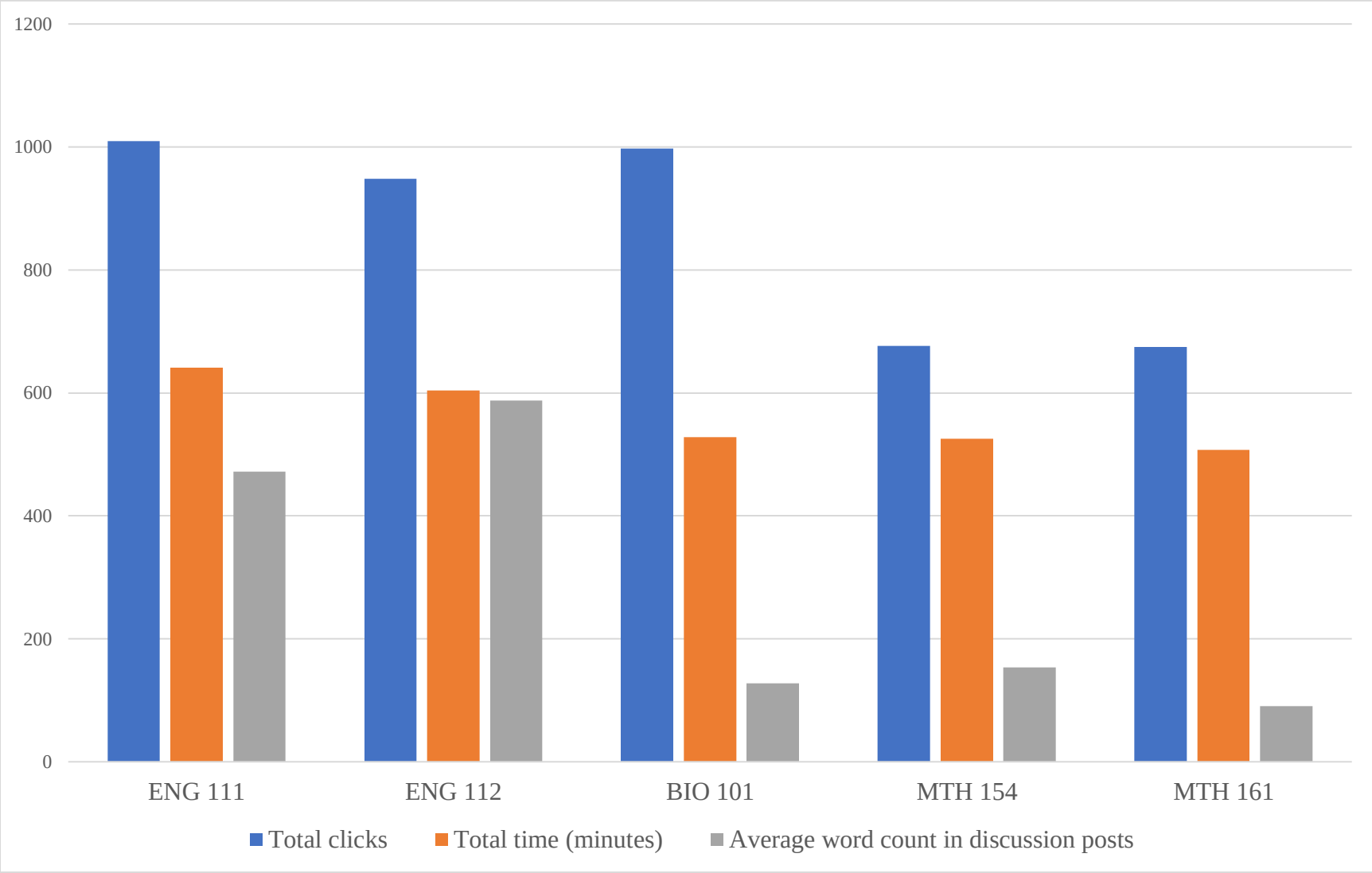


(Figure 4, continued)
Panel B: LMS predictor categories



Notes: each bar corresponds to a separate random forest prediction model using the set of predictors indicated on the x-axis, using observations from the course indicated by the bar color. The predictor categories on the x-axis are abbreviated names for the same categories represented in Table 3: "Course/subject + instructor" includes 238 predictors; "Non-course-specific + demo" includes 41 predictors; "Early-term target" includes 12 predictors; "Early-term target + concurrent" includes 21 predictors; "Prior early-term" includes 13 predictors; and "Prior full-term" includes 16 predictors.

Figure 5: Average values of top LMS predictors, by course



Note: each of these LMS predictors refers to the early-term (1st quarter) portion of the target course.

Appendix: comparison of model performance for online versus in-person observations.

While all VCCS courses can use Canvas' LMS features, online courses typically require more LMS interaction with the student.[1] We show the number of online versus in-person observations in Panel A of Appendix Table A5. The majority (73.8 percent) of the student-by-course section observations in our analytic sample are online, which is driven in some part by the inclusion of Fall 2020 and Spring 2021 in our analytic sample during which most coursework was still online due to the COVID pandemic. Indeed, 94.4 percent of observations in the validation set, which consists entirely of Spring 2021 observations, are online. Online enrollment in the validation set is over 99 percent for ENG 111, ENG 112, and BIO 101.

Panel B of Appendix Table A5 shows that for most (but not all) of the early target term LMS predictors, the online observations have considerably higher mean values. For example, the average total minutes spent logged in was 655 minutes for online observations and 279 for in-person observations. However, assignment submission data is available for more in-person observations (57.3 percent) compared to online (49.2 percent).

Given these differences, we explore whether the added value of LMS predictors differs for online versus in-person observations. To do so, we calculate separate c-statistics online versus in-person subsets of the validation sample. We present these results in Appendix Table A6. The c-statistics for the online observations closely mirrors the results in Figure 1. However, we observe a significant drop in the c-statistic for the LMS-only model for the in-person observations, equal to 0.647 for the 1st term sample and 0.708 for the 2+ terms sample. Interestingly, the in-person c-statistic is higher for Admin-only models and is only slightly lower for the Full predictor models (compared to Figure 1). These results suggest that LMS-only models are of significantly less value

[1] We classify all hybrid courses, which VCCS defines as having 50-99% of course instruction occurring online, as online courses.

for in-person observations; however, given the validation sample from Spring 2021 contains only 5.6% in-person observations, we caution against drawing strong conclusions from this particular comparison.

Because the training set contains a significantly larger share of in-person observations (31.5 percent for 2+ terms sample and 37.2 percent for 1st term sample), and because the computation of feature importance scores are not reliant on the validation sample, we build modality-specific models with the full set of predictors and compare the feature importance scores in Appendix Table A7. We find that the LMS predictors have higher feature importance for the online observations compared with the in-person observations. Comparing Panels A and B which show the top 30 predictors for the modality-specific models using the 2+ terms sample, respectively, we see that there are four LMS predictors in the top 10 predictors for online observations, but only two LMS predictors in the top 10 for in-person observations. Similarly, the top rated LMS predictor has a ranking of two (i.e. second most important feature) for online observations, but a ranking of seven for in-person observations. We find similar patterns when comparing Panels C and D which show the same set of results using the 1st term sample.

Appendix Table A1: Summary statistics for full VCCS population during analytic sample observation window

		2+ terms		1st term	
	Full population	SU19, FA19, SU20, FA20	SP21	SU19, FA19, SU20, FA20	SP21
	(1)	(2)	(3)	(4)	(5)
Panel A: Sample sizes					
Student x course x section observations	1,448,506	869,583	326,783	224,472	27,397
Unique students	251,846	183,186	99,355	77,717	9,749
Unique courses	2,856	2,467	2,130	1,536	1,039
Unique course sections	85,401	62,020	22,060	40,606	9,301
Panel B: Student characteristics					
White	52.7%	53.6%	52.4%	50.6%	51.1%
Black	18.8%	18.1%	18.0%	18.0%	22.4%
Hispanic	12.6%	13.4%	13.7%	12.9%	5.6%
Asian	7.9%	7.9%	7.9%	8.5%	7.6%
Other	5.4%	5.2%	5.4%	6.1%	6.0%
Female	58.7%	59.8%	60.2%	55.8%	57.8%
Age	24	24.5	24.3	21.3	27.6
Cumulative GPA (at start of the target term)	2.96	2.96	2.95	N/A	
Credits accumulated prior to target term	31	30.9	31.3	N/A	
Panel C: Course characteristics					
200-level	49.9%	48.4%	50.0%	39.3%	39.7%
Average course-level enrollment	170	172.3	166.3	289.5	306.3
Average section-level enrollment	16.3	16.8	15.5	19.7	19.9

(Table A1, continued)

Applied Technologies	19.0%	18.1%	17.7%	15.6%	14.4%
Arts	9.5%	9.9%	9.9%	10.2%	7.9%
Business/Finance	7.1%	7.1%	7.8%	8.3%	8.4%
Engineering	21.6%	20.8%	22.3%	20.6%	24.0%
Foreign Languages	2.9%	3.0%	3.2%	3.5%	3.7%
Humanities	6.8%	7.1%	7.4%	8.8%	9.1%
Mathematics	1.0%	1.2%	1.0%	1.6%	19.2%
Medical Sciences	19.0%	19.3%	16.9%	14.6%	11.8%
Natural Sciences	3.5%	3.6%	3.7%	4.2%	5.1%
Social Sciences	9.5%	9.9%	10.2%	12.6%	13.7%

Notes: student race and sex are averaged across unique students, while student age and prior academic history are averaged across unique student x term cells. Course characteristics are averaged at the course-level (with the exception of section-level enrollment, which is averaged at the course x section level). The unit of observation in the prediction model is student x term x course x section. For both the 1st term and 2+ terms samples, the Training set consists of data from the Summer 2019, Fall 2019, Summer 2020, and Fall 2020 terms; the validation set contains observations from the Spring 2021 term.

Appendix Table A2: Full list of predictors

Predictor description	Category	Sub-category	Available for 1st term observations
Average historical grade in the target course	Admin	Course-specific	X
Average historical grade in the concurrent courses	Admin	Course-specific	X
23 college indicators	Admin	Course-specific	X
Course meeting time is in the evening	Admin	Course-specific	X
Student is taking concurrent courses with historic grades available	Admin	Course-specific	X
Student took the target course's prerequisites (if applicable)	Admin	Course-specific	X
Target course is 200-level	Admin	Course-specific	X
Target course section is online	Admin	Course-specific	X
Average grade in target course's prerequisites	Admin	Course-specific	X
Enrollment in target course section	Admin	Course-specific	X
Target course is in a Summer term	Admin	Course-specific	X
Student has previously taken the target course	Admin	Course-specific	
Student's average prior grade in the target course (if repeating the course)	Admin	Course-specific	
Has taken prior Arts courses (target course = X subject)	Admin	Course-subject-specific	
Average grade in prior Arts courses (target course = X subject)	Admin	Course-subject-specific	
Has taken prior Business/Finance courses (target course = X subject)	Admin	Course-subject-specific	
Average grade in prior Business/Finance courses (target course = X subject)	Admin	Course-subject-specific	
Has taken prior Engineering courses (target course = X subject)	Admin	Course-subject-specific	
Average grade in prior Engineering courses (target course = X subject)	Admin	Course-subject-specific	
Has taken prior Foreign Languages courses (target course = X subject)	Admin	Course-subject-specific	
Average grade in prior Foreign Languages courses (target course = X subject)	Admin	Course-subject-specific	

(Table A2, continued)

Has taken prior Humanities courses (target course = X subject)	Admin	Course-subject-specific	
Average grade in prior Humanities courses (target course = X subject)	Admin	Course-subject-specific	
Has taken prior Medical Sciences courses (target course = X subject)	Admin	Course-subject-specific	
Average grade in prior Medical Sciences courses (target course = X subject)	Admin	Course-subject-specific	
Has taken prior Mathematics courses (target course = X subject)	Admin	Course-subject-specific	
Average grade in prior Mathematics courses (target course = X subject)	Admin	Course-subject-specific	
Has taken prior Applied Technologies courses (target course = X subject)	Admin	Course-subject-specific	
Average grade in prior Applied Technologies courses (target course = X subject)	Admin	Course-subject-specific	
Has taken prior Natural Sciences courses (target course = X subject)	Admin	Course-subject-specific	
Average grade in prior Natural Sciences courses (target course = X subject)	Admin	Course-subject-specific	
Has taken prior Social Sciences courses (target course = X subject)	Admin	Course-subject-specific	
Average grade in prior Social Sciences courses (target course = X subject)	Admin	Course-subject-specific	
Age at time of target course enrollment	Admin	Demographic	X
Instructor works full-time at VCCS	Admin	Instructor-related	X
Instructor has taught the target course in the past	Admin	Instructor-related	X
Average grade assigned by the instructor in the target course	Admin	Instructor-related	X
Instructor has been teaching at VCCS for 6+ years	Admin	Instructor-related	X
15 field of study indicators (2 digit CIPs)	Admin	Non-course-specific academic records	X
Enrolled in a transfer-oriented associate degree program	Admin	Non-course-specific academic records	X
Enrolled in an occupation-oriented associate degree program	Admin	Non-course-specific academic records	X
Enrolled in a certificate program	Admin	Non-course-specific academic records	X
Enrolled in any development courses in the target term	Admin	Non-course-specific academic records	X
# credits attempted in the target term	Admin	Non-course-specific academic records	X
% attempted credits during target term that are evening	Admin	Non-course-specific academic records	X
% attempted credits during target term that are the 200-level	Admin	Non-course-specific academic records	X
% attempted credits during target term that are online	Admin	Non-course-specific academic records	X
Total credits accumulated prior to target term	Admin	Non-course-specific academic records	
Cumulative GPA	Admin	Non-course-specific academic records	

(Table A2, continued)

Credits attempted in last term (prior to target term)	Admin	Non-course-specific academic records	
Slope of credits attempted in prior terms	Admin	Non-course-specific academic records	
Ever dually enrolled	Admin	Non-course-specific academic records	
Slope of term-level GPA in prior terms	Admin	Non-course-specific academic records	
Missing indicator for term GPA of the last term	Admin	Non-course-specific academic records	
Missing indicator for term GPA of the second-to-last term	Admin	Non-course-specific academic records	
# terms enrolled at VCCS prior to target term	Admin	Non-course-specific academic records	
% prior attempted credits completed	Admin	Non-course-specific academic records	
% prior attempted credits that were developmental courses	Admin	Non-course-specific academic records	
% prior attempted credits "Incomplete"	Admin	Non-course-specific academic records	
# stop-out terms between initial enrollment and target term	Admin	Non-course-specific academic records	
% prior attempted credits "Withdrawn"	Admin	Non-course-specific academic records	
Stddev of term-level credit completion rate	Admin	Non-course-specific academic records	
Term GPA of the last term prior to the target term	Admin	Non-course-specific academic records	
Term GPA of second-to-last term prior to the target term	Admin	Non-course-specific academic records	
# assignment submissions in the 1st quarter	LMS	Early-term target course	X
# assignment submissions in the 1st quarter data is available	LMS	Early-term target course	X
On-time assignment submissions in 1st quarter data is available	LMS	Early-term target course	X
% on-time assignment submissions in the 1st quarter	LMS	Early-term target course	X
Average session length in 1st quarter	LMS	Early-term target course	X
Stddev of session lengths in 1st quarter	LMS	Early-term target course	X
Total # clicks in the 1st quarter	LMS	Early-term target course	X
Total minutes spent in 1st quarter	LMS	Early-term target course	X
Average depth (position) of posts within a discussion forum thread (original post = 1) in 1st quarter	LMS	Early-term target course	X
Average # words per discussion forum thread in 1st quarter	LMS	Early-term target course	X
# original discussion forum posts created in 1st quarter	LMS	Early-term target course	X
# discussion forum replies in 1st quarter	LMS	Early-term target course	X
Average # assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	X

(Table A2, continued)

# assignment submissions in the 1st quarter is available for concurrent courses	LMS	Early-term concurrent	X
On-time assignment submissions in 1st quarter is available for concurrent courses	LMS	Early-term concurrent	X
% on-time assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	X
Average session length in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	X
Taking concurrent courses with LMS measures available	LMS	Early-term concurrent	X
Stddev of session lengths in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	X
Total # clicks in the 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	X
Total minutes spend in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	X
Average # assignment submissions in 1st quarter, prior courses	LMS	Prior early-term	
# assignment submissions in 1st quarter is available for prior courses	LMS	Prior early-term	
On-time assignment submissions in 1st quarter is available for prior courses	LMS	Prior early-term	
% on-time assignment submissions in 1st quarter, averaged across prior courses	LMS	Prior early-term	
Average session length in 1st quarter, averaged across prior courses	LMS	Prior early-term	
Prior courses taken by the student have 1st quarter LMS measures available	LMS	Prior early-term	
Stddev of session lengths in 1st quarter, averaged across prior courses	LMS	Prior early-term	
Total # clicks in the 1st quarter, averaged across prior courses	LMS	Prior early-term	
Total minutes spend in 1st quarter, averaged across prior courses	LMS	Prior early-term	
Average depth of posts within a discussion forum thread in 1st quarter, averaged across prior courses	LMS	Prior early-term	
Average # words per discussion forum thread in 1st quarter, averaged across prior courses	LMS	Prior early-term	
# original discussion forum posts created in 1st quarter, averaged across prior courses	LMS	Prior early-term	
# discussion forum replies in 1st quarter, averaged across prior courses	LMS	Prior early-term	
Average # assignment submissions in full term, prior courses	LMS	Prior full-term	
# assignment submissions in full term is available for prior courses	LMS	Prior full-term	
On-time assignment submissions in full term is available for prior courses	LMS	Prior full-term	
% on-time assignment submissions in full term, averaged across prior courses	LMS	Prior full-term	
Average session length in full term, averaged across prior courses	LMS	Prior full-term	
Prior courses taken by the student have full term LMS measures available	LMS	Prior full-term	
Stddev of session lengths in full term, averaged across prior courses	LMS	Prior full-term	

(Table A2, continued)

Total # days with any activity in full term, averaged across prior courses	LMS	Prior full-term
Total # weeks with any activity in full term, averaged across prior courses	LMS	Prior full-term
Total # clicks in thefull term, averaged across prior courses	LMS	Prior full-term
Total # sessions (i.e. logins) in the full term, averaged across prior courses	LMS	Prior full-term
Total minutes spend in full term, averaged across prior courses	LMS	Prior full-term
Average depth of posts within a discussion forum thread in full term, averaged across prior courses	LMS	Prior full-term
Average # words per discussion forum thread in full term, averaged across prior courses	LMS	Prior full-term
# original discussion forum posts created in full term, averaged across prior courses	LMS	Prior full-term
# discussion forum replies in full term, averaged across prior courses	LMS	Prior full-term

Appendix Table A3: Confusion matrices

Panel A: Admin Predictors

2+ terms				1st term			
Actual Grade	Pred(ABC)	Pred(DFW)	N	Actual Grade	Pred(ABC)	Pred(DFW)	N
A	108,416	6,418	114,834	A	8,658	1,386	10,044
B	54,697	8,379	63,076	B	3,197	956	4,153
C	24,183	8,489	32,672	C	1,592	646	2,238
D	7,850	4,484	12,334	D	689	402	1,091
F	12,680	16,683	29,363	F	2,253	1,470	3,723
W	4,339	14,046	18,385	W	395	1,536	1,931
N	212,165	58,499	270,664	N	16,784	6,396	23,180
% Observations with accurate prediction = 82.2%				% Observations with accurate prediction = 72.7%			

Panel B: LMS Predictors

2+ terms				1st term			
Actual Grade	Pred(ABC)	Pred(DFW)	N	Actual Grade	Pred(ABC)	Pred(DFW)	N
A	105,408	9,426	114,834	A	8,873	1,171	10,044
B	53,096	9,980	63,076	B	3,301	852	4,153
C	24,465	8,207	32,672	C	1,611	627	2,238
D	8,118	4,216	12,334	D	693	398	1,091
F	13,124	16,239	29,363	F	1,567	2,156	3,723
W	7,954	10,431	18,385	W	739	1,192	1,931
N	212,165	58,499	270,664	N	16,784	6,396	23,180
% Observations with accurate prediction = 79.0%				% Observations with accurate prediction = 75.6%			

Panel C: Full Predictors

2+ terms				1st term			
Actual Grade	Pred(ABC)	Pred(DFW)	N	Actual Grade	Pred(ABC)	Pred(DFW)	N
A	109,649	5,185	114,834	A	9,202	842	10,044
B	55,304	7,772	63,076	B	3,383	770	4,153
C	24,426	8,246	32,672	C	1,585	653	2,238
D	7,662	4,672	12,334	D	673	418	1,091
F	11,106	18,257	29,363	F	1,519	2,204	3,723
W	4,018	14,367	18,385	W	422	1,509	1,931
N	212,165	58,499	270,664	N	16,784	6,396	23,180
% Observations with accurate prediction = 83.7%				% Observations with accurate prediction = 79.0%			

Notes: each of the six groupings shows the confusion matrix for the prediction model that includes the set of predictors indicated by the column heading (Admin, LMS, Full), and the sample of observations based on timing (2+ terms, 1st term). Within a confusion matrix, each cell contains the number of observations in the validation sample who received a grade as indicated by the row labels, and was predicted to receive a grade as indicated by the column labels. Note that the column N contains the sum of observations within each row, while the row N contains the sum of observations within each column.

Appendix Table A4: Feature Importance for Course-specific models

Panel A: ENG 111

Ranking	Predictor	Category	Subcategory	FI Score
1	Total # clicks in the 1st quarter	LMS	Early-term	0.099
2	# original discussion forum posts created in 1st quarter	LMS	Early-term	0.066
3	Total minutes spent in 1st quarter	LMS	Early-term	0.063
4	# discussion forum replies in 1st quarter	LMS	Early-term	0.058
5	% on-time assignment submissions in the 1st quarter	LMS	Early-term	0.051
6	% prior attempted credits "Withdrawn"	Admin	Non-course-specific academic records	0.05
7	# credits attempted in the target term	Admin	Non-course-specific academic records	0.048
8	# assignment submissions in the 1st quarter	LMS	Early-term	0.039
9	Average position of posts in forum thread (original post = 1) in 1st quarter	LMS	Early-term	0.034
10	Average # words per discussion forum thread in 1st quarter	LMS	Early-term	0.027
11	Total # clicks in the 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.022
12	Cumulative GPA	Admin	Non-course-specific academic records	0.02
13	Total minutes spend in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.018
14	Stddev of session lengths in 1st quarter	LMS	Early-term	0.017
15	Age at time of target course enrollment	Admin	Demographic	0.016
16	Stddev of session lengths in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.015
17	Average session length in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.014
18	Average session length in 1st quarter	LMS	Early-term	0.014
19	Term GPA of the last term prior to the target term	Admin	Non-course-specific academic records	0.014
20	Average grade assigned by the instructor in the target course	Admin	Instructor-related	0.013
21	Average historical grade in the concurrent courses	Admin	Course-specific	0.012
22	Average # assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.012
23	% prior attempted credits completed	Admin	Non-course-specific academic records	0.012

24	Average historical grade in the target course	Admin	Course-specific	0.009
25	Enrollment in target course section	Admin	Course-specific	0.009
26	% on-time assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.009
27	Stddev of term-level credit completion rate	Admin	Non-course-specific academic records	0.007
28	Total credits accumulated prior to target term	Admin	Non-course-specific academic records	0.007
29	Credits attempted in last term (prior to target term)	Admin	Non-course-specific academic records	0.007
30	Term GPA of second-to-last term prior to the target term	Admin	Non-course-specific academic records	0.006

Panel B: ENG 112

Ranking	Predictor	Category	Subcategory	FI Score
1	% prior attempted credits "Withdrawn"	Admin	Non-course-specific academic records	0.068
2	# original discussion forum posts created in 1st quarter	LMS	Early-term	0.049
3	Total # clicks in the 1st quarter	LMS	Early-term	0.043
4	Cumulative GPA	Admin	Non-course-specific academic records	0.037
5	# discussion forum replies in 1st quarter	LMS	Early-term	0.036
6	Total minutes spent in 1st quarter	LMS	Early-term	0.035
7	# credits attempted in the target term	Admin	Non-course-specific academic records	0.035
8	Term GPA of the last term prior to the target term	Admin	Non-course-specific academic records	0.029
9	# assignment submissions in the 1st quarter	LMS	Early-term	0.027
10	Average position of posts in forum thread (original post = 1) in 1st quarter	LMS	Early-term	0.026
11	Average # words per discussion forum thread in 1st quarter	LMS	Early-term	0.023
12	Average grade in prior Humanities courses	Admin	Course-subject-specific	0.022
13	% on-time assignment submissions in the 1st quarter	LMS	Early-term	0.021
14	% prior attempted credits completed	Admin	Non-course-specific academic records	0.021
15	Stddev of term-level credit completion rate	Admin	Non-course-specific academic records	0.015
16	Average grade assigned by the instructor in the target course	Admin	Instructor-related	0.015
17	Term GPA of second-to-last term prior to the target term	Admin	Non-course-specific academic records	0.014
18	Total # clicks in the 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.014
19	Total minutes spend in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.014
20	Stddev of session lengths in 1st quarter	LMS	Early-term	0.013

21	Slope of term-level GPA in prior terms	Admin	Non-course-specific academic records	0.013
22	Average session length in 1st quarter	LMS	Early-term	0.013
23	Stddev of session lengths in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.012
24	Average session length in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.012
25	Average historical grade in the concurrent courses	Admin	Course-specific	0.011
26	Average # assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.01
27	Total # days with any activity in full term, averaged across prior courses	LMS	Prior full-term	0.01
28	Average grade in prior Social Sciences courses	Admin	Course-subject-specific	0.01
29	Total credits accumulated prior to target term	Admin	Non-course-specific academic records	0.01
30	Enrollment in target course section	Admin	Course-specific	0.01

Panel C: BIO 101

Ranking	Predictor	Category	Subcategory	FI Score
1	% prior attempted credits "Withdrawn"	Admin	Non-course-specific academic records	0.033
2	Total # clicks in the 1st quarter	LMS	Early-term	0.032
3	Cumulative GPA	Admin	Non-course-specific academic records	0.031
4	# credits attempted in the target term	Admin	Non-course-specific academic records	0.031
5	Total minutes spent in 1st quarter	LMS	Early-term	0.028
6	Term GPA of the last term prior to the target term	Admin	Non-course-specific academic records	0.028
7	# assignment submissions in the 1st quarter	LMS	Early-term	0.022
8	Average grade in prior Humanities courses	Admin	Course-subject-specific	0.021
9	Term GPA of second-to-last term prior to the target term	Admin	Non-course-specific academic records	0.019
10	% prior attempted credits completed	Admin	Non-course-specific academic records	0.018
11	% on-time assignment submissions in the 1st quarter	LMS	Early-term	0.018
12	Stddev of session lengths in 1st quarter	LMS	Early-term	0.018
13	Average historical grade in the concurrent courses	Admin	Course-specific	0.018
14	Total # clicks in the 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.017
15	Average grade assigned by the instructor in the target course	Admin	Instructor-related	0.017
16	Average session length in 1st quarter	LMS	Early-term	0.016
17	Total minutes spend in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.016

18	Stddev of session lengths in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.015
19	Slope of term-level GPA in prior terms	Admin	Non-course-specific academic records	0.015
20	Average session length in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.015
21	Average grade in prior Social Sciences courses	Admin	Course-subject-specific	0.015
22	Average historical grade in the target course	Admin	Course-specific	0.015
23	Stddev of term-level credit completion rate	Admin	Non-course-specific academic records	0.014
24	Age at time of target course enrollment	Admin	Demographic	0.014
25	Average # assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.014
26	Total credits accumulated prior to target term	Admin	Non-course-specific academic records	0.014
27	Slope of credits attempted in prior terms	Admin	Non-course-specific academic records	0.012
28	Credits attempted in last term (prior to target term)	Admin	Non-course-specific academic records	0.012
29	% prior attempted credits that were developmental courses	Admin	Non-course-specific academic records	0.011
30	# original discussion forum posts created in 1st quarter	LMS	Early-term	0.011

Panel D: MTH 154

Ranking	Predictor	Category	Subcategory	FI Score
1	% prior attempted credits "Withdrawn"	Admin	Non-course-specific academic records	0.071
2	Cumulative GPA	Admin	Non-course-specific academic records	0.059
3	Total # clicks in the 1st quarter	LMS	Early-term	0.058
4	Term GPA of the last term prior to the target term	Admin	Non-course-specific academic records	0.045
5	# credits attempted in the target term	Admin	Non-course-specific academic records	0.045
6	Total minutes spent in 1st quarter	LMS	Early-term	0.04
7	Term GPA of second-to-last term prior to the target term	Admin	Non-course-specific academic records	0.029
8	% prior attempted credits completed	Admin	Non-course-specific academic records	0.023
9	# assignment submissions in the 1st quarter	LMS	Early-term	0.023
10	Average grade in prior Humanities courses	Admin	Course-subject-specific	0.021
11	Total # clicks in the 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.018
12	Average grade in prior Social Sciences courses	Admin	Course-subject-specific	0.017
13	Total minutes spend in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.016
14	Stddev of session lengths in 1st quarter	LMS	Early-term	0.014

15	Average session length in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.013
16	# discussion forum replies in 1st quarter	LMS	Early-term	0.013
17	Stddev of session lengths in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.013
18	Slope of term-level GPA in prior terms	Admin	Non-course-specific academic records	0.013
19	Average grade assigned by the instructor in the target course	Admin	Instructor-related	0.013
20	Average session length in 1st quarter	LMS	Early-term	0.013
21	# original discussion forum posts created in 1st quarter	LMS	Early-term	0.012
22	Average # words per discussion forum thread in 1st quarter	LMS	Early-term	0.012
23	Age at time of target course enrollment	Admin	Demographic	0.012
24	Average historical grade in the concurrent courses	Admin	Course-specific	0.012
25	Average grade in prior Natural Sciences courses	Admin	Course-subject-specific	0.012
26	% on-time assignment submissions in the 1st quarter	LMS	Early-term	0.012
27	Average position of posts in forum thread (original post = 1) in 1st quarter	LMS	Early-term	0.011
28	Total credits accumulated prior to target term	Admin	Non-course-specific academic records	0.011
29	Enrollment in target course section	Admin	Course-specific	0.011
30	Stddev of term-level credit completion rate	Admin	Non-course-specific academic records	0.011

Panel E: MTH 161

Ranking	Predictor	Category	Subcategory	FI Score
1	% prior attempted credits "Withdrawn"	Admin	Non-course-specific academic records	0.074
2	# credits attempted in the target term	Admin	Non-course-specific academic records	0.051
3	Total # clicks in the 1st quarter	LMS	Early-term	0.041
4	Cumulative GPA	Admin	Non-course-specific academic records	0.034
5	# assignment submissions in the 1st quarter	LMS	Early-term	0.03
6	Total minutes spent in 1st quarter	LMS	Early-term	0.028
7	Term GPA of the last term prior to the target term	Admin	Non-course-specific academic records	0.024
8	Total # clicks in the 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.023
9	Average historical grade in the target course	Admin	Course-specific	0.02
10	Average grade assigned by the instructor in the target course	Admin	Instructor-related	0.02
11	Total minutes spend in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.019

12	% prior attempted credits completed	Admin	Non-course-specific academic records	0.019
13	Term GPA of second-to-last term prior to the target term	Admin	Non-course-specific academic records	0.018
14	Stddev of session lengths in 1st quarter	LMS	Early-term	0.017
15	Average session length in 1st quarter	LMS	Early-term	0.016
16	Average historical grade in the concurrent courses	Admin	Course-specific	0.016
17	Average session length in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.016
18	Stddev of session lengths in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.016
19	# discussion forum replies in 1st quarter	LMS	Early-term	0.015
20	Enrollment in target course section	Admin	Course-specific	0.015
21	Average grade in prior Humanities courses	Admin	Course-subject-specific	0.015
22	Average # assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.014
23	Slope of term-level GPA in prior terms	Admin	Non-course-specific academic records	0.013
24	Stddev of term-level credit completion rate	Admin	Non-course-specific academic records	0.012
25	Age at time of target course enrollment	Admin	Demographic	0.012
26	# original discussion forum posts created in 1st quarter	LMS	Early-term	0.012
27	Average position of posts in forum thread (original post = 1) in 1st quarter	LMS	Early-term	0.012
28	% on-time assignment submissions in the 1st quarter	LMS	Early-term	0.012
29	Total credits accumulated prior to target term	Admin	Non-course-specific academic records	0.011
30	Average grade in prior Social Sciences courses	Admin	Course-subject-specific	0.011

Notes: we calculate the FI (Feature Importance) Score using mean decrease in importance. The predictors that rank in the top 10 in all five course-specific models are highlighted in orange.

Appendix Table A5: Comparison of online versus in-person observations

Panel A: Number of observations

	Full Analytic Sample		2+ terms				1st term			
			Training Set		Validation Set		Training Set		Validation Set	
	Online	In-person	Online	In-person	Online	In-person	Online	In-person	Online	In-person
Student x course x section observations	866,136	301,933	475,104	218,389	255,306	15,286	113,611	67,204	22,115	1,054
Unique course sections	44,654	17,358	29,914	15,598	14,619	1,681	21,957	10,950	7,825	375

Panel B: Mean values of LMS predictors

	Full Analytic Sample		2+ terms				1st term			
			Training Set		Validation Set		Training Set		Validation Set	
	Online	In-person	Online	In-person	Online	In-person	Online	In-person	Online	In-person
# assignment submissions in the 1st quarter	8.52	7.47	8.32	7.44	8.42	6.41	8.52	7.47	9.57	7.07
# assignment submissions in the 1st quarter data is available	49.2%	57.3%	48.8%	58.2%	47.3%	49.4%	49.2%	57.3%	46.0%	44.4%
On-time assignment submissions in 1st quarter data is available	66.6%	54.5%	68.6%	56.4%	68.6%	62.6%	66.6%	54.5%	66.3%	67.0%
% on-time assignment submissions in the 1st quarter	34.8%	26.5%	35.4%	27.1%	35.0%	29.5%	34.8%	26.5%	33.6%	27.0%
Average session length in 1st quarter	15.89	8.54	15.11	8.80	12.29	7.08	15.89	8.54	14.20	7.82
Stddev of session lengths in 1st quarter	30.82	17.36	29.71	17.92	24.85	15.30	30.82	17.36	27.64	15.53
Total # clicks in the 1st quarter	1047	470	1034	514	930	611	1047	470	1039	478
Total minutes spent in 1st quarter	655	279	621	304	523	321	655	279	615	273

Average depth (position) of posts within a discussion forum thread (original post = 1) in 1st quarter	0.9091	0.2977	0.9114	0.2247	0.7559	0.1653	0.9091	0.2977	0.825	0.203
Average # words per discussion forum thread in 1st quarter	375	121	409	108	343	82	375	121	346	75
# original discussion forum posts created in 1st quarter	1.54	0.41	1.57	0.33	1.32	0.21	1.54	0.41	1.39	0.23
# discussion forum replies in 1st quarter	1.91	0.40	2.02	0.31	1.60	0.26	1.91	0.40	1.78	0.25

Appendix Table A6: C-statistics from validation sets restricted to observations in a particular modality

Sample	Predictor set	Online	In-person
2+ terms	Admin	0.8528	0.8632
2+ terms	LMS	0.7812	0.7081
2+ terms	Full	0.8705	0.8642
1st term	Admin	0.7238	0.7736
1st term	LMS	0.7812	0.6466
1st term	Full	0.8235	0.8085

Notes: each row corresponds to a separate random forest prediction model using the set of predictors indicated by the column Predictor Set, and observations from the sample of students based on academic history indicated by the column Sample.

Appendix Table A7: Feature Importance for Modality-specific models

Panel A: Model with 2+ terms observations that are online

Ranking	Predictor	Category	Subcategory	FI Score
1	% prior attempted credits "Withdrawn"	Admin	Non-course-specific academic records	0.08
2	Total # clicks in the 1st quarter	LMS	Early-term	0.041
3	Term GPA of the last term prior to the target term	Admin	Non-course-specific academic records	0.039
4	Cumulative GPA	Admin	Non-course-specific academic records	0.039
5	# credits attempted in the target term	Admin	Non-course-specific academic records	0.035
6	Total minutes spent in 1st quarter	LMS	Early-term	0.031
7	# original discussion forum posts created in 1st quarter	LMS	Early-term	0.027
8	% prior attempted credits completed	Admin	Non-course-specific academic records	0.023
9	# discussion forum replies in 1st quarter	LMS	Early-term	0.023
10	Term GPA of second-to-last term prior to the target term	Admin	Non-course-specific academic records	0.02
11	Average grade assigned by the instructor in the target course	Admin	Instructor-related	0.02
12	Average # assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term	0.019
13	Average historical grade in the target course	Admin	Course-specific	0.019
14	Stddev of term-level credit completion rate	Admin	Non-course-specific academic records	0.019
15	Average # words per discussion forum thread in 1st quarter	LMS	Early-term	0.015
16	Slope of term-level GPA in prior terms	Admin	Non-course-specific academic records	0.015
17	Average position of posts in forum thread (original post = 1) in 1st quarter	LMS	Early-term	0.015
18	% on-time assignment submissions in the 1st quarter	LMS	Early-term	0.014
19	Total # clicks in the 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.012
20	Stddev of session lengths in 1st quarter	LMS	Early-term	0.012
21	Total minutes spend in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.012
22	Total credits accumulated prior to target term	Admin	Non-course-specific academic records	0.012

23	Average historical grade in the concurrent courses	Admin	Course-specific	0.011
24	Average session length in 1st quarter	LMS	Early-term	0.011
25	Stddev of session lengths in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.01
26	Average session length in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.01
27	Total # clicks in thefull term, averaged across prior courses	LMS	Prior full-term	0.009
28	Slope of credits attempted in prior terms	Admin	Non-course-specific academic records	0.009
29	Total # days with any activity in full term, averaged across prior courses	LMS	Prior full-term	0.009
30	Enrollment in target course section	Admin	Course-specific	0.009

Panel B: Model with 2+ terms observations that are in-person

Ranking	Predictor	Category	Subcategory	FI Score
1	% prior attempted credits "Withdrawn"	Admin	Non-course-specific academic records	0.063
2	Cumulative GPA	Admin	Non-course-specific academic records	0.043
3	# credits attempted in the target term	Admin	Non-course-specific academic records	0.037
4	Term GPA of the last term prior to the target term	Admin	Non-course-specific academic records	0.036
5	Average historical grade in the target course	Admin	Course-specific	0.03
6	Average grade assigned by the instructor in the target course	Admin	Instructor-related	0.029
7	Total # clicks in the 1st quarter	LMS	Early-term	0.028
8	% prior attempted credits completed	Admin	Non-course-specific academic records	0.026
9	Average # assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term	0.026
10	Term GPA of second-to-last term prior to the target term	Admin	Non-course-specific academic records	0.025
11	Total minutes spent in 1st quarter	LMS	Early-term	0.022
12	Stddev of term-level credit completion rate	Admin	Non-course-specific academic records	0.021
13	Slope of term-level GPA in prior terms	Admin	Non-course-specific academic records	0.017
14	Average historical grade in the concurrent courses	Admin	Course-specific	0.016
15	Total credits accumulated prior to target term	Admin	Non-course-specific academic records	0.016
16	Total # clicks in the 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.015
17	Stddev of session lengths in 1st quarter	LMS	Early-term	0.015
18	Average session length in 1st quarter	LMS	Early-term	0.014

19	Total minutes spend in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.014
20	Average session length in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.013
21	Stddev of session lengths in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.013
22	Slope of credits attempted in prior terms	Admin	Non-course-specific academic records	0.012
23	Average # assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.012
24	Age at time of target course enrollment	Admin	Demographic	0.012
25	Credits attempted in last term (prior to target term)	Admin	Non-course-specific academic records	0.011
26	Enrollment in target course section	Admin	Course-specific	0.011
27	# stop-out terms between initial enrollment and target term	Admin	Non-course-specific academic records	0.01
28	% prior attempted credits that were developmental courses	Admin	Non-course-specific academic records	0.009
29	% on-time assignment submissions in the 1st quarter	LMS	Early-term	0.009
30	Target course is 200-level	Admin	Non-course-specific academic records	0.009

Panel C: Model with 1st term observations that are online

Ranking	Predictor	Category	Subcategory	FI Score
1	Total # clicks in the 1st quarter	LMS	Early-term	0.1
2	Total minutes spent in 1st quarter	LMS	Early-term	0.072
3	# credits attempted in the target term	Admin	Non-course-specific academic records	0.065
4	# discussion forum replies in 1st quarter	LMS	Early-term	0.057
5	# original discussion forum posts created in 1st quarter	LMS	Early-term	0.053
6	Average position of posts in forum thread (original post = 1) in 1st quarter	LMS	Early-term	0.042
7	Average historical grade in the target course	Admin	Course-specific	0.037
8	Average # words per discussion forum thread in 1st quarter	LMS	Early-term	0.036
9	% on-time assignment submissions in the 1st quarter	LMS	Early-term	0.036
10	Average # assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term	0.036
11	Average grade assigned by the instructor in the target course	Admin	Instructor-related	0.034
12	Total # clicks in the 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.034
13	Stddev of session lengths in 1st quarter	LMS	Early-term	0.03
14	Total minutes spend in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.029

15	Average session length in 1st quarter	LMS	Early-term	0.027
16	Average historical grade in the concurrent courses	Admin	Course-specific	0.025
17	Stddev of session lengths in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.024
18	Target course is in a Summer term	Admin	Course-specific	0.024
19	Average session length in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.023
20	Age at time of target course enrollment	Admin	Demographic	0.022
21	Enrollment in target course section	Admin	Course-specific	0.02
22	Target course is 200-level	Admin	Non-course-specific academic records	0.018
23	Average # assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.017
24	% on-time assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.013
25	% attempted credits during target term that are online	Admin	Non-course-specific academic records	0.013
26	% attempted credits during target term that are evening	Admin	Non-course-specific academic records	0.008
27	Enrolled in a transfer-oriented associate degree program	Admin	Non-course-specific academic records	0.007
28	Enrolled in any development courses in the target term	Admin	Non-course-specific academic records	0.007
29	Instructor works full-time at VCCS	Admin	Instructor-related	0.004
30	On-time assignment submissions in 1st quarter data is available	LMS	Early-term	0.004

Panel D: Model with 1st term observations that are in-person

Ranking	Predictor	Category	Subcategory	FI Score
1	Total # clicks in the 1st quarter	LMS	Early-term	0.09
2	# credits attempted in the target term	Admin	Non-course-specific academic records	0.081
3	Average historical grade in the target course	Admin	Course-specific	0.067
4	Total minutes spent in 1st quarter	LMS	Early-term	0.064
5	Average grade assigned by the instructor in the target course	Admin	Instructor-related	0.053
6	Average # assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term	0.05
7	Total # clicks in the 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.036
8	Total minutes spend in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.033
9	Stddev of session lengths in 1st quarter	LMS	Early-term	0.031
10	# original discussion forum posts created in 1st quarter	LMS	Early-term	0.03

11	Average historical grade in the concurrent courses	Admin	Course-specific	0.03
12	Average session length in 1st quarter	LMS	Early-term	0.028
13	Age at time of target course enrollment	Admin	Demographic	0.027
14	Stddev of session lengths in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.027
15	% on-time assignment submissions in the 1st quarter	LMS	Early-term	0.027
16	Average position of posts in forum thread (original post = 1) in 1st quarter	LMS	Early-term	0.026
17	Average session length in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.026
18	Average # assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.025
19	# discussion forum replies in 1st quarter	LMS	Early-term	0.023
20	Enrollment in target course section	Admin	Course-specific	0.022
21	Average # words per discussion forum thread in 1st quarter	LMS	Early-term	0.021
22	Target course is in a Summer term	Admin	Course-specific	0.017
23	Enrolled in any development courses in the target term	Admin	Non-course-specific academic records	0.014
24	Target course is 200-level	Admin	Non-course-specific academic records	0.014
25	% attempted credits during target term that are evening	Admin	Non-course-specific academic records	0.011
26	% on-time assignment submissions in 1st quarter, averaged across concurrent courses	LMS	Early-term concurrent	0.01
27	% attempted credits during target term that are online	Admin	Non-course-specific academic records	0.01
28	Enrolled in a transfer-oriented associate degree program	Admin	Non-course-specific academic records	0.008
29	Instructor works full-time at VCCS	Admin	Instructor-related	0.006
30	Instructor has been teaching at VCCS for 6+ years	Admin	Instructor-related	0.004

Notes: we calculate the FI (Feature Importance) Score using mean decrease in importance.