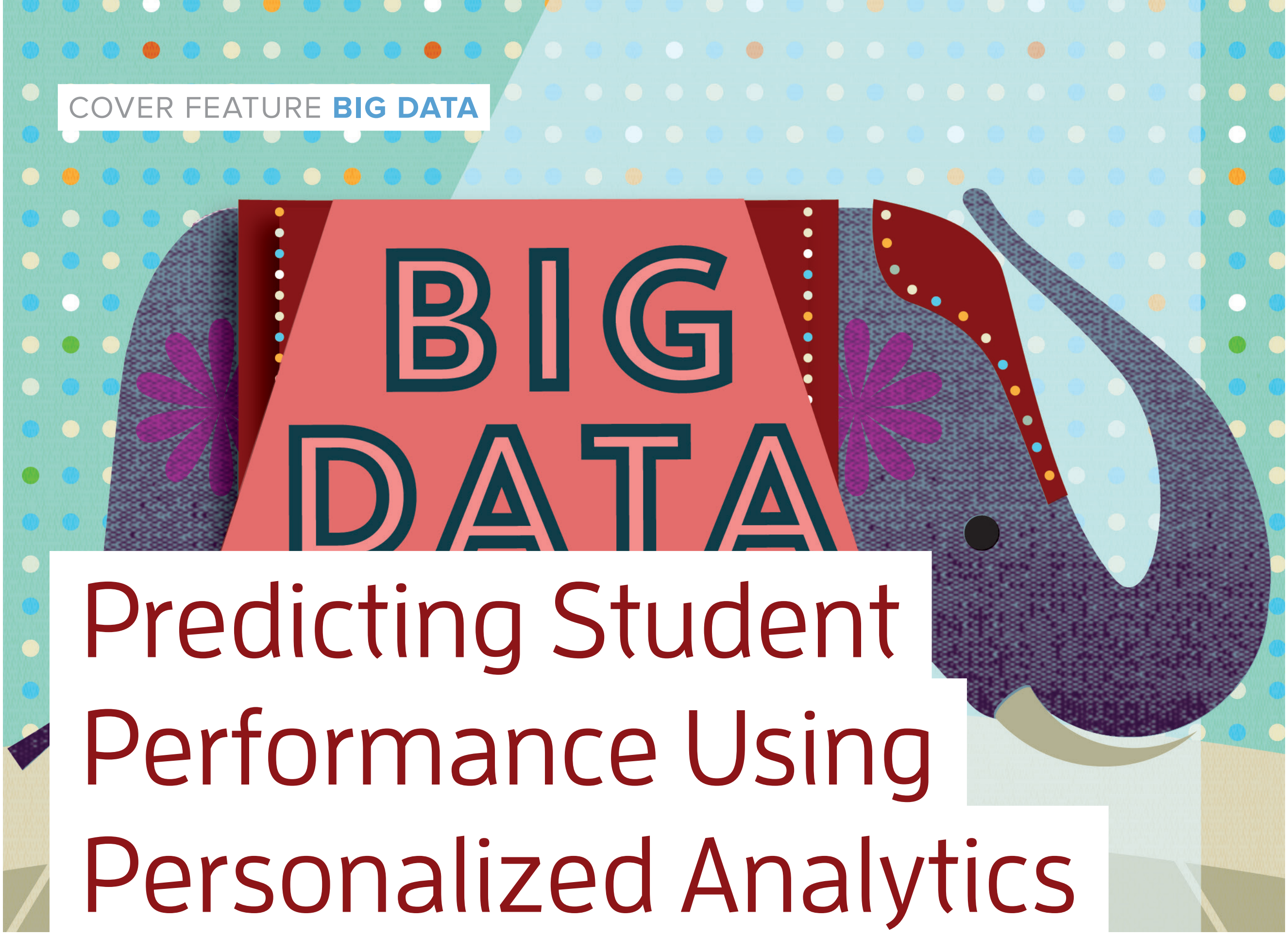

Predicting Student Performance Using Personalized Analytics

Asmaa Elbadrawy, University of Minnesota
Agoritsa Polyzou, University of Minnesota
Zhiyun Ren, George Mason University
Mackenzie Sweeney, George Mason University
George Karypis, University of Minnesota
Huzefa Rangwala, George Mason University

To help solve the ongoing problem of student retention, new expected performance-prediction techniques are needed to facilitate degree planning and determine who might be at risk of failing or dropping a class. Personalized multiregression and matrix factorization approaches based on recommender systems, initially developed for e-commerce applications, accurately forecast students' grades in future courses as well as on in-class assessments.

An enduring challenge in higher education is student retention.[1] There is a critical need to develop innovative approaches that ensure students graduate in a timely fashion and are well trained and workforce ready in their field of study.[2] As the volume and variety of data collected in both traditional and online university offerings continue to expand, new opportunities to apply big data analytics to grand challenges in higher education arise.

Many traditional educational institutions have deployed learning-management systems (LMSs) such as Blackboard (www.blackboard.com), Moodle (https://moodle.org), and Sakai (https://sakaiproject.org) to provide students with an online platform to access course content and to communicate and collaborate with instructors and peers. The data collected by LMSs can be used to track students' engagement and predict their future academic performance.[3] Such predictions can help students choose the most suitable majors and properly blend courses of varying difficulty in a semester's schedule; they can also provide early warnings to the students' instructors and academic advisors that they might need assistance.

In parallel, massive open online courses (MOOCs) have emerged as a popular low-cost technology to deliver distance education on a wide variety of topics to students across the world.[4] Given the minimal expense and lack of entry barriers such as prerequisites or skill requirements, many people enroll in MOOCs, but only a small fraction participate in the various course-related activities—viewing video lectures, studying course material, completing quizzes and homework, and so on—and many eventually drop out. Several researchers have analyzed usage logs to determine the factors associated with MOOCs' high attrition rate and to develop methods that predict how participants will perform and whether they are likely to drop out.

In response to these trends, researchers have deployed big data techniques to analyze information collected by traditional brick-and-mortar and online universities as well as MOOC providers. Several methods have been developed to tackle the problems of pre-registration grade prediction, within-class assessment prediction, GPA prediction, within class grade prediction, and drop versus no-drop prediction for MOOC students.[5–8] We present methods that draw on techniques from recommender systems to accurately predict students' next-term course grades as well as within-class assessment performance. These approaches ingest large amounts of heterogeneous and sparse data and can be used in all types of educational environments.

NEXT-TERM GRADE PREDICTION

These methods are designed to estimate the grades that a student is expected to obtain in the courses that he or she is considering taking for the next term. Students can use these estimated grades to select courses for which they are well prepared and expected to perform well in and also satisfy degree requirements, thereby allowing them to make progress towards graduation. The estimated grades also provide information as to the difficulty rating for each course, which can help students prioritize their studies and manage their study time. Moreover, course instructors and departments can also benefit by knowing how students registered for particular courses are expected to perform; this enables them to make adjustments including holding additional office hours, allocating more teaching assistants, and so on.

To predict the next-term grades we leverage four types of data: admissions records that include demographic information, high school scores, and SAT/ACT scores; grades in courses that were already taken by all the students; information about course content; and which instructors taught which courses. The historical student–course performance information is represented in a form of an $n \times m$ matrix G in which an entry g_{ij} stores the grade that student i obtained in class j. These grades are stored as a number between 0 and 4 by using a standard letter grade to number mapping F to 0 and A to 4. Since each student takes only a very small number of the courses being offered, the matrix G will be sparse and most of its entries will have no values associated with them.

We investigated two classes of methods for building the prediction models, both of which are used extensively in e-commerce recommender systems to predict consumer ratings and purchases. The first class builds these models by using linear regression approaches in which the grades obtained by the students in the courses that they have already taken and their associated characteristics are the predictor variables. The second class builds these models by using matrix factorization approaches to identify a low-dimensional space that jointly represents both the students and the courses such that a student's grade on a particular course is estimated as the inner product of their respective latent space representations. Given the nature of the domain, this latent space can correspond to the space of knowledge components.[8]

Regression-based methods

The first method that we investigated is the *course-specific regression* (CSpR), which predicts the grade that a student will achieve in a specific course as a sparse linear combination of the grades that the student obtained in past courses.[8] In order to estimate the CSpR model for a particular course, a course-specific subset of the data is used that consists of the students that have already taken that course along with the grades they achieved for that course and all the other courses they have taken prior to it. CSpR assumes that undergraduate degree programs are structured in such a way that courses taken by students provide the necessary knowledge and skills for them to do well in future courses. Consequently, the performance that a student achieved in a subset of the earlier courses can be used to predict how well he or she will perform in future courses.

The second method, *personalized linear multi-regression* (PLMR), uses a linear combination of k regression models, which are weighted on a per-student basis.[6] In this approach, the predicted grade $\hat{g}_{ij}$ for student i in course j is given by

$$\hat{g}_{ij} = w_0 + s_i + c_j + P_i W X_{ij},$$

where w_0 is a global bias, s_i is a bias term for student i, c_j is a bias term for course j, P_i is the $1 \times k$ vector of model weights for student i, W is the $k \times p$ matrix of regression coefficients, and X_{ij} is a feature vector encoding the various information associated with student i and course j. Specifically, the information encoded in X_{ij} includes the student and course numbers, the course instructor, the course level, and the department offering that course. The bias terms reflect the mean grades achieved by a student in the past and the mean grades for a given course in the past. These capture student-related factors (for example, skills, motivations, and learning habits) and course-related factors (for example, course material and difficulty). Note that setting the number of regression models in PLMR to 1 ($k = 1$) leads to a simple regression computation.

Matrix factorization–based methods

As a baseline method, we investigated a standard *matrix factorization* (MF) approach that approximates the observed entries of the student–course grade matrix G by the product of two matrices whose rank is at most k. In this approach, each student i and course j are represented via k dimensional latent feature vectors v_i and v_j, respectively, from which the inner product is used to predict the grade $\hat{g}_{ij}$ that student i will obtain on course j, that is,

$$\hat{g}_{ij} = \sum_{f=1}^{k} v_{i,f} v_{j,f} \quad .$$

One limitation of the standard MF method is that it ignores the sequence in which the students have taken the various courses and as such the latent representation of a course can potentially be influenced by the performance of the students in courses that were taken afterward. In order to address this problem, we also investigated *course-specific MF* (CSpMF) that relies only on the subset of the data used by CSpR in order to estimate an MF model that is specific to each course.

Finally, in order to exploit the additional information that we have available (admissions, course content, instructors, and the like) we also developed MF methods based on *factorization machines* (FM), which are general-purpose factorization techniques that can incorporate additional information while also leveraging the sparse student–course matrix.[7] Moreover, by using this additional information, FM is also able to make predictions about students that have not yet taken any courses, and for whom we have no past course performance information.

IN-CLASS ASSESSMENT PREDICTION

Predicting a student's performance on in-class assessments like quizzes and homework assignments can potentially provide the needed early intervention for students that are at risk of failing a course or dropping out.

Many researchers have used LMS and MOOC data to predict future academic performance, both to facilitate degree planning and to determine which students might be at risk of failing or dropping a class and would benefit from intervention. Rebecca Barber and Mike Sharkey applied standard linear regression analysis to aggregated LMS data and administrative records to develop a general predictive risk model indicating the likelihood of a given student failing a particular course.[9] Nguyen Thai-Nghe and his colleagues developed an MF method inspired by e-commerce recommender systems that considered individual students' performance on past LMS activities.[10] Other researchers have likewise applied MF variants to LMS data but ignored many of the students' interactions with the system, limiting the ability to provide even more granular forecasts.

Within the context of MOOCs, many student performance–forecasting approaches have been developed. To predict in real time when a student might stop engaging in a particular course, Sebastien Boyer and Kalyan

Veeramachaneni combined interaction data on the student from previous courses with data from previous weeks of the current course.[11] Tanmay Sinha and Justine Cassell tracked student activities across multiple MOOCs to predict grade sequences.[12] Zachary Pardos and colleagues adapted the cognitive modeling techniques of Bayesian knowledge tracing to predict student performance on homework, lecture sequences, and exams in a MOOC course.[5]

As part of our ongoing work on this problem, we have developed approaches that track the participation of a student within a class at a traditional university or a MOOC by analyzing the click-stream server logs and predicting the student performance on the next graded assessment. These approaches are based on estimating PLMR models to predict student performance on in-class assessments like quizzes and homework assignments within an LMS or MOOC. To do so, they utilize a wide range of features extracted from the server logs that track student interactions with the system, such as performance on previous assessments, number of logins, views of different course materials, number of attempts at a given assessment, and time between viewing material and attempting an assessment. Figure 1 shows examples of activities captured by a typical MOOC server log.

PLMR can also be used to analyze the relative performance of the different features for in-class assessment prediction. Model parameters should all have nonnegative values to ensure that they additively contribute to predicted grades. Such PLMR implementations can be incorporated into student monitoring and early warning systems to indicate when a student is having trouble or falling behind and thus might need assistance.

DATASETS

We trained and tested our performance prediction methods on four datasets: George Mason University (GMU) transcript data, University of Minnesota (UMN) transcript data, UMN LMS data, and Stanford University MOOC data.

GMU transcript data

This dataset consists of course records we obtained on 30,754 students enrolled at GMU from summer 2009 to spring 2014 (a total of 15 terms) who had declared in 1 of 144 majors, each of which belongs to 1 of 13 colleges. During this time period, the students took 9,085 courses, each classified as 1 of 161 disciplines and taught by 1 of 6,347 instructors. Discarding records with no grades (withdrawals) or grades that do not translate to the A–F scale (audits) yielded 894,736 student–course dyads. Of these dyads, 584,179 (65.29 percent) correspond to dyads involving transfer grades. All data was collected and anonymized in accordance with GMU institutional review board policies. The dataset includes various student, course, and instructor features, either categorical or real-valued.

For each student, we have demographic data from admissions records such as age, race, sex, zip code, high school ID and GPA, and SAT scores. For each student–course dyad, we have the student's declared major at the time and the grade earned. For each term in a student's transcript, we have the number of course credit hours attempted that term as well as the number attempted up to that term. To reflect the student's relative progress, we annotated each term with a value indicating the number of previous terms in which the student had taken courses.

Each course belongs to a particular discipline, is worth a fixed number of credit hours, and is assigned a particular course level. For each term, we have the aggregate student GPA for the course from the previous term as well as the cumulative aggregate student GPA of the course over all terms it was previously offered (in our dataset). We also have the number of students enrolled in all sections of the course during the current term, as well as the total number of students enrolled for all prior terms the course was offered. For each course we have the instructor's classification (adjunct, full-time, part-time, graduate research assistant, or graduate teaching assistant), rank (instructor, assistant professor, associate professor, eminent scholar, or university professor), and tenure status (term, tenure-track, or tenured). We mapped all transfer course records to GMU equivalents.

UMN transcript data

This dataset consists of course grades for 2,949 students that were part of UMN's computer science and engineering (CSE) and electrical and computer engineering (ECE) degree programs from fall 2002 to spring 2014. Both programs are part of the College of Science and Engineering (CS&E), which requires students to take a common set of core science courses during the first 23 semesters. We removed any courses that are not among those offered by CS&E departments and generally do not count toward degree requirements, as well as those taken pass/fail, leaving 2,556 courses. The resulting dataset contains 76,748 student–grade dyads.

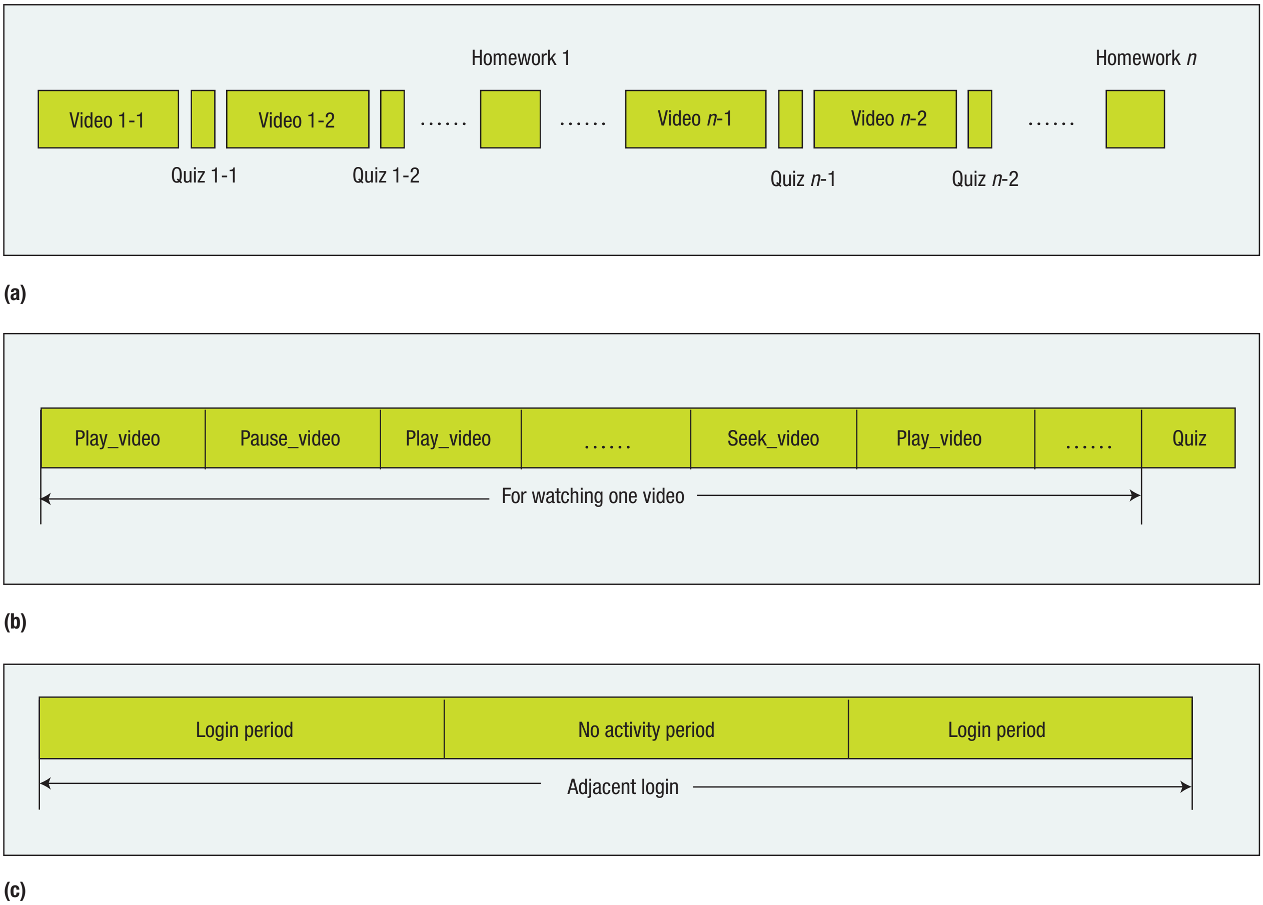

FIGURE 1. Different student activities within a typical massive open online course (MOOC). (a) Students view video lectures, each with an accompanying quiz, and have periodic homework assignments. (b) There are various video-viewing options including searching for a video and pausing/playing a video. (c) Students can have multiple login sessions, each of which might contain video watching, quiz attempts, and/or homework assignments.

UMN LMS data

The UMN dataset was extracted from the university's Moodle installation, one of the largest worldwide. It spans two semesters and includes 11,556 students and 832 courses belonging to 157 departments. Each student enrolled in at least four courses. The dataset includes 114,498 assignments and 75,143 quiz submissions, and 251,348 forum posts. Assignments and quizzes are referred to as *activities*, with grades scaled to be in the range from 0 to 1 such that the lowest grade that a student can get is 0 and the highest grade is 1. Each student–activity dyad (s, a) is associated with a feature vector f_{sa}. Features fall into three categories.

Student performance features. These describe the student's current GPA and the current grade achieved in the course's previous activities.

Activity and course features. These describe the activity type (assignment or quiz); the course level, which takes an integer value of 1, 2, 3, or 4 (4 being the most advanced); and the department to which the course belongs.

LMS features. Extracted from Moodle's server log files, these describe the student's interaction with the system prior to the activity due date and include the number of times the

TABLE 1. Next-term grade prediction results on George Mason University transcript data.

Method	Root-mean-square error (RMSE)	Mean absolute error (MAE)
Factorization machine (FM)	0.7423	0.52 ± 0.53
Personalized linear multi-regression (PLMR)	0.7886	0.57 ± 0.55
Random forest (RF)	0.7936	0.58 ± 0.54
Mean of means	0.8643	0.64 ± 0.58
Uniform random guessing	1.8667	1.54 ± 1.06

TABLE 2. Results for the top three methods from Table 1 for cold-start (CS) and non–cold-start (NCS) records.

Group	Dyad (percent)	Method	RMSE	MAE
NCS	48.60	FM	0.7423	0.5187 ± 0.5310
		PLMR	0.7890	0.5635 ± 0.5522
		RF	0.7936	0.5837 ± 0.5377
CS student only	42.31	RF	0.7381	0.5867 ± 0.4478
		FM	0.8112	0.6114 ± 0.5331
		PLMR	0.9917	0.7321 ± 0.6689
CS course only	01.75	FM	0.7456	0.5293 ± 0.5252
		RF	0.7776	0.5695 ± 0.5295
		PLMR	1.1771	0.7489 ± 0.9081
CS both	04.55	RF	0.8203	0.6603 ± 0.4867
		FM	0.8337	0.6614 ± 0.5075
		PLMR	1.2060	0.8829 ± 0.8215

student has posted a comment on a forum, read a forum discussion, viewed course material, added something (like a wiki page) to the course page, or accessed the course page. We measured each interaction multiple times: one, two, four, and seven days prior to the activity due date. We also included all interactions up to the due date of the previous assignment. We normalized the forum interaction features with respect to the total number of available forum discussions.

Stanford MOOC data

This dataset includes activity information about students who took the Stanford MOOC course "Statistics in Medicine" in summer 2014 and was obtained via a data-sharing agreement with the university (http://datastage.stanford.edu). Activities included viewing video lectures and articles, and attempting graded quizzes and homework assignments. The course had 9 learning units with 111 assessments including 79 quizzes, 6 homework assignments, and 26 single questions. Of the 13,130 students, 4,337 submitted at least 1 homework assignment, 1,262 completed part of all 6 homework assignments, 1,099 finished all homework assignments, 193 finished all quizzes and homework assignments, and 6,481 had video-related activity. To capture latent information behind the clickstream, for each student we extracted session-, quiz-, video-, homework-, time-, and interval-related features.

RESULTS

We used these datasets to evaluate our next-class grade prediction and in-class assessment-prediction (both LMS and MOOC) models.

Next-term grade prediction

Table 1 reports the next-term grade-prediction results on the GMU transcript data. FM produces the lowest-error predictions. PLMR is the next best, followed by random forest. Random Forests (RF) are classical supervised learning methods that combine a collection of decision trees trained on multiple samples of data. Each decision tree attempts to discover the most informative features that split the data into cohesive groups with lowest error with regards to the continuous target attribute value (specifically, course grade in our case). For comparison, the table includes prediction results from several other baseline methods: mean of means (mean of the student, course, and global means) and uniform random guessing.

Table 2 shows results for the top three methods for cold-start (CS)

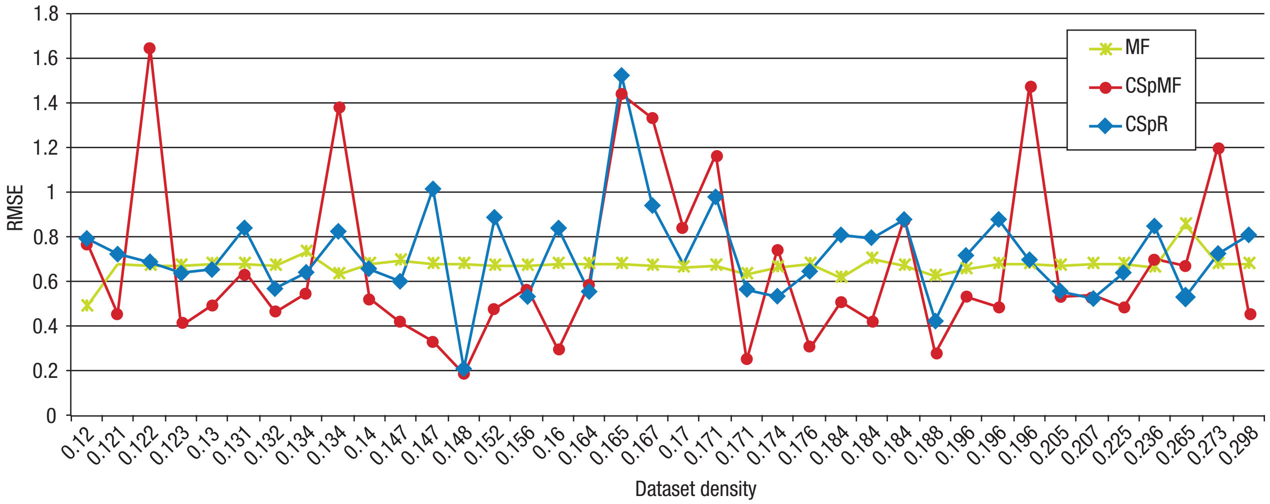

FIGURE 2. Next-term grade-prediction results on University of Minnesota (UMN) transcript data. Course-specific regression (CSpR) outperforms matrix factorization (MF) and course-specific MF (CSpMF) for most courses, achieving a root-mean-square error (RMSE) of 0.632.

and non–cold-start (NCS) records. CS records are student–course dyads lacking student or course information, or both, from any previous term, whereas NCS records contain student and course information from at least one previous term. RF outperforms FM when student information is absent; PLMR only performs well on NCS records. These results indicate that next-term grade prediction can be improved by substituting RF for FM when there is no prior student information. Doing so yields an overall root-mean-square error (RMSE) of 0.7443 compared to 0.7709 for FM and 0.7775 for RF.

Figure 2 shows next-term grade-prediction results on the UMN transcript data. It plots the RMSE per course achieved by three different methods—MF, CSpMF, and CSpR—with respect to the density of the student–course grade matrix of the course-specific dataset. The error rate for the course-specific techniques fluctuates more than that for MF, as a different model must be trained for each course. CSpR outperforms the other methods for most courses, achieving an RMSE of 0.632.

In-class LMS assessment prediction

Figure 3 shows in-class assessment-prediction results on the UMN LMS data, which was randomly split into training and test subsets, achieved by PLMR using three-feature combinations: activity and course features plus LMS features, student performance features plus activity and course features, and all feature sets. Prediction accuracy improves as the number of linear regression models increases, capturing more relations among features, especially student performance. Using 10 regression models and all three sets of features, the RMSE falls to 0.145.

Trained using all features, a baseline linear regression model gives an RMSE of 0.223 whereas PLMR with one linear model gives an RMSE of 0.168. In this context, PLMR improves on simple linear regression by capturing the general performance trends for each student and the course grade distributions. Further, the student-specific weighting function can adjust the regression result for each student.

We found that the features describing the student's cumulative GPA, cumulative grade, and viewing of course materials contributed most to grade-prediction accuracy. The features derived from forum activities contributed least, probably because they only appear with a small fraction (10 to 25 percent) of the training instances of each course.

In-class MOOC assessment prediction

Figure 4 shows in-class assessment-prediction results on the Stanford MOOC data using PLMR with a varying

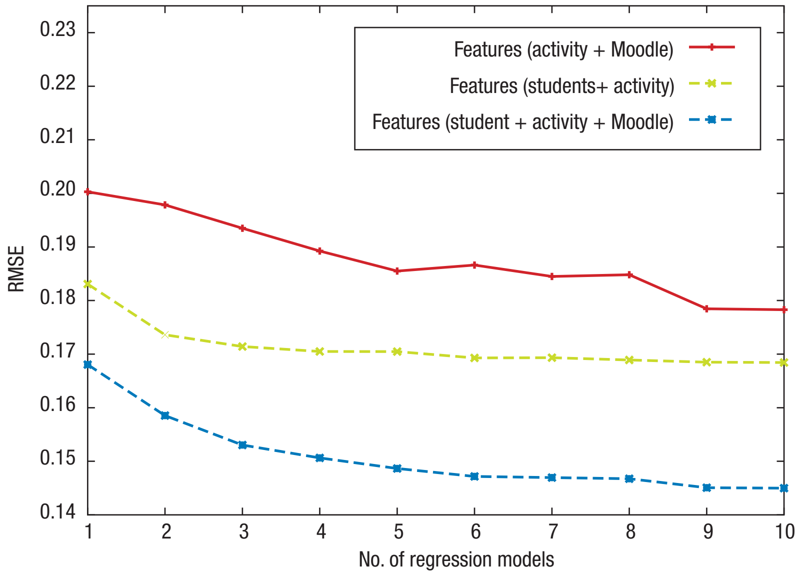

FIGURE 3. In-class assessment prediction results on UMN learning-management system (Moodle) data using personalized linear multiple regression (PLMR) models. Prediction accuracy improves as the number of linear regression models increases, capturing more relations among features, especially student performance. Using 10 regression models and all three sets of features, the RMSE falls to 0.145.

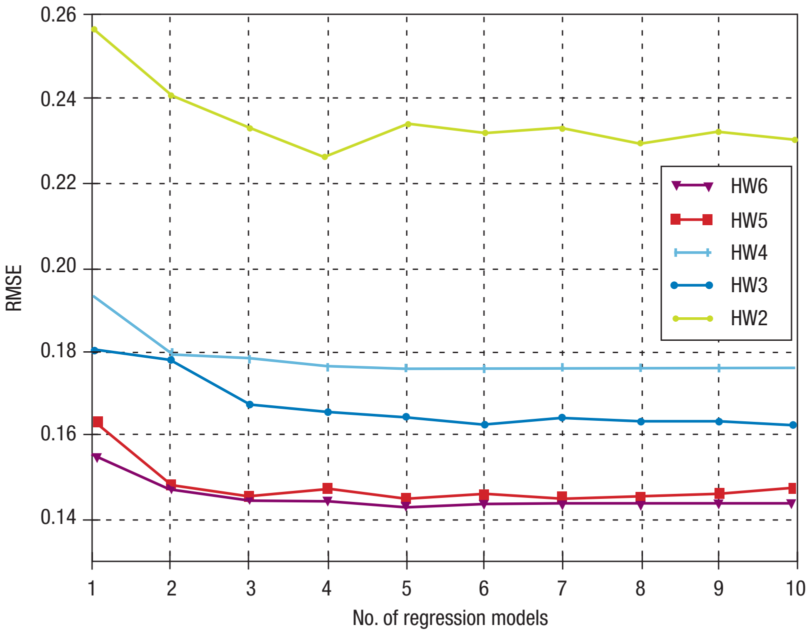

FIGURE 4. In-class homework (HW) assessment-prediction results on Stanford MOOC data using PLMR. RMSE decreases as the number of regression models increases, with an optimum of five regression models.

number of regression models. Given homework assignments, our goal was to predict the score a student achieved in each of the *n* assignments. To predict the score for a given student on a given homework referred by target homework, we extracted the features just before attempting that target homework. The training set consisted of all homework–student pairs where students attempted and received a score for a homework assignment before the target homework. RMSE decreases as the number of regression models increases, with an optimum of five regression models.

Recommender system–based personalized analytics are capable of forecasting student performance in a timely and accurate manner. Using only historical grade information coupled with available additional information such as transcript data, both PLMR and our advanced MF techniques can predict next-term grades with lower error rates than traditional methods. PLMR is also useful for predicting grades on assessments within a traditional class or online course by incorporating features captured through students' interaction with LMS and MOOC server logs. In the future, we plan to further refine these methods and incorporate them into degree-planning and early-warning systems to help solve the problem of student retention through graduation. C

ACKNOWLEDGMENTS

The research described in this article was supported by National Science Foundation Big Data grant nos. 1447489 and 1447788.

REFERENCES

1. S. Aud et al., *The Condition of Education 2013*, NCES report no. 2013-037, Nat'l Center for Education Statistics,

US Department of Education, 2013; http://nces.ed.gov/pubs2013/2013037.pdf.

2. *Nat'l Research Council, Building a Workforce for the Information Economy*, Nat'l Academies Press, 2001.
3. C. Romero, S. Ventura, and E. Garca, "Data Mining in Course Management Systems: Moodle Case Study and Tutorial," *Computers & Education*, vol. 51, no. 1, 2008, pp. 368–384.
4. L. Pappano, "The Year of the MOOC," *The New York Times*, 2 Nov. 2012; www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html?_r=0.
5. Z. Pardos et al., "Adapting Bayesian Knowledge Tracing to a Massive Open Online Course in edX," *Proc. 6th Int'l Conf. Educational Data Mining* (EDM 13), 2013; www.educationaldatamining.org/EDM2013/papers/rn_paper_21.pdf.
6. A. Elbadrawy, R.S. Studham, and G. Karypis, "Collaborative Multi-regression Models for Predicting Students' Performance in Course Activities," *Proc. 5th Int'l Conf. Learning Analytics and Knowledge* (LAK 15), 2015, pp. 103–107.
7. M. Sweeney, J. Lester, and H. Rangwala, "Next-Term Student Grade Prediction," *Proc. IEEE Int'l Conf. Big Data* (BigData 15), 2015, pp. 970–975.
8. A. Polyzou and G. Karypis, "Grade Prediction with Course and Student Specific Models," to be published in *Proc. 20th Pacific Asia Conf. Knowledge Discovery and Data Mining* (PAKDD), 2016.
9. R. Barber and M. Sharkey, "Course Correction: Using Analytics to Predict Course Success," *Proc. 2nd Int'l Conf. Learning Analytics and Knowledge* (LAK 12), 2012, pp. 259–262.
10. N. Thai-Nghe, T. Horváth, and L. Schmidt-Thieme, "Factorization Models for Forecasting Student Performance," *Proc. 4th Int'l Conf. Educational Data Mining* (EDM 11), 2011, pp. 11–20.
11. S. Boyer and K. Veeramachaneni, "Transfer Learning for Predictive Models in Massive Open Online Courses," *Artificial Intelligence in Education*, C. Conati et al., eds., LNCS 9912, Springer, 2015, pp. 54–63.
12. T. Sinha and J. Cassell, "Connecting the Dots: Predicting Student Grade Sequences from Bursty MOOC Interactions over Time," *Proc. 2nd ACM Conf. Learning@ Scale* (L@ S 15), 2015, pp. 249–252.

ABOUT THE AUTHORS

ASMAA ELBADRAWY is a PhD candidate in the Department of Computer Science and Engineering at the University of Minnesota Twin Cities. Her research interests include data mining, learning analytics, recommender systems, and the application of data-mining techniques within educational contexts. Contact her at asmaa@cs.umn.edu.

AGORITSA POLYZOU is a PhD student in the Department of Computer Science and Engineering at the University of Minnesota Twin Cities. Her research interests include learning analytics and data mining. Contact her at polyzou@cs.umn.edu.

ZHIYUN REN is a PhD student in the Department of Computer Science at George Mason University. Her research interests include data mining, recommender systems, and data analysis in education. Contact her at zren4@masonlive.gmu.edu.

MACKENZIE SWEENEY is an MS student in the Department of Computer Science at George Mason University. His research interests include data mining and learning analytics. Contact him at msweene2@masonlive.gmu.edu.

GEORGE KARYPIS is a professor and associate head of the Department of Computer Science and Engineering at the University of Minnesota Twin Cities. His research interests include data mining, bio- and cheminformatics, high-performance computing, information retrieval, learning analytics, collaborative filtering, and scientific computing. Karypis received a PhD in computer science from the University of Minnesota Twin Cities. Contact him at karypis@cs.umn.edu.

HUZEFA RANGWALA is an associate professor in the Department of Computer Science at George Mason University. His research interests include data mining and applications in learning sciences and bioinformatics. Rangwala received a PhD in computer science from the University of Minnesota Twin Cities. Contact him at rangwala@cs.gmu.edu.