MyJournalName manuscript No.
(will be inserted by the editor)

Grade Prediction with Course and Student Specific Models

Agoritsa Polyzou · George Karypis

Received: date / Accepted: date

Abstract The accurate estimation of students' grades in future courses is important as it can inform the selection of next term's courses and create personalized degree pathways to facilitate successful and timely graduation. This paper presents future-course grade predictions methods based on sparse linear and low-rank matrix factorization models that are specific to each course or student-course tuple. These methods identify the predictive subsets of prior courses on a course-by-course basis and better address problems associated with the *not-missing-at-random* nature of the student-course historical grade data. The methods were evaluated on a dataset obtained from the University of Minnesota, for two different departments with different characteristics. This evaluation showed that focusing on course specific data improves the accuracy of grade prediction.

Keywords Learning Analytics · Next-Term Grade Prediction · Course-Specific Models

1 Introduction

Data mining and machine learning approaches are being increasingly used to analyze educational- and learning-related datasets towards understanding how students learn and improving learning outcomes. This has led to the development of various approaches for modeling and predicting the success or failure of students in completing specific tasks in the context of intelligent tutoring systems [15, 19, 14, 18, 11, 8], building intelligent "early warning systems" that monitor the students' performance during the term [17, 2], predicting how well the students will perform by analyzing their activities with the learning management system (e.g., Moodle) [7, 16, 10], and predicting students' term and final GPA [13, 12, 1].

Our work focuses on developing methods that utilize historical student-course grade information to accurately estimate how well students will perform (as measured by their grade) on courses that they have not yet taken. Being able to accurately estimate students' grades in future courses is important as it can be used by them (and/or their academic advisers) to identify the appropriate set of courses to take during the next term, and create personalized degree pathways that enable them to successfully and effectively acquire the required knowledge to complete their studies in a timely fashion.

Existing approaches for predicting a student's grade in a future course [5, 6, 3] rely on neighborhood-based collaborative filtering methods. For each student whose grade needs to be predicted, a set of *similar* students are identified that have already taken that course and their grade is used to estimate the desired grade via some similarity-weighted aggregation function. Despite their relative simplicity, the estimations obtained by these methods are reasonably accurate indicating that there is sufficient information in the historical student-course grade data to make the estimation problem feasible.

A preliminary version of this work is published at PAKDD, 2016. Polyzou, A. and Karypis, G., 2016. Grade Prediction with Course and Student Specific Models. In Advances in Knowledge Discovery and Data Mining (pp. 89-101). Springer International Publishing.

Agoritsa Polyzou
University of Minnesota
E-mail: polyz001@umn.edu

George Karypis
University of Minnesota
E-mail: karypis@cs.umn.edu

In this paper we improve upon these methods by developing various future-course grade prediction methods that utilize approaches based on sparse linear models and low-rank matrix factorizations. These methods rely entirely on the performance that the students achieved in previously taken courses. A unique aspect of many of our methods is that their associated models are either specific to each course or specific to each student-course tuple. This allows them to identify and utilize the relevant information from the prior courses that are associated with the grade for each course and better address problems associated with the *not-missing-at-random* nature of the student-course historical grade data.

We experimentally evaluated the performance of our methods on a dataset obtained from the University of Minnesota that contained historical grades that span 12.5 years. Our results showed that the course specific models outperformed various competing schemes. Another conclusion was that the performance can significantly vary across different departments.

The remainder of the paper is organized as follows. Section 2 introduces the notation and definitions used. Section 3 describes the methods developed and Section 4 provides information about the experimental design. Section 5 presents an extensive experimental evaluation of the methods and compares them against existing approaches. Finally, Section 6 provides some concluding remarks.

2 Definitions and Notations

Throughout the paper, bold lowercase letters will denote column vectors (e.g., $\mathbf{y}$) and bold uppercase letters will denote matrices (e.g., $\mathbf{G}$). Individual elements will be denoted using subscripts (e.g., for a vector y_i, and for a matrix $g_{s,c}$). A single subscript on a matrix will denote its corresponding row. The sets will be represented by calligraphic letters.

The historical student-course grade information will be represented by a sparse matrix $\mathbf{G} \in \mathbb{R}^{n \times m}$, where n and m are the number of students and courses, respectively, and $g_{i,j}$ is the grade in the range of [0,4] that student i achieved in course j. If a student has not taken a course, the corresponding entry will be missing. The course, semester and student, whose grades need to be predicted will be called *target course*, *target semester*, and *target student*, respectively.

3 Methods

In this section we describe various classes of methods that we developed for predicting the grade that a student will obtain on a course that he/she has not yet taken.

3.1 Course-Specific Regression (CSR)

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. As a result, the performance that a student achieved in a subset of the earlier courses can be used to predict how well he/she will perform in future courses. Motivated by this, we developed a grade prediction method, called *course-specific regression* (CSR) that 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.

In order to estimate the CSR model for course c, we extract from the overall student-course matrix $\mathbf{G}$ the set of rows corresponding to the students that have taken c. For each of these students (rows), we keep only the grades that correspond to courses taken prior to course c. Let $\mathbf{G}^c \in \mathbb{R}^{n_c \times m}$ be the matrix representing that extracted information, where n_c is the number of students that took course c. In addition, let $\mathbf{y}^c \in \mathbb{R}^{n_c}$ be the grades that the students in $\mathbf{G}^c$ obtained in course c (y_i^c is the grade that was obtained by the student of the ith row of $\mathbf{G}^c$). Given this, the CSR model $\mathbf{w}^c \in \mathbb{R}_+^m$ for c is estimated as:

$$\underset{\mathbf{w}^c \geq 0}{\text{minimize}} \ \|\mathbf{y}^c - \mathbb{1} w_0^c - \mathbf{G}^c \mathbf{w}^c\|_2^2 + \lambda_1 \|\mathbf{w}^c\|_2^2 + \lambda_2 \|\mathbf{w}^c\|_1 , \tag{1}$$

where w_0^c is a bias term, $\mathbb{1} \in \mathbb{R}^{n_c}$ is a vector of ones, and λ_1, λ_2 are regularization parameters to control overfitting and promote sparsity. The model is non-negative because we assume that prior courses can only provide knowledge to future courses. The individual weights of $\mathbf{w}^c$ indicate how much each prior course contributes to the prediction and represent a measure of the importance of the prior course within the context of the estimated model. Using this model, the grade that a student will obtain in course c is given by:

$$\hat{y}^c = w_0^c + \mathbf{s}^T \mathbf{w}^c, \tag{2}$$

where $\mathbf{s} \in \mathbb{R}^m$ is the vector of the student's grades in the courses he/she has taken so far.

In this approach, prior to estimating the model using Equation 1, we first subtract from each $g_{i,j}^c$ grade

the GPA of the i-th student (GPA is calculated based on the information in $\mathbf{G}^c$). This centers the data for each student and takes into consideration a notion of student bias as it predicts the performance with respect to the current state of a student. Note that in the case of GPA-centered data, we remove the non-negativity constraint on $\mathbf{w}^c$. We found that by centering each student's grades around his/hers GPA leads to more accurate predictions (see Section 5.1).

3.2 Student-Specific Regression (SSR)

Depending on the major, the structure of different undergraduate degree programs can be different. Some degree programs have limited flexibility as to the set of courses that a student has to take and at which point in their studies they can take them (i.e., specific semester). Other degree programs are considerably more flexible and are structured around a fairly small number of core courses and a large number of elective courses.

For the latter type of degree programs, a drawback of the CSR method is that it requires the same linear regression model to be applied to all students. However, given that the set of prior courses taken by students in such flexible degree programs can be quite different, a single linear model can fail to capture the various prior course combinations. In fact, there can be cases in which many of the most important courses that were identified by the CSR model were simply not be taken by some students, even though these students have acquired the necessary knowledge and skills by taking a different set of courses. To address this limitation, we developed a different method, called *student-specific regression* (SSR), which estimates course-specific linear regression models that are also specific to each student.

The student specific model is derived by creating a student-course specific grade matrix $\mathbf{G}^{s,c}$ for each target student s and each target course c from the $\mathbf{G}^c$ matrix used in the CSR method. $\mathbf{G}^{s,c}$ is created in two steps. First, we eliminate from $\mathbf{G}^c$ any grades for courses that were not taken by the target student. Second, we eliminate from $\mathbf{G}^c$ the rows that correspond to the students that have not taken a sufficient number of courses that are in common with the target student s. Specifically, if $\mathcal{C}_s$ and $\mathcal{C}_i$ are the set of courses for student s and i, respectively, we compute the overlap ratio (OR) $= |\mathcal{C}_s \cap \mathcal{C}_i|/|\mathcal{C}_s|$ and if OR$< t$, then student i is not included in $\mathbf{G}^{s,c}$. The value of t is a parameter of the SSR method and high values ensure that the set of students forming $\mathbf{G}^{s,c}$ have taken many courses in common with s and have followed similar degree plans. Given $\mathbf{G}^{s,c}$, the SSR method proceeds to estimate the model using Equation 1 (with $\mathbf{G}^{s,c}$ replacing $\mathbf{G}^c$), and uses Equation 2 for prediction.

3.3 Methods based on Matrix Factorization

Low rank matrix factorization (MF) approaches have been shown to be very effective for accurately estimating ratings in the context of recommender systems [9]. These approaches can be directly applied to the problem of predicting the grade that a student will achieve on a particular course by treating the student-course grade matrix $\mathbf{G}$ as the user-item rating matrix.

The use of such MF-based approaches for grade prediction is postulated on the fact that there is a low dimensional latent feature space that can jointly represent both students and courses. Given the nature of the domain, this latent space can correspond to the space of knowledge components. Each course vector is the set of components associated with a course and each student vector represents the student's level of knowledge across these knowledge components.

By applying the common approaches of MF-based rating prediction to the problem of grade prediction, the grade that student i will obtain on course j is estimated as

$$\hat{g}_{i,j} = \mu + sb_i + cb_j + \mathbf{p}_i \mathbf{q}_j{}^T, \tag{3}$$

where μ is a global bias term, sb_i and cb_j are the student and course bias terms, respectively, and $\mathbf{p_i}$ and $\mathbf{q_j}$ are the latent representations for student i and course j, respectively. The parameters of the MF method ($\mu, \mathbf{sb} \in \mathbb{R}^n, \mathbf{cb} \in \mathbb{R}^m, \mathbf{P} \in \mathbb{R}^{n \times l}$, and $\mathbf{Q} \in \mathbb{R}^{n \times l}$) are estimated following a matrix completion approach that considers only the observed entries in $\mathbf{G}$ as

$$\begin{aligned} \underset{\mu,\mathbf{sb},\mathbf{cb},\mathbf{P},\mathbf{Q}}{\text{minimize}} & \sum_{g_{i,j} \in \mathbf{G}} \left(g_{i,j} - \mu - sb_i - cb_j - \mathbf{p}_i \mathbf{q}_j^T\right)^2 \\ & + \lambda(\|\mathbf{P}\|_F^2 + \|\mathbf{Q}\|_F^2 + \|\mathbf{sb}\|_2^2 + \|\mathbf{cb}\|_2^2), \end{aligned} \tag{4}$$

where λ is a regularization parameter and l is the dimensionality of the latent space, which is a parameter to this method.

The accurate recovery of the low rank model (when such a model exists) from a set of partial observations depends on having a sufficient number of observed entries, and on these entries be randomly sampled from the entries of the target matrix $\mathbf{G}$ [4]. However, in the context of student grade data, the set of courses that students take is not a random subset of the courses being offered as they need to satisfy their degree program requirements. As a result, such an MF approach may lead to suboptimal prediction performance.

In order to address this problem we developed a *course specific matrix factorization* (CSMF) approach that estimates an MF model for each course by utilizing a course specific subset of the data that is denser (in terms of the number of observed entries and the dimensions of the matrix). As a result, it contains a larger number of randomly sampled subsets of sufficient size.

Given a course c and a set of students $\mathcal{S}^c$ for which we need to estimate their grade for c (i.e., the students in $\mathcal{S}^c$ have not taken this course yet), the data that CSMF utilizes are the:

(i) the students and grades of the $\mathbf{G}^c$ matrix and $\mathbf{y}^c$ vector of the CSR method (Section 3.1), and
(ii) the students in $\mathcal{S}^c$ and their grades.

This data is used to form a matrix $\mathbf{X}^c \in \mathbb{R}^{(n_c+n_t)\times(m_c+1)}$, where n_c is the number of students in $\mathbf{G}^c$, $n_t = |\mathcal{S}^c|$, and m_c is the number of distinct courses that have at least one grade in $\mathbf{G}^c$ or $\mathcal{S}^c$. The values stored in $\mathbf{X}^c$ are the grades that exist in $\mathbf{G}^c$ and $\mathcal{S}^c$. The last column of $\mathbf{X}^c$ stores the grades $\mathbf{y}^c$ for the course c that were obtained from the students in $\mathbf{G}^c$. Thus, $\mathbf{X}^c$ contains all the prior grades associated with the students who have already taken course c and the students for which we need to have their grade on c predicted. Matrix $\mathbf{X}^c$ is then used in place of matrix $\mathbf{G}$ in Equation 4 to estimate the parameters of the CSMF method, which are then used to predict the missing entries of the last column of $\mathbf{X}^c$, which are the grades that need to be predicted.

4 Experimental Design

4.1 Dataset

The student-course-grade dataset that we used in our experiments was obtained from the University of Minnesota which has a very flexible degree program. It contains the students that have been part of the Computer Science and Engineering (CS&E) and Electrical and Computer Engineering (ECE) programs from Fall of 2002 to Spring of 2014. Both of these degree programs are part of the College of Science & Engineering. Students have to take a common set of core science courses during the first 2–3 semesters, but they can select more courses from different levels and departments.

Because of the nature of these departments, the curriculum coherence tends to be vertically aligned, i.e., what students learn in one lesson, course, or grade level is most likely going to be used by the next lesson, course, or grade level. Students select courses in order to learn the knowledge and skills that will progressively prepare them for more challenging, higher-level topics. However, we need to point out that this might not always be the case, as there are departments that are more horizontally aligned, where there do not exist such strong dependancies across different courses and levels.

While preprocessing the dataset, we removed any courses that are not part of those offered by departments in the college, as these correspond to various liberal arts and physical education courses, which are taken by few students and in general do not count towards degree requirements. Furthermore, we eliminated any courses that were taken as pass/fail. The initial grades were in the A–F scale, which was converted to the 4–0 scale using the standard letter-grade to GPA conversion. The resulting dataset consists of 2,949 students, 2,556 different courses, and 76,748 student-course grades.

We used this dataset to assess the performance of the different methods for the task of predicting the grades that the students will obtain in the last semester (i.e., the most recent semester for which we have data). For this reason, the dataset was further split into two parts, one containing the students that are still *active*, i.e., have taken courses in the last semester (D_{active}) and one that contains the remaining students ($D_{inactive}$). D_{active} contains 876 students, 19,089 grades, out of which 3,427 grades are for the 475 distinct classes taken in the last semester. $D_{inactive}$ contains 2,073 students and 57,659 grades.

These datasets were used to derive various training and testing datasets for the different methods that we developed. Specifically, for the CSR method we extracted the course specific training and testing datasets as follows. For each course c that was offered in the last semester, we extracted course-specific training and testing sets ($D^{c,\geq k}_{train}$ and $D^{c,\geq k}_{test}$) by selecting from $D_{inactive}$ and D_{active}, respectively, the students that have taken c, and prior to taken c, they also took at least k other courses. The reason that these datasets were parametrized with respect to k is because we wanted to assess how the methods perform when different amount of historical student performance information is available. In our experiments we used k in the set $\{5, 7, 9\}$. That information creates the grade matrix $\mathbf{G}^c$, where $g^c_{i,j}$ is the grade of the ith student on the jth course from the training set $D^{c,\geq k}_{train}$. Table 1 shows various statistics about the various course-specific datasets for different values of k.

For the CSMF method, the training dataset for course c was obtained by combining $D^{c,\geq k}_{train}$ and $D^{c,\geq k}_{test}$ into a single matrix after removing the grades that the target students achieved in course c.

For the MF method, the matrix $\mathbf{G}$ is constructed using data from all $\mathbf{X}^c$ matrices. It refers to the union

Table 1 Statistics for Course-Specific datasets.

	CS&E courses			ECE courses		
Prior courses	5	7	9	5	7	9
Average number of students in training set	386	325	258	414	377	332
Average number of students in test set	41	37	29	34	33	32
Average number of prior courses	178	176	173	158	156	155
Average number of grades	5,671	5,186	4,484	7,084	6,804	6,366
Courses predicted	24	24	24	25	25	25
Grades predicted	1,004	910	712	858	841	800

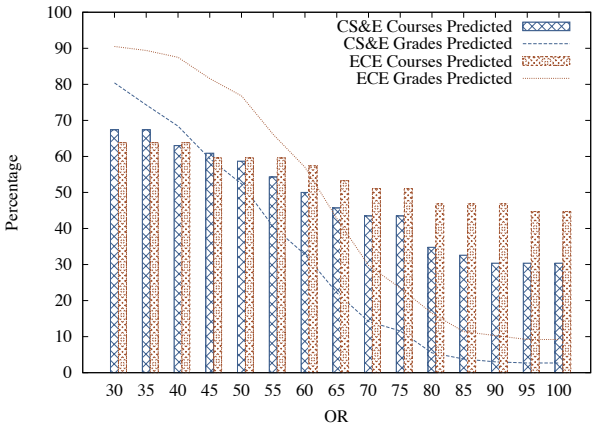

Fig. 1 Statistics of the datasets used in SSR w.r.t. overlap ratio.

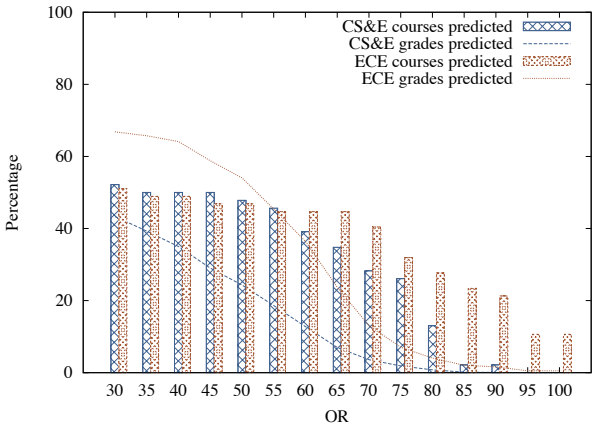

Fig. 2 Statistics of the common subset of datasets used in SSR and in course specific approaches w.r.t. overlap ratio.

of the sets $D_{train}^{c,\geq k}$ and $D_{test}^{c,\geq k}$ for every course to be predicted, after removing the grades that the active students achieved in the courses we want to predict. We formulated the dataset in this way in order to provide the same information for training and testing to all our models. Moreover, since we predict the grades for a specific semester, matrix $\mathbf{G}$ does not contain any grading information regarding following semesters.

In the SSR, the grade matrix $\mathbf{G}^{s,c}$ is created by selecting from $D_{train}^{c,\geq k}$ the set of courses that were also taken by student s and the set of students whose OR with s is at least t. Figure 1 shows some statistics about these datasets as a function of t, and Figure 2 shows only the common subsets that can be predicted by both course specific and SSR datasets. When the OR is more than 0.8, we cannot predict many grades because there are not enough students that had followed the same degree plan as the selected student.

Finally, we did not consider the courses that have less than 20 students in their corresponding dataset, as we consider them to have too few training instances for reliable estimation, or less than 4 test students, as we might not get valid results.

4.2 Competing Methods

In our experiments, we compared our methods with the following competing approaches.

1. **BiasOnly.** We only took into consideration local and global biases to predict the students' grades. These biases were estimated using Eqn. 4 by setting $l = 0$.
2. **Student-Based Collaborative Filtering (SBCF).** This method implements the approach described in [3]. For a target course c, every student i is represented by a vector whose non-zero entries are the grades that the student obtained on the courses taken prior to c. We compare the vector of a target student s against the vectors of the other students that have taken course c using the Pearson's correlation coefficient. We perform grade prediction while taking into consideration the positively similar students to s according to

$$\hat{g}_{s,c} = \bar{g}_s + \frac{\min(r, nbr)}{r} \frac{\sum_{i=1}^{nbr} (g_{i,c} - \bar{g}_i)\ \mathrm{sim}_{s,i}}{\sum_{i=1}^{nbr} \mathrm{sim}_{s,i}}, \quad (5)$$

where nbr is the number of students selected, r is a confidence lower limit for significance weighting, $\bar{g}_i$ is the average grade of the student prior taking c,

and $\text{sim}_{s,i}$ represents the similarity of target student s with i.

4.3 Parameters and Model Selection

For CSR, we let λ_1 take values from 0 to 40 in increments of 1 and λ_2 from 0 to 50 in increments of 1. For SSR, we let λ_1 take values from 0 to 10 in increments of 1 and λ_2 from 0 to 14 in increments of 2. For BiasOnly, MF and CSMF, we let λ take values from 0 to 16 in increments of 0.05. For SSR, the range of the tested values for overlap ratio is 0.3 to 1, in increments of 0.04 and for the confidence lower limit is 10 to 100, in increments of 10. For SBCF, we tested the number of neighbors to be from 10 to 100 with increments of 10. For MF and CSMF methods we tested the number of latent dimensions with the values 2, 5 and 8.

For SBCF, CSR and SSR, we used the semester before the target semester to estimate and select the best parameters. For BiasOnly, MF and CSMF, model selection was based on the performance of the validation set, which was a randomly selected 10% subset of the training data. For the CSMF model, the best-performing parameters were selected for each course.

4.4 Evaluation Methodology & Performance Metrics

We evaluated the performance of the different approaches by using them to predict the grades for the last semester in our dataset using the data from the previous semesters for training. We report the results for the courses belonging to CS&E and ECE departments.

We assessed the performance using the root mean square error (RMSE) between the actual grades and the predicted ones. Since the courses whose grades are predicted have different number of students, we computed two RMSE-based metrics. The first is the overall RMSE in which all the grades across the different courses were pooled together, and the second is the average RMSE obtained by averaging the RMSE values for each course. We will denote the first by RMSE and the second as AvgRMSE.

In order to get a better understanding of the quality of the predictions, we also report the distribution of the actual vs predicted letter grades. The grading system used by the University of Minnesota has 11 letter grades (A, A-, B+, B, B-, C+, C, C-, D+, D, F) that correspond to grades from 4 to 0 (4, 3.667, 3.333, 3, 2.667, 2.333, 2, 1.667, 1.333, 1, 0). After converting the predicted grades to their closest letter grade, we compute the percentage of grades that are within or more than x ticks away from their actual grades. A tick is defined as the difference between two successive letter grades (e.g., B vs B+ is one tick, A vs B is 3 ticks).

5 Experimental Results

5.1 Course-Specific Regression

Table 2 shows the performance achieved by the CSR and CSR-RC models when trained using the three different data sets discussed in Section 4.1. These results show that between the two models, CSR-RC, which operates on the GPA-centered grades, leads to considerably lower errors both in terms of RMSE and AvgRMSE, especially for the CS&E courses.

In terms of the sensitivity of their performance on the amount of historical information that was available when estimating these models (i.e., the minimum number of prior courses), we can see that the performance of the models does not change significantly for the CSR-RC method. CSR predicts CS&E courses better when using 5 prior courses, while it predicts better the ECE courses with 9 prior courses. This indicates that the model benefits from increased number of students that increased number of prior courses, because the students with 9 prior courses are only 67% of the students with 5 prior courses. The ECE department does not suffer from such low number of students left with 9 prior courses, as the corresponding percentage is 80% (statistics according to Table 1).

5.2 Student-Specific Regression

As one of the parameters for this problem was the overlap ratio between the courses of the target student and other students, Figure 3 presents the behavior of the model's RMSE (left) and AvgRMSE (right) as we vary the OR for $D_{test}^{c,\geq 9}(k = 9)$. When the OR is increased, the selected students have more courses in common with the target user and that leads to better performance.

In order to compare the performance of SSR against CSR-RC, Figure 4 shows the RMSE of the best CSR-RC and SSR models. The RMSE values were computed on the subsets of the test set that was predicted by both models for $D_{test}^{c,\geq 9}(k = 9)$. These results show that SSR leads to consistently worse predictions for the CS&E courses than the CSR-RC model. However, in the case of the ECE courses, SSR does better than CSR-RC when the OR is greater than 0.8. That might be related to the fact that the degree program of ECE is

Table 2 The performance achieved by Linear Course-Specific Regression per department.

	CS&E courses						ECE courses					
	RMSE			AvgRMSE			RMSE			AvgRMSE		
Prior courses	5	7	9	5	7	9	5	7	9	5	7	9
CSR	0.928	0.958	0.990	0.994	1.034	1.082	0.717	0.693	0.704	0.702	0.685	0.699
CSR-RC	0.727	0.725	0.722	0.726	0.726	0.716	0.634	0.632	0.634	0.651	0.646	0.651

The performance of the models trained on the different datasets were evaluated on the $D_{test}^{\geq 9}$ test set, which is the common subset among their respective test sets.

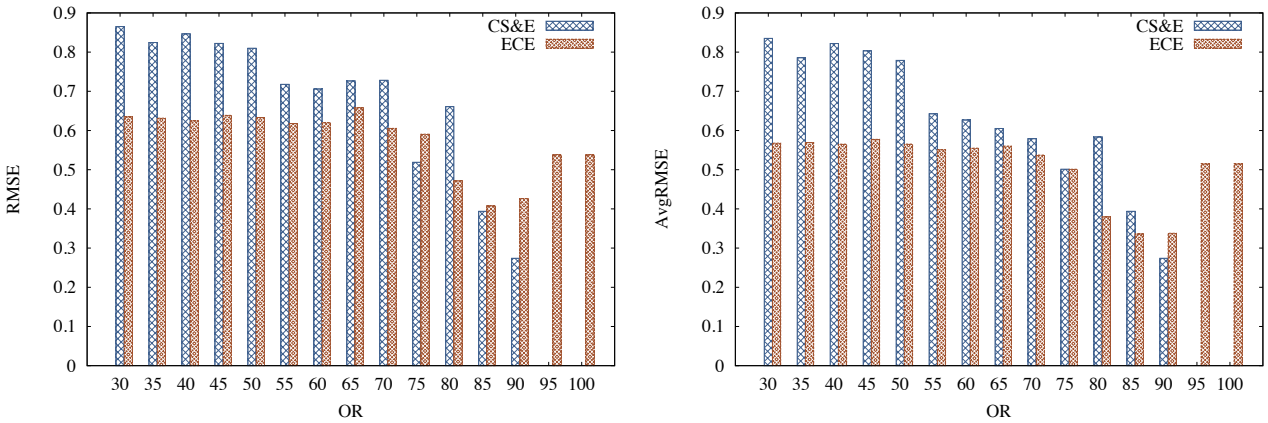

Fig. 3 The performance achieved by the SSR model w.r.t. overlap ratio.

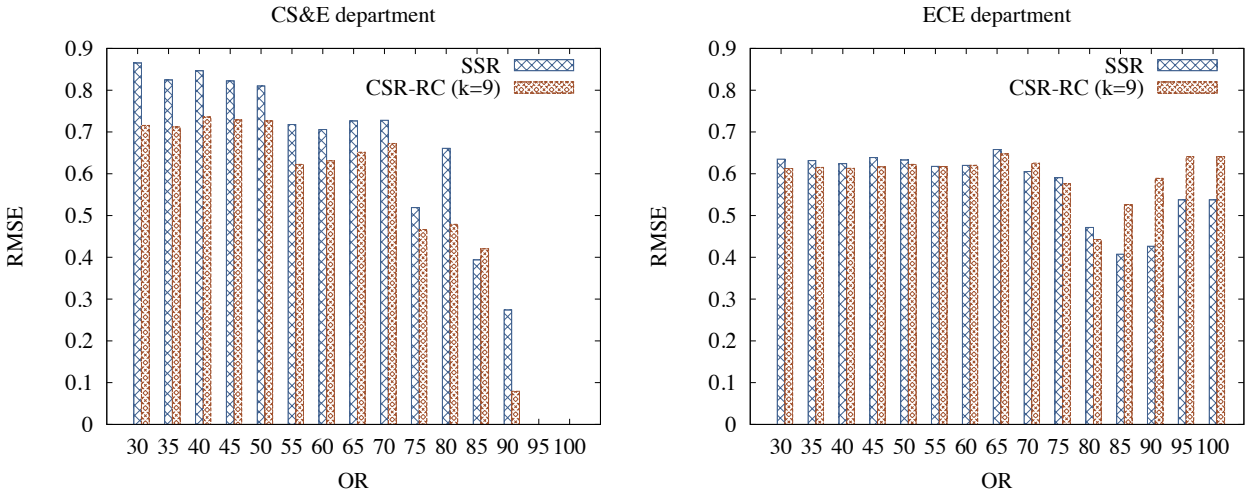

Fig. 4 Comparison of SSR model and CSR-RC with 9 prior courses w.r.t. overlap ratio. The other options for number of prior courses have similar behavior.

more structured than the CS&E degree program, giving some advantage to the SSR method. As shown in Figure 1, at such high OR values, the number of grades that can be predicted by SSR is small. For example, when OR is 0.8, the SSR model can predict less than 10% of the grades in the target semester.

5.3 Methods based on Matrix Factorization

The performance of the methods based on matrix factorization (Section 3.3) is shown in Table 3.

These results show that for the CS&E courses, CSMF performs the best in terms of RMSE and AvgMSE, for any number of prior courses. That confirms that by building matrix factorization models on smaller but denser course-specific sub-matrices, we can derive low-rank models that lead to more accurate matrix com-

Table 3 Errors per department for Matrix Factorization methods.

			CS&E courses		ECE courses	
Prior courses	Latent Factors		MF	CSMF	MF	CSMF
5	2	RMSE	0.740	0.734	0.603	0.606
	5		0.753	0.731	0.605	0.616
	8		0.735	0.734	0.596	0.602
	2	AvgRMSE	0.726	0.716	0.614	0.615
	5		0.732	0.717	0.608	0.628
	8		0.721	0.714	0.605	0.612
7	2	RMSE	0.741	0.739	0.606	0.615
	5		0.750	0.735	0.611	0.607
	8		0.744	0.734	0.598	0.601
	2	AvgRMSE	0.726	0.729	0.610	0.626
	5		0.720	0.711	0.607	0.617
	8		0.727	0.728	0.604	0.609
9	2	RMSE	0.740	0.735	0.604	0.603
	5		0.746	0.723	0.600	0.601
	8		0.751	0.733	0.597	0.598
	2	AvgRMSE	0.726	0.732	0.611	0.617
	5		0.721	0.714	0.601	0.611
	8		0.735	0.725	0.607	0.610

pletion. On the other hand, the performance of the ECE courses does not vary a lot. For that department, the best predictions are performed by MF, followed by CSMF with a RMSE difference of 0.002. A potential explanation for these results is that the ECE courses are part of a stricter degree program, whose structure is present even in the more general setting of MF. As a result, by selecting the course-specific sub-matrices does not provide any further insight to the data, as happens for the CS&E courses.

In order to see how the size of the training set associated with the different courses impacts the performance of the MF and CSMF methods, Figure 5 shows the cumulative AvgRMSE over the courses with increasing training size and the RMSE per course achieved from each method. Cumulative AvgRMSE is used to provide some insight to the impact that the training size has on the performance of our models. We can notice that for the ECE courses, MF model has an advantage against CSMF for relatively smaller courses. MF performs better for eight out of the ten smallest courses, indicating that it gains its accuracy by utilizing other data that are not included in the course specific datasets in order to compute better biases. Moreover, from the bottom part of the figure, we can confirm that the performance of both MF and CSMF is similar for the ECE courses in comparison to the CS&E courses.

In terms of the number of latent factors, we see that when we are using the smallest dataset for training (the one with 9 prior courses), the best performance is achieved for smaller number of latent factors compared to the datasets with 5 or 7 prior courses. In that case, the average number of grades per course is lower, which might not support a large number of latent factors.

5.4 Comparison with other methods

Table 4 compares the performance of the baseline approaches described in Section 4.2 (BiasOnly and SBCF) with the best-performing course-specific regression method (CSR-RC), the MF and CSMF methods. From these results we can see that CSR-RC leads to the best RMSE for the CS&E courses and MF leads to the best RMSE for the ECE courses, closely followed by CSMF (0.002 difference).

A summary of the comparison between every pair of methods tested can be found on Table 5. For each method, we count the courses for which a method wins, ties and losses in terms of RMSE against each other method tested. This analysis shows that for the CS&E courses, CSR-RC outperforms the other methods, except SBCF that is very close, in the majority of the courses, whereas for the ECE courses, the CSMF outperforms each one of the other methods (even MF method that has slightly better RMSE) in the majority of the courses.

5.5 Fine grain analysis of the predictions

In order to gain a better understanding as to the types of errors generated by the different methods and the real-world implication of the predictions Tables 6 and 7 analyze the performance achieved by the different

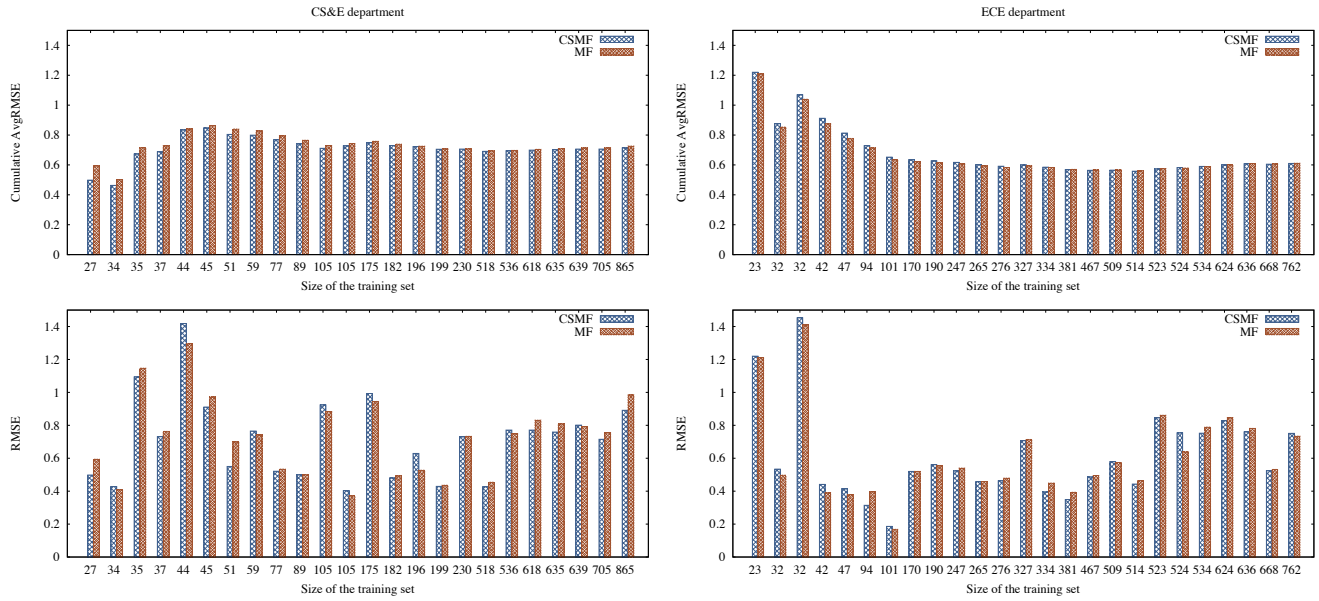

Fig. 5 Cumulative AvgRMSE w.r.t. increasing training size (top) and RMSE achieved per course (bottom) of CSMF and MF models for $D_{test}^{c,\geq 9}(k=9)$.

Table 4 Errors per department.

	CS&E courses						ECE courses					
	RMSE			AvgRMSE			RMSE			AvgRMSE		
Prior courses	5	7	9	5	7	9	5	7	9	5	7	9
BiasOnly	0.752	0.752	0.754	0.740	0.734	0.738	0.633	0.634	0.634	0.642	0.640	0.642
SBCF	0.733	0.733	0.732	0.713	0.713	0.710	0.619	0.619	0.619	0.621	0.620	0.620
CSR-RC	0.727	0.725	0.722	0.726	0.726	0.716	0.634	0.632	0.634	0.651	0.646	0.651
MF	0.735	0.741	0.739	0.726	0.726	0.726	0.596	0.598	0.597	0.605	0.604	0.607
CSMF	0.731	0.734	0.722	0.717	0.728	0.714	0.602	0.601	0.598	0.612	0.609	0.610

The performance of the models trained on the different datasets were evaluated on the $D_{test}^{\geq 9}$ test set, which is the common subset among their respective test sets.

Table 5 Wins/Ties/Losses for every pair of methods tested.

	CS&E courses					ECE courses				
	OnlyBias	SBCF	CSR-RC	MF	CSMF	OnlyBias	SBCF	CSR-RC	MF	CSMF
OnlyBias		7/1/16	7/1/16	6/5/13	7/2/15		8/2/15	10/3/12	7/6/12	6/2/17
SBCF	16/1/7		12/1/11	11/3/10	13/2/9	15/2/8		13/4/8	8/4/13	8/4/13
CSR-RC	16/1/7	11/1/12		13/2/9	12/4/8	12/3/10	8/4/13		7/4/14	6/3/16
MF	13/5/6	10/3/11	9/2/13		9/2/13	12/6/7	13/4/8	14/4/7		8/3/14
CSMF	15/2/7	9/2/13	8/4/12	13/2/9		17/2/6	13/4/8	16/3/6	14/3/8	

The cell (i, j) refers to the wins/ties/losses of the i-th method compared to the corresponding j-th method.

methods by focusing on grade ticks as opposed to RMSE values.

Table 6 shows the percentage of predicted grades that were close to the true grades, over all the instances predicted by a model. For the CS&E department, CSMF is the model with the most grades that are predicted to be within two ticks from their true values, while CSR-RC is the best model when focusing on exact predictions. For the ECE department, MF has the highest percentages, and CSMF can be better only for the case of 9 prior courses, within two letter grades from the actual grades.

Table 7 analyzes the performance of the models on the instances that they fail to accurately predict. We examine the difference between the grades over or under predicted, i.e., they are predicted to be more or less

than their real values respectively. In this case, the lower the percentage, the better the model is, as there are less inaccurate predictions. These results show that, compared the CS&E, ECE has less under predictions, but higher number of over predictions of more than one tick. Moreover, the best methods for the CS&E courses are the CSR-RC and CSMF, and for the ECE courses, are the MF and CSMF. Another finding is that CSR-RC has the highest percentages of under prediction errors for the ECE department. The reason this is happening is because a student might have not taken an important course, and its corresponding regressor will be missing while estimating their grade. As a result, we can see that in the case of this department, that has a stricter degree program, CSR-RC (that is a linear model) cannot handle the absence of an important prior course. However, CSR-RC is the only model that manages to lower the over prediction error while using more dense data (case of 9 prior courses).

Table 8 compares the RMSE per course for the methods of BiasOnly, SBCF, CSR-RC, CSMF and MF, for both the CS&E and ECE departments. Some statistical information per course is also included. This information suggests that if a course has a poor RMSE, then it is very likely that the standard deviation of the grades on the test set is quite high or higher that the standard deviation of the grades on the training set.

6 Conclusions

In this paper, we presented two course-specific approaches based on linear regression and matrix factorization that perform better than existing approaches based on traditional methods, assuming that the degree programs involved have a vertical structure. In that case, focusing on a course specific subset of the data can result in more accurate predictions. Moreover, the performance for different departments can significantly vary, as they may have different characteristics and structures. A student-course specific approach was also developed but its accuracy in grade prediction is limited by the diverse nature of degree plans. Overall, the course-specific methods can improve the performance of grade prediction over other methods tested for our dataset, while the degree of improvement depends on the department.

Acknowledgements This work was supported in part by NSF (IIS-0905220, OCI-1048018, CNS-1162405, IIS-1247632, IIP-1414153, IIS-1447788) and the Digital Technology Center at the University of Minnesota. Access to research and computing facilities was provided by the Digital Technology Center and the Minnesota Supercomputing Institute. http://www.msi.umn.edu

References

1. Al-Barrak, M.A., Al-Razgan, M.: Predicting students final gpa using decision trees: A case study. International Journal of Information and Education Technology 6(7), 528 (2016)
2. Arnold, K.E., Pistilli, M.D.: Course signals at purdue: using learning analytics to increase student success. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. pp. 267–270. ACM (2012)
3. Bydžovská, H.: Are collaborative filtering methods suitable for student performance prediction? In: Progress in Artificial Intelligence, pp. 425–430. Springer (2015)
4. Chen, Y., Bhojanapalli, S., Sanghavi, S., Ward, R.: Coherent matrix completion. arXiv preprint arXiv:1306.2979 (2013)
5. Denley, T.: Course recommendation system and method. http://www.google.com/patents/US20130011821, [Online; accessed 4 October 2015]
6. Denley, T.: Austin peay state university: Degree compass. EDUCAUSE Review Online. Available: http://www.educause.edu/ero/article/austin-peay-state-university-degree-compass (2012)
7. Elbadrawy, A., Studham, R.S., Karypis, G.: Collaborative multi-regression models for predicting students' performance in course activities. In: Proceedings of the Fifth International Conference on Learning Analytics And Knowledge. pp. 103–107. ACM (2015)
8. Hwang, C.S., Su, Y.C.: Unified clustering locality preserving matrix factorization for student performance prediction. IAENG International Journal of Computer Science 42(3) (2015)
9. Kantor, P.B., Rokach, L., Ricci, F., Shapira, B.: Recommender systems handbook. Springer (2011)
10. Luo, J., Sorour, E., Goda, K., Mine, T.: Predicting student grade based on free-style comments using word2vec and ann by considering prediction results obtained in consecutive lessons. Proceedings of the 8nd International Conference on Educational Data Mining pp. 396–399 (June 2015)
11. McKay, T., Miller, K., Tritz, J.: What to do with actionable intelligence: E 2 coach as an intervention engine. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. pp. 88–91. ACM (2012)
12. Ogunde, A., Ajibade, D.: A data mining system for predicting university students? graduation grades using id3 decision tree algorithm. Journal of Computer Science and Information Technology 2(1), 21–46 (2014)
13. Osmanbegović, E., Suljić, M.: Data mining approach for predicting student performance. Economic Review 10(1) (2012)
14. Pardos, Z.A., Heffernan, N.T.: Using hmms and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset. Journal of Machine Learning Research W & CP (2010)
15. Romero, C., Ventura, S., Espejo, P.G., Hervás, C.: Data mining algorithms to classify students. In: Educational Data Mining 2008 (2008)
16. Sorour, S.E., Mine, T., Goda, K., Hirokawa, S.: A predictive model to evaluate student performance. Journal of Information Processing 23(2), 192–201 (2015)
17. Starfish: Earlyalert. http://www.starfishsolutions.com/home/student-success-solutions/, [Online; accessed 4 October 2015]
18. Thai-Nghe, N., Drumond, L., Horváth, T., Schmidt-Thieme, L.: Using factorization machines for student modeling. In: UMAP Workshops (2012)

Table 6 Analysis of the accuracy of the predictions in terms of letter grades.

		5 prior courses					9 prior courses				
		BiasOnly	SBCF	CSR-RC	MF	CSMF	BiasOnly	SBCF	CSR-RC	MF	CSMF
CS&E	no error	23.73	23.87	**26.40**	25.84	24.85	25.84	24.16	**26.68**	24.58	24.43
	within one tick	62.63	63.59	63.45	**65.02**	62.47	62.08	63.32	63.31	63.04	**63.74**
	within two ticks	81.58	82.95	83.81	81.98	**84.22**	81.46	83.25	84.37	81.98	**84.52**
ECE	no error	30.38	28.38	26.37	**30.13**	26.38	30.98	28.11	26.27	**30.88**	27.40
	within one tick	66.62	65.39	64.89	**67.36**	66.24	65.74	64.98	65.52	**67.51**	66.00
	within two ticks	87.75	88.26	86.50	**90.12**	88.86	87.59	88.09	86.26	88.62	**89.73**

These numbers correspond to the percentage of the predicted grades that were exactly, within one tick or two ticks away from the true letter grade. One tick corresponds to a letter grade away from the true grade, i.e., we predict a grade of B while the student took B- in a course.
While comparing models, the higher the percentage the better it is for the grades predicted exactly, or less than one or two ticks away. For each case, the best percentage is in bold.

Table 7 Analysis of the error severity of the predictions in terms of letter grades.

		5 prior courses					9 prior courses				
		BiasOnly	SBCF	CSR-RC	MF	CSMF	BiasOnly	SBCF	CSR-RC	MF	CSMF
CS&E	underpredict (>1 tick)	20.34	18.93	**18.38**	18.79	21.34	20.76	19.22	**18.94**	19.37	19.22
	overpredict (>1 tick)	16.96	17.38	18.08	**16.10**	16.11	17.10	17.38	17.66	17.51	**16.96**
	underpredict (>2 ticks)	9.53	**7.56**	7.98	8.97	8.00	9.10	7.28	7.14	8.28	**7.00**
	overpredict (>2 ticks)	8.82	9.39	8.12	8.96	**7.70**	9.38	9.39	**8.40**	9.66	**8.40**
ECE	underpredict (>1 tick)	14.74	15.38	15.22	**13.73**	15.74	15.21	15.87	15.08	**12.60**	14.73
	overpredict (>1 tick)	18.60	19.16	19.83	18.84	**17.96**	**18.96**	19.04	19.31	19.83	19.22
	underpredict (>2 ticks)	4.88	3.75	4.99	**2.73**	4.50	4.86	3.75	5.47	**3.61**	3.63
	overpredict (>2 ticks)	7.33	7.92	8.45	7.08	**6.58**	7.86	8.05	8.18	7.71	**6.59**

These numbers correspond to the percentage of the predicted grades that were one or two ticks away from the true letter grade. One tick corresponds to a letter grade away from the true grade, i.e., we predict a grade of B while the student took B- in a course.
A model under or over predicts when the grade predicted is lower or higher, respectively, than the actual one.
While comparing models, the lower the percentage the better it is for the grades predicted more than one or two ticks away. For each case, the best percentage is in bold.

19. Toscher, A., Jahrer, M.: Collaborative filtering applied to educational data mining. KDD cup (2010)

Table 8 Errors per course for all methods for the case of 9 prior courses ($D_{test}^{c,\geq 9}(k=9)$).

Course	train	test	feat	nnz	offer	Mn Tr	StD Tr	Mn Te	StD Te	OnlyBias	SBCF	CSR-RC	MF	CSMF
CSCI2x	89	6	164	1345	23	3.180	0.596	3.000	0.577	0.492	0.357	0.454	0.499	0.499
CSCI2x	196	19	167	2796	23	2.697	1.109	2.825	0.670	0.511	0.540	0.531	0.527	0.629
CSCI2x	105	21	165	1843	6	2.686	1.011	3.032	0.625	0.390	0.466	0.416	0.373	0.403
CSCI3x	705	45	240	10272	21	3.135	0.813	3.104	0.852	0.738	0.651	0.803	0.755	0.715
CSCI4x	639	53	234	9774	23	2.847	0.821	2.799	0.912	0.795	0.809	0.788	0.792	0.801
CSCI4x	635	25	230	9147	23	3.035	0.906	2.747	0.981	0.787	0.812	0.756	0.809	0.759
CSCI4x	865	56	252	13502	23	3.099	0.924	3.018	1.191	0.952	1.005	0.852	0.984	0.891
CSCI4x	618	52	225	11379	19	3.530	0.630	3.141	0.904	0.857	0.815	0.736	0.831	0.771
CSCI4x	105	15	168	2136	20	2.797	1.048	3.400	0.762	0.928	0.916	0.879	0.884	0.924
CSCI4x	536	45	219	9593	21	3.173	0.784	3.015	0.886	0.826	0.747	0.757	0.749	0.771
CSCI4x	230	87	193	5198	4	3.229	0.826	3.134	0.849	0.760	0.732	0.738	0.733	0.730
CSCI4x	518	55	219	9448	20	3.094	0.813	3.345	0.422	0.513	0.433	0.502	0.454	0.427
CSCI5x	175	28	180	3409	13	3.154	0.773	2.738	1.146	0.942	0.947	0.980	0.945	0.993
CSCI5x	37	8	123	849	7	3.441	0.658	3.458	0.686	0.916	0.855	0.864	0.763	0.731
CSCI5x	59	15	132	1444	9	3.057	0.929	3.511	0.569	0.806	0.637	0.670	0.741	0.765
CSCI5x	34	10	96	804	5	3.167	0.901	3.767	0.300	0.400	0.372	0.524	0.408	0.428
CSCI5x	27	15	128	736	10	2.518	1.212	3.045	0.619	0.622	0.643	0.822	0.594	0.498
CSCI5x	182	65	231	4195	21	2.984	0.920	3.149	0.627	0.522	0.515	0.481	0.492	0.480
CSCI5x	51	9	131	1040	10	2.869	1.065	3.519	0.419	0.548	0.525	0.587	0.699	0.550
CSCI5x	45	15	119	1039	4	2.593	0.973	3.022	1.078	0.955	0.897	0.836	0.972	0.911
CSCI5x	35	15	135	866	5	2.771	1.053	3.089	1.380	1.161	1.102	1.086	1.147	1.094
CSCI5x	44	6	115	924	9	2.667	1.061	3.055	1.420	1.310	1.389	1.170	1.296	1.418
CSCI5x	77	17	122	1678	12	3.043	0.663	3.059	0.649	0.554	0.484	0.520	0.534	0.521
CSCI5x	199	30	167	4222	9	3.201	0.713	3.222	0.450	0.437	0.389	0.432	0.436	0.429
EE2x	334	19	111	4143	23	3.083	0.812	2.807	0.511	0.428	0.423	0.408	0.449	0.395
EE2x	467	33	185	6748	23	2.720	0.854	2.788	0.724	0.497	0.506	0.470	0.495	0.486
EE3x	509	19	139	7421	23	2.898	0.865	3.053	0.774	0.533	0.578	0.570	0.574	0.579
EE3x	624	43	181	11045	22	2.707	0.846	2.481	1.069	0.835	0.849	0.837	0.846	0.828
EE3x	32	5	92	629	13	3.552	0.739	3.800	0.400	0.578	0.482	0.647	0.497	0.534
EE3x	524	16	149	7758	22	3.357	0.689	3.813	0.333	0.736	0.681	0.766	0.638	0.754
EE3x	668	61	201	12230	21	3.564	0.565	3.404	0.392	0.651	0.602	0.607	0.532	0.524
EE3x	523	18	157	7812	22	2.683	0.926	2.537	1.112	0.866	0.856	0.823	0.861	0.846
EE3x	636	45	183	11597	23	2.759	0.879	2.644	0.970	0.777	0.768	0.860	0.781	0.760
EE3x	534	35	170	9141	22	2.917	0.835	2.638	0.889	0.783	0.749	0.788	0.789	0.752
EE4x	247	14	158	5070	11	2.831	0.931	3.071	0.402	0.541	0.457	0.554	0.540	0.523
EE4x	170	58	128	4631	3	2.998	0.782	3.052	0.729	0.521	0.535	0.564	0.520	0.520
EE4x	42	16	91	1161	3	3.786	0.674	3.958	0.110	0.451	0.368	0.420	0.389	0.441
EE4x	276	53	177	6679	10	2.992	0.897	2.918	0.666	0.452	0.479	0.468	0.479	0.463
EE4x	94	23	144	2269	10	3.784	0.643	3.942	0.188	0.501	0.422	0.496	0.396	0.314
EE4x	265	43	179	6485	11	2.782	0.794	2.876	0.810	0.444	0.446	0.443	0.458	0.458
EE4x	327	25	191	7237	21	3.002	0.815	3.213	0.810	0.706	0.725	0.672	0.714	0.707
EE4x	190	59	155	5029	5	2.942	0.739	2.780	0.701	0.550	0.567	0.584	0.554	0.561
EE4x	514	57	182	11080	10	3.243	0.652	3.398	0.644	0.470	0.463	0.462	0.464	0.443
EE4x	381	47	171	8313	13	3.802	0.446	3.901	0.167	0.541	0.460	0.500	0.393	0.349
EE4x	762	60	226	17340	21	3.627	0.449	3.706	0.713	0.855	0.812	0.774	0.734	0.751
EE4x	101	17	176	2680	9	3.865	0.345	3.882	0.196	0.208	0.174	0.187	0.168	0.186
EE5x	47	21	121	1270	3	3.702	0.675	3.905	0.426	0.455	0.447	0.504	0.378	0.414
EE5x	23	8	97	647	9	3.478	0.714	2.958	1.172	1.240	1.228	1.206	1.209	1.220
EE5x	32	5	116	743	10	3.219	0.762	3.133	1.572	1.427	1.435	1.662	1.411	1.453

feat = features, offer = offerings, Mn = Mean, Tr = Train, Te = Test, StD = Standard Deviation.
The second and third columns refer to the number of students in the training and test set, respectively.
From the course names, we can see the department and the academic level of the course.