A Data Mining Approach to Understanding Curriculum-Level Factors That Help Students Persist and Graduate

Paul Previde
Department of Computer Science
San Francisco State University
San Francisco, California, USA
pprevide@mail.sfsu.edu

Celia Graterol
Metro College Success Program
San Francisco State University
San Francisco, California, USA
celiag@sfsu.edu

Mary Beth Love
Metro College Success Program
San Francisco State University
San Francisco, California, USA
love@sfsu.edu

Hui Yang
Department of Computer Science
San Francisco State University
San Francisco, California, USA
huiyang@sfsu.edu

***Abstract*—This Research Full Paper describes the analysis of curriculum-level factors that affected the persistence and graduation outcomes of over 4,000 undergraduate students at San Francisco State University. This work addressed four questions: (1) how did the timing of students' Mathematics courses affect their performance and outcome; (2) whether students who progressed farther through the prescribed foundation course sequences of the university's long-duration learning community program exhibited higher persistence and graduation rates; (3) what were the most frequently-taken sequences of courses, and whether students who progressed farther through those sequences exhibited higher graduation rates; and (4) whether greater progress was more important than other demographic and academic factors for predicting persistence and graduation. We found that students who took their first non-remedial Math course in the second year showed higher fifth-term and seventh-term persistence than students who took it in the first year. Also, students who progressed farther through their chosen or prescribed sequences consistently exhibited higher persistence and graduation rates. Furthermore, a student's persistence was a more reliable predictor of graduation than other features. Overall, these findings can potentially inform an institution's strategies for maximizing persistence and graduation by emphasizing a student's progress through the curriculum.**

***Index Terms*—curriculum design, educational data mining**

I. Introduction

Higher education institutions face ongoing pressure to maximize their retention and graduation rates despite constrained resources. The graduation rate for first-time, full-time freshmen at 4-year institutions was 40.6% nationally for the 2010 starting cohort [1]. Retention also remains a challenge, with 75.3% of first-time, full-time freshmen retained for their second year for the 2015 starting cohort nationally [2].

Data mining [4], [28], [30] and machine learning [13], [27] have been employed to find the strongest predictors of a student's persistence and graduation outcomes based on various demographic and personal factors, but the impact of factors that operate at the curriculum level, rather than at the demographic or personal level, have been less extensively researched. Specifically, the impact of the ordered sequence of courses taken by a student on the likelihood of that student to persist and graduate has not been studied in depth [6]. Furthermore, while the significance of early Mathematics courses has been recognized [6], the effect of when the first Mathematics course is taken within the first two years has not been extensively studied.

The challenges of retention and graduation are especially acute for under-represented, under-privileged, and/or first-generation students. For example, national graduation rates for black and Hispanic first-time, full-time freshmen at 4-year institutions was 21.4% and 31.7% respectively for the 2010 starting cohort, compared to 45.2% for white students [1]. The Metro College Success Program ("Metro"), a long-duration learning community program at San Francisco State University, is endeavoring to narrow this gap, focusing its efforts on under-represented, Pell-eligible, and/or first-generation students. The Metro program is an evolution of the learning community model [33], [34] in that it provides cohorted enrollment in sequenced courses during the student's first two years, as well as enhanced access to timely support services throughout a student's college education. A novel feature of this program is that it incorporates a carefully-designed sequence of three scaffolded foundation courses, to be taken during the first two years of college [20].

Given the foregoing challenges and research landscape, this work explores the effect of curriculum-level factors, specifically the timing of students' Math courses and progress through course sequences, by analyzing a comprehensive dataset consisting of academic and demographic data for over 2,000 students in the Metro program from 2009 through 2016. The dataset also includes a matched group of over 2,000 nonparticipating students. Except where otherwise indicated, these two student groups are analyzed as a single, combined pool of under-represented, first-generation, and under-privileged students.

This work applies a multifaceted approach to the dataset. First, we examine the timing and performance of students' first non-remedial Math course to elucidate the effect, if any, of the timing of the Math course on performance, persistence, and graduation. Our findings revealed a dependence as to both which Math course was taken, and when it was taken.

Second, we analyze the sequences of scaffolded foundation courses and compare the persistence and graduation rates of students who make differing levels of progress through those sequences. Throughout this work, a student has positive persistence in a particular term (e.g., seventh-term persistence) if she registered for at least one class in that term, counted from the first term in which she was a first-time, full-time freshman. Our findings emphasized the importance of keeping students progressing through their chosen course sequences.

Third, to generalize the above inquiry and further elucidate the relationship between sequence progress and persistence/graduation outcomes, we apply sequential pattern mining to identify the sequences of courses taken by the largest numbers of students. We then compare the persistence and graduation rates of students who make different levels of progress through those sequences. This approach allowed the academic records of the students themselves, rather than any pre-set notions about what courses the students take, to dictate which course sequences to evaluate. The findings consistently revealed the greater importance of the extent of a student's progress through a sequence, as opposed to the content of the sequence itself.

Fourth, we applied different machine learning techniques to assess the relative influence of a student's fifth-term and seventh-term persistence against several demographic and academic factors for predicting graduation status. Our results were consistent with our earlier findings and led to novel insights about the importance of students' curriculum progress.

Overall, this study builds on the relatively sparse prior work that utilizes sequential pattern mining to look at sequences of courses to reveal previously-unknown trends. This work also reveals some of the potential benefits and challenges of sequential pattern mining in analyzing student data.

II. Related Work

Previous studies have employed sequential pattern mining [3] in analyzing various aspects of educational data, including the activities of students in massive open online courses [21] and in small problem-solving groups [24], among many others. Campagni et al. [7] applied sequential pattern mining to computer science students' completion of final examinations. Each student's academic career was represented as a sequence of examinations ordered by semester, and examinations taken during the same semester comprised an non-ordered set. In an ideal sequence, a student took the final exam in the same semester in which she took the associated course; however, students in this institution were not required to obey this ideal scenario. The work computed the bubble sort distance of each student's sequence from the ideal sequence, and found that greater deviation from the ideal sequence led to lower likelihood of graduation.

This work extends the aforementioned applications of sequential pattern mining to the courses taken across multiple semesters by San Francisco State University students to determine whether students who make greater progress through commonly-observed sequences exhibit better persistence or graduation rates. Part of this work resembles that of Gopalakrishnan et al. [13], where various feature ranking algorithms were employed to identify the most influential academic, demographic, and personal features of Metro students for predicting persistence and graduation outcomes. However, this work covers many more students over a longer period. Furthermore, this work also considers features that capture the extent to which students progress through their chosen course sequences.

III. Methods

This section describes the methodologies that we employed to address the aforementioned four problems, which are:

1) The timing of students' completion of certain Math courses, and its effect on persistence and graduation.
2) The effect of increased progress through the Metro program's sequence of three foundation courses on persistence and graduation.
3) The extraction of the most frequently-taken course sequences and the effect of increased progress through those sequences on persistence and graduation.
4) Determination of the relative importance of sequence progress as a feature, compared to demographic and personal features, in predicting graduation outcomes.

Figure 1 depicts the software modules used to answer these questions, each of which is discussed in turn below.

A. Data Collection and Preprocessing

First, we collected the relevant data from the Metro program and from the university's student records and institutional research departments. The general characteristics and features of the dataset are shown in Table I and broken down by student categories: Metro participants, and the matched student group. Students in the latter group were matched with Metro participants based on first-generation status, under-represented minority (URM) status, gender, Pell eligibility and needing remediation in Math and/or English. For each category, the cohorts include the entering classes from 2009 through 2016. In each student's record, the semesters were numbered such that the student's first semester as an entering freshman was semester 1, and so on. Semesters in which a student did not register for classes were still assigned the next number. Student records with missing values for third-, fifth-, and seventh-term persistence and graduation status were removed, because we did not impute missing values for these response variables. Otherwise, missing values were left in place but

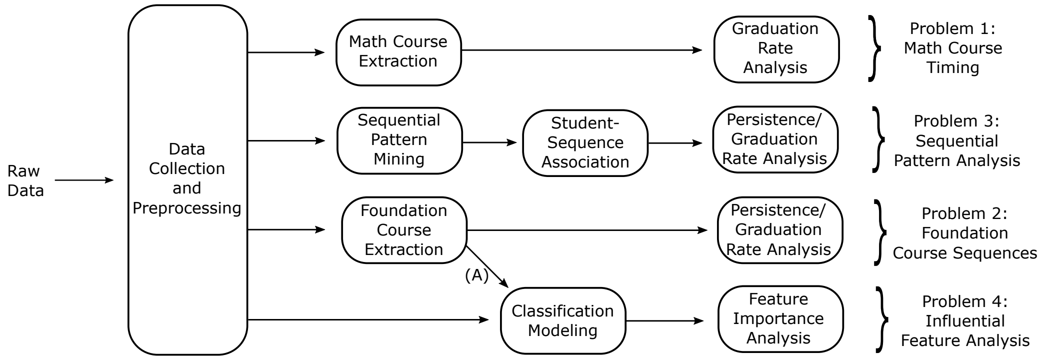

Raw Data
Data Collection and Preprocessing
Math Course Extraction
Sequential Pattern Mining
Foundation Course Extraction
(A)
Student-Sequence Association
Classification Modeling
Graduation Rate Analysis
Persistence/ Graduation Rate Analysis
Persistence/ Graduation Rate Analysis
Feature Importance Analysis
Problem 1: Math Course Timing
Problem 3: Sequential Pattern Analysis
Problem 2: Foundation Course Sequences
Problem 4: Influential Feature Analysis

Fig. 1. Major functional modules associated with each problem addressed in this work.

omitted for statistical calculations. Table I indicates the size of the resulting dataset.

We implemented the data collection and preprocessing steps using the Python programming language (v. 2.7.12) and the Pandas data manipulation and analysis library [22].

TABLE I
CHARACTERISTICS OF THE STUDENT DATASETS.

	Long-Duration Learning Community	Matched Student Group
Number of Students	2,281	2,276
Features	Race Mother's education Father's education Gender Pell eligibility Course registrations Course grades 3rd term persistence 5th term persistence 7th term persistence Graduation status ELM, EPT scores Cohort assignment	Race Mother's education Father's education Gender Pell eligibility Course registrations Course grades 3rd term Persistence 5th term Persistence 7th term Persistence Graduation status ELM, EPT scores

B. Math Course Analysis

The Metro program assigns its students to different academies and provides students with registration assistance, advising, tutoring, and many other services at no cost to the student [20]. Each academy provides a pathway of recommended courses (some of which are required, as described in the next section). Each academy's pathway of courses includes a recommended Mathematics course. Some of the Mathematics courses are precalculus or calculus courses oriented towards students interested in or majoring in STEM subjects, while other courses are less technical and geared towards non-STEM areas (Table II). For brevity, the calculus-related courses will be referred to as "STEM" and the others as "non-STEM."

TABLE II
MATH COURSES STUDIED IN THIS WORK.

Category	Course Name	Subject
STEM	MATH110	Business Calculus
	MATH199	Pre-Calculus
	MATH226	Calculus I
non-STEM	MATH124	Elementary Statistics
	ISED160	Data Analysis in Education
	PSY171	Quantitative Reasoning in Psychology

Using the set of student records described in the preceding subsection, we extracted the semester number in which each student in the combined pool of the Metro program and the matched student group took the recommended Math courses (Figure 1, Problem 1). We also extracted the student's grade in that course and fifth-term persistence, seventh-term persistence, and graduation outcome. We then evaluated the performance of the students in each of the Math courses in Table II. We also determined the persistence and graduation rates of students, broken down by whether they took a STEM or non-STEM Math course and whether they took that course in the first year (semesters 1 or 2) or second year (semesters 3 or 4). We used the Chi-square statistical test to validate whether there was a relationship between the year in which the student took the Math course and fifth-term persistence, seventh-term persistence, and graduation status for both the STEM and non-STEM courses. For this task we used the Scipy statistical computing library for Python [32].

Note that throughout this work, the 2009-2015 cohorts were considered when determining fifth-term persistence, 2009-2014 cohorts for seventh-term persistence, and 2009-2013 cohorts for graduation status.

C. Foundation Course Sequence Analysis

The Metro program requires each student to enroll in an academy-specific sequence of three carefully-designed and scaffolded foundation courses that emphasize community-building and social justice. A summary of the three-course sequence and when each course is taken is shown in Figure 2.

Let S represent the set of students in the Metro program, and $F = <f_1 \rightarrow f_2 \rightarrow f_3>$ the sequence of three foundation courses for a student s_j. The pseudo-code for analyzing their progress through F is shown in Figure 3.

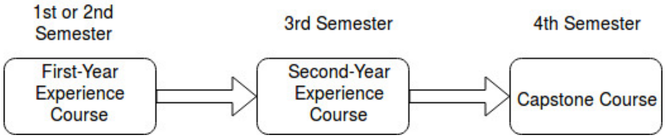
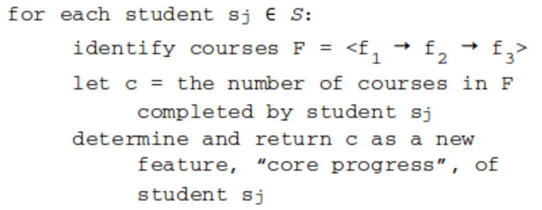


Fig. 2. Foundation course sequence of the Metro program.



```
for each student sj ∈ S:
     identify courses F = <f_1 → f_2 → f_3>
     let c = the number of courses in F
          completed by student sj
     determine and return c as a new
          feature, "core progress", of
          student sj
```

Fig. 3. Pseudo-code for foundation course sequence analysis.

Each student's progress through the three-course sequence of that student's academy is determined based on the number of courses that the student passed. As indicated by arrow (A) in Figure 1, that number of courses becomes a new feature (referred to hereafter as "core progress") for that student and is considered in the feature importance problem described in Section III-E. We aggregated students across all academies by their core progress values and computed the percentage of students at each core progress value (i.e., 1, 2, or 3) with positive fifth-term and seventh-term persistence and graduation outcomes.

D. Sequential Pattern Mining of Course Sequences

A brief description of the concept of sequential pattern mining is presented next [23]. Let S = $\{s_1, s_2, ... s_N\}$ be the collection of N student records, where each record s_i consists of an ordered collection of semesters, each of which in turn consists of the non-ordered set of k_j course(s) $\{c_1, c_2, ... c_{k_j}\}$ taken in semester j:

s_i = <$\{c_1, c_2 ... c_{k_1}\} \rightarrow \{c_1, c_2 ... c_{k_2}\}$ ...>

Moreover, semesters in which no courses were taken are represented as an empty set, as in the following example in which a student took no courses in the second semester:

s_i = <$\{c_1, c_2 ... c_{k_1}\} \rightarrow \{\} \rightarrow \{c_1, c_2 ... c_{k_3}\}$ ...>

Let the minimum support m$_{sup}$ represent the minimum proportion threshold of N student records that contain a given sub-sequence. Sequential pattern mining algorithms extract all sub-sequences that occur in at least N*m$_{sup}$ student records. Consider the following set of three student records:

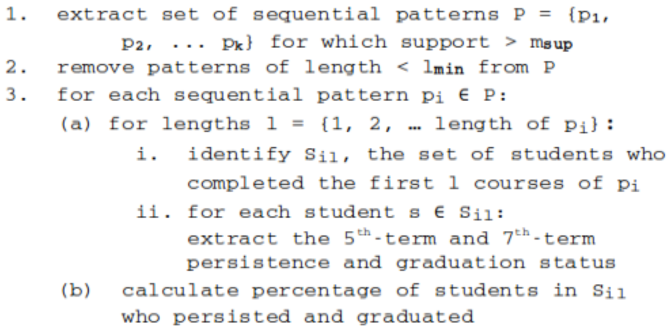
s_1: <{MATH100, ENG30} → {MATH101, ENG35} → {MATH102}>

s_2: <{MATH100, PHIL50} →{PHYS140}>

s_3: <{MATH100} → {MATH102}>

If m$_{sup}$ = 0.5, the sub-sequence <MATH100 → MATH102> would be the only length-2 sequence returned, where it spans over three semesters for Student s_1 and two semesters for Student s_3. The patterns <MATH100> and <MATH102>, each of length 1, would also be returned. Thus, in this work, given the set of student records and a certain m$_{sup}$ threshold, we extract the complete set of sequential patterns in the collection of student records.

Given the collection of course registrations per semester for all students, the user-specified input parameter m$_{sup}$, and minimum length l$_{min}$, then the pseudo-code associated with mining and processing the sequential patterns from the collection of student records in this study is shown in Figure 4. The sequence extraction was not limited to any particular courses, majors, or disciplines.



```
1.  extract set of sequential patterns P = {p_1,
       p_2, ... p_k} for which support > m_sup
2.  remove patterns of length < l_min from P
3.  for each sequential pattern p_i ∈ P:
    (a) for lengths l = {1, 2, … length of p_i}:
           i.  identify S_il, the set of students who
               completed the first l courses of p_i
           ii. for each student s ∈ S_il:
               extract the 5th-term and 7th-term
               persistence and graduation status
    (b) calculate percentage of students in S_il
        who persisted and graduated
```

Fig. 4. Pseudo-code for analysis of the most frequent course sequences.

As depicted in Figure 4, all patterns whose support exceeds m$_{sup}$ and whose length exceeds the specified minimum length l_{min} were extracted from the complete set of student records in steps 1 and 2, yielding a set P of qualifying sequential patterns. In step 3 each qualifying sequential pattern is used as a reference pattern, and the students whose academic records contained this reference pattern are identified and returned. We identified the ten length-3 patterns that were taken by the largest number of students in the dataset. In step 3, students whose records contained a contiguous subset of each of those ten sequential pattern were identified. For instance, if a reference pattern includes the courses <MATH100 → MATH101 → MATH102>, we identified not only students who took that entire pattern, but also students who took only the first n<3 of those courses.

To examine the effect of increased progress through sequences, we subjected the top ten most frequent sequences of length 3, and the top five most frequent sequences of length 5, to further analysis as depicted by Figure 4, 3(a)-3(b). For each length, these collections of most frequent sequences allowed at least ten distinct courses to be considered. For every student in the dataset, we determined how many of the courses in each sequence in the top ten (for length 3) and the top five (for length 5) that the student completed with a C- or better. For example, if a length-3 sequence includes courses <MATH110 → MATH120 → MATH130> and a certain student completed MATH110 and MATH120, that student is given a progress value of 2 for that sequence. Lastly, we aggregated the collections for each of the top ten length-3 sequences and computed the percentage of students at each progress level with positive fifth-term and seventh-term persistence and graduation outcomes. Likewise, for the top five length-5 sequences, the progress indicator ranges from 1 to 5.

We aggregated the collections for each of the top five length-5 sequences and determined the seventh-term persistence and graduation rates as a function of progress level.

We executed sequential pattern mining using the Sequential Pattern Mining Framework (SPMF), a Java-based library of data mining algorithms [10]. Specifically, we chose the PrefixSpan algorithm because of its efficient performance as reported in other studies [19], [23] and because it accepted string sequences [11].

E. Classification Models and Feature Ranking

While the tasks of the prior two subsections analyzed the extent of a student's progress through course sequences, this task weighs the relative importance of that progress against other features, such as race, gender, Pell eligibility, and ELM and EPT entrance scores. To carry out this task, we developed and trained classification models using the features identified in Table I for each of the associated pools of students indicated in the Table. (Note that, in order to maintain a focus on curriculum-level features rather than specific courses and to avoid making comparisons between cohort years, the following features were not considered for this task: individual course registrations or grades; or cohort.) To assess classification accuracy, we compared naive Bayes, logistic regression, and random forest classification models using the Metro program students and their matched group counterparts in a combined pool. Naive Bayes is a well-established and relatively simple probabilistic modeling technique that has demonstrated strong results despite its assumption of the independence of its features in making predictions of output classes [25]. Given a student with a certain set of input feature values, logistic regression models the probability that the student belongs to a given output class, and has been widely used in EDM [14], [27]. Random forest classifiers use decision trees as building blocks, and a random selection of the input features is used to split each node to grow each tree [5]. Moreover, the random forest algorithm can generate rankings of the influence of each input feature on the accuracy of the predictions, yielding the relative importance of each feature [14], [17].

To assess the relative importance of fifth- and seventh-term persistence compared to personal factors, such as race, gender, and entrance scores, we assessed the accuracy of each of the aforementioned models using ten-fold cross-validation for each model type, using the 1,302 available students in the 2009-2013 cohorts of the Metro and matched student groups for graduation analysis. We then used recursive feature elimination [12] in conjunction with the random forest classifier in order to identify the subset of features that gives rise to the best model. The steps of this algorithm are as follows:

1) Determine the importance of each feature for the full classification model, which uses all available input features.
2) Eliminate the feature that had the lowest Gini importance, and build a new classification model with the remaining features.
3) Determine the error rate of the new classification model.
4) Repeat steps 2 and 3 while the number of features k is greater than or equal to a chosen minimum (for example, 1).
5) Identify the set of k features leading to the lowest error rate.

Using the above steps, we started with a random forest model using all available features and then eliminated features with the lowest importance one at a time, and the reduced model with the lowest error rate was identified. We implemented the development, training, and tuning of each model, as well as recursive feature elimination, using the Scikit-Learn machine learning library for Python [29].

IV. Results and Discussion

This section presents and discusses the main findings and their implications for the four problems addressed in this work.

A. Math Course Timing

Table III presents the mean and standard deviation of the grades obtained by students who took each of the indicated Math courses, based on the semester in which the course was taken. For each course, the term in which students achieved the highest mean grade is shown in bold, with the median letter grade also shown.

TABLE III
Performance in Math courses by term taken (highest mean grade in each course in bold).

Category	Course	Mean/SD Grade Points and Median Letter Grade			
		Term 1	Term 2	Term 3	Term 4
STEM	Business Calculus	**2.48 ± 1.41 (B-)**	2.38 ± 1.35 (B-)	1.91 ± 1.38 (C)	2.18 ± 1.28 (C+)
	Precalculus	**2.78 ± 1.08 (B)**	2.57 ± 1.23 (B)	2.12 ± 1.32 (C+)	2.13 ± 1.34 (C)
	Calculus I	**2.77 ± 1.25 (B)**	2.36 ± 1.19 (C+)	2.12 ± 1.27 (C)	2.20 ± 1.28 (C+)
non-STEM	Elementary Statistics	**2.61 ± 1.26 (B)**	2.50 ± 1.21 (B)	2.18 ± 1.33 (C+)	2.31 ± 1.27 (C)
	Data Analysis in Education	**2.90 ± 1.02 (B)**	2.87 ± 1.26 (B)	2.78 ± 0.95 (B)	2.79 ± 1.18 (B)
	Quantitative Reasoning in Psychology	2.82 ± 0.78 (B)	**3.17 ± 0.81 (B+)**	3.05 ± 1.25 (B+)	2.78 ± 1.48 (B+)

For each Math course, students who took the Math course in the first year (terms 1 or 2) had the highest mean grade. The effect was most pronounced in the Calculus I course, where the mean grade of 2.77 for students who took it in term 1 was 0.65 or 0.57 grade points higher than the students in term 3 or term 4 respectively. Overall, the extent of the advantage for students to take the Math course in the first year was greater for STEM courses than for the non-STEM courses.

We then examined whether taking the Math course in the first versus second year of a student's curriculum led to higher fifth-term persistence, seventh-term persistence, or graduation rates (Table IV). The study revealed statistically significant evidence ($\alpha = 0.05$) that the fifth-term persistence and seventh-term persistence of students who took their STEM or non-STEM Math course in the second year (i.e., third or fourth semester) were higher than that of students who took the Math course in their first year. However, graduation rates were not significantly different. These findings can potentially inform the advising and curriculum-planning strategies utilized for entering students, to the effect that delaying the onset of the Math courses until the second year affords students time to develop the studying and learning skills needed for the Math course, and thus can lead to higher fifth-term and seventh-term persistence. However, these results also highlight the difficulty in attempting to predict a student's long-term outcome based on the performance in a single course taken in the first two years, since Table III indicates that students get higher Math grades in the first year. These results suggest that performance in a single course is too narrow a view of the student's academic curriculum for predicting persistence or graduation.

Accordingly, the next section expands the scope of the analysis of student academic data to the curriculum level, examining the extent of progress through the foundation course sequences rather than focusing on a single course.

TABLE IV

PERSISTENCE AND GRADUATION OF STUDENTS IN RELATION TO THE TIMING OF MATH COURSES.

STEM Math Courses			
	Taken in first year	Taken in second year	p
Fifth-term persistence	79.6%	93.0%	< 0.001
Seventh-term persistence	77.5%	88.7%	0.003
Graduation	43.7%	52.0%	0.113

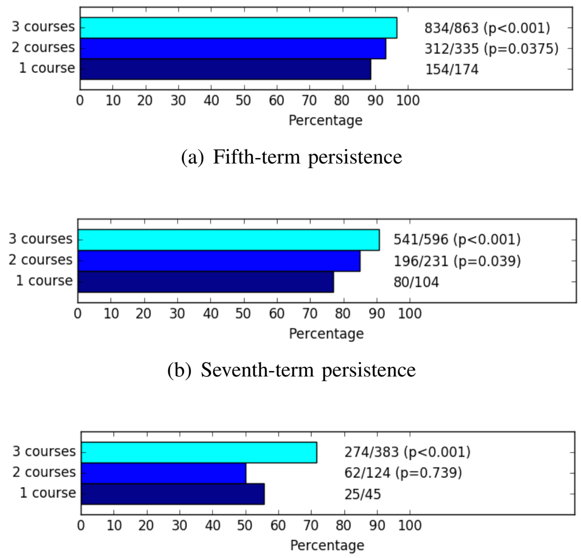
Non-STEM Math Courses			
	Taken in first year	Taken in second year	p
Fifth-term persistence	81.5%	89.6%	< 0.001
Seventh-term persistence	71.7%	81.5%	< 0.001
Graduation	54.9%	58.5%	0.299

B. Foundation Course Sequences

Each Metro student is required to take a 3-course foundation sequence during the first two years (Figure 2). We determined each student's progress through this sequence and aggregated the results by this "core-progress" value across all students.

The results for the foundation course sequence analysis are shown in Figure 5. The analysis revealed statistically significant evidence that fifth-term and seventh-term persistence percentages are higher for students who completed all three of the foundation courses, compared to those who complete only two. Likewise, students who complete two of the courses have higher fifth-term and seventh-term persistence than those who complete only one. For graduation, students who complete the entire course sequence had the highest graduation rate, while students who completed only one had a higher rate than those who completed two. The latter is an interesting finding, though it is worth noting that the sample size of the 1-course group for the graduation analysis (45 students total) is relatively small.

The foregoing results demonstrate that the extent of a student's progress through a prescribed course sequence offers potential for predicting a student's graduation likelihood. In the next subsection, we extend the examination to encompass the course sequences taken by the largest numbers of students. We thereby assess whether increased progress through the course sequences extracted via sequential pattern mining exhibits a similar effect on persistence and graduation rates.



(a) Fifth-term persistence

(b) Seventh-term persistence

(c) Graduation

Fig. 5. Fifth and seventh-term persistence and graduation rates by extent of foundation course sequence completion. Each bar is annotated with the raw data that yielded the percentage, and with the p-value for the hypothesis that the percentage represented by that bar is greater than the one below it.

C. Sequential Patterns

The number of distinct sequences that were mined from the student records from (1) the participants in the Metro program, and (2) the participants of the program along with the matched student group, are shown in Tables V and VI respectively. As expected, the number of distinct sequences increases as the required minimum support was decreased. Likewise, increasing the minimum sequential pattern length generally leads to a decrease in the number of distinct sequences extracted.

We aggregated the top ten most frequently-taken course sequences of length 3 as detailed in the Methods section. Figure 6 shows the fifth-term persistence, seventh-term persistence, and graduation percentages for each progress value for the Metro students and matched students collectively. Likewise, Figure 7 shows the seventh-term persistence and graduation percentages for each progress value, aggregated across the top five length-5 sequential patterns for the Metro and matched students.

Figure 6 reveals that students who make greater progress through their chosen sequences have statistically significantly higher persistence and graduation rates than those who make less progress through those same sequences. The same bar graphs (not shown due to space constraints) for each of the ten individual length-3 sequences revealed the same trend.

TABLE V
NUMBER OF DISTINCT SEQUENCES, METRO PROGRAM PARTICIPANTS.

Minimum sequence length	Number of sequences, $m_{sup} = 0.01$	Number of sequences, $m_{sup} = 0.02$	Number of sequences, $m_{sup} = 0.05$	Number of sequences, $m_{sup} = 0.10$
2	14949	4518	801	239
3	40549	7220	764	125
4	33162	2634	109	5
5	13240	238	2	0
6	3431	21	0	0

TABLE VI
NUMBER OF DISTINCT SEQUENCES, METRO PROGRAM PARTICIPANTS AND MATCHED STUDENT GROUP.

Minimum sequence length	Number of sequences, $m_{sup} = 0.01$	Number of sequences, $m_{sup} = 0.02$	Number of sequences, $m_{sup} = 0.05$	Number of sequences, $m_{sup} = 0.10$
2	4173	1383	243	51
3	4598	872	38	4
4	1604	38	0	0
5	138	0	0	0
6	0	0	0	0

The analysis for the length-5 sequential patterns, shown in Figure 7, reveals a similar trend as the length-3 patterns. For seventh-term persistence, students who make increased progress through the sequences (e.g., 5 courses versus 4, and so on) have higher persistence percentages than those who make less progress. For graduation, the analysis revealed that students who make the full 5-course progress through their sequences enjoyed higher graduation rates than students who took four, and likewise for four versus three.

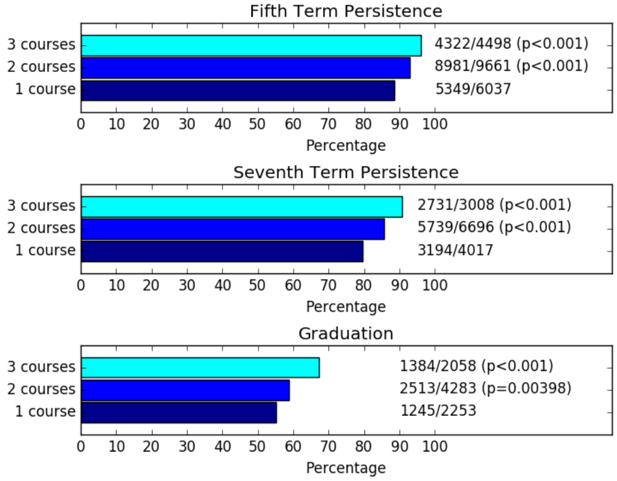


Fig. 6. Persistence and graduation rates by the extent of completing the top ten length-3 course sequences. Bars are annotated as in Fig. 5.

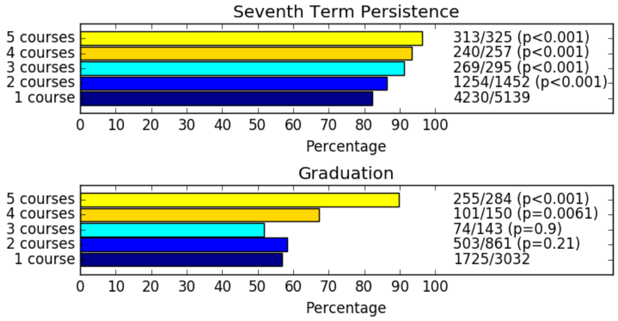


Fig. 7. Seventh-term persistence and graduation rates by extent of completing the top five length-5 sequences. Bars are annotated as in Fig. 5.

D. *Feature Importance and Recursive Feature Elimination*

We first assessed the accuracy of naive Bayes, logistic regression, and random forest models for predicting the graduation status of students using all available features (Table VII) for the pool consisting of the Metro participants and their matched counterparts. Boxplots for the accuracy scores are shown in Figure 8. The accuracy scores are within one standard deviation of each other, with no one model showing significantly better performance than the others (Table VIII). The random forest model's Gini importance values for the top six most important features (Figure 9) indicate that the seventh-term and fifth-term persistence were the most influential features, followed by the parents' education levels.

TABLE VII
ALL FEATURES VS. FEATURES RESULTING FROM RFE.

All Features	Features after RFE
Fifth-term persistence Seventh-term persistence Race Mother's education Father's education Gender Pell Eligibility ELM score EPT score EOP status	Seventh-term persistence Mother's education Father's education ELM score EPT score

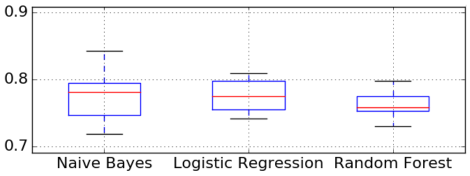


Fig. 8. Classification accuracy of the three models.

Recursive feature elimination (RFE) reduced the features used in the classification models to those listed in Table VII, right column. The accuracy scores of each model type with the reduced feature set is shown in Figure 10. The accuracy of the random forest model improved the most significantly

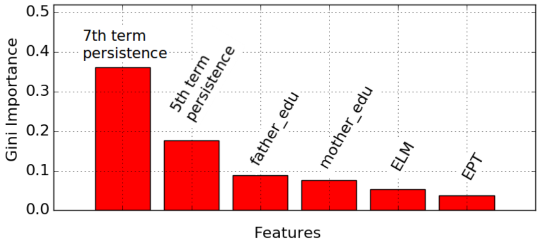


Fig. 9. Top six features from Random Forests using all available features.

after reduction of the feature set. Meanwhile, seventh-term persistence had the highest Gini importance of the features that remained after recursive feature elimination (Figure 11). Thus, for both the full and the RFE-reduced feature sets, it can be observed that seventh-term persistence outweighs other personal factors such as race, gender, or entrance scores, for predicting graduation.

TABLE VIII
CLASSIFICATION ACCURACY BEFORE AND AFTER RFE

	All Features	After RFE
Naive Bayes	77.5% ± 3.4%	74.8% ± 3.2%
Logistic Regression	77.6% ± 2.5%	74.8% ± 3.3%
Random Forest	76.7% ± 2.0%	78.7% ± 1.6%

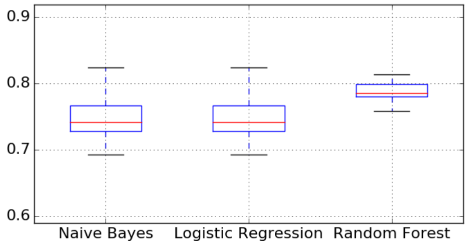


Fig. 10. Accuracy of different classifiers after recursive feature elimination.

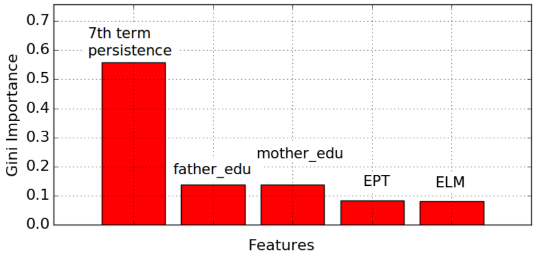


Fig. 11. Top ranked features after recursive feature elimination.

While the importance of seventh-term persistence in predicting graduation may seem unsurprising, we found that students with positive seventh-term persistence are still at substantial risk of not graduating. The random forest classifier's accuracy was 78.7% with seventh-term persistence as its most important feature (Table VIII). We also found that out of the students with positive seventh-term persistence from the 2009-2013 cohorts, 76.5% graduated as of the time this study was conducted. This attrition highlights the need for additional research as to why students who have persisted for seven semesters are unable to complete their degrees.

V. CONCLUSIONS AND FUTURE DIRECTIONS

This work has explored how certain curriculum-level factors, including progress through a student's sequence of courses as well as the timing of the first Mathematics course, affect the student's likelihood of persistence and graduation. This work has also assessed the utility, challenges, and potential future directions of sequential pattern mining in the analysis of student course sequences.

We found that students who take their first Math course in the second year, as opposed to the first year, have statistically significantly higher fifth-term and seventh-term persistence, but not a higher graduation rate. Future research should explore Math course timing for students in specific disciplines. We also found that increasing progress through the sequence of courses that students choose to take consistently leads to higher graduation rates. Future work can leverage sequential pattern mining in many ways to expand on these findings. For example, restricting the pattern mining to the course sequences of a particular major, wherein it is expected that most students will take the sequence of courses as dictated by each course's prerequisites and the department's requirements for that major, will likely circumvent the minimum support challenges revealed in this work. Sequential pattern mining can also help to improve student advising and curriculum design for the major, by elucidating when and how much students deviate from the expected course sequence and which related courses (i.e., courses that are useful but not required for the major) students tend to take. Likewise, any benefits or drawbacks from those deviations or related courses can be exposed. Moreover, exploring whether student performance in a Math course is affected by certain courses taken immediately before or concurrently with it could lead to curriculum improvement.

Finally, this work has shown that sequential pattern mining has the capacity to reveal new curricular features that can inform when intervention at the advising level is necessary. For example, as we show in this work, an interruption in a student's progress through a course sequence leads to a decrease in that student's graduation likelihood. Thus, prevention of such an interruption can be viewed as a priority for academic counselors (and, indeed, the Metro program in this work provides services geared to keep students progressing through their chosen curriculum). Future EDM research with sequential pattern mining may find additional features that can impact the quality of advising as well as curriculum design.

REFERENCES

[1] National Center for Education Statistics, "Graduation rate from first institution attended for first-time, full-time, bachelor's degree-seeking students at 4-year postsecondary institutions, by

race/ethnicity, time to completion, sex, control of institution, and acceptance rate: Selected cohort entry years, 1996 through 2010". https://nces.ed.gov/programs/digest/d17/tables/dt17_326.10.asp.

[2] National Center for Education Statistics, "Retention first-time degree-seeking undergraduates at degree-granting postsecondary institutions, by attendance status, level and control of institution, and percentage of applications accepted: Selected cohort entry years, 1996 through 2010". https://nces.ed.gov/programs/digest/d17/tables/dt17_326.30.asp.

[3] R. Agrawal, R. Srikant, "Mining Sequential Patterns," Eleventh Intl. Conf. on Data Eng'g., pp. 3-14, 1995.

[4] R. Asif et al., "Analyzing undergraduate students' performance using educational data mining," Comp. & Edu., v.113, pp.177-194, 2017.

[5] L. Breiman, "Random forests," Machine Learning, v.45-1, pp.5-32, 2001.

[6] S. Bhaskaran et al., "A data mining approach for investigating students' completion rates," Higher Education in the Twenty-First Century II. Taylor & Francis Group, 2016, pp. 105-116.

[7] R. Campagni et al., "Data mining models for student careers," Expert Systems with Applications, vol. 42, pp. 5508-5521, 2015.

[8] Y. Cao et al., "Orderness Predicts Academic Performance: Behavioral Analysis on Campus Lifestyle," arXiv preprint arXiv:1704.04103, 2017.

[9] W. Chen et al., "Principles for Assessing Adaptive Online Courses," 2017. http://www.andrew.cmu.edu/user/cjoewong/VelocityChess_EDM.pdf

[10] P. Fournier-Viger et al., "The SPMF Open-Source Data Mining Library Version 2," PKDD 2016 Part III, Springer LNCS 9853, pp. 36-40.

[11] P. Fournier-Viger, "Mining Frequent Sequential Patterns Using the PrefixSpan Algorithm". http://www.philippe-fournier-viger.com/spmf/PrefixSpan.php

[12] R. Genuer et al., "Variable selection using random forests,", Pattern Recognition Letters, vol. 31, no. 14, pp. 2225-2236, 2010.

[13] A. Gopalakrishnan et al., "A Multifaceted Data Mining Approach to Understanding what Factors Lead Students to Persist and Graduate," Comp. Conf. 2017, pp. 372-381.

[14] G. James et al., An Introduction to Statistical Learning, Springer, 2013.

[15] Z. Jiang et al., "A Novel Cascade Model for Learning Latent Similarity from Heterogeneous Sequential Data of MOOC," Proceedings of the 2017 Conf. on Empirical Methods in NLP, pp. 2768-2773, 2017.

[16] Z. Kolenovic et al.,"Improving Student Outcomes via Comprehensive Supports: Three-Year Outcomes From CUNY's Accelerated Study in Associate Programs," Comm. Coll. Rev., v.41-4, pp. 271-291, 2013.

[17] G. Louppe, "Understanding Random Forests: From Theory to Practice,", arVix:1407.7502, 2014.

[18] R. Martinez et al., "Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop," Proceedings of the 4th Intl Conf on Edu. Data Mining, pp. 111-120, 2011.

[19] D.S. Maylawati et al., "Comparison between BIDE, PrefixSpan, and TRuleGrowth for Mining of Indonesion Text," Journal of Physics: Conference Series, vol. 801, no. 012067, 2017.

[20] Metro Academies College Success Program, "About the Metro College Success Program" https://metro.sfsu.edu/about-metro-academies.

[21] M. Munk et al., "Quantitative and Qualitative Evaluation of Sequence Patterns Found by Application of Different Educational Data Preprocessing Techniques," IEEE Access, vol. 5, pp. 8989-9004, 2017.

[22] W. McKinney, "Data Structures for Statistical Computing in Python," Proceedings of the 9th Python in Science Conference, pp. 51-56, 2010.

[23] J. Pei et al., "Mining Sequential Patterns by Pattern Growth: The PrefixSpan Approach," IEEE TKDE, v.16-10, pp. 1424-1440, 2004.

[24] D. Perera et al., "Clustering and Sequential Pattern Mining of Online Collaborative Learning Data," IEEE TKDE, v. 21-6, pp. 759-772, 2009.

[25] I. Rish, "An empirical study of the naive Bayes classifier," IJCAI 2001 Workshop on Empirical Methods in AI, vol. 3, no. 22, pp. 41-46, 2011.

[26] C. Romero, S. Ventura, "Educational Data Mining: A survey from 1995 to 2005," Expert Sys. with Appl., vol. 33, pp. 135-146, 2007.

[27] C. Romero, S. Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE Trans. on Sys., Man, and Cybernetics Part C, Applications and Reviews, vol. 40, no. 6, pp. 601-618, 2010.

[28] A. Abu Saa, "Education Data Mining & Students' Performance Prediction," Intl J. of Adv. Comp. Sci. and Apps, v.7-5, pp. 212-220, 2016.

[29] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol 12, pp. 2825-2830, 2011.

[30] A.M. Shahiri et al., "A Review on Predicting Students' Performance using Data Mining Techniques," Procedia Computer Science, vol. 72, pp. 414-422, 2015.

[31] C. Sommo, A. Ratledge, "Bringing SUNY Accelerated Study in Associate Programs (ASAP) to Ohio: Early Findings from a Demonstration in Three Community Colleges," 2016. https://files.eric.ed.gov/fulltext/ED569162.pdf.

[32] E. Jones et al., SciPy: Open Source Scientific Tools for Python. http://www.scipy.org

[33] V. Tinto, "Classroom as Communities: Exploring the Educational Character of Student Persistence," J. Higher. Ed., v.68-6, pp. 599-623, 1997.

[34] C.M. Zhao, G.D. Kuh, "Adding Value: Learning Communities and Student Engagement," Res. in Higher Ed., v.45-2, pp. 115-138, 2004.