Mining University Registrar Records to Predict First-Year Undergraduate Attrition

Lovenoor Aulck
University of Washington
laulck@uw.edu

Dev Nambi
F.H. Cancer Research Center
dnambi@fredhutch.org

Nishant Velagapudi
University of Cal-Berkeley
nishray@berkeley.edu

Joshua Blumenstock
University of Cal-Berkeley
jblumenstock@berkeley.edu

Jevin West
University of Washington
jevinw@uw.edu

ABSTRACT

Each year, roughly 30% of first-year students at US baccalaureate institutions do not return for their second year and billions of dollars are spent educating these students. Yet, little quantitative research has analyzed the causes and possible remedies for student attrition. What's more, most of the previous attempts to model attrition at traditional campuses using machine learning have focused on small, homogeneous groups of students. In this work, we model student attrition using a dataset that is composed almost exclusively of information routinely collected for record-keeping at a large, public US university. By examining the entirety of the university's student body and not a subset thereof, we use one of the largest known datasets for examining attrition at a public US university (N = 66,060). Our results show that students' second year re-enrollment and eventual graduation can be accurately predicted based on a single year of data (AUROCs = 0.887 and 0.811, respectively). We find that demographic data (such as race, gender, etc.) and pre-admission data (such as high school academics, entrance exam scores, etc.) - upon which most admissions processes are predicated - are not nearly as useful as early college performance/transcript data for these predictions. These results highlight the potential for data mining to impact student retention and success at traditional campuses.

1. INTRODUCTION

Student attrition has long been a topic of great interest in higher education research, with government reports on attrition dating back over 100 years [31]. This interest stems from the fact that students who do not graduate are a lost investment on many fronts. For higher education institutions, limiting attrition is central to their financial sustainability as they devote scarce resources towards classes and services for non-completing students [17]. In particular, it is estimated that 30% of United States (US) first-year students do not return for their second year of post-secondary education with US taxpayers spending nearly $2 billion annually on educating non-returning first-year students alone [28]. Institutions are also concerned with attrition rates because they are central to estimates of institutional effectiveness, thereby affecting funding opportunities and government support [14]. Highlighting the impact of attrition at the institutional level also says nothing of its impact on students, who devote time, effort, and finances towards unfinished educational pursuits. Leaving college drastically alters career trajectories for students and those without college degrees face continually declining job growth and worsening job prospects [9].

In light of this, understanding motivations for students to drop out and possible remedies thereof is of great importance [12]. Empirical evidence to build student attrition theory has traditionally focused on survey-based research [30, 8]. However, survey instruments are often costly to implement, time-consuming for data collection, and produce results that are not always generalizable across institutions due to vastly different student profiles [34, 7, 8]. Institutional data that is routinely collected at colleges and universities (e.g. student application and transcript data) can provide an alternative data source and a way to supplement survey-based measures [8]. Leveraging data sources already in existence can add a means to more efficiently examine the student attrition problem and help institutions remedy the issue of attrition. One field that is primed to take advantage of this institutional data is educational data mining (EDM) and its focus on data-intensive techniques in educational settings [26, 4].

EDM is an emerging field with much of its research on attrition centered on massive online open courses (MOOCs) and other online environments (e.g. [35, 13]). Studying attrition in MOOCs and other online settings lends itself to expansive data collection opportunities and a detailed monitoring of students [23]. This limits the extent to which this work can be generalized to more traditional campus settings (i.e. campuses where learning is primarily on-campus, in-classroom). Meanwhile, EDM-centric work on predicting attrition at traditional campuses has been scarce and usually limited to small, homogeneous subsets of students rather than the entirety of a college student population. Additionally, the focus when predicting attrition is usually on how well it can be predicted and less so on what type of data is best for these predictions.

In this work, we predict the attrition of a large number of un-

dergraduate students (N = 66,060) using only their first year of academic data. The students we examine are not from a single department or major within a university. Rather, they span the entirety of a student body, thereby comprising a dataset with heterogeneous aspirations, backgrounds, and goals. In addition, we rely almost entirely on data that is routinely collected at institutions of higher education. With this data, we seek to answer two questions: to what extent can undergraduate student attrition be predicted using a limited amount of data from registrar records and what types of data from registrar records are most useful in predicting attrition. The first of these has been explored in the past while using smaller and/or homogeneous student populations; the second has not been systematically examined in the literature to our knowledge.

To answer the above questions, we mine the institutional data records at a large, public university in the US and engineer features for predictions. We then create numerous machine learning models using the engineered features and compare the performance of these models to each other. Then, we create separate machine learning models using only groups of features and not the entirety of the feature space to compare the predictive power of different subsets of institutional data. This work is an extension of our previous work on modeling student attrition using a limited amount of data [3] but where we previously focused on using the *first term's* data in generating features for prediction, we use the *first year's* in this work. We also extend our previous work to build additional machine learning models, predict attrition as defined according to two different definitions (overall graduation and re-enrollment after students' first year), and examine the types of feature subsets most useful in predictions. In so doing, we present two key findings, both of which have many implications for administrative policy in higher education:

- We demonstrate that the graduation and second-year re-enrollment of students can be predicted using data that is routinely gathered at institutions of higher education.
- We show that demographic and pre-entry features have less predictive power than data on student academics.

2. RELATED WORK

There are many examples of predicting attrition at traditional campuses. Most of these focus on small, homogeneous subsets of students. Moseley predicted the graduation of 528 nursing students using rule induction methods, obtaining high accuracies but not controlling for the number of terms/semesters examined for each student [21]. Dekker et al looked at only the first semester grades of 648 students in the Electrical Engineering department at the Eindhoven University of Technology and were able to predict dropout with 75-80% accuracy [10]. Kovačić used tree-based methods on a similarly-sized dataset of 453 students at the Open Polytechnic of New Zealand, finding ethnicity and students' course taking patterns to be highly useful in prediction [18]. Bayer et al. looked at 775 applied informatics students at the Czech Republic's Masaryk University across three years [5]. Without limiting the amount of information available for each student, they found that including features related to students' social behavior can boost prediction accuracy by over 10% for some models. These and similar studies, however, focus on relatively small (e.g. N < 2,000) subgroups of students with similar academic pursuits/foci. In addition, there is little consistency with respect to the timeframes across which data is examined for each student. Other approaches to predict attrition at traditional campuses include early alert systems, which are often labor intensive and poorly funded [29]. These alert systems have been shown to positively benefit students (e.g. [16]), but usually rely on data gathered in the midst of a course or an academic term (e.g. [27, 15]), which may not always be feasable.

The work we present more closely relates to a subset of literature looking at student attrition in the context of the heterogeneity of students across an entire campus and not just a subset thereof. Our work also deals with much larger student populations than those described above and, in this sense, it more closely resembles a more recent body of literature. Delen used 8 years of institutional data on over 25,000 students at a large, public US university, predicting whether the students would return for their second year [11]. However, due to class imbalances, Delen re-sampled the majority class and ultimately used only 6,454 students for predictions. Ram et al. used data on about 6,500 freshmen at a large, public US university to predict whether students would drop out after their first semester, and for those that did not, whether they will drop out after an additional term [25]. Ram et al. supplemented data from institutional databases with student smart card transactions to infer social integration. More recently, Nagy and Molontay predicted the dropout of 15,825 students from the Budapest University of Technology and Economics using only their information prior to college entry with some success [22].

There are a few ways in which our work contributes to this body of literature. Firstly, we use a much larger dataset than has been previously examined specifically for attrition (66,060 students). We examine the entirety of a large university's student body and we do not limit the extent of heterogeneity of the students in the dataset. Additionally, we also address the question of what types of features are most useful in predicting student attrition. In particular, previous works have generally used all available data sources concurrently in determining which students will attrite. In this work, we explore what types of routinely-collected institutional data fare best when predicting attrition by comparing performance using different data subsets in isolation. Finally, we concurrently compare predictions for two different definitions of "attrition," highlighting the degree to which operationalizing the term can impact results.

3. METHODS

We describe the methods for this work by first detailing the data used in the project. We then give relevant operational definitions with respect to how we define attrition. Thereafter, we discuss the data subsets used in the predictions and the features generated. Lastly, we describe the setup of the machine learning experiments.

3.1 Data Description

We collected psuedonomyized, de-identified data from the University of Washington (the University) data stewards in 2017. The University is a traditional campus setting where a vast majority of instruction is in person and face-to-face. No

personally identifiable information was collected for the students; instead, students were referenced using unique identifying keys. Table 1 shows the tables that were pulled from the registrar databases. In general, the data included information on students' demographics, complete transcript records at the University, and information from applications to the University. We did not have any information on students' financial aid status or economic status other than that which was derived from their ZIP code, as described below. Socioeconomic factors can play a large role in the student attrition process [6], however, we did not have access to student finances for use in this work. We also did not have access to any exit surveys from students who had either left the University or had graduated.

Table 1: Data pulled from registrar databases

Table	Description
Application Data	Information from student applications to the University including high school coursework
Guardian Data	Information on student guardians as pulled from student applications to the University
Demographic Data	Information on student demographics including date of birth, race, ethnicity, gender, etc.
Major Data	Information on majors declared by students on a term-by-term (quarter-by-quarter) basis
Test Score Data	Information on student standardized test results
Transcript Data	Information on student coursework and grades on a term-by-term (quarter-by-quarter) basis

We restricted data to high school graduates who first enrolled at the University as matriculated, baccalaureate-degree-seeking undergraduate students between 1998 and 2010 without previously attending another post-secondary institution full-time. These students are henceforth referred to as "freshmen." The dataset included students who were in a college in high school program but excluded those who attended junior/community college full-time after high school and then transferred to the University. Because the data was pulled in 2017, we used the year 2010 as a cutoff to allow for six full years of visibility on student academics at the University before labelling a student as a "non-completion," as defined in Section 3.2. In total, the dataset consisted of 66,060 unique freshmen entrants. We then further limited the data for each student to information through one calendar year from each student's first enrollment at the University. This data was limited to one calendar year for all students, regardless of the number of courses they took/passed, their grades, or their backgrounds.

After joining tables of interest using the unique student identifiers, we created features for the prediction experiments by either pulling them directly from the raw data or engineering them for each student. The features were grouped in 7 groupings, which are described in Section 3.3; a comprehensive list of features and descriptions thereof is available upon request but was not provided in this writing in the interest of space. In total, there were 1,405 features and all features were generated for each student without exception.

3.2 Definitions

Ambiguity with respect to operational definitions of dropout in literature on student attrition can make it difficult to compare results across studies [24, 33]. There are numerous ways in which attrition has been defined in existing literature, be it students dropping out from a particular course (e.g. [21]), re-enrolling after their first term (e.g. [1]), re-enrolling after their first year (e.g. [11]), graduating on time (e.g. [3]), or reaching some other relevant milestone (e.g. [10]). In this work, we defined attrition in two ways and analyze both. We examined attrition from students' first year to their second ("re-enrollment" and "non-re-enrollment") as well as looking at whether a student graduated on time ("graduate" and "non-completion"). We do not examine attrition on a term-by-term basis because of the relatively few students who leave the University after only a single term, as discussed in Section 4.1. We operationally defined non-completion and re-enrollment as described below.

3.2.1 Non-Completion

We defined "non-completion" as any freshman student who did not graduate with a baccalaureate degree from the University within 6 calendar years of first entry to the University. We defined a "graduate" as a freshman who graduated from the University with a baccalaureate degree within 6 calendar years of first enrollment. The University uses a quarter term system and we used the span of four consecutive academic quarters as a measure of one calendar year. Six calendar years for graduation was thus the span of 24 consecutive academic quarters. This definition of non-completion only accounted for students' first baccalaureate degree and did not take into account double-majors or double degrees. For example, if a student was simultaneously pursuing two baccalaureate degrees but only graduated with one in five years, they would be a graduate; alternatively, if the student had graduated with both degrees but during their seventh year, they would be considered a non-completion. Because we focused on registrar records from a single institution, defining non-completion in this manner does not take into account students' academic progression after leaving the University. This is because we only had access to registrar records from a single institution and did not track students across multiple institutions - they could have very well transferred from the University and graduated in good standing.

We accounted for students who took part in a college in high school program by converting their transferred credit total to a count of academic quarters completed while assuming typical full-time enrollment at the University. For example, if a student completed 30 credits in a college in high school program, we converted this credit total to a count of terms completed at the University (in this case, 2, as students typically take 15 credits per term). We rounded the result from this conversion where appropriate. We then deducted this number when determining whether the student had graduated within an appropriate amount of time.

3.2.2 Re-Enrollment

We defined "re-enrollment" as a student who completed at least one additional course within one calendar year of the end of their first calendar year at the University (i.e. within 4 academic quarters from the end of their first year). "Non-re-enrollments" were students who were not re-enrollments. In this work, the definitions of graduation and re-enrollment were treated mutually exclusive in that all graduates were not necessarily re-enrollments. It should be noted that the University requires students who do not enroll for two consecutive terms without an excused leave to be re-admitted at the discretion of the University.

3.3 Feature Groupings

For every student, we engineered the subsets of features that are described below. For all student grades, we calculated a grade percentile and a z-score by comparing each students' grades to the grades of all undergraduate students who had taken the same course at the same time. References to grades include the student's GPA (on a 4.0 scale), their percentile score (from 0-100), and their z-score for courses (representing the number of standard deviations from the mean, assuming a normal grade distribution). References to "performance" for the feature groupings include grades and credits earned, at the least. In some cases, references to performance may also include the number of graded credits earned (versus courses taken pass-fail) and the number of credits attempted. A brief description of each of the feature subsets is provided in Table 2.

Table 2: Data subsets used in predictions

Subset	Description
Base Data	Year and quarter of University entry (included with every other data subset)
Demographic Data	Non-academic data prior to entry to the University, including demographics
Department-level Data	Measures of performance aggregated by course department
First-Year Summary Data	Aggregated measures of academic performance during first year
Grouped Course Data	Measures of performance aggregated by course number and STEM gatekeepers
Major Data	Counts of majors declared on a term-by-term basis
Pre-Entry Data	Academic data prior to entry to the University.

3.3.1 Base Data

Base data consisted of only three features and was included in the feature space when making predictions using every other data subset described. The base data included students' calendar year of entry to the University, their quarter of entry to the university (i.e. which of the four academic quarters was a student's first; ranging from 1 to 4, with 1, 2, 3, and 4 corresponding to winter, spring, summer, and autumn academic quarters, respectively), and a quarter-year variable which consisted of students' year of entry multiplied by 4 and added to the quarter of entry to create a relative time scale. These features were included to account for any time-related variation in graduation rates.

3.3.2 Demographic Data

Demographic data consisted of student's non-academic information prior to entry to the University. This included, but was not limited to, students' gender, race, ethnicity, age at college enrollment, veteran status, and student athlete status. We also included information from students' application to the University, such as information on the students' high schools (excluding high school grades), parents' educational attainment, and students' ZIP (postal) code, which was either pulled from their high school information or, when unavailable, from their university application. We joined students' ZIP codes with 2015 US census data[1] to find the average income and educational attainment in each ZIP code. We also included the distance from the University to each student's home ZIP code. Features derived from ZIP codes were the only features from sources external to the University's registrar databases.

3.3.3 Department-level Data

Department-level data consisted of student performance in course offerings grouped by course prefix. For example, this included performance in all BIOL (biology) courses grouped together, performance in all HIST (history) courses grouped together, etc. We excluded course prefixes wherein at least 10 students from the dataset did not take a course. In all, this included 200 unique course prefixes and 1000 features, with GPA, percentile grade, z-score, credits earned, and graded credits earned calculated for each prefix. We used department-level data instead of individual course data after preliminary modeling using individual courses did not yield strong results. The expansive feature space when engineering features across individual courses also significantly increased the requisite computational power/time for modeling and we decided against pursuing this further.

3.3.4 First-Year Summary Data

First-year summary data consisted of aggregate measures of students' first year at the University. This included, among other things, students' course performance, credits taken, number of courses failed, number of quarters enrolled, and enrollment in a freshman seminar courses. The first-year summary data also included aggregate measures of students' performance in their first, second, third, and fourth quarters as well as student performance in the last academic quarter for which they were enrolled during their first year (regardless of which quarter it was). We also included differences between students' performance in successive quarters.

3.3.5 Grouped Course Data

Grouped course data consisted of student course performance grouped either by course number or by performance in "STEM gatekeepers." To group courses by course number, we aggregated performance across all courses that were numbered below 100, from 100-199, from 200-299, from 300-399, and 400+. The course numbering generally reflected whether the course was designed to be taken by lowerclassmen or upperclassmen and, in some cases, also indicated

[1]From the US Census Bureau's American Fact Finder

during which year students typically took the course. STEM gatekeepers refer to introductory science, technology, engineering, and math (STEM) courses which often function as pre-requisites for STEM majors and degrees. These gatekeeper courses tend to be highly competitive and performance in these courses is a key determinant of whether a student will be accepted into any of the highly competitive STEM majors. We grouped the performance in STEM gatekeepers by course department and topic (e.g. the calculus series, the general chemistry series, the organic chemistry series, etc.) as well as across all STEM gatekeepers.

3.3.6 Major Data

Major data consisted of counts of students' major declarations during their first academic year. In most cases, students entered the University with a "pre-major" designation before declaring their major(s) of interest some time during their first or second year. These pre-major designations varied based on field of interest (e.g. pre-engineering, pre-nursing, pre-health, etc.). Students' majors were recorded on a per-quarter basis by the University (once per quarterly transcript record) and we tallied the counts of major declarations for each student across the entirety of their first year. For example, a student who declared a math major in their first two quarters only to switch to geography in their third quarter and then add a history double major in their fourth quarter would have the values 2, 2, 1 in the math major, geography major, and history major features, respectively.

3.3.7 Pre-Entry Data

Pre-entry data consisted of students' academic information prior to attending the University. This included, among other things, students' entrance exam scores, high school GPA, high school coursework, and college in high school program participation and performance. We did not include any information on students after their enrollment at the University in the pre-entry data.

3.4 Machine Learning and Predictions

We randomly divided the students into training and test sets using a 80-20 split (N in training = 52,848; N in test = 13,212). We used the same test set when evaluating the predictive performance of each of the models to allow for direct comparisons to be made. The data was highly skewed with graduates and re-enrollments comprising 78.5% and 93.1% of all the data, respectively. Graduates and re-enrollments comprised 78.0% and 92.9% of the test data, respectively. Though dealing with class imbalances is of great interest when examining freshmen attrition [32], we did not use any balancing techniques as we wanted to work with the data in its original, unaltered form. We scaled the training data by subtracting the median of each feature and dividing by the respective feature's interquartile range. We subsequently scaled the test data using the scaling values for each feature from the training data.

We used five different machine learning models to predict each student's graduation and re-enrollment: regularized logistic regression (LR), K-Nearest Neighbors (KNN), random forests (RF), support vector machines (SVM), and gradient boosted trees (XGB). We trained each model across the entirety of the training data and used the same training instances to train each of the models. We trained each model separately to predict graduation and re-enrollment. We tuned model hyperparameters for each model using 5-fold cross validation on the training data, after which the models were re-trained on the entirety of the training data using the tuned hyperparameters. We report final error metrics and performance on the test set, which was consistent across all models, regardless of whether predicting graduation or re-enrollment.

After developing predictive models using all features, we created regularized logistic regression models using each of the 6 feature subsets highlighted in Section 3.3 in isolation. The base data (see Table 2) was included in the feature space for each data subset. The rationale behind using regularized logistic regression for these models is further discussed in Section 4.3. We understand that an alternative approach would be to test all the models listed above for each of the data subsets to find the best performing model/subset combinations. That said, we believe our approach was still suitable for comparing different data subsets. When modeling using data subsets, we used the same observations as before to train each of the models and, as before, we developed a separate model for predicting graduation and re-enrollment for each of the data subsets. As such, the *training instances* were the same across models but the *training features* differed depending on the feature subset used. We tuned the regularization strength for these regularized logistic regression models using 5-fold cross validation on the training dataset and we report results on the test set.

4. RESULTS AND DISCUSSION

4.1 Student Characteristics

We show the number and proportion of graduates and re-enrollments in Figure 1. In all, 78.5% of students were labelled graduates while 93.1% of students were labelled re-enrollments. These proportions were verified with the University's office of institutional analysis. Such highly skewed data towards graduates and re-enrollments can be expected in a large, tier-1 research university setting where there has been considerable, long-standing effort to improve the overall attrition rate over time. That said, it must also be noted that at an institution with such a large student population, even small fractions of the student body represent hundreds of students on an annual basis. Across the timeline of the dataset (13 cohorts), 14,196 non-completions and 4,593 non-re-enrollments represent 1,092 and 351 students on an annual basis, respectively.

We show the cumulative percentage of students who either graduated or left the University across time in Figure 2. We used the first year as a cutoff for the data because, historically, a large number of students decide whether they will continue with their higher education pursuits during and immediately after their first year [28]. As such, developing models that can predict whether students will re-enroll for a second year and whether they are on a trajectory towards successful graduation could help administrators and academic advisors more effectively develop and deliver interventions directed towards students in need of assistance. When examining the data, 27.5% of all non-completions leave the university prior to the start of their 2nd year, 51.9% of non-completions leave the University between their 2nd and 6th

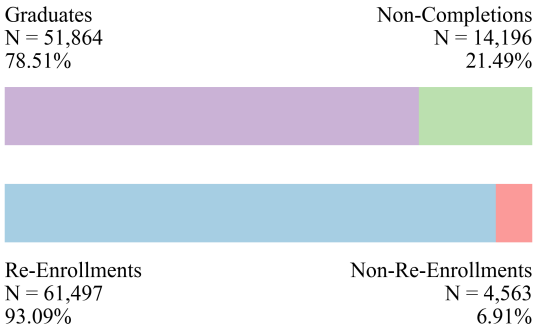

Figure 1: Counts and percentages of classes in the dataset. Definitions are provided in Section 3.2.

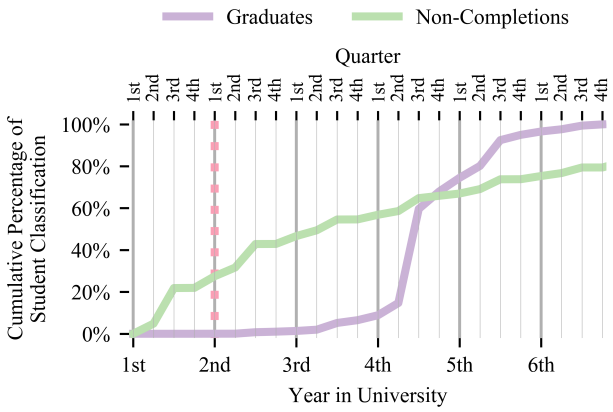

Figure 2: Cumulative graduation and non-completion curves of students. Years and quarters are relative to the time of first enrollment. The dotted line indicates the point to which data is limited for each student. Only students' first six years are shown, per the definition of "graduate."

year, and 20.6% continued to be enrolled at the University after their 6th year. The difference in number between non-completions who did not return for their 2nd year and non-re-enrollments can be attributed to non-re-enrollments who later returned to the University and graduated on time. Less than 5% of non-completions and less than 15% of non-re-enrollments left the University after only one term, leading us to not examine attrition after the first and second terms. In settings where attrition rates are higher after students' first and second terms, it may be more relevant to examine the performance of classifiers after one or two terms.

Figure 2 also shows that a majority of graduates (65.6%) completed their degrees during their fourth year at the University. The mean and median completion time for all graduates was 16.6 and 15.0 calendar quarters, respectively, from first enrollment. This is particularly apparent due to the near-sigmoidal shape of the cumulative graph for graduates, with a sharp rise during students' fourth year. We also see that there is a relative lack of students who graduated prior to the start of their third year. This highlights the difficulty in predicting graduation based on students' first year - a student typically does not graduate until several years later, during which a host of influences can shape an academic trajectory, be they personal, financial, or academic.

4.2 Predictions Using Different Algorithms

Table 3: Prediction results using all data features. Baseline values are based on test set.

	Graduation		Re-Enrollment	
Model	Accuracy	AUROC	Accuracy	AUROC
Baseline	78.0%	0.500	92.9%	0.500
LR	83.2%	0.811	95.0%	0.882
RF	83.1%	0.806	95.3%	0.887
XGB	83.0%	0.806	95.1%	0.885
KNN	82.5%	0.798	94.8%	0.876
SVM	78.0%	0.780	92.9%	0.862

We show the performance of each of the models using the entirety of the feature space in Table 3. The baseline measure in the Table refers to the majority class compositions in the test set. Generally speaking, most of the models had a similar comparative performance for each prediction task (i.e. predicting either graduation or re-enrollment). This hints at an effective ceiling with respect to predictive power from the types of features being used (i.e. ones pulled from registrar records) and that additional representations of the student experience (be they academic or social) should be incorporated. Alternatively, a more complex predictive model (e.g. deep neural networks) may also fare better in making these predictions. That said, given the data used, the models are able to predict the eventual graduation and re-enrollment of students fairly successfully, as evidenced by the relative improvements over baseline values for both prediction tasks.

For predicting graduation, logistic regression was the best-performing model, followed by random forests. When predicting re-enrollment, random forests performed the best, followed by gradient boosted trees and logistic regression. These results are generally in line with our previous work on similar tasks, where we found that logistic regression tends to work well compared to other models for predicting graduation and STEM attrition [2]. When examining the worst-performing models, the SVM model made predictions that consisted entirely of the majority class when predicting both graduation and re-enrollment, as seen by the models' accuracy being the same as the baseline values. Such results are typical of classifiers without much predictive strength on a dataset consisting of highly disproportionate classes. In this specific case, it may be remedied by using alternate kernels for the model, which we did not explore in this work.

We show the ROC curves for the models in Figure 3. These curves further illustrate the lack of differentiation with respect to model performance. For the same prediction task, the resulting ROC curves across the models were nearly identical with little difference in curvature. The more notable difference was when comparing the ROC curves for predicting graduation with those for predicting re-enrollment, as the curves for predicting re-enrollment were more prominently convex compared to those for predicting graduation. These curvatures, along with the metrics shown

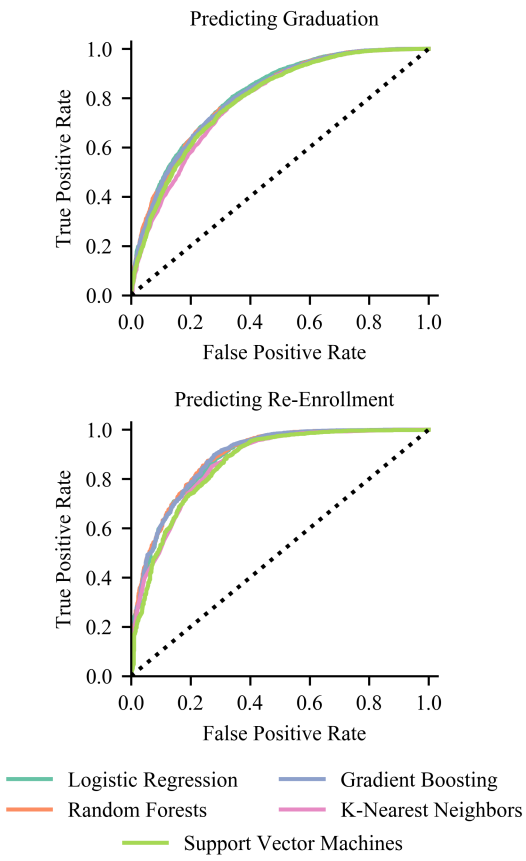

Figure 3: Receiver operating characteristic curves when using different machine learning models.

in Table 3, demonstrate that predicting students' eventual graduation is a more difficult task than predicting students' re-enrollment. We expected this as the cutoff for the data used in the predictions (i.e. students' first year) was near the point at which a student is classified as a re-enrollment (after their second year) but was much earlier than when a student was classified as a non-completion (after their sixth year). This helps highlight the degree to which differing operational definitions of attrition can vastly alter the perceived predictive strength of these classifiers. For other scenarios, alternate definitions of attrition may be more appropriate and the effectiveness of efforts to build predictive models will be colored by these definitions and institutional contexts.

We show the confusion matrices for the best models for predicting graduation and re-enrollment (logistic regression and random forests, respectively) in Figure 4. These matrices show a lower rate of false negatives for the models but a higher rate of false positives (i.e. students incorrectly classified by the models as having graduated or re-enrolled). To better understand this higher rate of false positives, we examined the complete transcript records of students who were classified accordingly. Across the false positives, we found numerous instances of non-completions and non-re-enrollments who had left the University with relatively strong grades in comparison to their graduating and

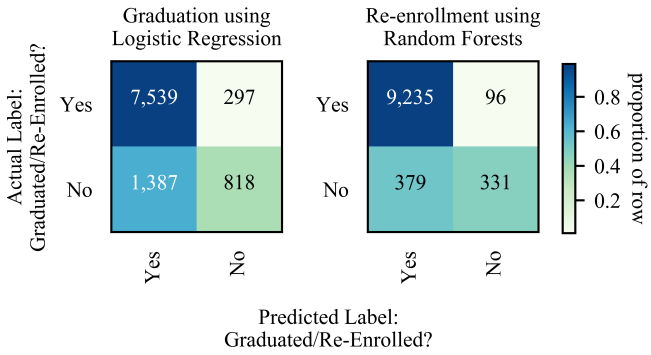

Figure 4: Confusion matrices when examining the top performing algorithms for predicting graduation (LR, left) and re-enrollment (RF, right).

re-enrolling peers. These students also often appeared to be pursuing very competitive majors and/or appeared to have rigorous post-graduation plans (e.g. pre-medical and pre-dental students). Many of these students remained in a pre-major state prior to their departure, indicating that though they had relatively strong grades, they likely were not able to enter into their degree program(s) of choice for various reasons and had to leave the University to pursue these ambitions as a result. Unfortunately, the University does not have a centralized major application database for admissions and rejections to specific majors. Having so could shed light on much of the motivation behind these students' desire to leave the University and if it was, in fact, motivated by not getting into competitive majors. That said, the fact that many of these students were academically similar to their graduating and re-enrolling counterparts further illustrates why there appears to be an effective ceiling with respect to predictive power using the given data, as seen in Table 3.

From a practical perspective, it should be noted that the classification thresholds for these models were not tuned with respect to either sensitivity or specificity. In practice, when developing institutional systems to identify students at-risk of leaving, it may be useful to raise the classification threshold when predicting whether a student will graduate or re-enroll, thus favoring lower recall at the expense of higher precision. This would effectively reduce the number of students who are predicted to graduate but in actuality do not (i.e. false positives) at the expense of more false negatives, which could be more acceptable when developing an alert system for students at risk of dropping out.

4.3 Predictions Using Different Data Subsets

After examining the results from predicting graduates and re-enrollments using all features, we used regularized logistic regression to predict graduation and re-enrollment using subsets of the data. We used logistic regression after we saw that it performed very well relative to other models for both prediction tasks (see Section 4.2) and because it had relatively fast training times due to having fewer hyperparameters to tune. This allowed us to more efficiently train the 12 different models that were needed when examining the performance of specific data subsets (i.e. separately modeling graduation and re-enrollment while using 6 different

Table 4: Prediction results using specific data subsets. Baseline values are based on test set.

Subset	Graduation Accuracy	Graduation AUROC	Re-Enrollment Accuracy	Re-Enrollment AUROC
Baseline	78.0%	0.500	92.9%	0.500
All	83.2%	0.811	95.0%	0.882
FY-Sum.	83.0%	0.795	94.9%	0.855
Department	82.3%	0.788	94.6%	0.847
Grouped	82.5%	0.781	94.6%	0.845
Major	79.9%	0.661	94.2%	0.768
Demo	78.0%	0.634	92.9%	0.643
Pre-Entry	77.3%	0.630	92.9%	0.616

data subsets in isolation for each).

We show the results when using data subsets in Table 4 alongside the performance of the logistic regression classifier from Section 4.2. Transcript-based features tended to perform better than information on students' prior to their enrollment at the University. More specifically, demographic data and pre-entry information did relatively poorly in predicting both graduation and re-enrollment. Intuitively, this is not a surprise as the admissions process at highly-competitive universities tends to be fairly selective with an emphasis on supporting and sustaining a successful yet diverse student body. Additionally, such institutions may already have efforts in place to reduce demographic disparities for student success. Meanwhile, when looking at transcript-based data subsets, first-year summary data performed the best with performance that was similar to using the entirety of the data. This is particularly noteworthy as the first-year summary data contained fewer features than the other transcript-based data subsets but was centered on summaries of performance across time rather than aggregations across course departments/numberings.

These findings are particularly interesting in light of work by other researchers. For instance, Nagy and Molontay found that attrition could be accurately predicted using what we outline as demographic and pre-entry features alone [22]. However, we do not see similar success here. We believe this could be due to vastly different educational settings and student profiles (e.g. here, most students tend to graduate/re-enroll while Nagy's student population primarily dropped out). In earlier work, Dekker et al. found that transcript-based features tend to have more predictive strength than pre-entry features, but examined this across rather limited data subsets [10]. Our results echo this finding. Recently, Manrique et al. found that attrition could be predicted using student performance in a few key courses [20]. Here, we find that aggregates across the first year tend to work better than more fine-grain representations of course-taking (e.g. grouping classes by course prefix and numbering). As discussed in Section 3.3.3, we decided against using individual course representations in this work.

We show the ROC curves for the regularized logistic regression models using each of the data subsets as well as the entire feature space in Figure 5. The fact that demographic and pre-entry data gave generally worse performance than

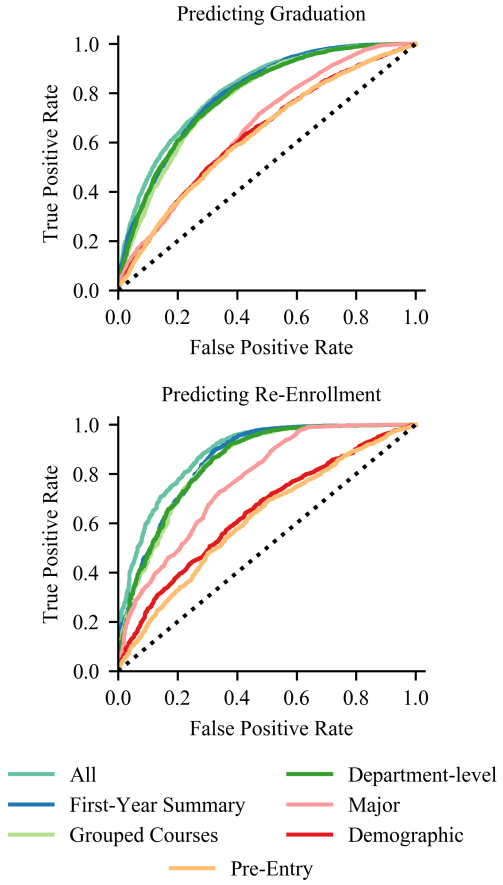

Figure 5: Receiver operating characteristic curves when using different subsets of data.

transcript-based features is very much apparent from the ROC curves. Data on majors, meanwhile, tended to perform worse than other transcript-based features but better than demographic and pre-entry data. The fact that using data on majors did not yield particularly strong results likely relates to the fact that most students in the dataset were in a pre-major state across their first year and formally declared their major of interest later in their undergraduate careers. As noted above, a centralized major application system was not available, else it could have been leveraged in addition to data on majors to draw a more clear picture of student academic interest. The other transcript-based datasets, meanwhile, had very similar curvatures for the ROC curves when predicting both graduation and re-enrollment.

We show confusion matrices from using the best-performing data subset in Figure 6. The best-performing data subset for both prediction tasks was first-year summary data. By comparing these confusion matrices to those shown in Figure 4, it can be seen that using just a limited subset of features tends to classify the data similarly to models built on the entirety of the data. This is true not only in terms of how effective the models are in making predictions, but also with respect to the relatively high rate of false positives seen across all four matrices.

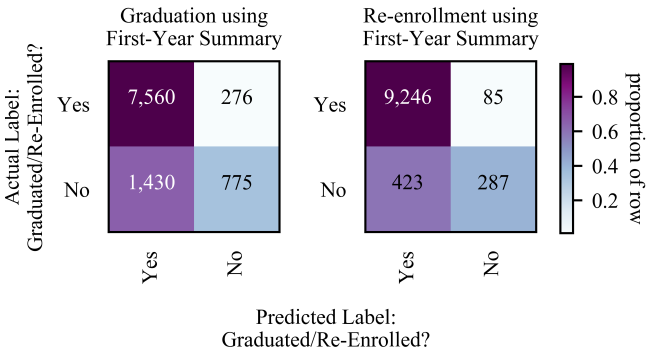

Figure 6: Confusion matrices when examining the top performing data subset for predicting graduation (left) and re-enrollment (right). The top performing data subset was the same for both tasks (first-year summary data).

5. FUTURE DIRECTIONS

We believe the findings regarding the data subsets have wide-ranging policy implications, particularly for identifying students at risk of dropping out in large, public universities. In such settings, there may be longstanding effort to decrease demographic disparities with respect to attrition and, as a result, transcript records may be more viable as features in predictive models than pre-entry/demographic information. Furthermore, these settings may also be resource-constrained with respect to time available for staff to hand engineer features. In such settings, knowing which features would be most predictive of attrition without the need to hand-engineer features across the entirety of data available to institutions could save time and effort in building models. We have had conversations with administrators at the University for better interpreting our results and improving the processes for identifying students in need of assistance.

Another direction of interest is better understanding the features used in predicting attrition. This includes not only further examining key individual determinants of attrition, as we have done in previous work [3, 2], but also finding the best combination of features across the subsets. We would like to examine this "minimum viable feature space" in the context of data available in registrar databases as well as investigate the degree to which these features relate to established theory on student attrition [12].

6. CONCLUSIONS

In this work, we use data from the registrar databases of a large, public US university to predict both graduation and re-enrollment using information limited to students' first calendar year at the university. We do this using a dataset of students that spans the entirety of the university student body and is thus much larger than previous studies predicting student attrition (N = 66,060). In so doing, we demonstrate that both graduation and re-enrollment can be effectively predicted using features generated from data that is routinely collected at institutions of higher education. Additionally, we also examine the degree to which specific subsets of registrar data can be useful in predicting attrition, finding that transcript-based features tend to outperform features based on student histories prior to college. This implies that effective strategies for intervention can be outlined based on registrar records.

Predicting re-enrollment after students' first year was a much more tractable task than predicting graduation. This can be attributed to the fact that predicting graduation necessitates predicting academic success years into the future from the point to which data was limited whereas predicting re-enrollment is within a much shorter timeframe. Considering the unpredictable influences that cause students to leave college prior to graduating (e.g. financial limitations, personal hardships, etc.), a more reliable prediction task may be to examine whether a student will return on a term-by-term basis. This could be particularly useful to develop alert systems to identify students at risk of dropout. However, this was not explored in this work due to the relatively few students who left the University after a single term.

We found that there appears to be an upper limit for predictive power for our dataset. This demonstrates the limitations when relying solely on registrar data and shows the need for additional features on the student experience to improve predictive power. Some potential features of interest include measures of social integration on campus and of financial aid. Better understanding student aspirations beyond simply using declared majors could also be of interest, especially using alternate representations of student course-taking behavior, as shown recently by Luo and Pardos [19].

Lastly, we show that features generated from transcript records, particularly aggregates and summaries of students' academics, perform better for predictions than demographic and pre-entry data. Much of this is likely due to the selectivity of the University and its admissions policy. Nevertheless, it demonstrates how useful transcript data can be for such prediction tasks in contrast to information on students prior to college. We demonstrate that using subsets of data from registrar databases (in this case, aggregates of students' first year) can be nearly as effective for predictions as hand-generating a wide swath of features from different institutional data sources.

7. ACKNOWLEDGMENTS

The authors would like to thank the data stewards at the University of Washington for their assistance in obtaining the data used in this work.

8. REFERENCES

[1] E. Aguiar, N. V. Chawla, J. Brockman, G. A. Ambrose, and V. Goodrich. Engagement vs performance: using electronic portfolios to predict first semester engineering student retention. In *Proceedings of the 4th International Conference on Learning Analytics And Knowledge*, pages 103–112. ACM, 2014.

[2] L. Aulck, R. Aras, L. Li, C. L'Heureux, P. Lu, and J. West. STEM-ming the tide: Predicting STEM attrition using student transcript data. *SIGKDD's Machine Learning for Education Workshop*, 2017.

[3] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West. Predicting student dropout in higher education. *ICML's Machine Learning in Social Good Applications Workshop*, 2016.

[4] R. S. Baker and P. S. Inventado. Educational data mining and learning analytics. In *Learning Analytics*, pages 61–75. Springer, 2014.

[5] J. Bayer, H. Bydzovská, J. Géryk, T. Obsivac, and L. Popelinsky. Predicting drop-out from social behaviour of students. In *Proceedings of the 5th International Conference on Educational Data Mining*, 2012.

[6] A. F. Cabrera, A. Nora, and M. B. Castaneda. The role of finances in the persistence process: A structural model. *Research in Higher Education*, 33(5):571–593, 1992.

[7] A. F. Cabrera, A. Nora, and M. B. Castaneda. College persistence: Structural equations modeling test of an integrated model of student retention. *The journal of higher education*, 64(2):123–139, 1993.

[8] A. L. Caison. Analysis of institutionally specific retention research: A comparison between survey and institutional database methods. *Research in Higher Education*, 48(4):435–451, 2007.

[9] A. P. Carnevale, N. Smith, and J. Strohl. Recovery: Job growth and education requirements through 2020. 2013.

[10] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers. Predicting students drop out: A case study. *International Working Group on Educational Data Mining*, 2009.

[11] D. Delen. Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice*, 13(1):17–35, 2011.

[12] C. Demetriou and A. Schmitz-Sciborski. Integration, motivation, strengths and optimism: Retention theories past, present and future. In *Proceedings of the 7th National Symposium on Student Retention, Charleston, SC*, pages 300–312, 2011.

[13] S. Halawa, D. Greene, and J. Mitchell. Dropout prediction in MOOCs using learner activity features. *Experiences and best practices in and around MOOCs*, 7, 2014.

[14] D. Hossler. Managing student retention: Is the glass half full, half empty, or simply empty? *College and University*, 81(2):11–14, 2006.

[15] W. E. Hudson Sr. Can an early alert excessive absenteeism warning system be effective in retaining freshman students? *Journal of College Student Retention: Research, Theory & Practice*, 7(3):217–226, 2005.

[16] S. M. Jayaprakash, E. W. Moody, E. J. Lauría, J. R. Regan, and J. D. Baron. Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1):6–47, 2014.

[17] N. Johnson. The institutional costs of student attrition. *Delta Cost Project at American Institutes for Research*, 2012.

[18] Z. J. Kovačić. Early prediction of student success: mining students enrolment data. In *Proceedings of Informing Science & IT Education Conference (InSITE)*, pages 647–665. Citeseer, 2010.

[19] Y. Luo and Z. A. Pardos. Diagnosing university student subject proficiency and predicting degree completion in vector space. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[20] R. Manrique, B. P. Nunes, O. Marino, M. A. Casanova, and T. Nurmikko-Fuller. An analysis of student representation, representative features and classification algorithms to predict degree dropout. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 401–410. ACM, 2019.

[21] L. G. Moseley and D. M. Mead. Predicting who will drop out of nursing courses: a machine learning exercise. *Nurse education today*, 28(4):469–475, 2008.

[22] M. Nagy and R. Molontay. Predicting dropout in higher education based on secondary school performance. In *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*, pages 000389–000394. IEEE, 2018.

[23] D. Niemi and E. Gitin. Using big data to predict student dropouts: Technology affordances for research. In *Proceedings of the International Association for Development of the Information Society (IADIS) International Conference on Cognition and Exploratory Learning in Digital Age*, 2012.

[24] T. J. Pantages and C. F. Creedon. Studies of college attrition: 1950–1975. *Review of educational research*, 48(1):49–101, 1978.

[25] S. Ram, Y. Wang, F. Currim, and S. Currim. Using big data for predicting freshmen retention. 2015.

[26] C. Romero and S. Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010.

[27] S. Sadati and N. A. Libre. Development of an early alert system to predict students at risk of failing based on their early course activities. 2017.

[28] M. Schneider. Finishing the first lap: The cost of first year student attrition in America's four year colleges and universities. *American Institutes for Research*, 2010.

[29] J. M. Simons. *A national study of student early alert models at four-year institutions of higher education.* Arkansas State University, 2011.

[30] W. Spady. Dropouts from higher education: Toward an empirical model. *Interchange*, 2(3):38–62, 1971.

[31] J. Summerskill. Dropouts from college. In *The American College*. Wiley, New York, 1965.

[32] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, 41(2):321–330, 2014.

[33] V. Tinto. Defining dropout: A matter of perspective. *New Directions for Institutional Research*, 1982(36):3–15, 1982.

[34] V. Tinto. *Leaving college: Rethinking the causes and cures of student attrition.* University of Chicago Press, 1987.

[35] D. Yang, T. Sinha, D. Adamson, and C. P. Rosé. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop*, volume 11, page 14, 2013.