Using Common Academic Indicators to Predict Graduation Rates at CSUN, 2005-2014

J. Martinez, A. Miller, R. Wolff,
B. Shapiro, and C. Shubin

Abstract - The median time to graduation at California State University, Northridge (CSUN) is five years and fewer than fifty percent of first-time freshmen graduate in fewer than eight years. We used data mining and predictive analytics to determine some of the key academic indicators of success at CSUN. The most important indicators that we found were (a) which math course the student was placed in (or took first); (b) student grade point average (GPA) at the end of each of the first two terms in residence; and (c) successful completion of a freshman experience seminar course (UNIV 100). When all three are considered simultaneously, we can correctly identify over two thirds of the students who will drop out without graduating (fallout $F \approx 0.3$), while misidentifying approximately one-fifth of students who ultimately graduate as at-risk of not graduating (recall $R \approx 0.8$).

Keywords : logistic regression; confusion matrix; ROC curve; data visualization; optimal unit load

Mathematics Subject Classification (2010) **:** 62-07; 62-09; 68

1 Introduction

California State University, Northridge (CSUN) is a comprehensive university offering 133 disciplines, 84 master's degree options, and two doctoral programs. Located in the San Fernando Valley of northwestern Los Angeles with over 40,000 students, CSUN is one of the largest single campus universities in the United States. A total of 6814 baccalaureate and 1913 graduate degrees were awarded during the 2012-2013 academic year.

The median time to graduation at CSUN is five years, and the long-term graduation rate (students who graduate less than eight years after their first matriculation) is below fifty percent (Figure 1). There is a very clear distinction between the grade point average (GPA) of students who drop out after several terms (or years), even when they are very close to completion, and those who actually graduate (Figure 2). We mined a ten-year database of student records to determine if there are any significant academic indicators of early student drop-out or long-term success. Here we present the results of this study including some predictive analytics.

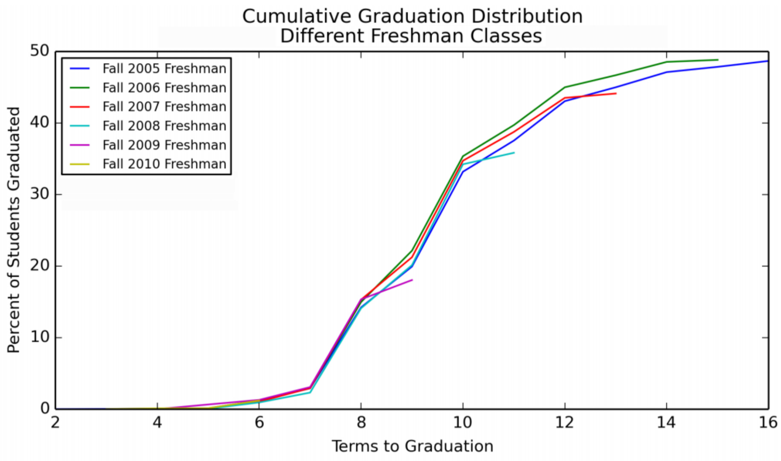

Figure 1: Graduation rates for first-time freshmen at CSUN. The curves show the cumulative graduation rate, as of the end of the 2012-2013 academic year, for each incoming freshman class from 2005 through 2010. As a large number of students take six or more years to graduate, the 2011 freshman class is the last for which we can reliably interpret graduation rates. The cumulative graduation rate gives the percentage of students who have graduated up until a given term.

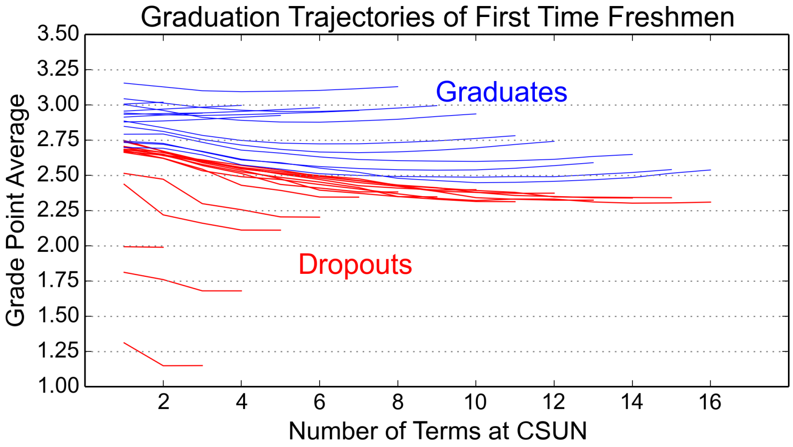

Figure 2: Average grade point average (GPA) as a function of number of regular terms (fall and spring semesters) at CSUN. Each individual line represents the subset of students who attended for a particular number of contiguous semesters before they dropped out (red) or graduated (blue).

2 Methods

A student records database was provided, courtesy of the Office of the Provost, from the Office of Institutional Research and stored in an SQLite database. [5] This database contained the student academic record (courses taken, when taken, and grade received) and academic assessment results [e.g. SAT, Entry Level Mathematics Test (ELM), Mathematics Placement Test (MPT)] of every student matriculated at CSUN from 2005 through 2014. We extracted various factors, individually and in combination, and used logistic regression to compare their relative importance in being able to predict student success at CSUN.

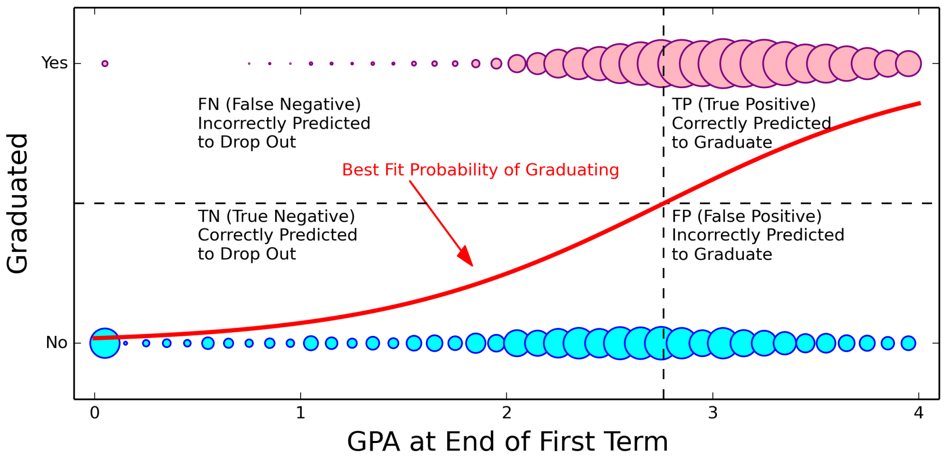

Figure 3: Logistic regression fit of data using first term GPA as a predictor of graduation. The red curve gives the estimator of the probability of graduation based purely on the GPA. This probabilistic model shows the errors inherent with the conceptual model. Wherever the threshold for acceptance (saying that a student is likely to graduate), there will be some false predictions. These are indicated by either False Positives (FP) or False Negatives (FN) in the figure. The sizes of the bubbles are directly proportional to the number of students that fell into that GPA bin.

Logistic Regression is primarily a data classification technique; it can be used to classify data between two or more classes. In the present study, we classified students as either "graduated" or "did not graduate." If there are two classes C_1 and C_2, then the posterior probability for C_1, given explanatory data vector $\mathbf{x}$ and feature vector $\mathbf{y}$ is

$$p(C_1|\mathbf{x}) = \sigma(f(\mathbf{x}; C_1, C_2)) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{y}) \tag{1}$$

where $\sigma(x) = 1/(1+e^{-x})$ is the saturating (or s-shaped) logistic sigmoid function, and

$$f(\mathbf{x}; C_1, C_2) = \ln \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)} \tag{2}$$

and the unknown parameter vector $\mathbf{w}$ is found by maximum likelihood estimation [1]. The method is similarly extended to multiple classes.

Each data set was divided up into a *training set* consisting of 80% of the data, and a *test set* consisting of the other 20% of the data. The probability distribution was fit to the training set, and then the method was evaluated using the test set. The standard measure of fit is the confusion matrix:

$$\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix} \tag{3}$$

where

TP=True Positives, i.e., number of students predicted to graduate who actually graduate;

FP=False Postives, i.e., number of students predicted to graduate who did not graduate;

FN=False Negatives, i.e., number of students predicted to not graduate who do graduate;

TN=True Negatives, i.e., number of students predicted to not graduate who did not graduate.

and its derived measures,

$$P = \text{Precision} = \frac{TP}{TP+FP}$$

the fraction of students graduated out of those predicted to graduate;

$$R = \text{Recall} = \frac{TP}{TP+FN}$$

the fraction of students predicted to graduate out of those who graduated,

$$F = \text{Fallout} = \frac{FP}{TN+FP}$$

the fraction of students who were incorrectly predicted to graduate out of those who did not graduate, and

$$ACC = \text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

the total accuracy of the prediction.

One way to evaluate a classifier is the Receiver Operating Characteristic curve, which plots the recall against the fallout as all parameters in the model vary [6]. (One would get a different point on the curve, for example, if 50% is not chosen as the probability boundary between the two classes.) In an ideal predictor, we would would want to have very low fallout and very high recall. Low fallout means that $1 - F$ is very close to 1, or the proportion of students who are predicted to drop out is predicted very accurately. We want high recall because R gives the proportion of graduates who were correctly predicted to graduate. Thus, F is plotted on the horizontal axis and R on the vertical axis, the optimal (F, R) point is in the upper left hand corner of the plot. One way of quantifying this with a single number, when plotting R vs F, is to calculate the total area under the curve. The better the classifier, the closer the area under the curve (AUC) will be to one [4] . If we had a classifier that put our point at precisely the upper left hand corner then the area would be precisely one [the ROC curve would connect the origin to (0,1) then (1,1)].

Analyses and plotting were performed in Jupyter notebooks using numpy and scipy. [3]

3 Results

3.1 Predictors of Success.

We used logistic regression to compare the following as predictors of student success for students who began as freshmen at CSUN, measured as whether or not the student will ever graduate: SAT Math Score; SAT Verbal Score; the combination of all three SAT scores in Math, Verbal, and Writing (as a vector); Grade Point Average (GPA) at the end of the first term at CSUN; GPA at the end of the each of the first two terms, as vector of two parameters; and the grade in UNIV 100, the "Freshman Experience" Course at CSUN.

A typical logistic regression result is shown in Figure 3. This figure shows the modeled probability density function for graduating as a function of a single variable, the student's grade point average after their first term. The circles illustrate the students who either graduate or drop out, divided into bins of width 0.1 on a 4.0 GPA scale, with the size of the circle proportional to the number of students in each bin. Thus a circle centered at 3.05 represents all the students with GPAs between 3.0 and 3.1. The plot illustrates the inherent difficulties with this method. Although the students who graduate do tend to have higher grade point averages overall than the students who do not, as illustrated by the large cluster of pink circles on the top right of the plot, there are many students with high GPAs during their first term who do not graduate (the blue circles in the bottom right quadrant of the plot) and many other students with low first term GPAs who actually do graduate (upper left quadrant). Predicting success or not requires an arbitrary GPA threshold, shown by the intersection of the four quadrants in the plot, indicating GPAs below which students are predicted not to graduate.

A better prediction can be made, indicated by improved accuracy and area under the curve (AUC) of the receiver operating characteristic (ROC) curve, by adding additional variables and finding a multivariate probability density. These functions are difficult to demonstrate visually, especially if they have more than two dependent variables. The confusion matrices are summarized in Table 1 and the ROC curves are shown in Figure 4. The main message from Table 1 is that these variables give us good predictors of students who will succeed but very few good predictors of students who will drop out.

Predictor	n	F	1-F	R	1-R	P	ACC	AUC
First Math + Two GPA + U100	1473	0.29	0.71	0.78	0.22	0.73	0.75	0.84
First Math + Two Terms GPA	8493	0.44	0.56	0.85	0.15	0.71	0.72	0.80
First Math + One Term GPA	8493	0.44	0.56	0.85	0.15	0.71	0.72	0.78
Two Term GPA	16031	0.54	0.46	0.91	0.09	0.74	0.75	0.76
One Term GPA	16031	0.62	0.38	0.91	0.09	0.72	0.72	0.72
First Math class	8493	0.58	0.42	0.83	0.17	0.65	0.65	67
All SAT	2397	0.15	0.85	0.23	0.77	0.51	0.60	0.61
UNIV 100	1524	0.81	0.19	0.96	0.04	0.53	0.57	0.58
Math SAT	6530	0.86	0.14	0.90	0.10	0.57	0.56	0.56
Verbal SAT	6528	0.95	0.05	0.96	0.04	0.56	0.56	0.55

Table 1: Results of various fits using logistic regression. $F = FP/(TN+FP)$, the percentage of dropouts who were incorrectly predicted to graduate; $1 - F = FP/(TN + FP)$, the percentage of dropouts who were correctly predicted to dropout; $R = TP/(TP + FN)$, the percentage of graduates who were correctly predicted to graduate; $1 - R = FN/(TP + FN)$, the percentage of graduates who were incorrectly predicted to drop out; ACC=accuracy = $(TP + FN)/(TP + TN + FP + FN)$, the percentage of overall correct predictions. P = precision = $TP/(FP + TP)$, the fraction of those predicted to graduate who actually graduate; AUC = area under the (ROC) curve. The sample size is the number of students in the test set. The training set was four times as large, and the total data set was five times as large.

3.2 Optimal Student Load.

Most students enroll in twelve to fifteen units of courses for credit (Figure 5). To graduate in four years, attending only during the regular fall and spring semesters, a student must successfully complete on average fifteen units per semester. To qualify for financial aid, a student need only register for twelve units. This means that a student who only takes the minimum course load will require at least five years to graduate. Although a small fraction of students begin with fifteen units, after two or three semesters they tend to reduce their load to twelve units.

We examined whether there was an optimal student load during freshmen year by plotting the number of units successfully completed as a function of the number of units attempted. We found no significant difference between students who choose to enroll in fifteen versus twelve units (Figure 6).

3.3 First Math Course.

To determine the relative effect of the first math course completed upon graduation, we examined each math class that had more than 500 enrollments during the reporting

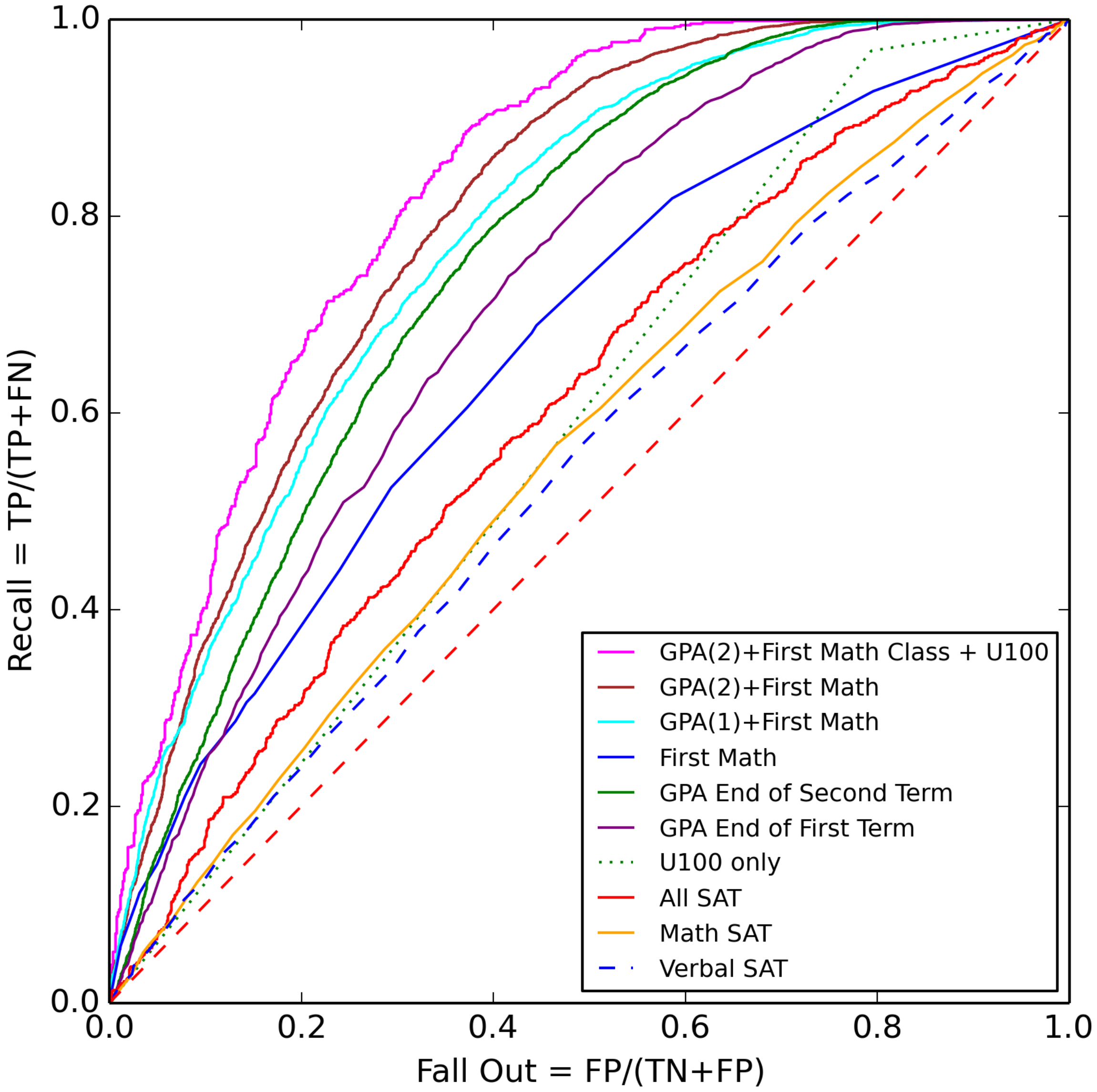

Figure 4: ROC curves for logistic regression for various features. The larger the area under the curve (AUC), the better the prediction made by the logistic fit. The dashed orange line is along the main diagonal of the plot for reference.

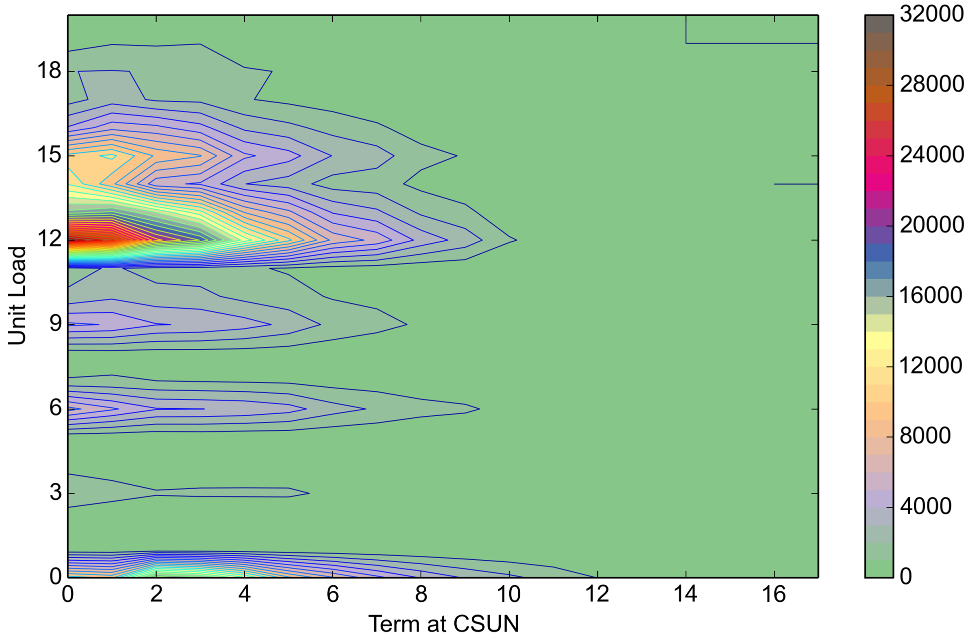

Figure 5: Topographic map of student load as a function of time at CSUN.

period. There were fourteen classes, as summarized in Table 2, roughly comprised as developmental math (92, 93, 94, 94A, 95); college algebra (102); terminal non-major courses (business math 103; math ideas 131; freshman stats 140); precalculus and calculus (105, 150A, 150B); and courses for elementary education (210, 310). In every case, there are some students who managed to graduate even if they failed the course the first time. Per Table 2, students who started in developmental math had a significantly lower probability of graduating than their non-developmental counterparts.

Students who started in Math 92 had a 40% chance of graduating if they passed the class; this increased to 50% if they started in Math 93 and pass that. Students who started in 92 or 93 and fail the first time have only a 9% or 16% change of graduating, respectively. By comparison, students who took the earlier Math 94 course (no longer offered) and passed it had a much higher rate of success at CSUN, with a 63% graduation rate. Students who failed Math 94 on their first attempt still had a better than 50% chance of graduating.

Students who started in Math 210 or Math 310 were the most likely to graduate among this population. Despite their higher numbers, these are not advanced courses in the sense that their only prerequisite is Math 93 or equivalent. However, the difference in graduation rate is probably due to the different student population that they are drawn from, which is almost entirely Liberal Studies majors with an intent to become an elementary or middle school teacher. The other courses all have much wider campus

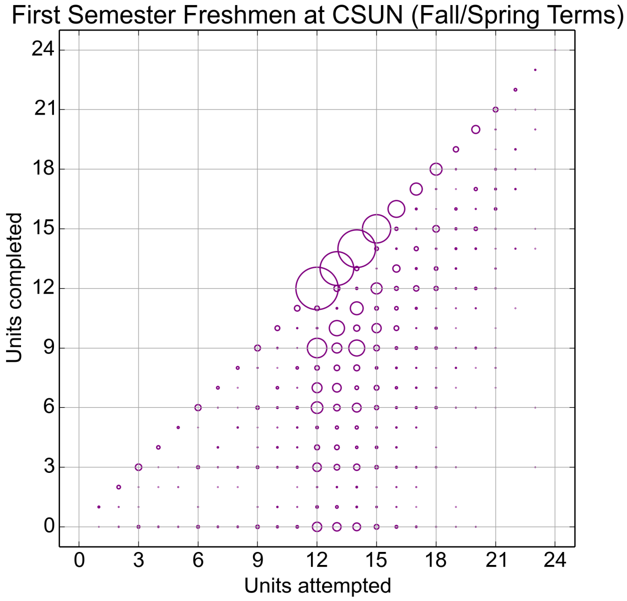

Figure 6: Units completed by first-time freshmen as function of units attempted. The sizes of the markers are proportional to the number or students.

draw. In addition, students may not take these course during their first year at CSUN, as they are classified as sophomore and junior level courses, so the audiences may be more mature, though this was not investigated.

3.4 Choice of Writing Course.

To see if the dependence on **primary course selection** that was demonstrated in math courses holds in other fields, we investigated the correlation between graduation rate and the choice of "Analytical Reasoning and Expository Writing" class. Unlike math, where there is a progression of courses that depend on one another as prerequisites, students must select an analytical reasoning course from among sixteen different options in the humanities. We quantified the results by graduation rates (in percent) among students who either passed or failed the course on their first attempt (Figure 7). We found no significant variation in the graduation rates due to choice of courses.

3.5 Most Commonly Failed Courses.

We next examined the most commonly failed courses taken in the last term before dropping out, to see if there was any relationship between course selection in the final term and decision to drop out.

Four of the five most commonly failed courses among all students in the final term

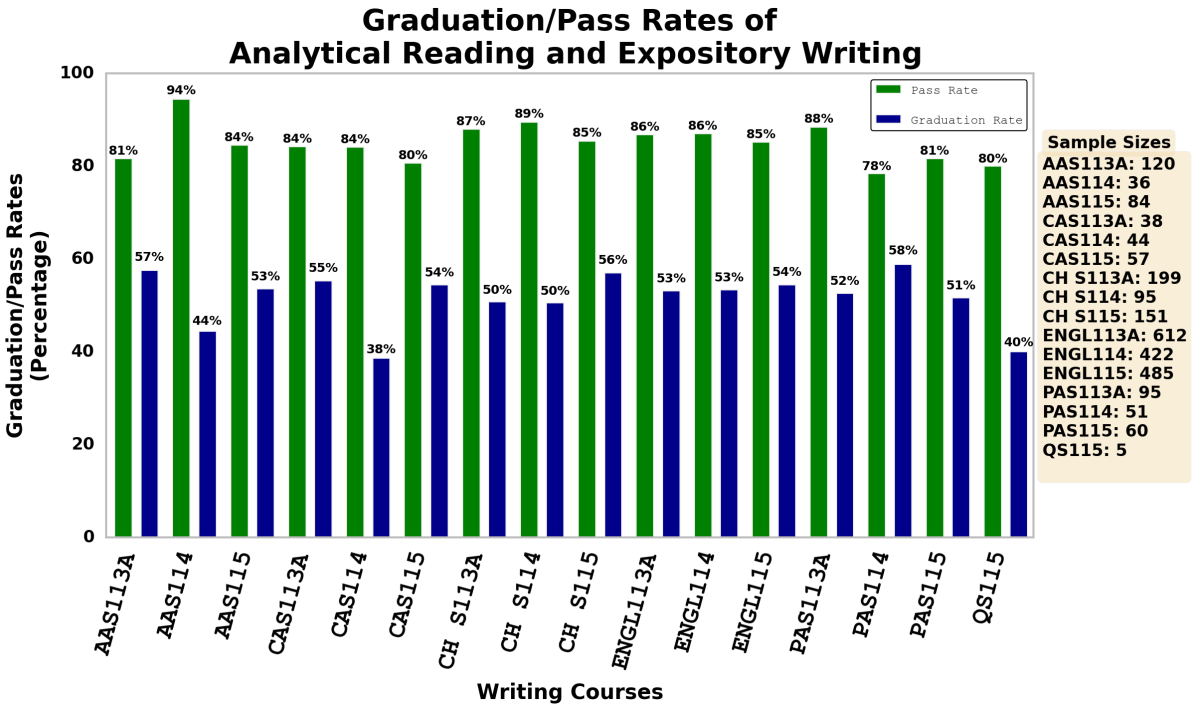

Figure 7: Graduation rates as a function of whether or not a student passed or failed their "Analytical Reasoning and Expository Writing" on their first attempt.

			Did Not	Passed	Passed	Failed	Failed	$P(G\mid P)$	$P(G\mid F)$
Class	Students	Graduate	Graduate	Graduated	Did Not	Graduated	Did Not		
					Graduate		Graduate		
MATH92	5335	30.25	69.75	27.27	40.41	2.98	29.33	40.3	9.2
MATH93	6443	40.26	59.74	35.34	34.66	4.92	25.08	50.4	16.4
MATH94A	2339	61.69	38.31	51.95	29.33	9.75	8.98	63.9	52.1
MATH94	1856	66.70	33.30	54.26	23.28	12.45	10.02	70.0	55.4
MATH95	1099	69.43	30.57	49.23	18.38	20.20	12.19	72.8	62.4
MATH102	5284	56.42	43.58	36.51	15.12	19.91	28.46	70.7	41.2
MATH103	2856	64.29	35.71	48.32	16.35	15.97	19.36	74.7	45.2
MATH105	1064	69.74	30.26	55.17	17.67	14.57	12.59	75.7	53.6
MATH131	3299	61.93	38.07	54.53	23.76	7.40	14.31	69.7	34.1
MATH140	4775	65.17	34.83	48.82	14.43	16.36	20.40	77.2	44.5
MATH150A	731	61.70	38.30	48.70	19.70	13.00	18.6	71.2	41.1
MATH150B	593	66.44	33.56	52.28	14.67	14.17	18.89	78.1	42.9
MATH210	1331	77.01	22.99	60.11	13.15	16.90	9.84	82.1	63.2
MATH310	1597	88.17	11.83	63.68	5.26	24.48	6.57	92.4	78.8

Table 2: Graduation rates (in percentages) broken down for most commonly taken first math course. The last two columns give the probability of graduation given the student passes ($P(G|P)$) and given the student fails ($P(G|F)$). Students still attending CSUN at the end of the study period were excluded from the count. The numbers give the percentages of students who eventually graduated or left the university for another reason, per the first math course they took at CSUN. This table shows all first math classes completed by at least 500 students who subsequently left the university.

were math courses. The failed courses, in order, were Math 102, Math 93, Math 92, Soc 150, and Math 140. As illustrated in Figure 8, in two of these courses (Math 102 and Math 140) the rate of failure was approximately the same for all levels of academic maturity. Thus additional experience on campus did not provide these students with the academic preparation they needed to succeed. On the other hand, the failure rate for Soc 150 dropped, indicating either a lower rate of enrollment by upperclassmen or better student preparation.

Figure 9 shows the most commonly failed courses the last term before graduating among students who started at CSUN as first-time freshmen. The most common classes failed remain, in order, Math 93, Math 102, Math 92, and Soc 150, but a number of other courses are also significant. The preponderance of failures occur when the student is classified as a freshman.

The most commonly failed courses among transfer students are shown in Figure 10. While Math 102 and Math 140 remain the second and sixth most common courses failed the term before dropping out, this group is dominated by failed upper level courses, primarily in College of Business and Economics (CoBaE). Among the top fifteen courses, ten are in CoBaE. Besides the two math classes mentioned, there are FCS 340 (”Marriage and Family Relations”, 10th most common), Chem 333 (12th), and Art 315 (14th).

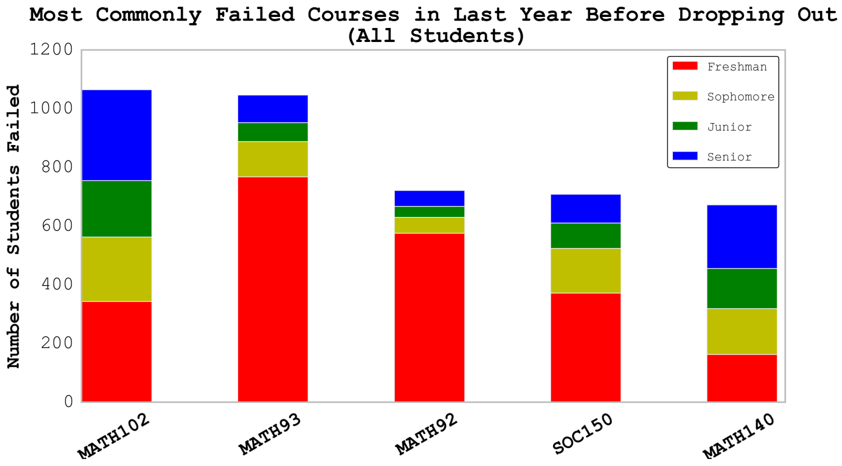

Figure 8: Most common classes failed before dropping out.

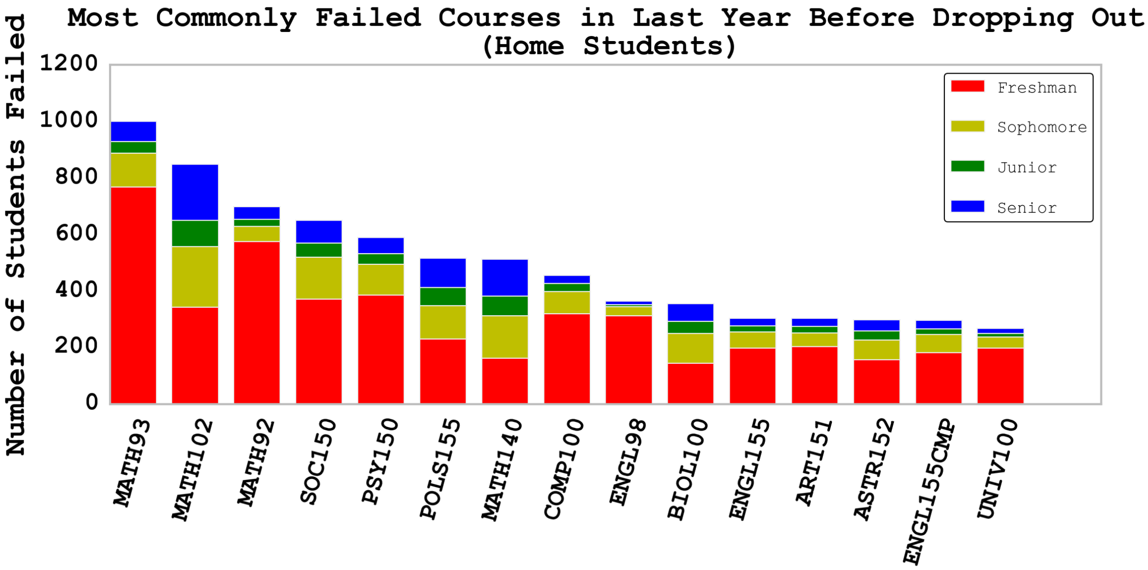

Figure 9: Most common classes failed before dropping out for First-Time Freshmen.

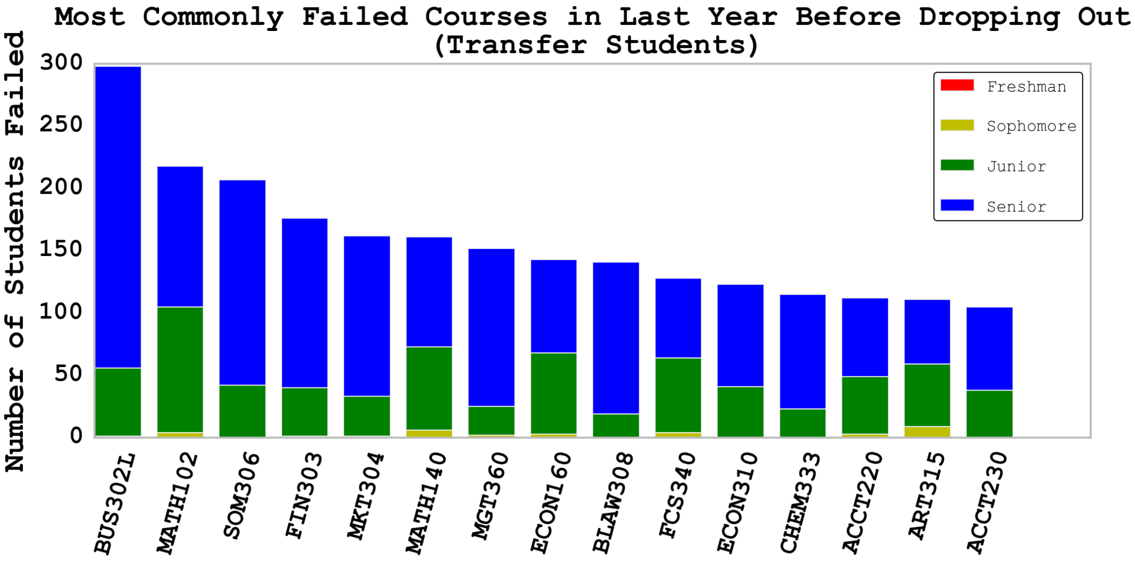

Figure 10: Most common class failed before dropping out (transfer students only)

4 Discussion

This paper has been a preliminary attempt to understand what factors cause students to fail to complete college. Anecdotal evidence suggests that students leave school because of factors such as conflicts with work schedules, changing off-campus employment or familial responsibility, difficulty commuting to work, and financial insecurity. Up to ten percent of students who leave CSUN may actually be transferring to other universities (CSUN Office of Institutional Research). University academic policy changes introduce a variety of additional latent variables that we have not taken into account. Nevertheless, a great deal of additional information can still be found by examining the academic databases. While some early warning and course recommender systems exist, we are not aware of any system for CSUN students that indicates course paths or major changes that help at-risk students with specific profiles.

Acknowledgments

We would like to acknowledge the help of Provost Harry Hellenbrand (Emeritus); the Office of the Provost at CSUN; and the Office of Institutional Research at CSUN.

References

[1] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[2] E. Jones, E. Oliphant, P. Peterson, *et al.*, SciPy: Open Source Scientific Tools for Python, available online at the URL: `http://www.scipy.org/`

[3] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing, *et al.*, Jupyter Notebooks—a publishing format for reproducible computational workflows, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, (2016), 87–90.

[4] K. Murphy *Machine Learning: A Probabilistic Perspective.* MIT Press, 2012.

[5] R. Hipp, SQLite, available online at the URL: `http://www.hwaci.com/sw/sqlite/`

[6] T. Fawcett, An introduction to ROC Analysis, *Pattern Recognition Letters*, **27** (2006), 861–874.

Jorge Martinez
California State University, Northridge
18111 Nordhoff St., Northridge, CA 91330
E-mail: `jorge.martinez.530@my.csun.edu`

Andrew Miller
University of California San Diego
10760 Big Bend Ave., Sunland, CA 91040
E-mail: `awmiller@ucsd.edu`

Richard Wolff
Columbia University
312 Prospect Ave., New York, NY
E-mail: `rtw2123@columbia.edu`

Bruce Shapiro
California State University, Northridge
Mail Stop 8313, 18111 Nordhoff St., Northridge, CA 91330
E-mail: `bruce.e.shapiro@csun.edu`

Carol Shubin
California State University, Northridge
18111 Nordhoff St., Northridge, CA 91330
E-mail: `carol.shubin@csun.edu`

Received: December 28, 2018 **Accepted**: February 27, 2019
Communicated by Mona Mocanasu