Analysis of Student Success

Jorge Martinez, Andrew Miller

Sponsored by Dr. Bruce E. Shapiro, Department of Mathematics

California State University, Northridge

Abstract

We present a study of student success at CSUN. The goal of this analysis of student records was to identify and pinpoint which, if any, student behaviors might lead to his or her eventual success or failure, potentially as early as their first semester. A student that is 'at risk' of not graduating could manifest at any time in their academic career regardless of GPA, SAT score, or how close they are to graduating. By identifying 'at risk' students they can be offered the resources that they might need to succeed, thereby preventing them from dropping out. Questions that we wanted to address include: What are the early predictors of student success (failure) (e.g, specific course grades, GPA, study skills, work/study, etc)? Can we identify students early enough in their academic careers to make a difference by developing models using predictive analytics? Do transfer students perform better than first time freshmen students? What are the demographics of student drop-outs? Can we identify the students in the middle that could go either way? Can we build an automated system that screens through CSUN students for 'at risk' students?

Background

The California State University, Northridge (CSUN) is a comprehensive university offering 68 bachelor's degrees, 47 credentials, 58 master's programs, and two doctoral programs. Located in the San Fernando Valley of north-western 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.

Methods

A student records database was provided, courtesy of the Office of the Provost, from the Office of Institutional Research and stored in a SQLite database. 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 used the iPython Notebook (IDE) to perform data mining and computational processes. We used the Python API (application programming interface) Pandas to extract

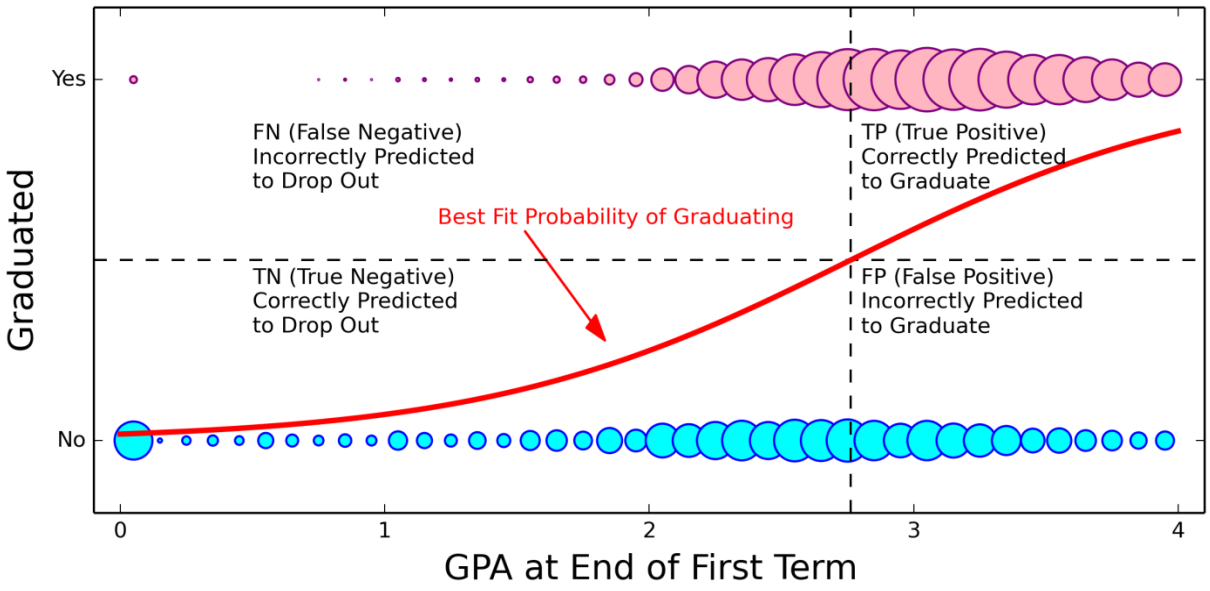
various factors, individually and in combination. Then we used another Python API, scikitlearn's logistic regression model to compare their relative importance in being able to predict student success at CSUN.

Logistic Regression is primarily a data classification technique; it can be used to separate data points into one of two or more classes. In the present study, we used it to identify students as either "graduated" or "did not graduate." To perform logistic regression, we first randomly separated the data into two smaller data sets: a "training" set containing 80% of the original data, and a "test" set composed of the remaining 20% of the original data. It then fits the training set to the explanatory values (e.g., GPA) to find a best fit probability function (e.g., the probability of graduating as a function of GPA). Then we assigned a threshold to make predictions, based on the test set, of whether or not the student will be successful, e.g., we required that the probability of graduation must be greater than 50 percent. To evaluate the predictions, the model generates a confusion matrix based on the test set:

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

As depicted in **figure 1**, the **confusion matrix** is made up of four categories that our predictions can fall into. They can either be: True Positives (TP), False Positives (FP), True Negatives (TN), or False Negatives (FN). Using the confusion matrix, we evaluated the correlation between the explanatory variables (e.g. SAT scores, GPA) and the independent variable (e.g. graduation rate). These four categories are plugged into formulas in order to get the **Accuracy**, **Precision**, **Fallout**, and **Recall** of the model. We used the following equations:

Figure 1: Shows a regression line fitting CSUN students GPA to their Graduation Status. The blue and red bubbles are a histogram using bubbles instead of bars that represents the distribution of students clustered by their GPA and whether or not they graduated. All students fall into one of the following four groups: True Positives, False Positives, True Negatives, or False Negatives. This graph also shows the limitations of logistic regression; it is highly likely that there will always be outliers that don't fit the model.



$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

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

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

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

To visualize the data, we used Matplotlib and D3JS.

What are predictors of student success?

Useful predictors can be identified by having high accuracy, recall, and precision alongside with low fallout. Predictors can be good at identifying success but not failure, or they can be good at identifying failure but not success. The best predictor would be able to identify both success and failure simultaneously.

Results

Good predictors that we found include a student's GPA, the grade received in their First Math Course (FMC), and whether they passed Uni100 (a university preparation course). The strongest stand alone predictor that was found in the study was GPA; however, using all these predictors in combination created the best models.

Bad predictors include Entry Level Math (ELM) placement score, SAT math score, and a student's respective zip code. Interestingly, our results show that the SAT, an exam widely believed to predict college success, is terrible at predicting graduation. Upon removing these predictors from the models, Accuracy, Precision, Fallout, and Recall improved considerably.

Can we identify students early enough in their academic careers to make a difference?

The short answer is yes. Immediately

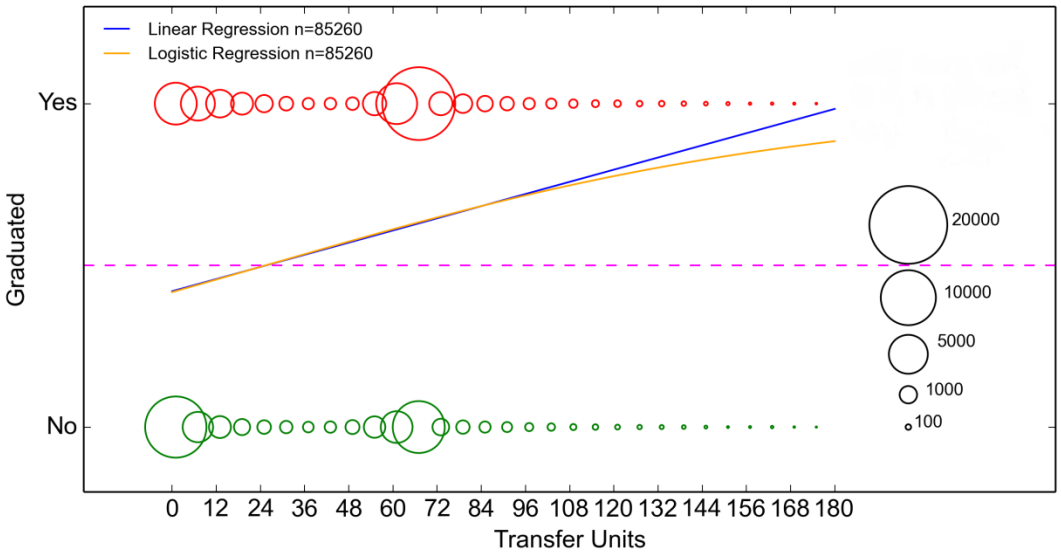


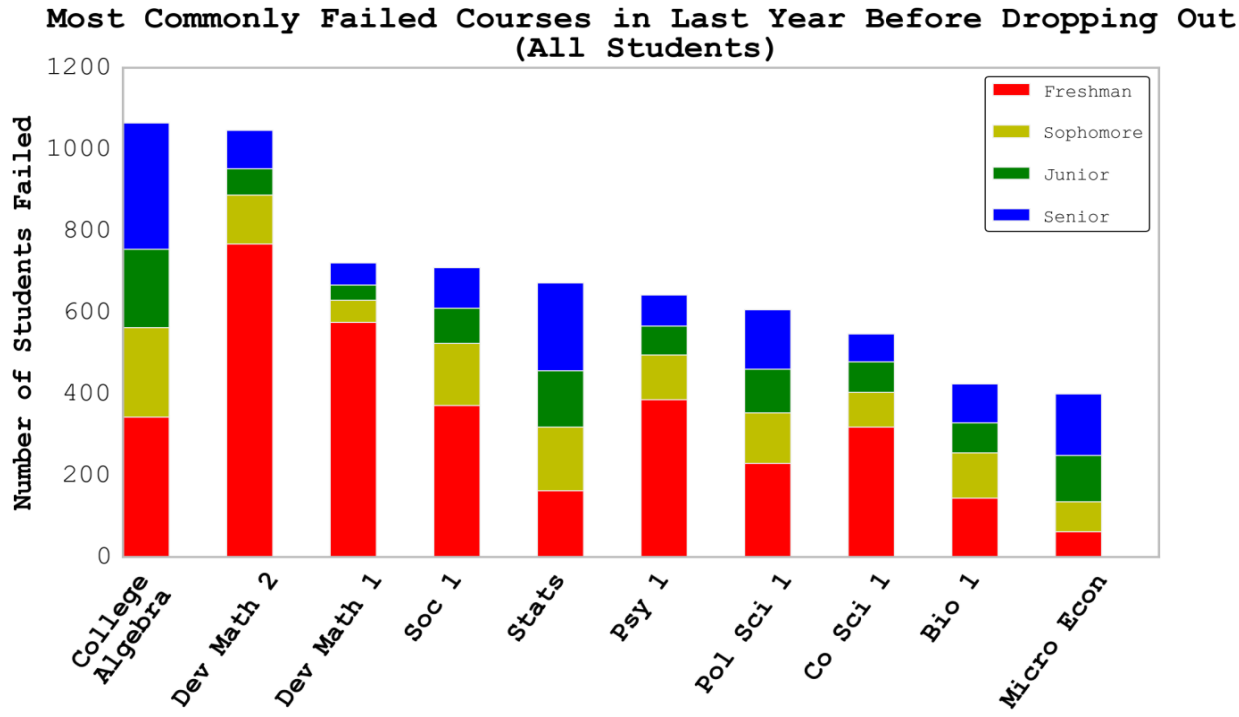
Figure 2: Shows a logistic and linear regression of Graduation Status as a function of Transfer Units.

following a student's first semester, using the student's GPA we can predict within 72% accuracy whether that student will graduate or fail. Using their second semester GPA in combination with their first semester GPA we can predict within 75% accuracy whether or not the student will graduate or fail.

Transfer Students from community colleges perform better and have a higher chance of graduating. We confirmed this by looking at graduation rates between first time freshmen at CSUN and transfer students. As shown in **figure 2**, students with 0 transfer units have a huge cluster of non-graduates whilst students coming in with 60 transfer units have a huge cluster of graduates. This is likely due to many first time freshmen being weeded out in their first semesters. It is important to note that these same phenomena could have occurred at the community college level. However, the records that we have are limited to CSUN and we were unable to investigate that.

Often when people think of college dropouts, they think of the freshmen year of college as being notorious for having a high drop out rate. Urban logic attributes this to weed-out courses like elementary mathematics and general education classes. From **figure 3,** we see that freshmen, in fact, represent 49.2% of classes failed before dropping out, but sophomores, juniors, and seniors represent 17.6%, 14.0%, and 19.2%, respectively. Of the top five failed courses, four are math courses. Interestingly, the most commonly failed math course, College Algebra, is more or less equally distributed among the class groups. Introductory Statistics, nominally a freshman class, showed a very similar trend, but with seniors being slightly larger. Ultimately, this shows that math courses are bottlenecks alongside

Figure 3: Shows the most commonly failed courses during a student's last year before the student drops out. The different colors represent what school year the students were a part of before they dropped out. The freshmen numbers although large were not unexpected; however, there was a surprising amount of senior drop outs that only raises further questions.



with some 100 level general education courses.

Discussion

Students in the Middle

While we can predict many of the students who are likely to either succeed or fail with a high degree of accuracy, there are a large number of "students in the middle." These are students whose outcomes are far less certain. Many of these students fall into the false positive and false negative categories. In our ongoing work we are attempting to further elucidate the root causes of these students' success and/or failure.

Although these students are the most difficult to understand and identify, their identification is crucial because many of these students in middle who fail look almost identical, academically, to successful students, except that they don't graduate. This leads us to believe that they have the potential to succeed if we could only identify the correct resource to offer to them. The models that we used have helped in identifying good predictors and not so good predictors. However, an inherent problem with the regression models is that we will always have some predictions that are incorrect. To minimize this, research can look into more advanced computational methods of analysis.

Predictive Analytics: Looking Ahead with Machine Learning

Currently there is no machine learning system in place that is actively searching for students that are 'at risk' of dropping out. An automated system that screens CSUN students in real time can be used to find students that are 'at risk' of not being successful in their academic career. The only systems currently in place are those in which queries such as "Tell me all the students with GPAs in a given range who failed certain classes" must be made manually by the user. These systems are not generally available to academic advisors, they are only available to deans and counselors, and do not use any sort of predictive or smart analytics.

New Explanatory Variables

There is more data that can be mined for explanatory variables that we would like to examine but this data has not yet been made available. One possible set of predictors we would like to investigate is the number of hours a student works per week in combination with the number of units that are being attempted in the current semester. We would conjecture that as the appropriately weighted sum of these two numbers increases the likelihood of graduating in a reasonable amount of time may decrease.

Acknowledgments

We give our deepest thanks to Dr. Bruce Shapiro and Dr. Carol Shubin for allowing us to sunbathe under the glowing sun that is their intelligence. We'd also like to thank the STEM department at LAMC and especially thank Dr. Mike Fenton for making this opportunity possible for us. We'd like to thank the LAMC Math

department for all the help in various forms that they provided. This also would have not been possible without Prof. Emil Sargsyan and all the crash courses on Linear Algebra, Calculus, and Statistics.