Forecasting Student Outcomes at University-Wide Scale Using Machine Learning

Drew Wham PhD.
Penn State Data Scientist
3C Shields Building
University Park
1-(843)-327-5278
Fcw5014@psu.edu

ABSTRACT

Elements of applied statistics and computer science are quickly integrating and being applied to a diverse set of problems in academia and industry. Here I explore the potential value of this multi-disciplinary approach to applications in higher education by applying it to forecasting course level outcomes for individual students at all of Penn State's campuses. Utilizing hundreds of data sources on individual students, ranging from past performance to current course engagement, I demonstrate the potential accuracy of forecasting techniques at identifying high risk students early in the course term. Our preliminary results suggest that %50 of students that earned a D or F in 2015 could have been identified prior to the start of the course.

CCS Concepts

Computing methodologies~Machine learning approaches • *Computing methodologies~Supervised learning by regression* • **Applied computing~Education**

Keywords

Machine Learning; Student Success; Feature Finding; Modeling; Scalability

1. INTRODUCTION

In 2015 Penn State hosted more than 800,000 unique student course enrollments across it's main, branch and online campuses. Of those unique course enrollments, %6.4 resulted in a course withdrawal and %7.1 resulted in a grade of D or F. Both of these results (GPA < 2.0 or withdrawing) result in "regret" for both the student and the institution since students earning a D, F or W fail to make progress toward graduation. Early and accurate identification of these students would provide opportunities for advisors to recommend better course sequencing and/or direct students to better utilize learning resources when they are identified as "at risk".

A model was therefore developed to predict student outcomes prior to the start of the semester (day-zero). Others have previously used logistic regression [2,3] and classification and regression trees [3] to predict at-risk students prior to enrollment by modeling end of semester GPA. The approach presented here is similar in that it is trained on a number of the same features that range from the student's previous academic achievements, the student's demographic characteristics and current campus involvement. However, the model presented here is more granular, because it was designed to predict both the probability of withdrawal and the GPA of every student for every enrolled class. To this end, it was also trained on historical data from the individual course and instructor. Here we describe the development of this model and its resulting relative accuracy.

LAK '17, March 13-17, 2017, Vancouver, BC, Canada
ACM 978-1-4503-4870-6/17/03.
http://dx.doi.org/10.1145/3027385.3029467

2. METHODS

2.1 Forecast Model Development

The primary data model was developed from data in Penn State's data warehouse. This data model comprised all transcript records from Fall 2010 to Summer 2016 totaling over 5.5 million unique records from over 250,000 unique students. The primary data model also contained a number of predictive features ranging from student's previous academic achievements (high school GPA, high school class rank, SAT scores, cumulative GPA etc.), the student's demographic characteristics (gender, age, etc.) campus involvement (number of enrolled credits, athlete status, application type etc.) and historical data from the individual course and instructor (average grade, course level etc.). This Primary data model was then split into a training and test set, with all semesters from Fall 2010 to Summer 2015 going into the training set and Fall 2015, Spring 2016 and Summer 2016 going into the test set.

Two models were then developed utilizing the statistical programing language R [4] and Python using the gradient boosted tree based machine learning algorithm [1] XGBoost [5]. Both models were trained on 5 years of data (FA2010-SU2015) and tested on one year of data (FA2015-SU2016). The "Day-Zero" withdrawal model was trained using a logistic regression objective function and a log-loss evaluation metric. After training the withdrawal model, all students that withdrew from classes and did not earn a grade were removed from the dataset. We then trained the "Day-Zero" grade model using a linear regression objective function against the earned GPA of the students on a 4.0 scale using a root mean squared error evaluation metric.

3. RESULTS

3.1 Day-Zero Model

The models were evaluated both on the training data as well as the test data. The day-zero withdrawal model had a log-loss of 0.174 on the training data and 0.196 on the test data. The

day-zero grade model had a root mean squared error of 0.725 on the training data and 0.757 on the test data.

Figure 1 shows the relationship between prediction error and the frequency of that error in the day-zero grade model on the test data. Here error was calculated as the predicted grade minus the observed grade. In this model most errors are relatively small (near zero) with a slight bias towards under predicting. The mean absolute error of the day-zero grade model is 0.533. The model, however, makes more "large" errors in the positive direction than in the negative direction. That is, there are more students that do not do as well as expected than students that do better than expected. This unbalance distribution at the extremes likely explains the slight negative bias which is likely due to the evaluation metric giving additional weight to large errors.

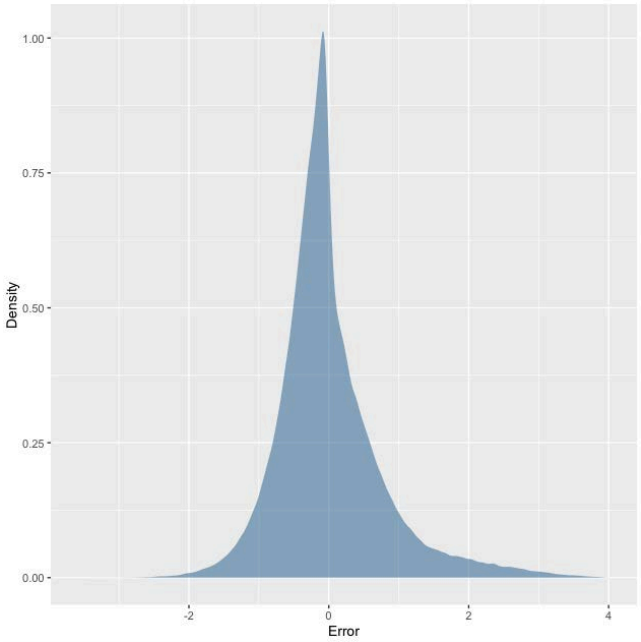

Figure 1. Error of Day Zero Model

To better understand which courses tended to have relatively high and relatively low prediction accuracy we calculated the mean absolute error for every unique course (n=9754) we then filtered out courses for which there were less than 500 student records and then subset out the 25 courses with the most accurate and 25 courses with the least accurate average predictions.

Table 1. Subset of the Courses with the most accurate average prediction

Course	Number of Students	Prediction Error
CHEM_200	635	0.330
ENGL_200	4896	0.427
IST_400	828	0.252
ACCTG_400	556	0.450

Table 2. Subset of the courses with the least accurate average prediction

Course	Number of Students	Prediction Error
MATH_100	4302	0.701
THEA_100	814	0.814
IST_100	2657	0.801
STAT_100	1102	0.727

Table 1 shows a subset of the courses with the most accurate predictions and table 2 shows a subset of the courses with the least accurate predictions. The courses with the most accurate predictions spanned a wide range of disciplines including Accounting, Biology, Chemistry, Finance, Kinesiology, Nursing and Spanish. The courses with the least accurate predictions also spanned a wide range of disciplines including Art, Chemistry, Economics, Math and Theater. Surprisingly, some disciplines, such as Chemistry, had courses that appeared on both extremes of prediction accuracy. This suggests that model accuracy was not highly linked to any particular discipline. Rather, an emergent pattern was that courses that tended to have the least accurate predictions seemed to be more commonly taken early in a students career whereas courses that tended to have the most accurate predictions tended to be taken later in the academic career. This is perhaps an unsurprising result because the amount of data that is available for making a prediction increases over the students time at the institution so it follows that better predictions would be available to 3rd and 4th year students than 1st and 2nd year students.

4. REFERENCES

[1] Friedman, J.H. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29, 1189-1232.

[2] Gansemer-Topf, A. M., Compton, J., Wohlgemuth, D., Forbes, G., & Ralston, E. 2015. Modeling Success: Using Preenrollment Data to Identify Academically At-Risk Students. *Strategic Enrollment Management Quarterly*, *3*(2), 109-131.

[3] Kaleita, A. L., Forbes, G. R., Ralston, E., Compton, J. I., Wohlgemuth, D. 2016. "Pre-Enrollment Identification of At-Risk Students in a Large Engineering College." *International Journal of Engineering Education* 32(4) 1647.

[4] R Core-Team. 2015. R: A language and environment for statistical computing. https://www.R-project.org

[5] XGBoost. 2016. https://xgboost.readthedocs.io