Predicting Graduation at a Public R1 University

Henry Anderson
University of Texas at Arlington
henry.anderson@uta.edu

Afshan Boodhwani
University of Texas at Arlington
afshan.boodhwani@uta.edu

Ryan Baker
University of Pennsylvania
rybaker@upenn.edu

ABSTRACT: In this **poster**, we build a set of high-performance machine learning models to predict 6-year graduation for university undergraduate students, a critical metric for state and federal reporting and university evaluation, using Linear Support Vector Machines, Decision Trees, Logistic Regression, and Stochastic Gradient Descent binary classifiers. We use a data set of over 14,000 students from six Fall cohorts, containing 104 features, drawn from pre-existing university data. This minimizes sparsity and data collection time, while improving coverage of the student body and student activities. Our models achieve high performance, and identify GPA and completed credit hours as the most important predictors.

Keywords: graduation, predictive modelling, first time in college, machine learning

1 INTRODUCTION

For many universities, graduation is an important measure of institutional effectiveness, particularly in an era when some institutions are criticized for very low graduation rates. Researchers have thus sought to understand and predict students' graduation, frequently using machine learning and data mining techniques (Raju and Schumacker, 2015; Kuh et al., 2008; Karamouzis and Vrettos, 2008), and achieving high predictive accuracies. This poster reports on early, but promising, results of our own such efforts to predict 6-year graduation for first time in college (FTIC) undergraduate students in a public, four-year university.

2 DATA

We used a data set taken from a publicly-funded, four-year state research university in the southern United States, which serves a diverse population, and is a federally designated Hispanic-Serving Institution. The data set includes 14,706 FTIC students admitted in the Fall terms of 2006-2012 (inclusive). Only data from a student's first academic year were included, since prior research has shown that this early period of a student's college career is the most critical for retention and graduation outcomes (Tinto, 2006; Arnold and Pistilli, 2012).

We only used data that the university collects as part of its routine reporting efforts, which allows us to make use of a large number of features for each student, while minimizing the data's overall sparsity. Compared to features only available for a smaller number of students—e.g. surveys, interviews—this makes our resulting predictions more reliable given our choice of modeling algorithms.

3 METHODOLOGY

We extracted 104 features related to students' first academic year, in order to capture a broad view of students' experiences and activities. Through prior reporting work (e.g., to state and federal agencies), a number of variables had been identified by the university that provided our data as likely predictors of student graduation. We use these variables, along with several closely related measures, as features in our models. These features fall into four major categories: *academic performance* (e.g. GPA, credit hours completed), *financial information* (e.g. scholarships, unmet need), *pre-admission information* (e.g. SAT/ACT scores, high school rank), and *extra-curricular activities* (e.g. involvement in Greek Life or Athletics).

The target for classification was defined as the binary variable: *did the student graduate from this university within 6 years of first enrolling?* Using this definition, 46% (6,787) of the students in the data set were assigned a label of "true" (graduated). This does not distinguish different types of failure to graduate—students who drop out, transfer to another institution, or graduate in more than six years are all assigned a "false" value for classification. While these do represent very different student outcomes, each still represents a student who is not being fully served by their university, and whom we wish to identify early in their academic career.

We trained a set of four binary classifiers on the data set to predict FTIC students' 6-year graduation, using the scikit-learn 0.20.0 (Pedregosa et al., 2011) implementations: Linear-kernel Support Vector Machine (SVM), Decision Tree, Logistic Regression, and Stochastic Gradient Descent classifier (SGD). All predictor variables were scaled to zero mean and unit variance when training and evaluating the SVM, Logistic Regression, and SGD models. We held out 20% of the data for testing, and performed 5-fold cross-validation on the remaining 80% to tune model parameters. Models with the highest AUC-ROC score during cross-validation were evaluated on the held-out data.

4 RESULTS AND DISCUSSION

Table 1: AUC-ROC and F1 scores for each model, evaluated on the held-out testing set.

Model	AUC	F1
Decision Tree	0.786	0.785
Linear SVM	0.801	0.795
Logistic Regression	0.814	0.810
SGD Classifier	0.824	0.822

The scores on the testing set are reported in Table 1. As that table shows, each of the four classifiers performed approximately equally well on the held-out data.

The models' feature weights are not directly comparable, making it difficult to identify the most important predictors overall. To account for this, we compute an approximate "overall importance" metric. For each model, we sort features by the absolute value of their assigned weight, then calculate each feature's average rank across the four models. Total credit hours completed, cumulative GPA at the end of the first academic year, out-of-major GPA, and the percent completed credit hours (the student's completed credit hours as a percent of the credit hours they enrolled for) were consistently the highest-ranked features (both in the overall ranking and within each model), which is in keeping with much of the prior work on graduation prediction that finds GPA and credit hours to be the most important predictors.

5 FUTURE WORK

Our current results are encouraging, though they only represent an early analysis of the data. The predictions and feature rankings need to be tested experimentally, to investigate whether they are useful for guiding student interventions and changes in university policy. The models may also be suppressing the effects of lower-ranked features, which may be more directly useful for informing interventions or instructional practices; this merits further investigation, e.g. by re-fitting the models using only a subset of the available features. Given the possible applications in interventions and policy decisions, these models should also be thoroughly tested for algorithmic bias, e.g. lower performance for specific student races or ethnicities. We see this as the most pressing, avenue of future work.

REFERENCES

Arnold, K. E., & Pistilli, M. D. (2012, April). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267-270). ACM. https://doi.org/10.1145/2330601.2330666.

Karamouzis, S. T., & Vrettos, A. (2008). An artificial neural network for predicting student graduation outcomes. In *Proceedings of the World Congress on Engineering and Computer Science* (pp. 991-994).

Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. *The journal of higher education*, 79(5), 540-563.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct), 2825-2830.

Raju, D., & Schumacker, R. (2015). Exploring student characteristics of retention that lead to graduation in higher education using data mining models. *Journal of College Student Retention: Research, Theory & Practice*, 16(4), 563-591.

Tinto, V. (2006). Research and practice of student retention: What next?. *Journal of College Student Retention: Research, Theory & Practice*, *8*(1), 1-19.