A Decision-Support System for Student Retention Based on Machine Learning

J.D Jayaraman
Associate Professor
Department of Finance
New Jersey City University

Julian Garcia
Research Analyst
Institutional Effectiveness
New Jersey City University

Sue Gerber
Assistant Vice President
Institutional Effectiveness
New Jersey City University

NJCU
NEW JERSEY CITY UNIVERSITY

Introduction

Undergraduate student dropout is a major challenge at American universities with the average 6 year graduation rate at around 59% (NCES, 2017). The National Center for Educational Statistics pegs median income levels for young adults with a bachelor's degree at 64% higher than those with only a high school diploma . Thus, dropping out of college not only impacts the financial well-being of individuals but the economy as a whole. Low retention rates also adversely affect the reputation of the educational institution and could lead to potential loss of funding and inability to compete for quality students. Thus, improving student retention is of paramount importance at institutions of higher education.

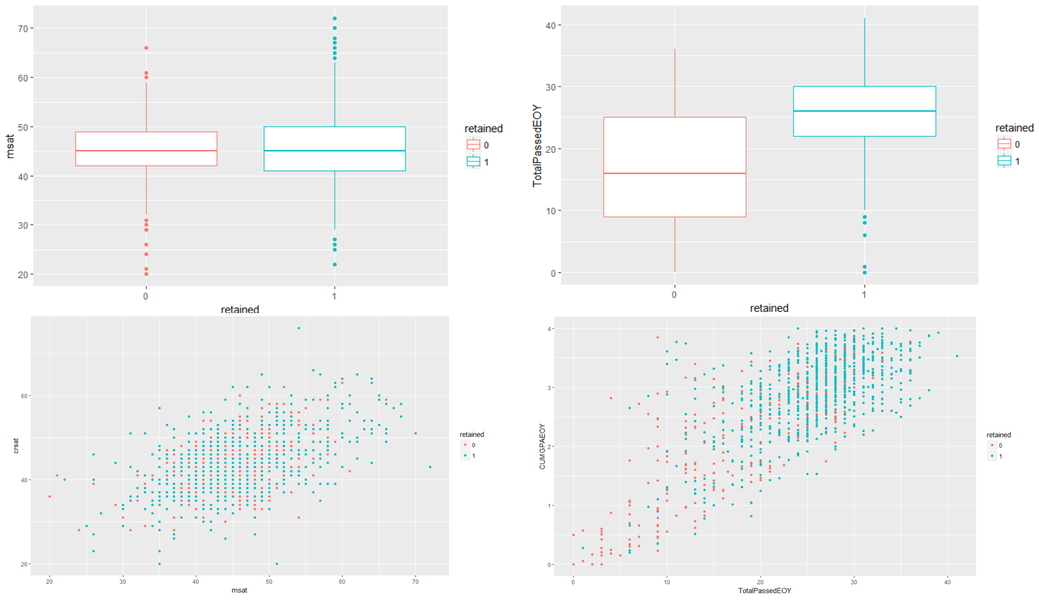
A crucial factor in increasing student retention is the ability to accurately identify at-risk students in a timely manner, so that relevant interventions can be provided. This requires a detailed understanding of the factors affecting retention rates and the ability to apply sophisticated statistical techniques to accurately predict student dropout. One such sophisticated area of predictive modeling is machine learning. Machine learning techniques have been widely employed in predictive modelling in a wide range of fields with remarkable accuracy. This paper describes a machine learning approach to predicting college student dropout. This study will employ machine learning techniques to build models to predict transfer student and native student dropout using a comprehensive dataset from a Hispanic serving four year university with a large percentage of transfer students and a fairly high dropout rate.

Literature Review

- Theoretical Framework
  - Tinto (1975)
    - Academic difficulty,
    - Adjustment problems
    - Lack of clear academic goals
    - Lack of commitment
    - Inability to integrate with the college community
    - Uncertainty, incongruence an isolation
- Machine learning based approaches
  - Delen (2010)
    - Dataset consisting of 39 variables such as SAT score, high school GPA
    - Compared different techniques and found support vector machines produced 80% accuracy
  - Lauria, et al. (2012)
    - Support Vector Machines Performed better than decision trees

Data Overview

- Six years of entering first-time and transfer cohort data (2011 – 2016)
- 11,241 students in the sample
- 60% female and 40% male
- Hispanic, 22% is Black, 25% is White and 8% is Asian
- 28 features which fell into the following categories: pre-college characteristics, college performance and progress, financial aid, community engagement and demographics.



Feature	Importance
Total Credits Passed EOY	64.60
Cumulative GPA EOY	56.16
Term GPA	40.93
Total acc transfer credits	31.96
Term Credits Passed	29.75
Basic Arithmetic	25.08
Fafsa Need	24.34
Basic Reading	24.14
Math SAT	24.04
Total SAT	23.70
Loans total	23.56
Parent Adjusted Gross Income	23.51
Grant Total	23.49
Housing	22.26
Critical Reading SAT	22.20
Credits Enrolled	20.95
Mother Highest Grade Level	18.84
Basic English	18.19
Basic Algebra	17.94
Father Highest Grade Level	17.41
Ethnicity	16.50
HS Percentile	14.02
Adjusted Gross Income	12.76
Scholarship Total	6.62
Work-Study Total	4.85
Total Test	2.74
Total Transfer Credits	2.42
Planned or Declared Major	1.44

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1
glm (Logistic Regression)	0.81	0.86	0.74	0.85	0.86	0.85
gbm (Boosted Trees)	0.84	0.91	0.74	0.85	0.91	0.88
SVM (radial kernel)	0.84	0.89	0.77	0.87	0.89	0.88
randomForest	0.86	0.92	0.75	0.86	0.92	0.89

Model Overview

Data Partition

- 70% for training 30% for testing (prediction results are based on the testing sample).

Models

- Logistic Regression
- Support Vector Machine (Radial Kernel)
  - a classifier defined by an optimal separating hyperplane that gives the largest separation or margin between classes. The SVM can also be used as a nonlinear classifier by incorporating nonlinear kernel functions.
- Boosted Trees
  - Ensemble of decision trees where boosting is used
  - Trees are grown sequentially and trained using information from previous trees.
- Random Forest
  - Ensemble of decision trees where bagging is used with a slight tweak that decorrelates data and reduces variance.
  - As opposed to boosting, bagging continually trains resampled datasets then averages the results.
- Imbalance Correction
  - With average 1st year retention rates at 70% this created a class imbalance.
  - Synthetic Minority Oversampling Technique (SMOTE) was used to address this.

Findings

- It is evident that pre-college characteristics were poor predictors in comparison to college performance.
- The most important features used in the models were related to student performance. Particularly cumulative GPA and total credits passed at the end of the first year. The box and scatter plots illustrated a comparison of the most important pre-college characteristics, SAT scores.
- While all models performed with over 80% accuracy randomForest models performed the best with 86% accuracy on our testing sample.
- Support Vector Machines produced the most balanced results in terms of sensitivity and specificity but when F1 was calculated using the precision and recall of each model randomForest still yielded the strongest results.
- We are currently working on collecting more features, in particular we plan to integrate unstructured data from advisor notes stored in the EAB system and classify them using sentiment analysis.

Conclusions

- Machine learnings models can accurately predict student dropout
- Practical implication
  - Study demonstrates that educational institutions can use their existing databases that contain routine student data to accurately predict at-risk students.
- Institutions can cost effectively deploy these machine learning models as a decision aid to enrollment management, student advising personnel and faculty