Predicting Student Retention in STEM Majors

Andrew Sage
Dan Nettleton
Cinzia Cervato
Craig Ogilvie

Iowa State University

e-mail: ajsage@iastate.edu

August 10, 2015
Objective: Use information available early in the first semester to identify first-year ISU undergraduate students most likely to leave STEM majors. We seek to . . .

- Estimate the probability of each student leaving STEM
- Identify important predictors of a student leaving STEM
We consider STEM students who stayed at ISU for at least 1 year
- 5,247 students from 2011-12, 2012-13
- 537 (10.3%) left STEM during first year

36 explanatory variables
- High school academic performance (GPA, rank, courses, Regent Admission Index)
- Standardized test scores (ACT®, SAT®)
- MapWorks® survey factors
- First semester ISU courses and learning community participation
- Gender
Random Forest Methodology

- Predictions based on ensembles of decision trees (Breiman, 2001)
- Useful in classification and regression problems
- Handles large number of explanatory variables
- Handles nonlinear relationships and complicated interactions
- Provides measure of variable importance
- Implemented in R packages *randomForest* (Liaw & Wiener), *randomForestSRC* (Ishwaran & Kogalur), *party* (Hothorn et al.)
  - Conditional inference trees in *party* perform best when predictors vary in numbers of values/categories
A Decision Tree

n=number of students in node  
y=proportion leaving STEM
Random Forests

Growing a Forest from Training Data

- Many trees (1,000 in our case)
- Each tree grown from a different bootstrap sample
- Random subset of predictor variables considered for each split
- Trees grown until nodes are homogeneous

Predicting New Cases

- Run new case through each tree in a forest
- Probability of a response class is estimated by the proportion of trees “voting” for that class
Identifying At-Risk Students

- Predict 2012-13 cases using 2011-12 as training data
- Consider 500 students most likely to leave STEM to be at-risk
- 20.2% of at-risk students left STEM, compared to 8.1% of others

![Predictive Performance Graph]

- Classification: At-Risk vs. Not At-Risk
- Proportion Leaving STEM: 0.20 for At-Risk, 0.10 for Not At-Risk
- Marginal distribution of estimated probabilities for 2012-13 students
- Forest grown using 2011-12 data
• Marginal distribution of estimated probabilities for 2012-13 students
• Forest grown using 2011-12 data
Effect of Analytical Skills Self-Assessment

- Marginal distribution of estimated probabilities for 2012-13 students
- Forest grown using 2011-12 data
Assessing Variable Importance

Permutation Importance

1. Make predictions for out-of-bag cases for each tree
2. Compute misclassification rate
3. Randomly permute values for an explanatory variable and re-predict
4. Compute new misclassification rate
5. Large increase in misclassification rate indicates variable is important
Important Variables

Ten Important Predictors

- Regent Admissions Index
- LC Member
- Self-Efficacy*
- Analytical Skills*
- Biology Units
- Sex
- HS GPA
- HS Rank
- Chemistry Units
- MapWorks: Environment

* variable from MapWorks® self-assessment survey.
Comparison with Logistic Regression

- Performance comparable to logistic regression model obtained using backwards selection
Future Work

- Incorporate data from 2013-14, 2014-15
- Additional explanatory variables (ALEKS®, ACT® inventory survey)
- Identify students likely to drop out of ISU altogether
- Ongoing adaptive predictions
- Examine differences between STEM majors


H Ishwaran and U.B. Kogalur. Random Forests for Survival, Regression and Classification (RF-SRC), R package 1.5.4. 2014.

