Technical Report: Analyzing Factors Influencing Student Academic Success and Dropouts in Higher Education

Duc Nguyen and Linh Vu

2024-05-10

Introduction

The transition from secondary to higher education often poses significant challenges for students, leading to academic difficulties and dropout. This phenomenon critically affects students' future prospects and institutions' reputations and financial stability. Our research investigates the factors influencing student outcomes in higher education, utilizing statistical learning methods to predict student outcomes based on variables such as demographic, socio-economic, and academic information. Specifically, we asked: "What factors are the most predictive of student graduate outcomes in higher education?"

Previous studies primarily focus on predicting academic success using statistical methods. Beaulac and Roosenthal use random forests to predict whether students will complete their program based solely on course grades in college, with 78.84% accuracy.[1] In addition, Martins et al., whose dataset we will use, achieved a higher accuracy rate for boosting algorithms than random forests, but were yet to incorporate student performance in their first semesters.[2] Our research will take a step further by adding in early college performance predictors and significantly more observations into our analysis.

Methods

Dataset Description

The data set of this study was sourced from the UC Irvine Machine Learning Repository, containing records from the Polytechnic Institute of Portalegre, Portugal. The dataset spans the school years 2008-09 and 2018-19, consisting of 4424 student records across various undergraduate degrees. The data features 36 predictors and a categorical response variable indicating student outcomes: `dropout`, `enrolled`, and `graduate`.

Data Processing

The data was loaded into RStudio using UCI's URL. Initial data wrangling involved checking for missing values (none found) and recoding categorical variables from numerical to factor types. Further modifications included collapsing the levels of complex categorical variables like parental qualifications and occupations into more general categories for analytical and modeling purposes.

Exploratory Data Analysis

Summary Statistics and Data Visualizations

[1] Cédric Beaulac and Jeffrey S. Rosenthal, "Predicting University Students' Academic Success and Major Using Random Forests," in *Research in Higher Education* 60, no.7 (2019): 1048-64, doi.org/10.1007/s11162-019-09546-y.

[2] Mónica V. Martins et al., "Early Prediction of student's Performance in Higher Education: A Case Study," in *Trends and Applications in Information Systems and Technologies* 1 (2021): 166-75, doi.org/10.1007/978-3-030-72657-7_16.

Key variables were selected based on their potential influence on student outcomes, focusing on factors like parental background, previous academic achievements, and early college performance. Data visualizations included histograms to explore distributions and boxplots to examine relationships between predictors and student outcomes. A correlation analysis was performed to check for multicollinearity, ultimately leading to the combination of variables on early college performance, from semesterly into yearly, in order to reduce redundancy and enhance model interpretability.

Statistical Methods and Model Diagnostics

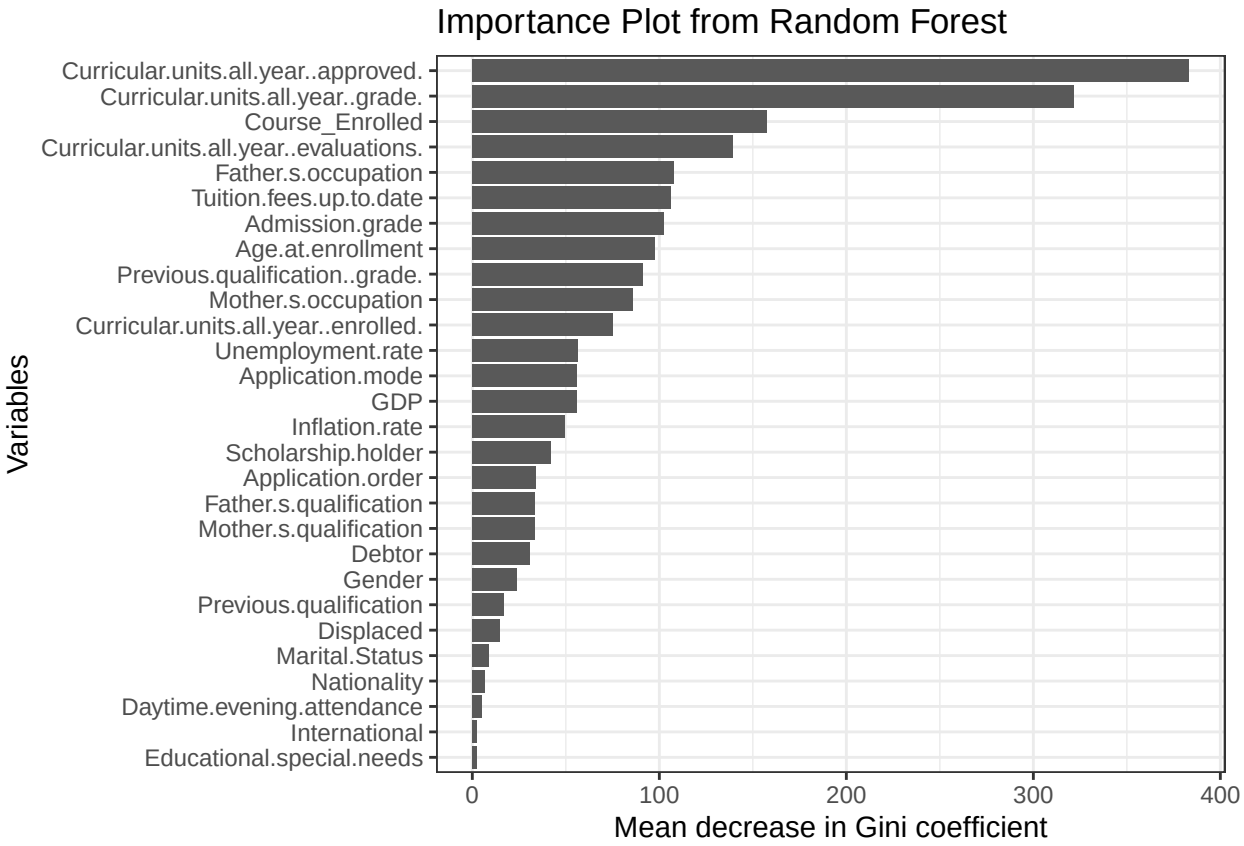
The exploratory analysis utilized forward selection to identify significant predictors. This phase included testing potential variables against the response variable using both visual and statistical methods to ensure only the most relevant predictors were retained. Selected variables were then used to construct a baseline model, planning to employ tree-based classification methods for their robustness in handling non-linear relationships and interactions among predictors.

Results

For the purpose of our study, we will use two classification models that are able to show the relative predictive power of the independent variables: random forests and boosted trees. Random forest is a tree-based model using ensemble methods. The random forest consists of many trees that are constructed using a random sample of predictors at each split. Random forest is suitable because it automatically performs feature selection. Another advantage of it is that the `randomForest` package allows us to glimpse into the importance of the predictors in our classification. Below is the importance plot of a random forest model with 100 trees and 5 predictors considered at each split.

```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:randomForest':
##
##     margin
```



As can be seen, the variables relating to current academic standing and course enrolled in college are the most significant in assessing a student's chance of dropping out.

We will attempt boosted trees during the weekend using `XGBoost`. Boosted trees are another tree-based ensemble method but with learning feature. That is, each tree is built with the consideration of the previous tree's error. Introducing learning rate into the model has the potential to increase our accuracy rate significantly. Besides, boosted trees can also give us information on the importance of predictors.