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 (73%) than random forests (72%), 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 dataset 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.

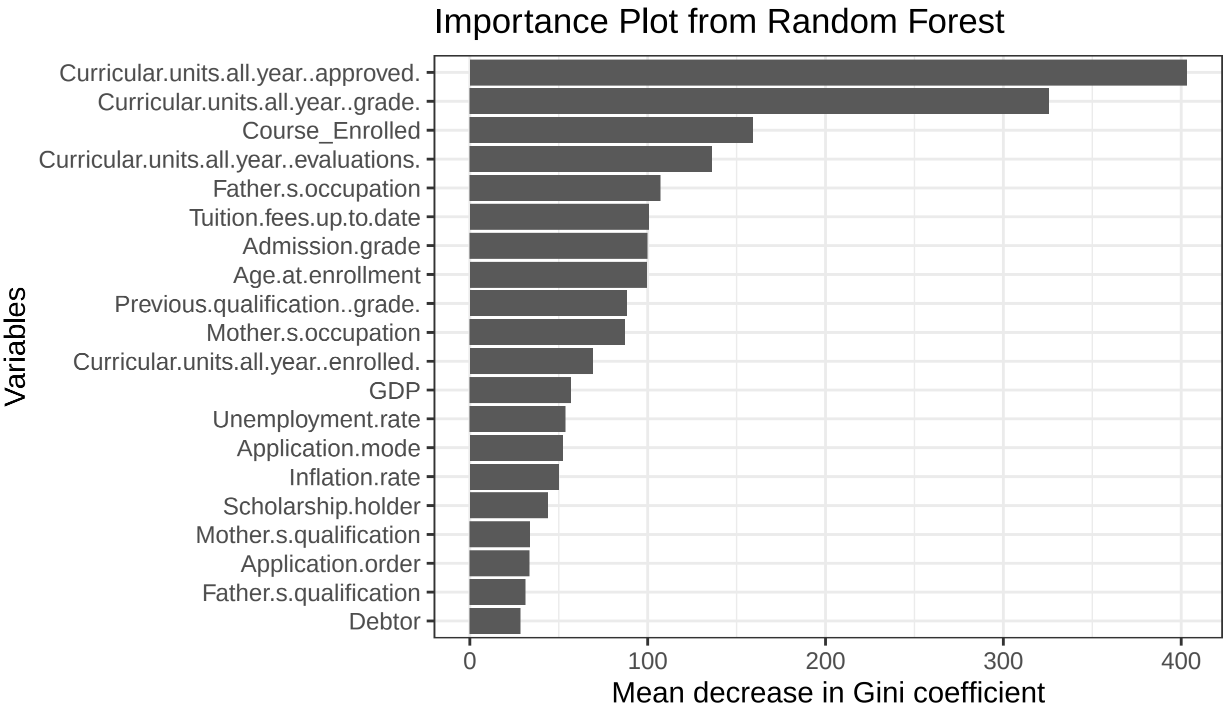
Key variables were selected based on their potential influence on student outcomes, focusing on factors like parental background, previous academic achievements, and early college performance. Data visualizations included histograms to explore distributions and boxplots to examine relationships between 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

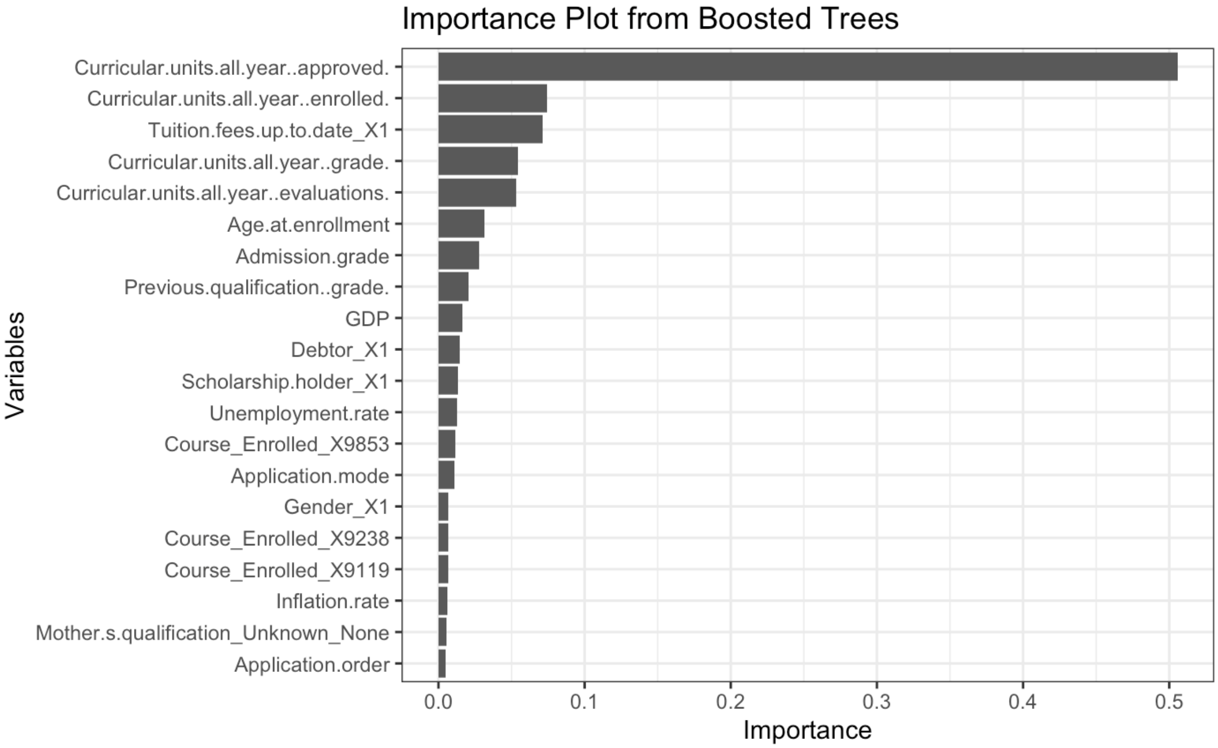
For the purpose of our study, we will use two classification models that are able to show the relative predictive power of the independent variables: random forests and boosted trees. Random forest is a tree-based model using ensemble methods. The random forest consists of many trees that are constructed 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 250 trees and 5 predictors considered at each split.



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. Specifically, the average number of curricular credits approved per semester in the first year was the most predictive, followed by student's average grade in the first year and the course they enrolled (i.e., the major/concentration at the Portuguese institution). Other variables in the top 10 most important include father's and mother's occupation, whether the tuition fees were paid up to date, and admission grade. For reference, the cross-validated accuracy rate of this random forest is 77.48%, which is roughly equal to the reviewed literature's findings, which ranges between 70% and 80%. That said, we are primarily concerned with the variable importance.

```
## [1] 0.7747855
```

We then performed boosted trees using the `tidymodels` framework and `xgboost` engine. Boosted trees are another tree-based ensemble method but with learning feature. That is, each tree is built in the consideration of the previous tree's error. Introducing learning rate into the model has the potential to increase our accuracy rate significantly. Moreover, boosted trees can also give us information on the importance of predictors. Below is the importance plot of a boosting model with 250 trees, a tree depth of 5, a minimum node size of 10 observations, and learning rate of 0.4, which are our best parameters. The plot was generated using the `vip` package, from a separate .Rmd file. Boosting algorithms are computationally intensive, so we used Grinnell's computer instead of the server.



Note that the measurement unit of variable importance is not mean decrease in Gini index, but rather the Shapley value due our code `vip()`.[3] As shown above, 4 out of 5 most important predictors are academic data in the first year of college, with the average number of credits approved per semester stood out remarkably. There is a significant similarity between random forest and boosted trees in terms of the 10 best variables (7 matches out of 10). However, boosted trees enlisted whether the student's family was in debt (`Debtor`) and previous qualification grade as important. For reference, the cross-validated accuracy rate is 77.82%, which is minimally higher than that of our random forest model.

Discussion

A review of the results generated from the model and synthesis with the context from which the data was generated or observed, a restatement of research objective and an answer to the original research question, a

[3] The Comprehensive R Archive Network, *vip: Variable Importance Plots*, https://cran.r-project.org/web/packages/vip/index.html

discussion limitations of the study as well as areas for further research.

Objective: investigate the importance of predictors –> Our research is an extension to Martins et al. (2021), where we confirm that the addition of initial college performance is crucial to identifying potential dropouts, based on the importance plot. In addition, our boosted trees accuracy at roughly 78% is higher than the authors' boosting model without college performance (73%), further proving the usefulness of early performance.

Contribution: we have identified those variables as highly predictive of student's chance of dropping out. We recommend early intervention accordingly.

Limitations: Random forests and boosted trees are intensive models. We might not have found the best parameters for either of our models (i.e., our models might not be optimal).

Another challenge is class imbalance, in which the number of records is widely disparate among the classes. This is often the case in our context because it is likely that the minority of students underperform and drop out. Class imbalance may lead to a higher error in identifying dropouts, which is contrary to our research' implication for early academic intervention.

Future research: should look into more types of data about student's first year, not just about studies but perhaps social life. Also, might want to replicate our findings at a different instituion.

Code Appendix

Bibliography

Beaulac, Cédric, 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. https://doi.org/10.1007/s11162-019-09546-y.

Martins, Mónica V., Daniel Tolledo, Jorge Machado, Luís M. Baptista, and Valentim Realinho. "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. https://doi.org/10.1007/978-3-030-72657-7_16.

The Comprehensive R Archive Network. *vip: Variable Importance Plots* (n.d.). https://cran.r-project.org/web/packages/vip/index.html