Technical Report

Duc Nguyen and Linh Vu

2024-05-10

Introduction

ipsum loren

Methods

ipsum loren

Exploratory Data Analysis

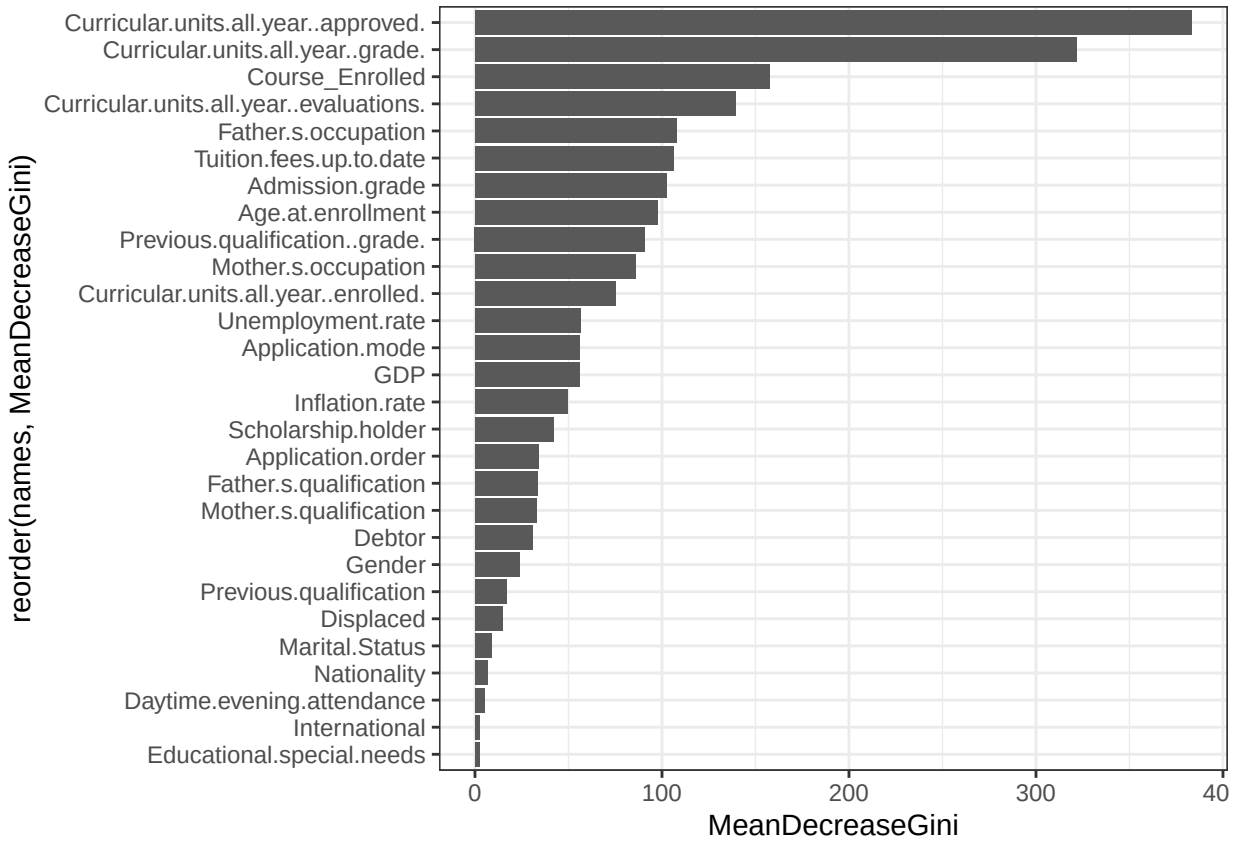
ipsum loren

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.