Analyzing Factors Influencing Academic Success and Dropouts in Higher Education

Duc Nguyen Linh Vu

Grinnell College

May 16, 2024

Table of Contents

Table of Contents

Motivation

Transitioning from secondary to higher education poses a significant challenge to students

Motivation

Transitioning from secondary to higher education poses a significant challenge to students

- potentially leading to academic failure and dropouts

Motivation

Transitioning from secondary to higher education poses a significant challenge to students

- potentially leading to academic failure and dropouts
- negatively affecting students, the school, and parents

Research question

What factors are the most predictive of student graduate outcomes in higher education?

Literature review

- Beaulac & Rosenthal (2019) predict whether students will complete their program based on college course grades with random forests

Literature review

- Beaulac & Rosenthal (2019) predict whether students will complete their program based on college course grades with random forests
  Accuracy of 78.84%

Literature review

- Beaulac & Rosenthal (2019) predict whether students will complete their program based on college course grades with random forests
  Accuracy of 78.84%
- Martins et al. (2021) predict student dropouts based on data collected at enrollment

Literature review

- Beaulac & Rosenthal (2019) predict whether students will complete their program based on college course grades with random forests
  Accuracy of 78.84%
- Martins et al. (2021) predict student dropouts based on data collected at enrollment
  Accuracy for boosting (73%) is higher than for random forests (72%)

Table of Contents

Data set

From UC Irvine Machine Learning Repository

Data set

From UC Irvine Machine Learning Repository

- collected and donated by Martins et al.

Data set

From UC Irvine Machine Learning Repository

- collected and donated by Martins et al.
- with 4424 observations, 36 predictors, 1 response variable

Data set

From UC Irvine Machine Learning Repository

- collected and donated by Martins et al.
- with 4424 observations, 36 predictors, 1 response variable

```{r}
colnames(student_data)
```

```
 [1] "Marital Status"                                 "Application mode"
 [3] "Application order"                              "Course_Enrolled"
 [5] "Daytime/evening attendance"                     "Previous qualification"
 [7] "Previous qualification (grade)"                 "Nationality"
 [9] "Mother's qualification"                         "Father's qualification"
[11] "Mother's occupation"                            "Father's occupation"
[13] "Admission grade"                                "Displaced"
[15] "Educational special needs"                      "Debtor"
[17] "Tuition fees up to date"                        "Gender"
[19] "Scholarship holder"                             "Age at enrollment"
[21] "International"                                  "Curricular units 1st sem (credited)"
[23] "Curricular units 1st sem (enrolled)"            "Curricular units 1st sem (evaluations)"
[25] "Curricular units 1st sem (approved)"            "Curricular units 1st sem (grade)"
[27] "Curricular units 1st sem (without evaluations)" "Curricular units 2nd sem (credited)"
[29] "Curricular units 2nd sem (enrolled)"            "Curricular units 2nd sem (evaluations)"
[31] "Curricular units 2nd sem (approved)"            "Curricular units 2nd sem (grade)"
[33] "Curricular units 2nd sem (without evaluations)" "Unemployment rate"
[35] "Inflation rate"                                 "GDP"
[37] "Target"
```

Data processing

Our data preparation steps include:

- Checking for missing values

Data processing

Our data preparation steps include:

- Checking for missing values
- Converting categorical variables to factor type

Data processing

Our data preparation steps include:

- Checking for missing values
- Converting categorical variables to factor type
- Simplifying complex categorical variables into fewer levels

Table of Contents

Goal

- Familiarize with the data

Goal

- Familiarize with the data
- Visualize relationships between features and target

Goal

- Familiarize with the data
- Visualize relationships between features and target
- Potentially form assumptions about the strongest predictors

Method

- Data visualization:
    - Histograms
    - Boxplots, Barplots

Method

- Data visualization:
  - Histograms
  - Boxplots, Barplots
- Statistical Methods:
  - Five-number summaries
  - Correlation matrix

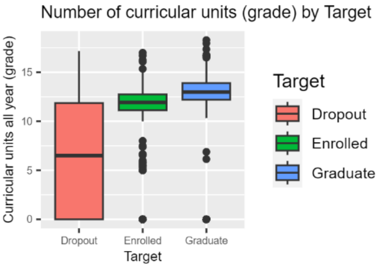
Results

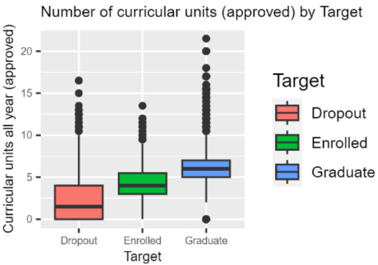
- Combine early college performance metrics from semester-based to yearly

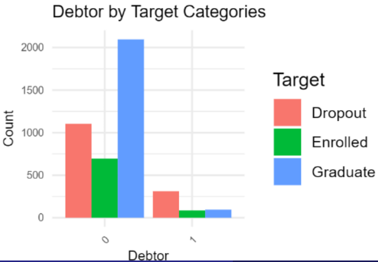
Results

- Combine early college performance metrics from semester-based to yearly
- Identify some potential strong predictors among the features

Results (cont.)







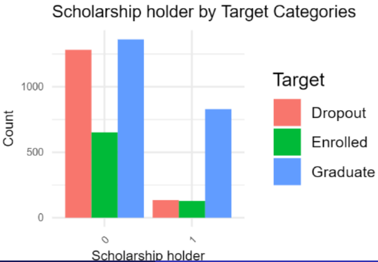


Table of Contents

Model building

- Our dependent variable is a three-class categorical variable

Model building

- Our dependent variable is a three-class categorical variable
  - `Dropout`, `Enrolled`, and `Graduate`
- We are also concerned with the task of inference

Model building

- Our dependent variable is a three-class categorical variable
  - `Dropout`, `Enrolled`, and `Graduate`
- We are also concerned with the task of inference
- We will use two classification models:

Model building

- Our dependent variable is a three-class categorical variable
  - `Dropout`, `Enrolled`, and `Graduate`
- We are also concerned with the task of inference
- We will use two classification models:
  - random forests
  - boosted trees

Random forest

- We use `randomForest` package

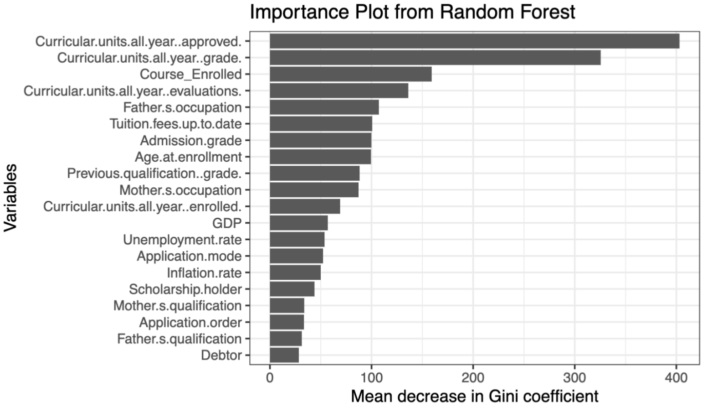
```
```{r}
library(randomForest)
set.seed(2024)
rf_mod <- randomForest(Target ~., train_student, ntree = 250, mtry = 5)
rf_mod
```
```

```
Call:
 randomForest(formula = Target ~ ., data = train_student, ntree = 250,      mtry = 5)
               Type of random forest: classification
                     Number of trees: 250
No. of variables tried at each split: 5

        OOB estimate of  error rate: 22.69%
Confusion matrix:
         Dropout Enrolled Graduate class.error
Dropout      872       82      178  0.22968198
Enrolled     152      202      270  0.67628205
Graduate      59       55     1638  0.06506849
```

Random forest (cont.)

- `ntree = 250, mtry = 5`



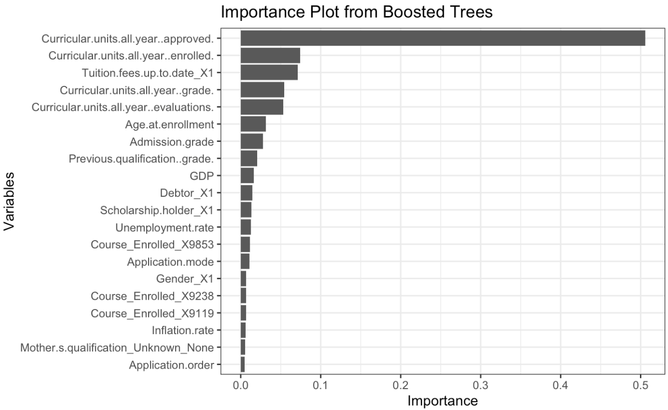
- `accuracy = 77.48%`

Boosted trees

- We use `tidymodels` package
- set engine "`xgboost`"

Boosted trees (cont.)

- `trees = 250, min_n = 10, tree_depth = 5, learn_rate = 0.4`



- `accuracy = 77.82%`

Table of Contents

Conclusions

- Strongest predictors of college dropouts: first-year performance metrics (the number of curricular units approved, enrolled, grades, and evaluations), tuition fee payment status, age at enrollment.

Conclusions

- Strongest predictors of college dropouts: first-year performance metrics (the number of curricular units approved, enrolled, grades, and evaluations), tuition fee payment status, age at enrollment.
- Significantly strong: the average number of credits approved per semester

Conclusions

- Strongest predictors of college dropouts: first-year performance metrics (the number of curricular units approved, enrolled, grades, and evaluations), tuition fee payment status, age at enrollment.
- Significantly strong: the average number of credits approved per semester
  $\rightarrow$ Build upon Martins et al. (2021), confirming the importance of first-year college performance metrics

Conclusions

- Strongest predictors of college dropouts: first-year performance metrics (the number of curricular units approved, enrolled, grades, and evaluations), tuition fee payment status, age at enrollment.
- Significantly strong: the average number of credits approved per semester
  → Build upon Martins et al. (2021), confirming the importance of first-year college performance metrics
  → Our accuracy (78%), surpasses those of the authors' models (73%)

Limitations

- Random Forest and Boosted Trees are computationally intensive

Limitations

- Random Forest and Boosted Trees are computationally intensive
- Potentially non-optimal parameter settings

Limitations

- Random Forest and Boosted Trees are computationally intensive
- Potentially non-optimal parameter settings
- Limited data which affects generalizability

Limitations

- Random Forest and Boosted Trees are computationally intensive
- Potentially non-optimal parameter settings
- Limited data which affects generalizability
- Class imbalance for the response variable, which might lead to higher error

Areas for Further Research

- Inclusion of aspects related to social life on campus

Areas for Further Research

- Inclusion of aspects related to social life on campus
- Replicate the study at various institutions to enhance representativeness

References

Beaulac, C., & Rosenthal, J. S. (2019). Predicting University Students' Academic Success and Major Using Random Forests. In *Research in Higher Education*, *60*(7), 1048-1064. https://doi.org/10.1007/s11162-019-09546-y

Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M., & Realinho, V. (2021). Early Prediction of Student's Performance in Higher Education: A Case Study. In *Trends and Applications in Information Systems and Technologies*, *1*, 166-175. https://doi.org/10.1007/978-3-030-72657-7$_1$6