Important Variables in Predicting grade Outcomes in intro courses

The Data:

The data consists of freshman admitted between Fall 2016 and Fall 2019, who are California residents. Some variables we are interested in are if they took AP tests for:

* Biology
* Calculus
* Chemistry
* Stats
* Physics
* Comp Sci

The introductory courses I picked to study are as follow:

* MAT\_021A,
* MAT\_021B
* BIS\_002A,
* BIS\_002B,
* CHE\_002A,
* CHE\_002B,
* MAT\_016A,
* MAT\_016B
* ECN\_001A,
* HIS\_017A,
* STA\_013,
* PSC\_001,
* PHY\_007A,
* ANT 2
* ANS 1

I chose basic demographics for controls in addition to creating dummy variables for AP scores, and ethnicity.

MODELS:

Running a fancy function to compute a bunch of different models, I found that the random forest models had the lowest error, and highest R2 value, although linear regression and gradient boosted algorithms performed well.

WHAT I FOUND: [NOTE: I need to check accuracy using some sort of metric…]

There were to facets of this analysis that stuck out:

1. I found that the models predicted the grades in some classes better than others. For example, the test data of ANT 2 only had a R2 value of .10 (training data dim was 2499), whereas the model for ANS 1 (dim = 791) had an R2 value of 0.45.
2. The “important variable” varied depending on the type of class. In particular, there seems to be a divide between STEM and non-STEM classes, and the different disciplines. For example, it seems GPA and APPASSED is more important for non-STEM classes, and for STEM classes either the AP test in the subject or SAT scores are more important.
3. The models for the non-STEM classes seemed less predictive with lower R2 scores. Note: I should check the OOB scores, since they might be a better indicator of the model performance.

Analysis of individual courses…

MAT\_021A

OOB score for random forest: 0.29

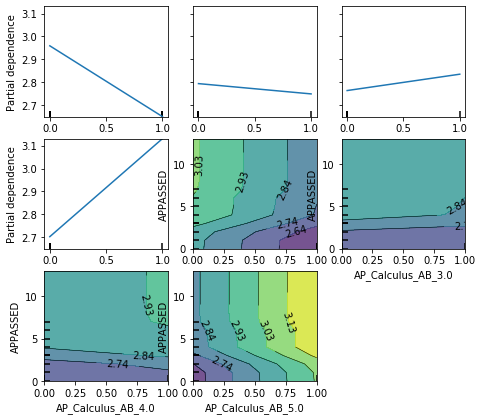
R2 score for test data: 0.288

Important variables (using permutation importance):

|  |  |
| --- | --- |
| 0.1032 ± 0.0167 | SAT\_TOTAL\_OFFICIAL\_SQ |
| 0.0882 ± 0.0124 | AP\_Calculus\_AB\_0.0 |
| 0.0688 ± 0.0045 | GPA |
| 0.0630 ± 0.0102 | AP\_Calculus\_AB\_5.0 |
| 0.0457 ± 0.0046 | APPASSED |
| 0.0271 ± 0.0038 | LOWINCOME |
| 0.0256 ± 0.0017 | AP\_Calculus\_BC\_0.0 |
| 0.0238 ± 0.0014 | ATOG |
| 0.0220 ± 0.0016 | LOW\_INCOME |
| 0.0182 ± 0.0014 | ATOGE |
| 0.0152 ± 0.0014 | FIRST\_GENERATION |
| 0.0152 ± 0.0032 | TRIPLE |
| 0.0139 ± 0.0009 | ATOGG |
| 0.0118 ± 0.0008 | ATOGA |
| 0.0111 ± 0.0019 | AP\_Calculus\_BC\_5.0 |
| 0.0106 ± 0.0021 | AP\_Chemistry\_0.0 |
| 0.0100 ± 0.0010 | ATOGF |
| 0.0096 ± 0.0008 | ETH\_MX |
| 0.0092 ± 0.0034 | AP\_Calculus\_AB\_4.0 |
| 0.0088 ± 0.0009 | AP\_Calculus\_AB\_3.0 |

Partial Dependency Plots:

Reading from left-to-right AP\_CALC\_AB\_not\_taken, AP\_CALC\_AB\_3.0, AP\_CALC\_AB\_4.0, AP\_CALC\_AB\_5.0,



MAT\_021B

BIS\_002A,

BIS\_002B,

CHE\_002A,

CHE\_002B,

MAT\_016A,

MAT\_016B

ECN\_001A,

HIS\_017A,

STA\_013,

PSC\_001,

PHY\_007A,

ANT 2

ANS 1