StudentPerformance

gvant

2023-06-25

This dataset was downloaded from Kaggle, and gives total reading, writing, and math scores for standardized testing of high school students, as well as the demographic information of each student. In this project, I break down and analyze more closely some of the factors which contribute to success on standardized testing, and also fit different models and compare the results. I also split the model into an 80/20 testing/training subset.

```
## [1] 1000
```

```
##   gender race.ethnicity parental.level.of.education        lunch
## 1 female        group B           bachelor's degree     standard
## 2 female        group C                some college     standard
## 3 female        group B             master's degree     standard
## 4   male        group A          associate's degree free/reduced
## 5   male        group C                some college     standard
## 6 female        group B          associate's degree     standard
##   test.preparation.course math.score reading.score writing.score agg
## 1                    none         72            72            74 218
## 2               completed         69            90            88 247
## 3                    none         90            95            93 278
## 4                    none         47            57            44 148
## 5                    none         76            78            75 229
## 6                    none         71            83            78 232
```

The above represents a small sample of the data set and that it contains 1000 rows.

Random Forest Regression

We are starting off with a simple random forest regression since it takes no underlying assumptions about the data or its scale. As such, it serves as a jumping off point to compare other analytical methods.

```
rf.reg <- randomForest(agg ~ gender + race.ethnicity +
                         parental.level.of.education + lunch +
                         test.preparation.course, data=train,
                       ntree=150, maxnodes=30)

# most important factors
importance(rf.reg, type=2)
```

```
##                              IncNodePurity
```

```
## gender                          26293.74
## race.ethnicity                  39172.17
## parental.level.of.education     35758.43
## lunch                           65689.44
## test.preparation.course         65128.22
```

The "IncNodePurity" column is essentially a Gini coefficient statistic which is a rough measure of variable importance. The higher the value, the more that variable contributes to the overall results.

The test preparation course is the second most important variable according to this model, which lines up with intuition. The most important variable, however, is the "standard / reduced" lunch option. This could be a rough indicator of poverty, but it is not known what the criteria is to qualify for reduced lunch.

We now assess the accuracy of the model using the "test" data set.

The random forest model predicted the majority of scores with a 10% margin of error. This is a decent result, but also indicates that the model may need more relevant variables in order to make a proper prediction.

Linear Regression

We now fit a linear regression using dummy variables for the categories. Linear regression works here because even though the variable importance has been assessed, it would be beneficial to understand the numerical effect of each variable on the total score.

```
##     gender         race                      ped               lunch
##  female:518   group A: 89   associate's degree:222   free/reduced:355
##  male  :482   group B:190   bachelor's degree :118   standard    :645
##               group C:319   high school       :196
##               group D:262   master's degree   : 59
##               group E:140   some college      :226
##                             some high school  :179
##         prep          math           reading          writing
##  completed:358   Min.   :  0.00   Min.   : 17.00   Min.   : 10.00
##  none     :642   1st Qu.: 57.00   1st Qu.: 59.00   1st Qu.: 57.75
##                  Median : 66.00   Median : 70.00   Median : 69.00
##                  Mean   : 66.09   Mean   : 69.17   Mean   : 68.05
##                  3rd Qu.: 77.00   3rd Qu.: 79.00   3rd Qu.: 79.00
##                  Max.   :100.00   Max.   :100.00   Max.   :100.00
##      total
##  Min.   : 27.0
##  1st Qu.:175.0
##  Median :205.0
##  Mean   :203.3
##  3rd Qu.:233.0
##  Max.   :300.0
```

```
##
## Call:
## lm(formula = total ~ gender.fm + race.B + race.C + race.D + race.E +
##     ped.h + ped.sc + ped.a + ped.b + ped.m + lunch.s + prep.n,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -148.133  -25.208    1.348   28.800   83.962
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  180.530      5.515  32.733  < 2e-16 ***
## gender.fm     13.038      2.715   4.802 1.88e-06 ***
## race.B         3.914      5.436   0.720  0.47170
## race.C         4.995      5.140   0.972  0.33148
## race.D        13.873      5.189   2.674  0.00766 **
## race.E        18.150      5.787   3.136  0.00178 **
## ped.h         -4.882      4.394  -1.111  0.26683
## ped.sc         9.263      4.281   2.164  0.03078 *
## ped.a         11.514      4.305   2.674  0.00765 **
## ped.b         20.389      5.087   4.008 6.70e-05 ***
## ped.m         23.984      6.238   3.845  0.00013 ***
## lunch.s      -23.430      2.802  -8.361 2.79e-16 ***
## prep.n        22.984      2.827   8.129 1.67e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.15 on 791 degrees of freedom
## Multiple R-squared:  0.226,  Adjusted R-squared:  0.2143
## F-statistic: 19.25 on 12 and 791 DF,  p-value: < 2.2e-16
```

We find similar results to the random forest model in terms of variable importance. Having reduced lunch, for instance, has an effect of subtracting 23 points from the total score. There is a significant gendered effect of +13 points with female advantage. In addition, the disparity between children with parents completing "some high school" vs a master's degree is another 24 points. To put this in perspective, taking the preparation course grants students about 23 points.

The combined effect of a master's degree parent, non-reduced lunch, and course preparation gives a total of 70 point increase against a student who has reduced lunch, a parent with "some high school", and who has not taken the preparation course. This is a 1.64 standard devation disparity.

Let's trim the original model using only significant variables.

```
##
## Call:
## lm(formula = total ~ gender.fm + race.D + race.E + ped.sc + ped.a +
##     ped.b + ped.m + lunch.s + prep.n, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -144.329  -25.412    0.936   28.960   85.132
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  181.511      3.150  57.624  < 2e-16 ***
## gender.fm     13.373      2.699   4.955 8.82e-07 ***
## race.D        10.137      3.144   3.224 0.001316 **
## race.E        14.183      4.023   3.525 0.000447 ***
## ped.sc        11.891      3.609   3.295 0.001027 **
## ped.a         14.357      3.617   3.970 7.85e-05 ***
## ped.b         22.987      4.542   5.061 5.19e-07 ***
## ped.m         26.750      5.789   4.621 4.46e-06 ***
```

```
## lunch.s      -23.555      2.799  -8.415  < 2e-16 ***
## prep.n        23.239      2.815   8.255 6.33e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.13 on 794 degrees of freedom
## Multiple R-squared:  0.224,  Adjusted R-squared:  0.2152
## F-statistic: 25.47 on 9 and 794 DF,  p-value: < 2.2e-16
```

The final model gives an adjusted R2 statistic of 21.5%, suggesting that these factors explain only a small chunk of the total variance in students' scores.

Accuracy Comparison

Table 1: Random Forest

accuracy05	accuracy15	accuracy25
0.26	0.612	0.837

Table 2: Linear Model

accuracy05	accuracy15	accuracy25
0.25	0.663	0.857

The two models performed similarly, with the linear model slightly edging out the random forest model.