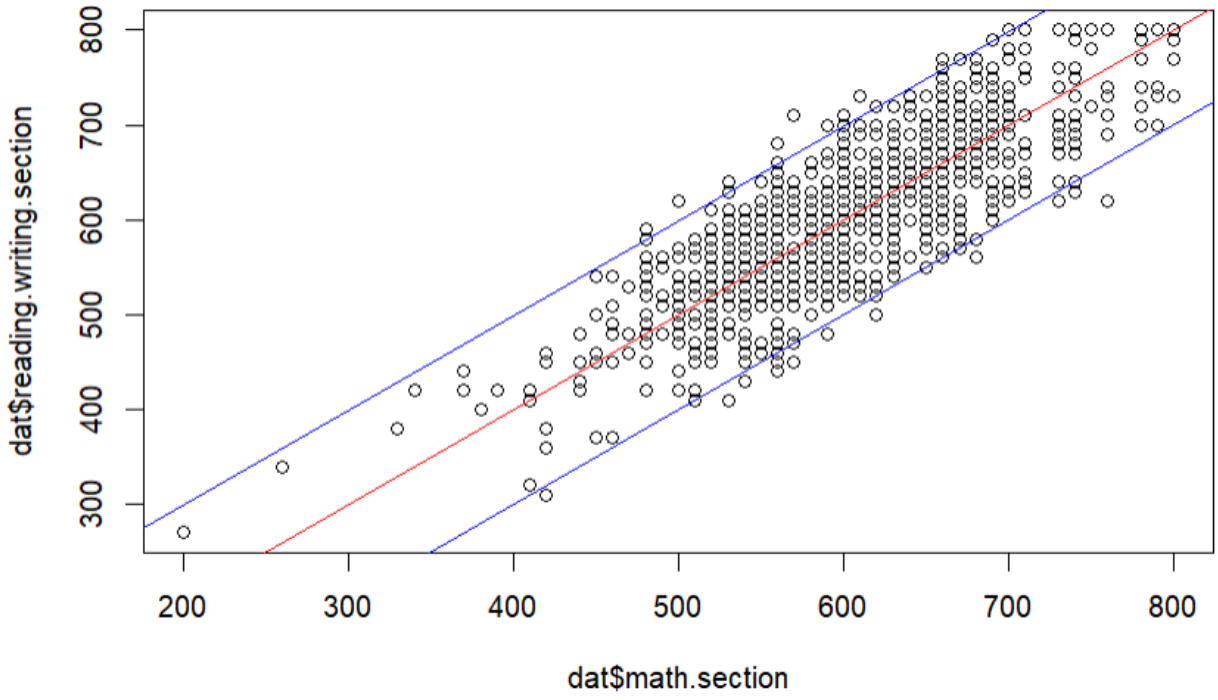
Abstract and Problem Statement

The SAT standardized testing system is an important metric that colleges use to gauge the reading, writing, and mathematics skills of prospective students. The SAT was invented in the 1920's and was first used in 1933 by the Harvard University to determine scholarship recipients and over time has been popularized, updated and transformed to its final form today[2]. Today, almost every student who seeks higher education must take the SAT or another standardized test to be admitted. When a student takes the exam, they fill out some basic demographic information, and the goal of this project is to define a model that can predict an individual's score based on their demographic information.

Related Work

Some others have attempted to answer questions related to the current question of predicting SAT score, but their attempts to accurately predict scores falls flat of a thorough study. The issue can be shown through this graph, which plots math section scores and reading and writing section scores for every student.



As can be seen by this graph, students tend to perform roughly the same in both sections, with most students scores being within 100 points of one another. Now, when performing predictions on math scores, others left in the highly correlated reading and writing scores, so all they really proved was that students that performed well on the rest of the exam performed well on the math section as well. These conclusions are not very helpful, so in this project, scores will be combined into a final score and then the categorical predictors will be used to predict score.

Methods

The methods section will be broken into two parts. The first will describe the data, how SAT Scoring works, and the transformations needed to get the data ready for analysis. From there, the second part will focus on predicting SAT score based on demographic information.

Variables, SAT Scores, and Data Transformation

Variables

Currently, the dataset contains five predictor variables (race/ethnicity, parental level of education, discounted lunch, test preparation course taken, and sex), and three response variables (math percentage, reading score percentage, writing score percentage) laid out in the following way:

- Race/ethnicity: Group A, B, C, D, or E
- Parental Level of Education: some high school, high school, some college, associate's degree, bachelor's degree, master's degree
- Lunch: Standard, free/reduced
- Test preparation completed: none, completed
- Sex: M, F

The three outcome variables, math percentage, reading score percentage, writing score percentage are all decimals from 0 to 1. To get a final score, an operation must be performed on the data.

SAT Scores

The SAT gives students a score from 400 to 1600, and that score is broken into two sections, with 800 points being determined by the math score and the other 800 from a combined reading and writing score. To obtain the reading and writing test score, then number of correct questions is scaled to be out of 40 for both reading and writing; then, round to integers and add the scores and multiply by 10 to get a section score out of 800[2]. Now that we have a math section score and a reading and writing section score, add the scores to get a total score. Using the example grading scale, the data will now be transformed.

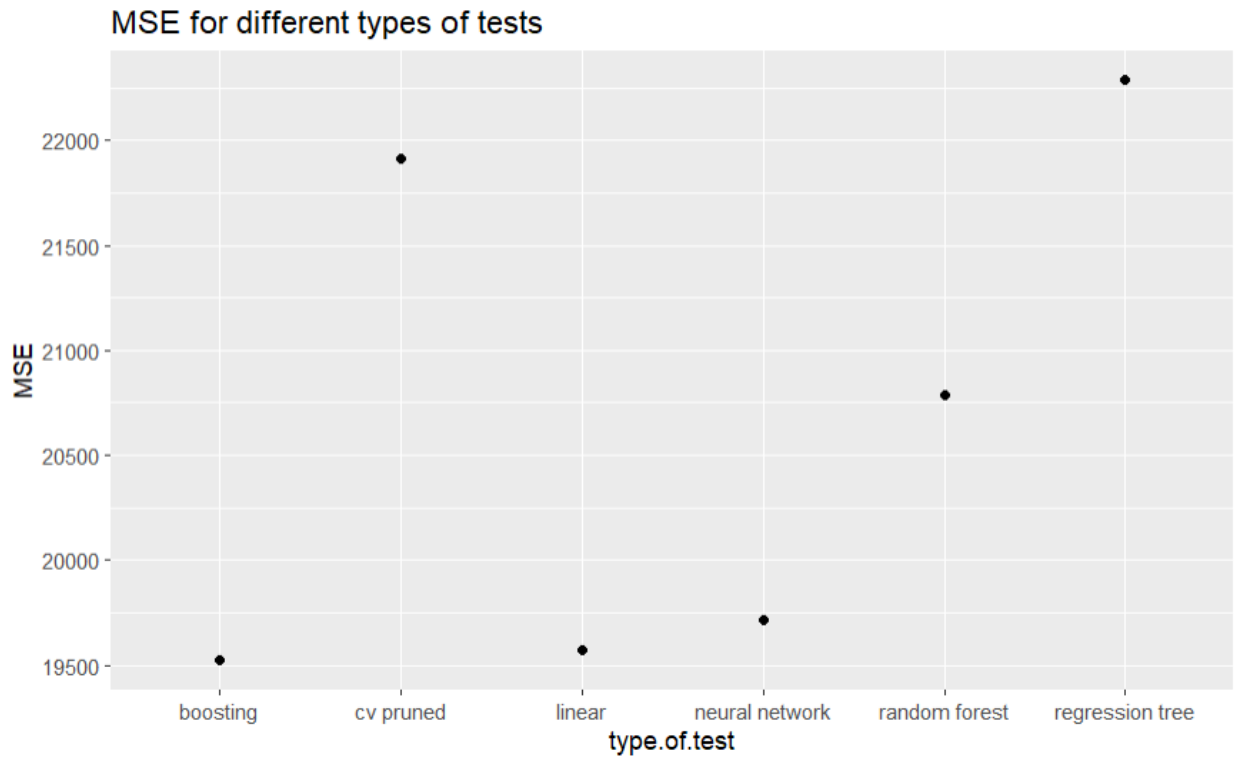
Data Transformation

Now to transform the data. Using the percentage of correct answers, the percentages were scaled to raw scores. Then, using the raw scores and the accompanying conversion table[2], total scores were calculated and assigned to each student.

Prediction

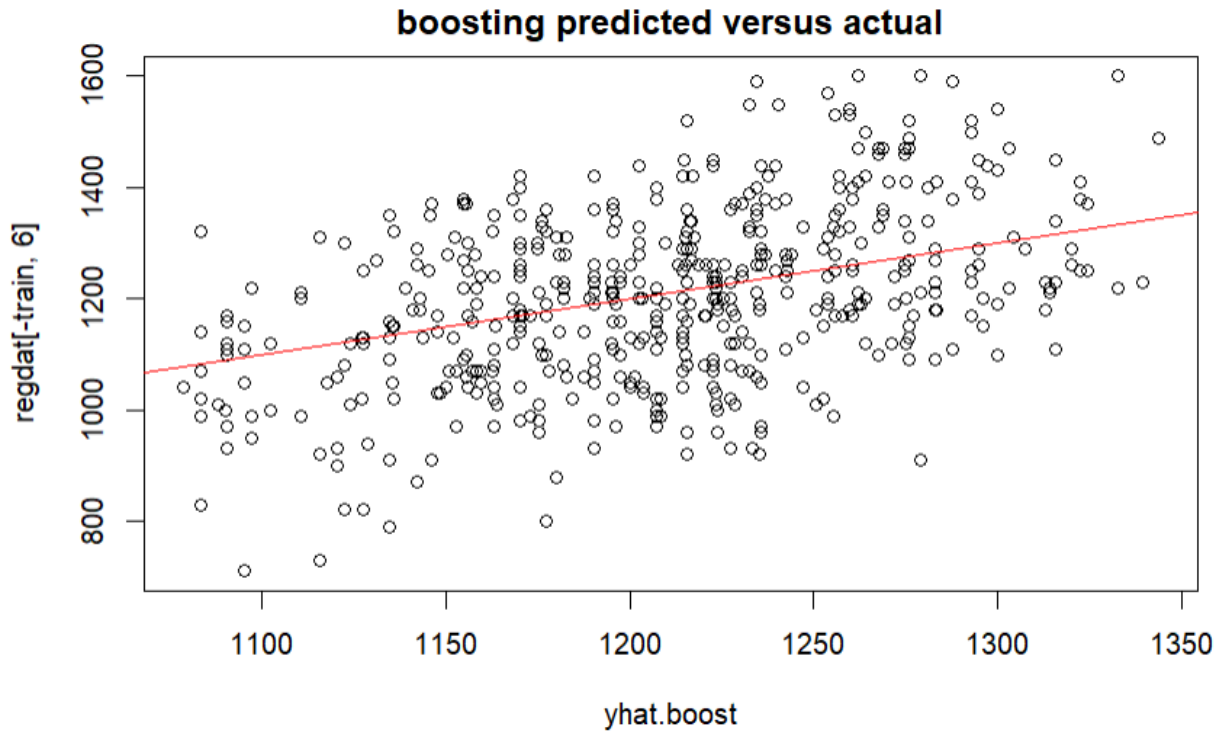
Due to the largely categorical nature of predictor values, tree methods are the best option to use when predicting the score of students. To find the best fit, the following methods were tested: regression tree, pruned regression tree with cross validation, random forest, boosting, linear regression, and a neural network, the python sklearn.neural_network MLPRegressor. All testing error was measured using the same train and test set, with half of the data being training and the other half being testing. Overall, the final testing accuracy for each method is listed in the graphic below. Testing accuracy is the metric of choice to measure model effectiveness.

Results

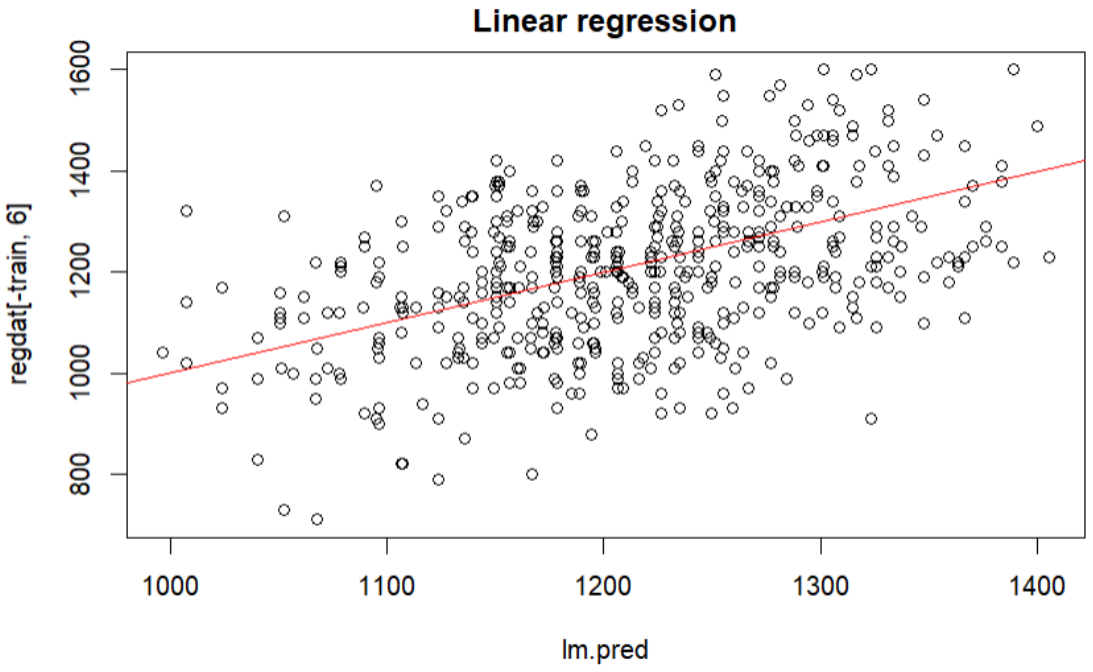


type.of.test <chr>	MSE <dbl>
regression tree	22288.30
cv pruned	21912.34
random forest	20785.44
boosting	19522.30
linear	19573.54
neural network	19718.50

As seen above, boosting performed the best, with the linear model and the neural network close behind. The fact that the linear model performed very well is surprising due to the lack of non-factor variables. The reported testing mean squared error for the boosting model is 22288.3, with the following testing plot, plotting predicted versus actual, the red line is a perfect prediction.



Overall, this prediction does a fairly descent job, but has some major shortcomings. For starters, the model only predicts values from around 1100 to 1350, while the actual range of SAT scores ranges from around 800 to 1600. Similar to boosting, linear regression had a low testing mean squared error value, but the same difficulty with predicting performance as boosting. Linear regression had a testing MSE of 19573.5, with the following accuracy plot:



This was the output of the linear regression model:

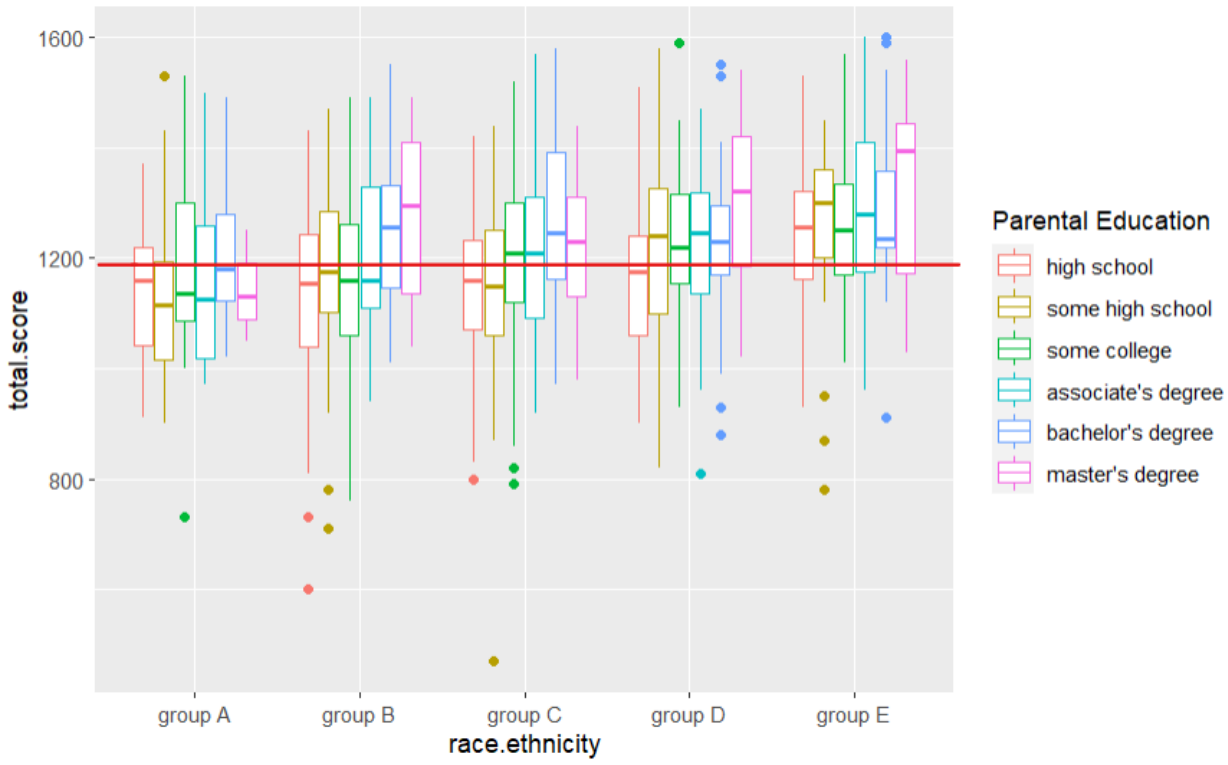
```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             1171.61      27.76  42.204  < 2e-16 ***
R.Egroup B                43.72      25.61   1.707 0.088395 .
R.Egroup C                54.26      24.09   2.253 0.024730 *
R.Egroup D                82.41      24.85   3.316 0.000981 ***
R.Egroup E               112.00      26.74   4.189 3.32e-05 ***
P.L.Ebachelor's degree    22.28      22.54   0.988 0.323546
P.L.Ehigh school         -76.44      18.49  -4.134 4.21e-05 ***
P.L.Emaster's degree      16.48      27.96   0.590 0.555701
P.L.Esome college        -20.12      18.30  -1.100 0.271974
P.L.Esome high school    -65.57      19.20  -3.414 0.000693 ***
Lstandard                 99.68      12.55   7.944 1.37e-14 ***
T.P.Cnone                -82.00      12.85  -6.381 4.11e-10 ***
SM                       -16.74      12.35  -1.355 0.175896
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 135.8 on 487 degrees of freedom
Multiple R-squared:  0.2587,    Adjusted R-squared:  0.2404
F-statistic: 14.16 on 12 and 487 DF,  p-value: < 2.2e-16
```

Overall, this model does not perform very well, the R-squared statistic is only about 0.25, not a very good model, mainly due the inability to separate data. This output does also lead to some conclusions that can be drawn about the data most notably:

- Being in racial group D or E most likely helped increase score
- If a student’s parents did not attend college (high school or some high school), their score is likely to be lower
- Being on standard lunch increased score as opposed to being on free / reduced lunch
- Students who did not take the practice course performed worse, those that did performed better

Discussion and Conclusion

This project attempted to predict a student’s SAT score based solely on their race, parental level of education, income level (free / reduced lunch), test prep course, and sex. Looking back, one thing should have neem changed: when working with tuning parameters, the data should have been split into three parts, one for training, one for validating and tuning, and one for testing Without a doubt, this was very difficult to model due to two main reasons: one, test performance was not accessible like previous math score predictions, as students who performed well on the test were likely to perform well on math, and two, the data has large, overlapping sections of scores. Despite there being visible differences in the distribution of scores for every group, the range of scores withing each group made separating them almost impossible. Rendering all combinations is difficult, but the following graphic can sum up the difficulty of separation:



This graph only shows ethnicity and education, but the point is clear. This red line shows that a score of just below 1200 would land in the IQR for all but two combinations of race and parental education. Additionally, one can see that due to the large variance within each group, well performing students in the combinations of factors that should lead to the worst scores can perform as well or better than the worst students whose factors should lead them to have the best scores (maximum of the worst possible factors was a 1350 and the minimum in the best combination of factors was a 1230). Overall, due to the inability to separate data due to its variability, more factors would be needed to accurately predict a student's SAT score based on factor variables about them.

Sources

DATA: Students Performance in Exams | Kaggle

[1] History of the SAT: A Timeline, *PBS*,

https://www.pbs.org/wgbh/pages/frontline/shows/sats/where/timeline.html

[2] Scoring Your SAT Practice Test #1, *College Board*,

https://collegereadiness.collegeboard.org/pdf/scoring-sat-practice-test-1.pdf

[3] scikit-learn: Machine Learning in Python, https://sklearn.org,