Capstone 2 Project Report

Problem Statement

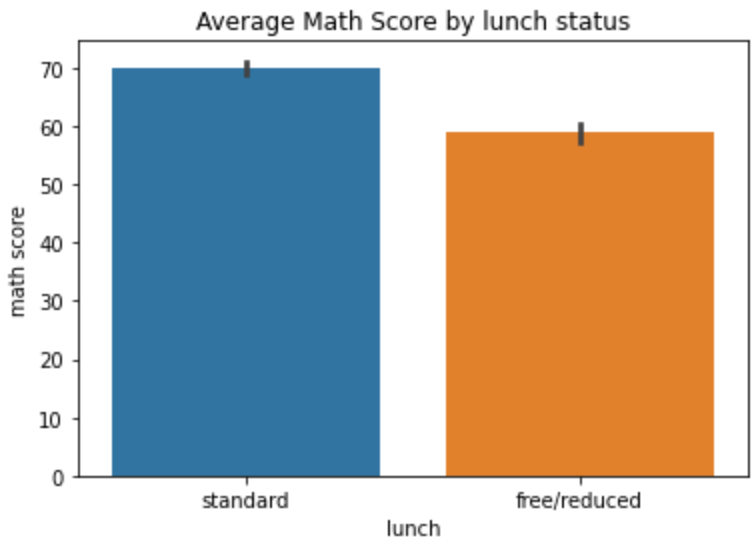
At a high school of interest, there is concern regarding below overall test scores on 3 standardized tests, in the subjects of math, reading and writing. At this particular school, an alarmingly high # of students receive a “failing” grade below a score of 70. Our goal is to improve these scores, and most importantly increase the total number of students that pass the test in all 3 areas. We hope to do this by analyzing the different features that describe each of our students and how they correlate with success on the test. Through this process, we intend to identify characteristic trends in students that may be at risk of failing, as well as any areas we can focus on for improving scores with the tools we have.

We have the standardized test scores of 1,000 students at a particular high school. We also have 5 identifying traits for each student. These include each student’s gender, ethnic category, and parent’s level of education, as well as whether they qualify for free/reduced lunch, and whether they completed the optional test preparation course prior to the tests. I intend to explore the data to locate trends, and then utilize different machine learning models to predict individual scores based on student characteristics. This will help to identify how much weight and importance each characteristic holds in determining test performance.

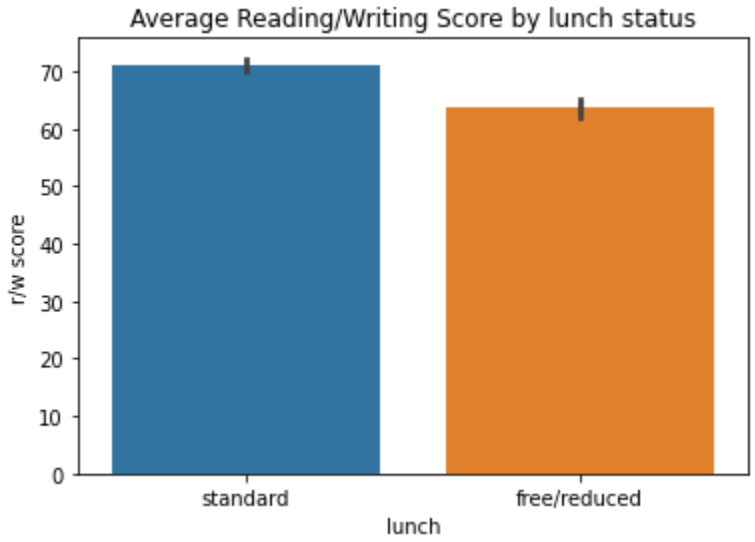
Considering 70 as the mandatory score for passing the test, 59% of the students scored less than 70 on the math test and 50.4% on the reading/writing tests. The school wants as many of its students to pass as possible, but in the first step, we hope to at least improve to where over half the students passed the math test, while at least 60% of the students pass the reading/writing sections of the test.

Data Exploration Findings

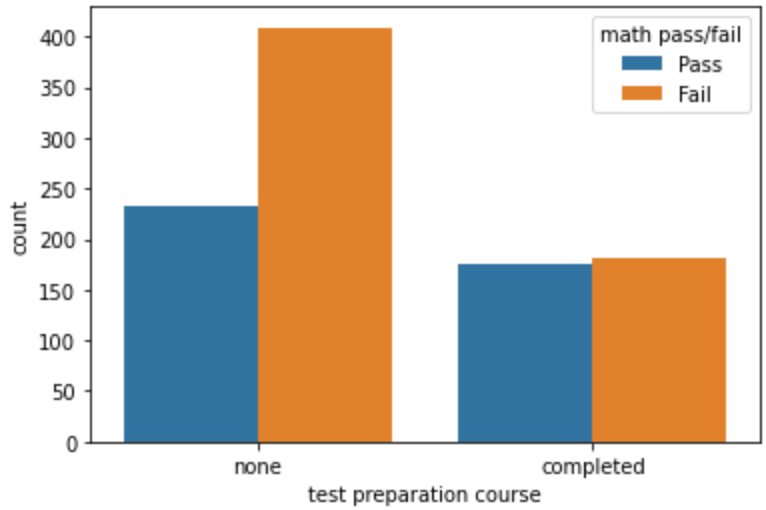
On average, those on free/reduced lunch score at least 10% lower on both sets of scores. The largest difference is seen in Math, where the mean score of those on free/reduced lunch was 58, while the mean score for those on standard lunch was 70.

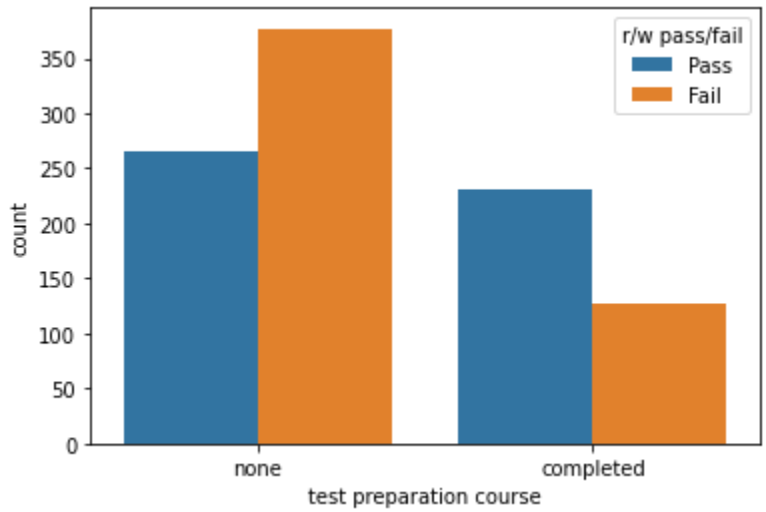


For Reading/Writing, the average score for those on free/reduced lunch was 63, compared to 71 for those on standard lunch.



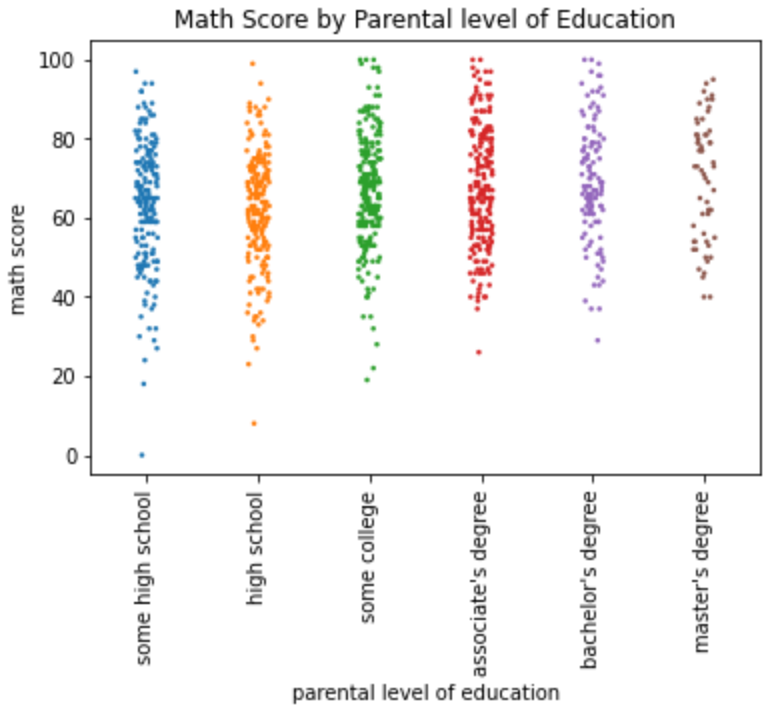
A Test Preparation course was offered as an option to all students, but across the board, most students did not take it. In the end only 358 students out of 1,000 completed the test prep course. This is unfortunate, as there was a much higher pass rate among those who took the test preparation course.

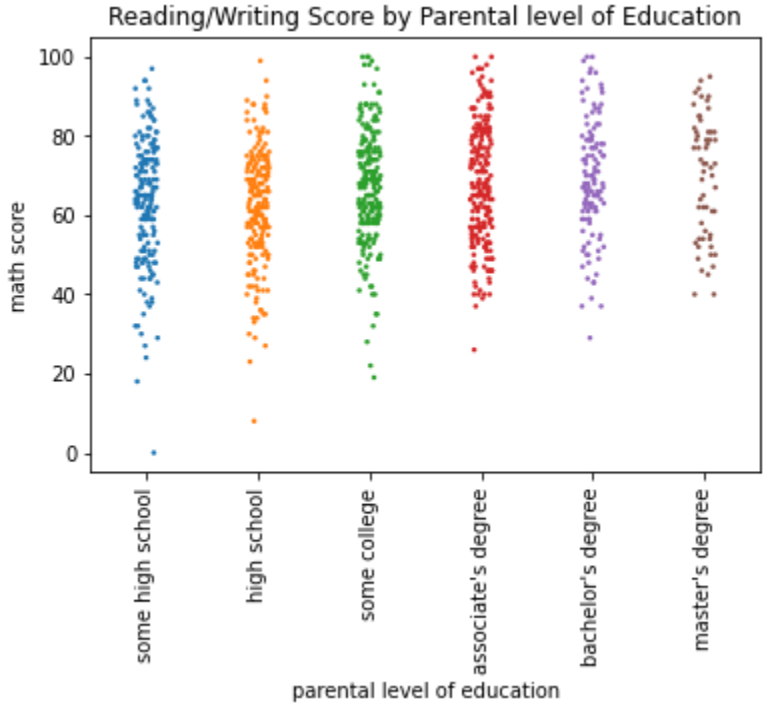




For both tests, pass rates are considerably higher for those that took the preparation course. Just over 36% of the students that did not take the prep course passed the math section, while the pass rate of those who took the prep course was 49%. On the reading/writing side, the pass rate was 41% for those with no test prep, but over 64% for those who did take the prep course!

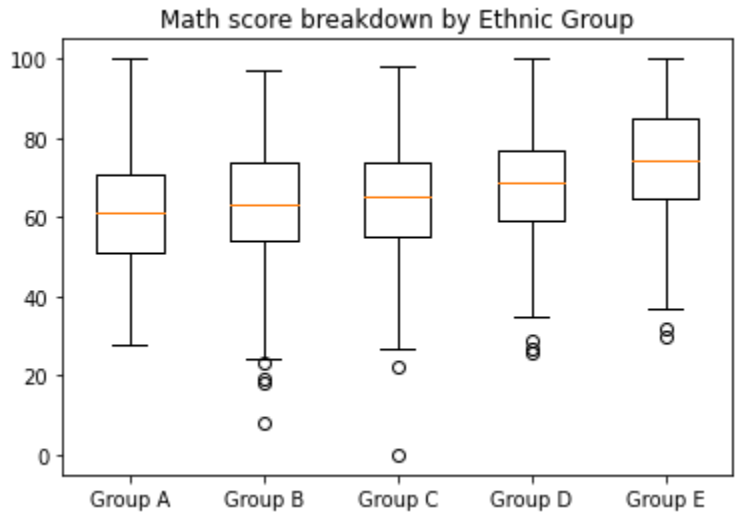
The next features I looked at were trends in scores based on parental level education, and ethnicity group, which has been broken down into 5 different categories. On parental level education, we can see a slight trend moving upwards in terms of score as we move up the ladder, and on the “lower” ends of the education spectrum are weighed down by some extreme outliers on the negative end.

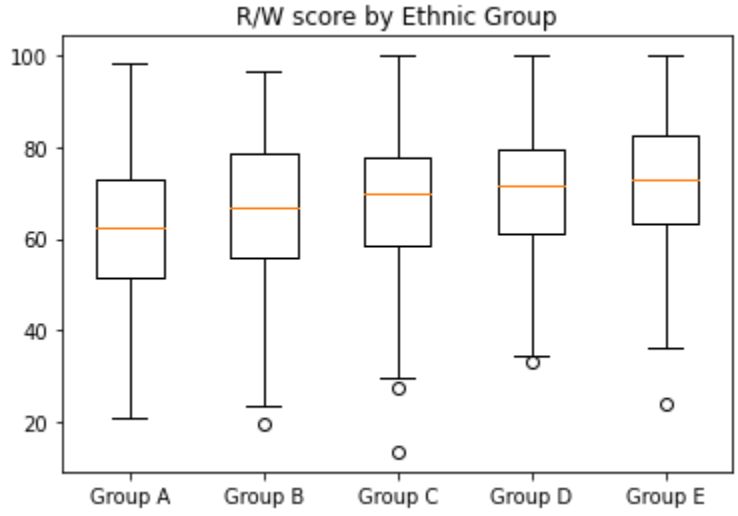




These patterns do indicate that parental level education is something to consider in our predictions as well, though it will likely not hold as much weight in our models as the first 2 features.

Looking at ethnic groups, the average scores and distribution also seem to indicate a slight increase as we move from A to E. Again, there is enough of a trend to consider this feature as well when predicting scores.





Group E seems to outperform all other categories, and there is a particularly wide gap between Group E's math scores from Group D and the other groups.

The final feature is simply gender. I explored the data a little further to see if there were any trends that might help in modeling, and found a noticeable difference on both tests when it comes to performance by gender. It seems the boys in the school on average

performed better on the math sections, with the girls bettering the boys in the reading and writing sections. This should certainly be considered for any model that we fit.

	Math average	**Math pass %**	**R/W average**	**R/W pass %**
Females	63.63	34.4%	72.54	61.0%
Males	68.73	47.9%	64.39	37.3%

In the end, it looks like all 5 of our features play a factor in terms of test scores, and we may very well expect to use all 5 when creating the models.

Pre-Processing

Since all of our independent variables are categorical, I adjusted the dataframe for modeling purposes by using Pandas “get dummies” function on 4 of the independent variables. Parental education I considered to be an ordinal category, moving from the least amount of education, “some high school”, towards the highest level of Master’s degree. To help a model interpret that, I simply mapped the values according to that order, with 0 for “some high school” through 5, for “Master’s degree”.

Because the data was categorical, no scaling was necessary.

In order to measure whether a student passed or failed the test, I had to create a new feature in my dataframe based on score. With 70 and above being the standard for passing a test, all scores of 70 or greater were defined as a “pass”, with anything below deemed a “fail”.

Modeling

Regression Models

I first attempted to see how well a few different types of regression models might be able to predict the 2 scores independently, applying them first to math and then to reading/writing. The table below shows the R2 values for each model.

Regression Model	Reading & Writing R2	Math R2
Ordinary Least Squares	0.28	0.25
Linear Regression	0.23	0.19
Ridge Regression	0.23	0.19
Lasso Regression	0.18	0.13

Using straight regression models, it was determined on all models that predicting the reading and writing scores was easier than predicting math scores. None of the models could explain more than 28% of total variance, but this may very well be due to the fact that I am using categorical data to try to predict a continuous variable like a specific score. This may be particularly difficult because 3 of our categories are also only binary. The models did work best with all 5 categories.

Classification Models

So, my next step was to focus on establishing a categorical variable for the dependent, target variable. As mentioned before, I used “pass” and “fail” as my 2 categories, with 70 as the dividing line. Next I tried some different classification models to see how well we can predict whether a student passed or failed.

	Math Scores				**Reading/Writing Scores**			
	F1	**Acc.**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Prec.**	**Rec.**
Random Forest	0.48	0.61	0.5	0.46	0.61	0.62	0.62	0.60
Logistic Regr.	**0.57**	**0.68**	**0.60**	**0.55**	**0.68**	**0.69**	**0.70**	0.66
KNeighbors	0.47	0.59	0.47	0.46	0.59	0.59	0.59	0.59
Decision Tree	0.42	0.61	0.50	0.36	0.56	0.60	0.61	0.52
SVM	**0.58**	**0.68**	0.59	0.56	**0.68**	0.68	0.68	**0.68**

Like with the regression models, the classification models had better success predicting the reading & writing results. The SVM and Logistic Regression models performed best on both accounts, considerably better than the other 3 models. Like with the regression models, using all 5 features made the model perform best. With the math scores, however, we see much greater variance in the weight of each characteristic, with the Lunch variable, ethnic category, and test prep variables having considerably more

influence than parental level of education or gender. On the reading and writing test, the test preparation variable played the largest role in prediction, but the distribution in the weight of each feature was far less spread out.

Conclusions

From both data exploration and modeling, we were able to see how each of the 5 categorical features of our students contributed to their scores. It is clear that those on free and reduced lunch perform considerably lower than their peers on both tests, and that test preparation does indeed improve scores among all groups.

Using the classification models, the best models performed as high as 68% on pass or fail predictions for the reading and writing test, which is high enough to certainly provide value in recognizing which students may be at risk of failure going into the next standardized tests. This can allow the school to pinpoint early those that may need some extra attention and monitoring over the course of the year.

The data on the test preparation variable also validates having a test preparation course, as there is clear improvement in scores for both subjects among those who participated. It would certainly be of great value if the school were able to increase the # of students who take the test prep course, and I would suggest taking certain actions to ensure that, by either making the course mandatory, or at the least providing encouraging outside incentives.