Tennessee Graduation Rates

Moises Figueroa

Context

Can a model accurately predict graduation rates based on demographic information and school financial data?

If a model can be created to predict this, at-risk schools can be identified and special intervention can take place to raise future graduation rates. Graduation rates in this context means the percentage of students that graduate on time.

It may also be possible that model schools can be identified. What is meant by model schools is that even though the predictive model predicted a lower graduation rate for a particular school, the school ended up doing far better than anticipated. Other schools can adopt their approach and hopefully achieve similar results.

Data

Data was sourced from the Tennessee Department of Education's website. Only data from 2018 - 2019 was used. The data used were: Chronic absenteeism, Demographics, Finance, and Graduation rates. These files are in Excel format and contain aggregate data. Each row of these files have a School ID and District ID that can be used to identify the school.

Data Wrangling

Most of the columns in these files had to be dropped to filter out only relevant columns. After this, all the files were merged, using the District and School ID, into a single Pandas DataFrame that contained these columns:

1. 'ECONOMICALLY_DISADVANTAGED', % of students that are Economically Disadvantaged
2. 'H_Female', % of Students that are Hispanic Females
3. 'AA_FEMALE', % of Students that are African American Females

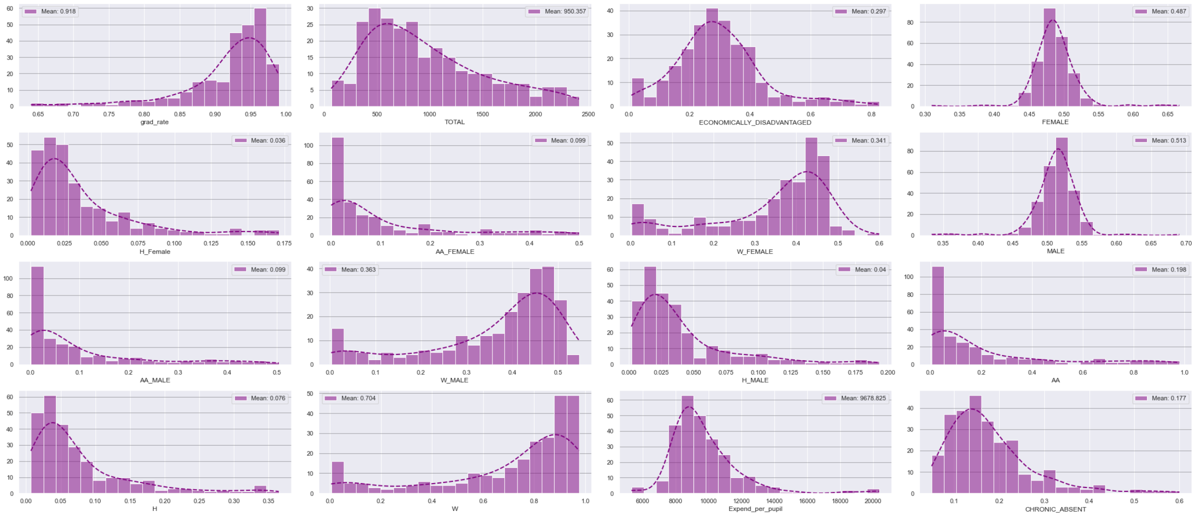
4. 'W_FEMALE', % of Students that are White Females
5. AA_MALE' , % of Students that are African American Males
6. 'W_MALE', % of Students that are White Males
7. 'H_MALE', % of Students that are Hispanic Males
8. 'AA' , % of Students that are African American
9. 'H' % of Students that are Hispanic
10. 'W', % of Students that are White
11. 'Expend_per_pupil', The amount of money the school allocates per pupil
12. ‘CHRONIC_ABSENT', % of Students that are Chronically Absent

Data Preprocessing

The preprocessing steps taken were:

1. Removing outliers
2. Splitting into train/test
3. Scaling the train/test feature columns

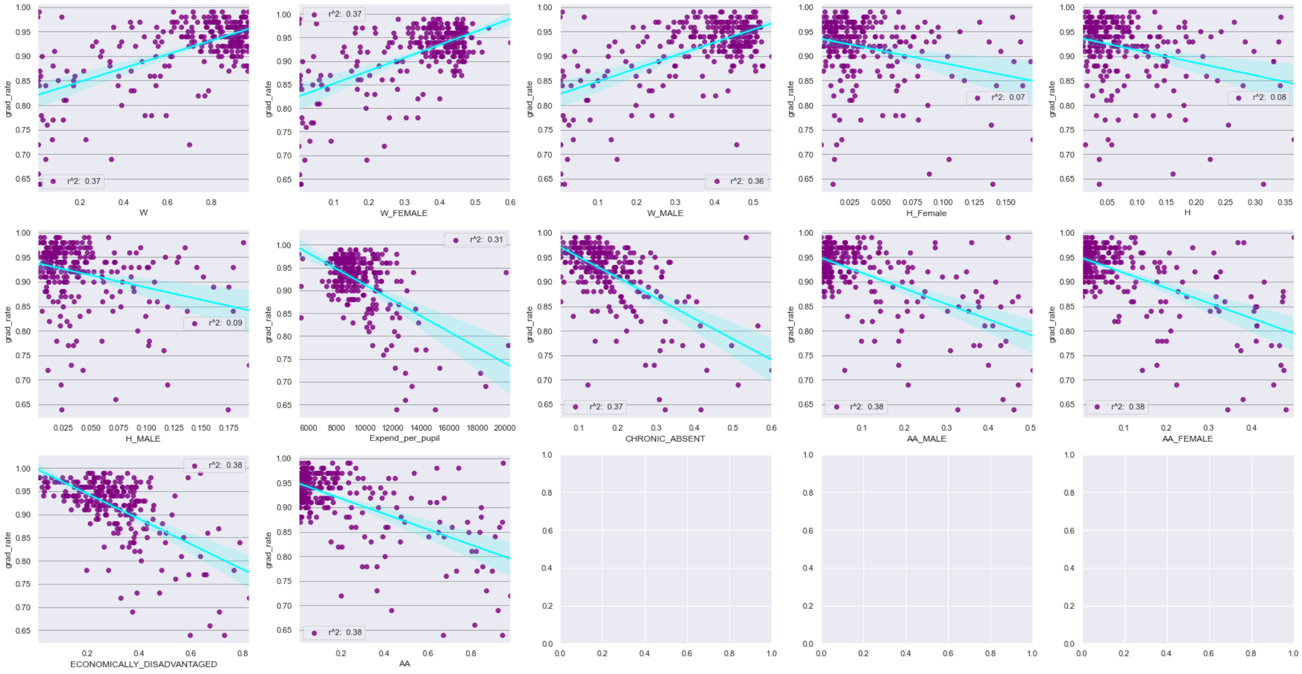
Exploratory Analysis



Distribution of all columns

Larger Version

The demographics data closely resemble the demographics of Tennessee. This would make sense, and it signifies that the data we obtained is accurate.

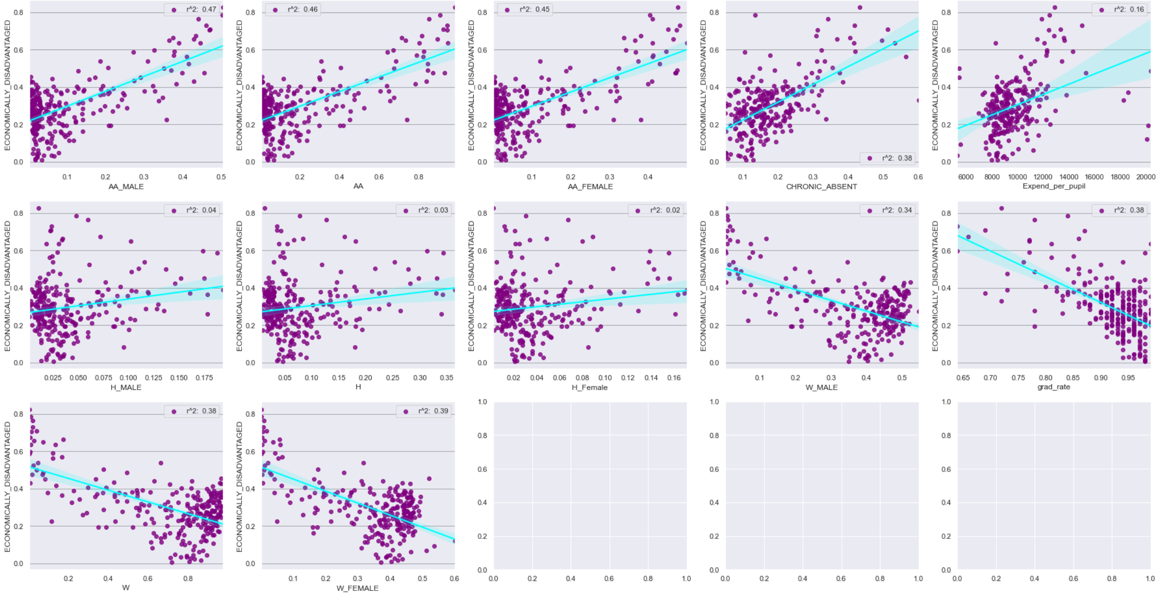


r^2 of graduation rates vs all columns

Larger Version

It would seem that if a school has more proportionally more historically disadvantaged individuals, graduation rates tend to be lower. This likely also ties in with the proportion of students that are economically disadvantaged.

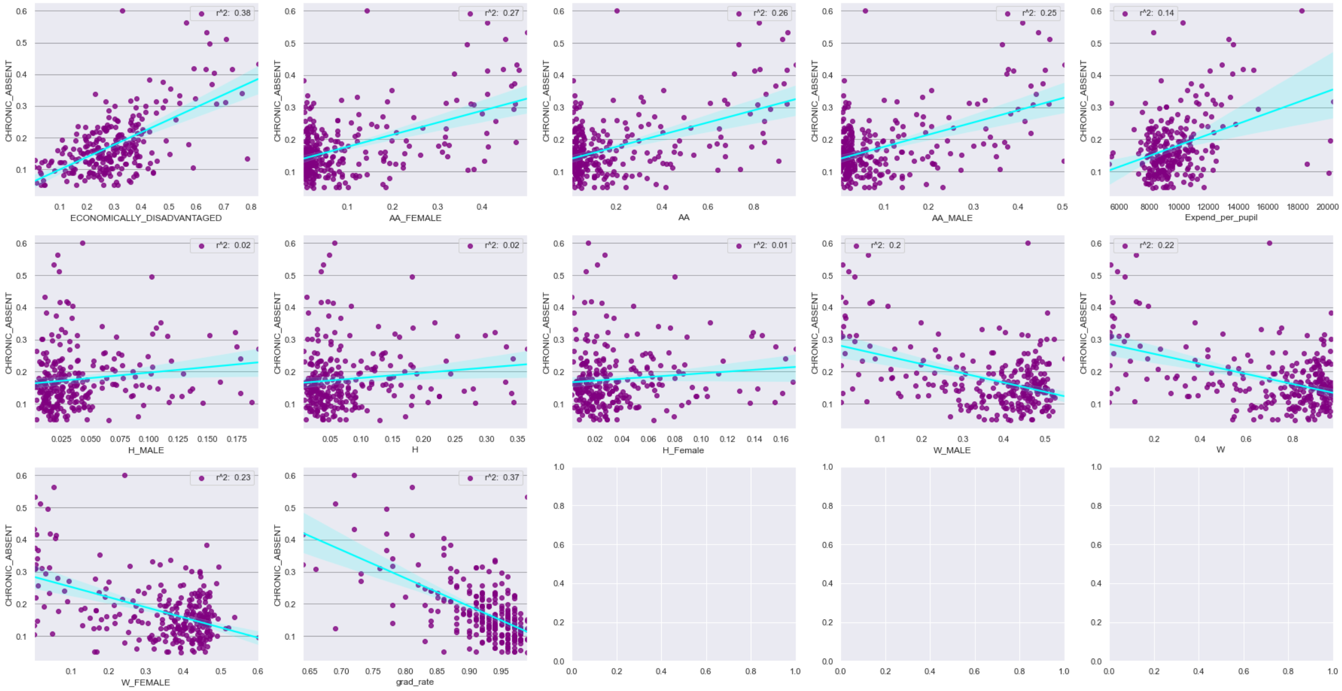
Something of interest to note is that expenditure per pupil is negatively correlated with graduation rates. This could mean a few things. This could mean that siphoning more money into schools does not mean improved outcomes. It could also mean that struggling schools receive the most funding, but since they are already struggling, the money simply is not enough to raise graduation rates all that much.



r^2 of Economically disadvantaged vs all columns

Larger Version

The general trend seems to be that if schools have a higher proportion of historically disadvantaged individuals, then they also have a higher proportion of economically disadvantaged individuals, which means they have a lower graduation rate.



Larger Version

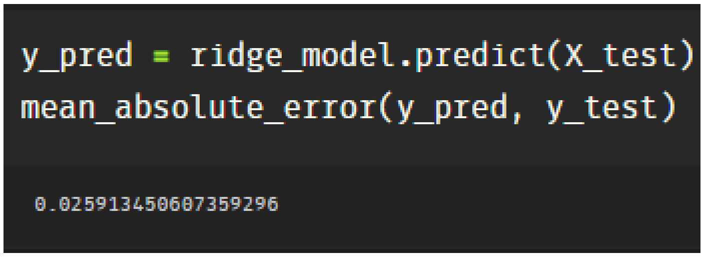
r^2 of Chronically absent vs all

In this context, chronically absent means that a student has missed 10% or more days that the student is enrolled. Obviously if a large proportion of students are missing class, the outcomes will not be good. The students that seem to miss the most class are the economically disadvantaged ones.

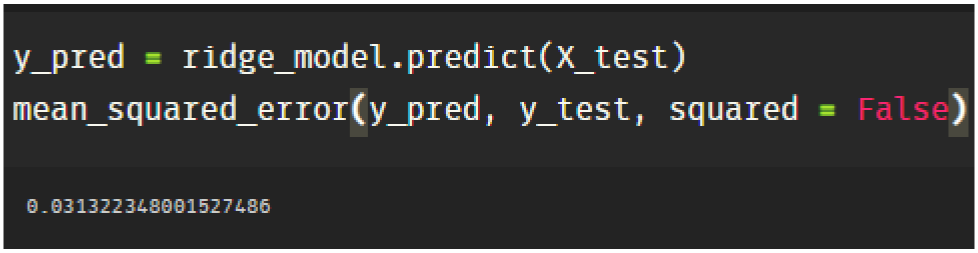
Method

The method I used was to build four models: Multiple Linear Regression, Ridge Regression, Lasso Regression, and Random Forest and choose the one that had the smallest mean absolute error. Ridge and Lasso were hyperparameter-tuned using GridSearch. Random Forest was hyperparameter-tuned using RandomizedSearchCV, since it has many tunable parameters that could have many different values. The scoring used for tuning was Mean Absolute Error.

Results

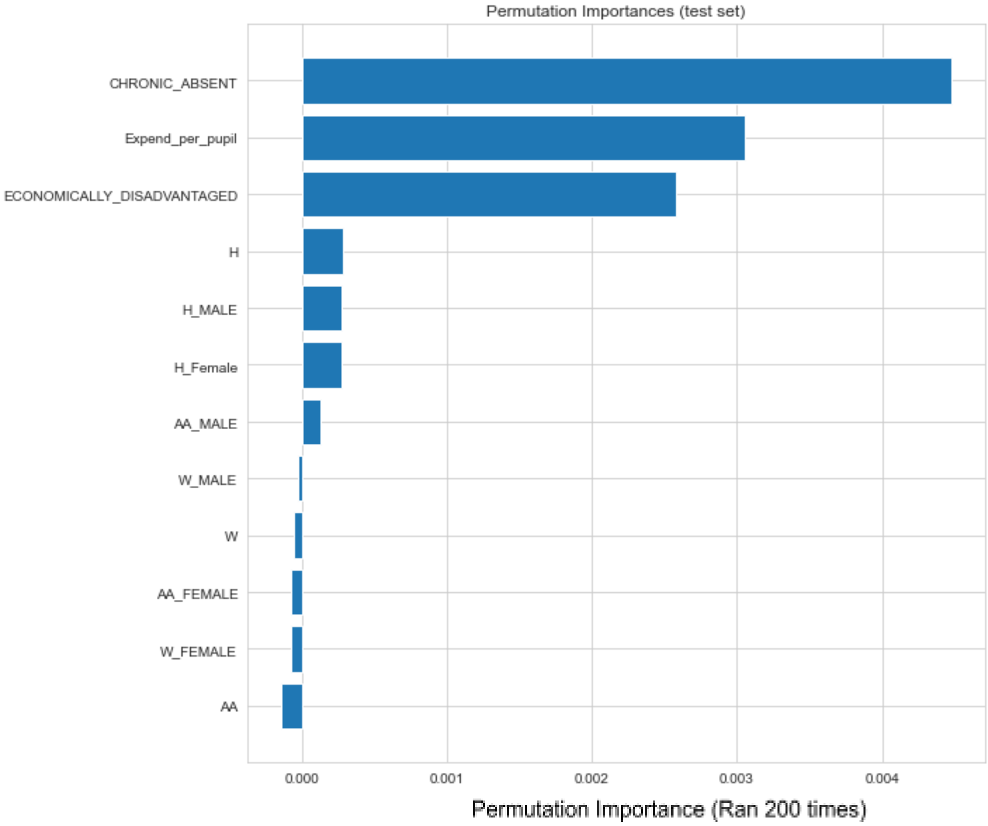


```
y_pred = ridge_model.predict(X_test)
mean_absolute_error(y_pred, y_test)

 0.025913450607359296
```



```
y_pred = ridge_model.predict(X_test)
mean_squared_error(y_pred, y_test, squared = False)

 0.031322348001527486
```

Both the MAE and RMSE are small, which means the model predicted very accurately. RMSE penalizes large errors heavily and in this case it is still small.



Running permutation importances shows that the most important features are proportion of chronically absent students, expenditure per pupil, and proportion of economically disadvantaged. The race of the student seems to play very little into how the model makes predictions. Inclusion of some races even makes the model perform less accurately.

Conclusion

Graduation rates for Tennessee High School can be accurately predicted using only a few features. Further work should be done to see how exclusion of some features affect the overall performance of the model. Further work should also be done to see if there are any High Schools that “beat the odds”; where the model predicts a way lower value than what is actually seen.