Virginia State SOL Score Analysis

Nicholas Newton
Undergraduate Computer Science
University of Virginia
Charlottesville, Virginia
ncn6mq@virginia.edu

Christian Kinzer
Undergraduate Computer Science
University of Virginia
Charlottesville, Virginia
cfk5ax@virginia.edu

John Fishbein
Undergraduate Computer Science
University of Virginia
Charlottesville, Virginia
jhf5my@virginia.edu

***Abstract*—In this paper, we examine many different variables from disparate data sets to predict standardized testing performance for a variety of demographics and subtests for each school in Virginia. We generate these predictions from factors that would be difficult for the school to control, so that we can separate out the variance in test scores due to controllable factors and thus give more context to the performance of schools on standardized tests. Secondarily, we identify the feature importance from the model used to generate these predictions, which can provide to school districts insight into the reasons a school might be performing poorly. A link to our project GitHub and dataset is provided here**

I. Introduction

Extensive research has been performed regarding the potential uses of machine learning in education. Machine learning can be used to grade students, improve student retention, predict student performance, and test students [1]. Teaching methods are being refined using machine learning to optimize education [2]. Machine learning techniques are applied to data collected by online educators in an effort to analyze the learning process [3]. Researchers have also used machine learning to predict whether students will drop out of online courses [4]. We add to this body of research through a machine-learning based analysis of standardized testing scores in the state of Virginia.

Under the 2002 No Child Left Behind Act, each state in the US created standardized tests designed to evaluate student performance in a variety of subject areas. Though this act was repealed in 2015, most states including Virginia still maintain a standardized testing program. The standardized tests Virginia created are referred to as the Virginia Standards of Learning, or the SOL tests. There are SOL tests in the subjects of English, Math, Science, and History/Social Sciences. These tests are critically important, as they are used by the state to evaluate schools, teachers, administrators, and students.

However, the raw scores on tests alone are not a fair estimate of performance. Financial and demographic variables can affect a school's performance, but are outside of its power. A school with chronically low test scores could in fact be succeeding when the disadvantages of its student body are taken into account. The reverse is true as well; it is possible for a school perceived to have acceptable scores to actually need review. Thus, instead of by raw scores, a school should be judged by its ability to elevate its students above the performance expected by circumstance through controllable factors such as creating an effective learning environment.

Unfortunately factors such as the effectiveness of a learning environment are difficult to measure. However, we reason that variance in raw scores can be described as $ControllableVariance + UncontrollableVariance$. Thus, if we create a model to explain the variance contributed by uncontrollable factors and subtract its predictions from the raw scores, the difference should approximate the variance due to controllable factors. So, in this mode, we are able to estimate the true performance of Virginia schools indirectly by predicting the pass rates from uncontrollable financial and demographic data. Our datasets are described in section II. The model we use to generate these predictions is an XGBoost model, the training of which is also described in section II.

With these predictions, we can eliminate variance in a school's performance on SOL tests contributed by uncontrollable factors such as an underprivileged school body. By comparing actual performances on SOLs to our predicted performances, we hope to provide principals, superintendents, and state representatives with a valuable tool to guide decisions around resource allocation and personnel. Through the use of a random forest model, we are also able to report the importance of each variable in estimating a school's performance. Thus, we can provide additional insight into what might disadvantage a school.

II. Method

As stated in the previous section of this paper, our goal in this project is to predict SOL pass rates based on teacher salaries and various demographic data, the former two both from the Virginia Department of Education, and the latter from the American Community Survey. These datasets were joined on school district labels. This resulted in a dataset containing information on demographic variables including income, ethnicity, and education as well as the salaries of teachers. However, after modeling this data, we determined it was insufficient. The problem was that the data gave information only on the level of school district. To more fully explain uncontrollable variance school performance, we needed a higher resolution of data.

Unfortunately, we were unable to locate publicly available data on the school level. Our solution was to scrape that information from a website designed to inform parents on the quality of schools. This website displays information on demographic variables such as income level and ethnicity for the school as well as the number of students enrolled and the teacher student ratio. We were able to scrape this using the python libraries “requests” and “json.” The code for this process is available HERE. Because this dataset contained some missing values, imputation was necessary. After examination, it was apparent that the missing values were uniformly distributed along our target variable of pass rates, so we chose to impute missing values with the mean. Then, after we finished cleaning the data, we joined the scraped dataset with the previous school district dataset. This was possible because both datasets contained the names of schools. However, around 20% of the schools in the scraped data set had a different name than that listed in the school district level dataset. After removing punctuation and spacing to eliminate simple discrepancies such as “T.J. Middle” not matching with “TJ Middle,” we were left without matches for 16% of the data. We determined that it would be too time consuming to match each of the 2,095 schools by hand, and our model would not be significantly improved by this extra data, so we simply left out data for which we did not have a match. This is why the school district level dataset has over 70,000 rows while the final data set has only 60,000 rows.

Because our final dataset contains the pass rates we are trying to predict, this is a supervised machine learning problem. We also determined it is a regression problem because pass rates are continuous. There are many regression techniques in machine learning, and in the next section of this paper, we discuss our implementations of a Multiple Linear Regression, Random Forest Regression, and XGBoost Regression. Linear regressions compute a linear relationship of best fit between our input variables (teacher salaries, teacher to student ratios, and demographic data) and our output variable (pass rate). Both Random Forest and XGBoost are decision-tree based algorithms, which are known to work well for tabular data such as ours.

III. Experiments

Over the course of this project, we have tried three different machine learning techniques to create our model. We set aside 80% of the school data for training and left out 20% for subsequent testing. Though, at first we thought we might not need to separate test data because we were no Virginia schools outside of our dataset, and thus interpolation was unnecessary, we determined that the extremely low errors that resulted were a product of our tree-based models memorizing the data. To avoid this phenomenon we re-introduced a test set. We tried a variety of models on our training data. First, we attempted a basic Multiple Linear Regression as implemented by the sklearn library. After training the model on the training set, the model was able to predict the train set with a RMSE of 9.154, and a MAE of 6.441. Given that the average pass rate for all of the schools in the 2018-2019 school year is 73.28, this RMSE is only about 12.5% away from the average.

Following this, we moved on to try a more powerful regression model. We attempted to predict the pass rates of schools using a Random Forest Regression. We performed grid search cross validation to tune a set of several hyperparameters for the Random Forest regressor. In doing this, we identified that the best hyperparamters for the Random Forest regressor were $n\_estimators = 150$, $max\_features = "sqrt"$, and $max\_depth = None$. Using these hyperparameters, we evaluated the model using 3 fold cross validation and we achieved an average RMSE of 8.185. This model performed slightly better than the basic Linear Regression on our train data. Furthermore, when trained on all of the train data, the Random Forest Regression achieved an RMSE of 3.44 on the same train set.

Finally, we moved on to attempt a XGBoost Regression. We trained this model with 10000 boost rounds with a squared error objective and achieved better results than with the Random Forest model. After 10000 training boost rounds, the XGBoost model achieved an RMSE of 7.884 on the validation set, and an RMSE of 6.150 on the train set. Thus, the XGBoost model was able to predict a given Virginia school's pass rate within about 10.7%. We believe this error rate indicates that our XGBoost model is fairly close to explaining all the variance from uncontrollable factors in school performance.

From this point we moved on to feature analysis. Using the linear regression model discussed above, we again ran our data through a 3-fold cross validation training process, but this time we included Lasso regularization. Using the added L1 norm in the loss function that is included in Lasso, this process has the effect of identifying and eliminating unimportant coefficients while prioritizing more important ones. As a result of this process, we identified the Lasso coefficient corresponding with each feature. This linear regression with Lasso regularization achieved an average RMSE of 9.16 on the 3-fold cross validation.

Furthermore, we also examined the relative importance of each feature using the Random Forest model. In random forest models, the relative importance of a given feature can be estimated according to the average depth of that feature split over all of the individual decision trees. The higher a given feature appears on average, the more important the feature is in the model. The feature importances derived from both the lasso regression and the random forest model are reported in the following section.

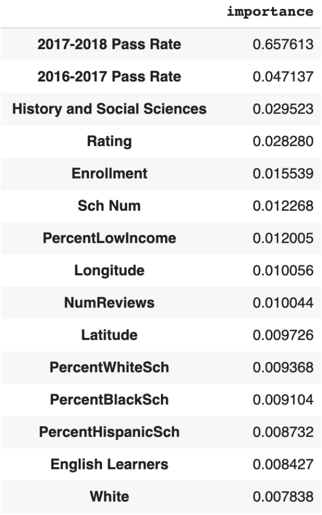
Next, we performed the same ML techniques on a subset of the data. In the previous predictions, we included features like pass rates from previous years, and school rating. In this round of predictions, we removed any feature that was within the school's control. Thus, after removing these features, we were left with the uncontrollable aspects that could influence a given school's overall pass rate. Interestingly, with this subset of data, we were still able to predict the pass rates with a similar RMSE in each model. In the linear regression we recorded an RMSE: 12.541 and an MAE: 9.324 on the train set. In the Random Forest model we recorded a cross validation RMSE: 9.261 and an MAE: 6.273. In the XGBoost model, a validation

RMSE: 8.449. We also reexamined the feature importance when using this subset of data.

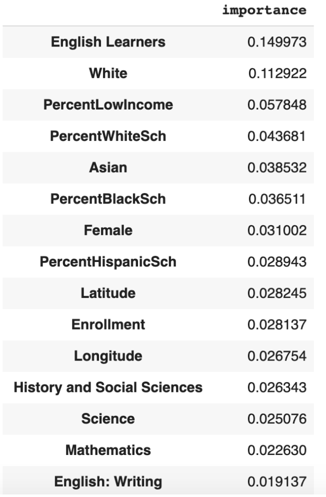
IV. Results

Using our test data set, we evaluated the performance of each model. On the complete set of unseen data, the linear regression achieved a RMSE of 9.278 and a MAE of 6.470; the Random forest model achieved a RMSE of 8.124 and a MAE of 5.405; and the XGBoost model achieved an RMSE of 7.883 and MAE of 5.336. Based on these results, the XGBoost model was able to explain the most variance in the data from the variables.

In the case of the models trained on only the subset of uncontrollable features, we recorded similar test error. The linear regression achieved an RMSE: 12.652 and MAE: 9.371, the Random forest model achieved an RMSE: 9.051 and MAE: 6.262 and the XGBoost achieved an RMSE: 8.450 and MAE: 5.943. These reported errors are remarkably close to the original predictions even though only a subset of data was used during training.



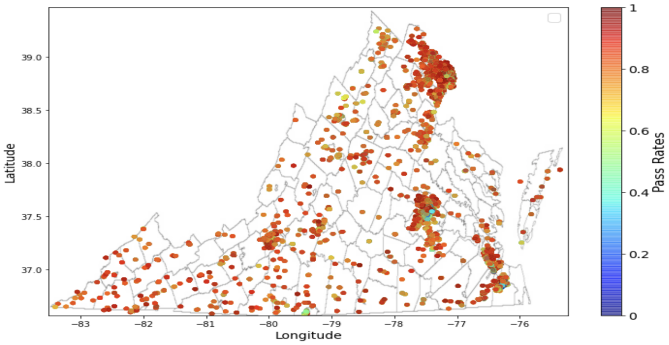
	importance
2017-2018 Pass Rate	0.657613
2016-2017 Pass Rate	0.047137
History and Social Sciences	0.029523
Rating	0.028280
Enrollment	0.015539
Sch Num	0.012268
PercentLowIncome	0.012005
Longitude	0.010056
NumReviews	0.010044
Latitude	0.009726
PercentWhiteSch	0.009368
PercentBlackSch	0.009104
PercentHispanicSch	0.008732
English Learners	0.008427
White	0.007838



	importance
English Learners	0.149973
White	0.112922
PercentLowIncome	0.057848
PercentWhiteSch	0.043681
Asian	0.038532
PercentBlackSch	0.036511
Female	0.031002
PercentHispanicSch	0.028943
Latitude	0.028245
Enrollment	0.028137
Longitude	0.026754
History and Social Sciences	0.026343
Science	0.025076
Mathematics	0.022630
English: Writing	0.019137

The image above displays a comparison between the feature importances determined from the Random Forest model when using the full dataset(left) and when using the specified subset(right). As shown, the features from each are different, even though the models are similar in predictability. Clearly, after ignoring features in the dataset such as pass rates from previous years, the truly important features can be identified. On both lists also are the Latitude and Longitude of each school as well. To examine this further, we generated an plot of where exactly each school is in the state.

Next, we examined the coefficients of each feature created using Lasso Regularization, we found that 14 of our 41 features had a zero coefficient. During the training of the model using Lasso Regularization, it was determined that the features with highest magnitude Lasso coefficients were White (-25.743), English Learners (-22,888), Female (13.631), Asian (10.656), and English Reading (-6.520).



V. Conclusion

Through this project, we learned that a disturbing amount of variance in schools' test scores are due to uncontrollable factors. We were able to predict a school's test score within 10.7% using only the uncontrollable variables. That implies that a remarkable amount of a school's performance is outside of its control. Still, the results of this experiment should be useful to anyone interested in learning the true quality of schools in Virginia. To summarize that true quality, we created a custom metric we call TruePerformance. The metric is designed to give an easily interpretable measure of a school's ability to teach independent of circumstance.

We were able to produce this metric by using the XGBoost model discussed in section III. We used the model to predict the pass rates of every school in Virginia, and in fact we were able to further predict the pass rate of each sub-demographic for each school. Then, the TruePerformance metric was generated by simply subtracting those predictions from the raw pass-rate to identify the residuals. Those residuals should approximate the variance due to factors that schools can control. Of course, there will be some unknowable amount of uncontrollable variance that the model didn't account for still remaining in the TruePerformance metric, so it should be used with caution and in conjunction with other measures of performance. The full dataset we used for this project with the TruePerformance metric included is available here. A dictionary of the features is to be found in the repository's README file.

Additionally, our analysis of feature importance found in section IV gives some insight into the types of factors that most affect the performance of a school or subgroup on the SOL tests.

VI. Contributions

Everyone did an equal share in this project. Nick did all the data collection, cleaning, and joining. John did the modeling with the help of Nick and Christian, and Christian wrote up the document with the help of Nick and John.

References

[1] Danijel Kučak, Vedran Juričić, Goran Dambić. Machine Learning in Education - A Survey of Current Research Trends. In *29th DAAAM International Symposium on Intelligent Manufacturing and Automation*, pages 406–410.

[2] Xiaojin Zhu. Machine Teaching: An Inverse Problem to Machine Learning and an Approach Toward Optimal Education In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 4083–4087.

[3] Ciolacu et al. Education 4.0 — Fostering student's performance with machine learning methods. *2017 IEEE 23rd International Symposium for Design and Technology in Electronic Packaging (SIITME)*, 2017.

[4] Lykourentzou et al. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers Education*, Volume 53, Issue 3, November 2009, Pages 950-965.