Predictive Models for Math Assessment Proficiency and Reading Assessment Proficiency in Iowa K-12 Schools

**Shrey Agrawal**

**Abstract**

Proficiency assessments are utilized in schools to test student’s competency in specific skills, but there has been previous research showing that proficiency scores are correlated with factors like poverty. If so, it is important to recognize that these factors exist so that I can find a way to reduce their influence on student success. In this project, I seek to make models using multiple regression and random forests that could predict the percentage of 8th-grade students who were proficient in math and reading assessments in Iowa public schools. I found that in general, our models were not great at predicting the percentage of students that would be proficient in assessments, with low R^2 values and relatively high RMSE values. I did find that economic variables and demographics had some importance in determining assessment proficiency, but our models cannot really be applied more widely, and our data has limitations.

**Background**

This analysis explores different models that can predict math and reading proficiency in public schools, using data from the Iowa Department of Education. It has already been shown that economic factors and factors from the transition to high school is an important predictor for schools’ graduation rates (Ritter 2015). Because I wanted to focus on student success at an earlier age, I decided to look at proficiency testing rather than graduation rates. While educators have raised concerns about the influence that high stakes testing in schools has on learning behavior and the work environment (Kubow and Debard 2000), proficiency tests have still been shown to be a valid measurement of competency in the tested skills (Lanese 1992). It has been previously demonstrated, however, that poverty and proficiency scores are correlated (Gallagher 1993). In order to measure student success at an important time in their education (just before the transition to high school) and rely on a valid standard measurement, I look for predicting variables of 8th grade math and reading proficiency percentages. What model can most strongly predict the math and reading proficiency assessment scores in Iowa public schools?

**Methods and Results**

I first cleaned the data to select only the desired variables so I was able to work with them appropriately. I compiled various data sets from the Iowa Department of Education differentiated by school district for a five-year period from 2013 to 2018: free and reduced-price lunch, enrollment demographics for PK-12, proficiency assessments, unilateral removals (suspensions and expulsions), teacher salaries, attendance data, and graduation and dropout data. I removed the data from school districts that did not have eighth grade reading and math proficiency scores, because the number of school districts that did not report proficiency data was very low compared to the total number of school districts in Iowa. When cleaning the data, I noticed that the data for percent proficiency in reading and math was slightly skewed. I attempted to transform the data to remove the skewness of the data (log(x), 1/x, x2, etc.). Although these transformations did reduce the skewness of the data, they did not meaningfully improve the accuracy of the models, so the data was left in its original form to minimize unnecessary complexity in the models.

The data was then split into training and testing datasets. I selected 70% of the data for the testing dataset because our sample of 1726 school districts is relatively small, so I wanted to maximize the proportion of the data in the training dataset to create the best model. Because the training and testing datasets both had similar means and standard deviations, I proceeded with a best subsets analysis to find the best 5 variable model for predicting reading and math proficiency scores. I created a pairs plot and histograms for the five explanatory variables for each model (Reading proficiency data: FreeReducedLunch, Percent.White, Average.Salary, K12.ADA.Rate, and StuTeacRatio, Math proficiency data: FreeReducedLunch, Percent.White, Average.Salary, K12.ADA.Rate, Percent.Hispanic). These charts showed that the data for these variables was not as skewed as the proficiency data, except for the Percent.White and Percent.Black variables. These variables were skewed because most districts in Iowa have a very high white student percentage and a very low percentage of black students. Transforming Percent.White and Percent.Black did not improve the model, so they remained in their original form. The rest of the data was also not transformed. I was also able to determine that the percent proficiencies on the reading and math assessments did not change much over time, so I did not need to account for changes in scores each year.

The five explanatory variables from the best subsets model were used to create multiple linear regression models for reading and math proficiency. I continued to use the data without the outliers because the removal of the outliers improved the R squared of the regression model. Then, I squared all the explanatory variables in the model to determine if there were any important interactions between two explanatory variables. I found that there were not any interactions between variables which improved the model, so I chose not to add any interactions in order to simplify our final linear models. The R squared value for the reading proficiency model was 0.2479 and the R squared for the math proficiency model was 0.3207. The plots of the residuals for the explanatory variables did not show any patterns, meaning that I did not need to transform any of the explanatory variables in the model. The plots of the residuals for the fitted values did not show a pattern, so I determined that although our r squared values were low for each model, a linear model was the best regression model for the training data.

I used the random forests method of creating a model in order to produce a reasonable model for the training data while avoiding overfitting the data. I was able to include a greater number of variables in the random forest model when compared to the linear regression model. All the non-numerical columns were removed, specifically the district name and county name. The remaining quantitative variables were used to create the random forest model. The random forest models for both reading and math were optimized to have the best possible RMSE of 5.76 for math and 5.52 for reading. Through our visualizations I found that there were outliers in the data, specifically in the average daily attendance, which were making the data more difficult to fit. By removing the four outliers, I was able to decrease the RMSE of the random forest models.

After creating the potential models for the reading and math proficiency scores, I used the testing dataset to determine the accuracy of our models. The linear regression model on the testing data had an R squared value of 0.2204 for math proficiency and 0.216 for reading proficiency. The R squared value decreased by about 30 percent for the math proficiency model, indicating that the linear model might be overfit. On the other hand, the reading proficiency R squared is very similar between the training and testing datasets. This indicates that the reading proficiency linear regression is a reasonable model. The RMSE of the regression tree model when the testing data was used was 7.39 for math proficiency and 6.47 for reading proficiency. The increase in RMSE for the reading data was not very large, so it is likely that the regression tree model was not overfit. For the math data, the regression tree had a larger testing RMSE compared to the training data, which means that it is possible that the model is slightly overfit.

**Discussion**

The model which was the best at predicting the math proficiency scores was the random forests model. This is because the linear regression model and the random forests model for math proficiency were both slightly overfit. However, the R squared value for the linear regression indicates that the linear model does not fit the data particularly well. On the other hand, the random forests model had an RMSE which was comparable to the RMSE for the reading proficiency. The best model for reading proficiency was the linear regression model because although the linear regression model had a poor R squared value, the R squared only decreased by about 10 percent when tested. This indicates that the model is likely well fit to the data, but that the variables that I have used to create the model are poor predictors of student reading scores. For both models, the difficulty of predicting reading and math proficiency is evident in the low R squared values and relatively high RMSE values for even the most optimized models made with the Iowa public school data.

While this analysis is a good start in this area, I recognize the limitations of the data I am working with and the ability of our models to be applied more widely. Because I was working with data collected at the district level, it would be interesting to see if the model was still applicable at the individual school level and how the accuracy would change. Nevertheless, using data from the district level may have reduced some confounding variables between individual schools that would have less of an influence on the district wide data. I built these models using data from the state of Iowa public schools, and the models would likely lose accuracy in a different state or at a national level. I would be interested in comparing these models with models created using national data or data from another state. Additionally, application of these models to private schools or charter schools would be influenced by differences in the economic factors I used, average teacher salary and percent of students qualifying for free and reduced lunch. Another limitation to account for within our model is interconnected variables - for example the percentage of students who are eligible for free and reduced lunch at a school and teacher salaries are economic variables that are connected. While there is evidence that proficiency assessments are valid testing methods to assess competency in the tested subject, I should not assume that standardized test scores are an accurate measurement of general success for students. Applying these models more widely would not be accurate because of so many differences between education departments and school systems across the country- but it is interesting to start developing models that can predict test scores based on variables of school district.

**References**

Gallagher, Michael P. “Proficiency Testing and Poverty: Looking within a Large Urban District,” April 1993.

Kubow, Patricia K., and Robert DeBard. 2000 “Teacher Perceptions of Proficiency Testing: A Winning Ohio Suburban School District Expresses Itself.” *American Secondary Education* 29, no. 2 : 16–25. <http://www.jstor.org/stable/41064422>.

Lanese, James F. “Statewide Proficiency Testing: Establishing Standards or Barriers?” April 1992.

Ritter, Barbara. “Factors Influencing High School Graduation,” 2015. *Washing Student Achievement Council*

Skybo, Theresa and Jacalyn Buck. 2007. "Stress and Coping Responses to Proficiency Testing in School-Age Children." *Pediatric Nursing* 33 (5) (Sep): 410-8. <https://grinnell.idm.oclc.org/login?url=https://www.proquest.com/scholarly-journals/stress-coping-responses-proficiency-testing/docview/199494155/se-2?accountid=7379>.

**Appendix A: Multiple Regression on Testing Data**

**Figure A1: Math Proficiency**

Table

Description automatically generated

**Figure A2: Reading Proficiency**

Text, table

Description automatically generated

**Appendix B: Random Forests on Testing Data**

**Figure B1: Math Proficiency**

**Text, letter

Description automatically generated**

**Figure B2: Reading Proficiency**

**Text

Description automatically generated**

**Appendix C: Best Subsets Analysis**

**Figure C1: Math Proficiency**

Chart

Description automatically generated

**Figure C2:** **Reading Proficiency**      Chart

Description automatically generated

**Appendix D: Pairs Plots**

**Figure D1**

A picture containing text

Description automatically generated

**Figure D2**

A picture containing text, screenshot

Description automatically generated

**Appendix E: Interaction Plots**

**Figure E1:**

Chart, scatter chart

Description automatically generated

**Figure E2:**

Chart, scatter chart

Description automatically generated