Machine Learning for Virginia: Using Machine Learning to Predict Standardized Testing Scores

CS4991 Capstone Report, 2025

Jerry Gu
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
ncq9fn@virginia.edu

ABSTRACT

The Virginia Standards of Learning (SOL) examinations are designed to measure academic standards for each grade level. To investigate what factors were the most impactful, we built a pipeline to clean and process data taken from a public dataset. That dataset was then used to train various machine learning models which we then used to predict the scores based on various economic, demographic and geographic factors. We utilized different models such as random forest regression, k-means and linear regression to build the best predictive model and analyzed which factors contributed to the most accurate model. Our best predictive model was the XGBoost implementation that had an R^2 =
.7683 and RMSE = 7.1986 and weighted factors such as End of Year Average Daily Membership and School Level Expenditures Per Pupil the highest, while factors such as Title 1 Code, Poverty Level, and Male/Female ratio were weighted the least. In the future, more categorical data and more feature engineering could be used to tease out more complex relationships such as school density to account for differences in rural and urban centers.

1. INTRODUCTION

According to the National Literacy Institute, 21% of adults in the United States are considered to have Low English Literacy. Two-thirds of these adults are born in the United States. (Memdova & Pawlowski, 2019). It cannot be overstated how important it is for children to receive a proper education in English, not only as an educational topic itself, but as a gateway to acquire and learn new skills in other subject areas.

However, the disparities in education are not randomly distributed, and there are patterns and correlations of poor educational quality. Many factors can contribute to this, and we can use data to pinpoint schools that may be falling behind other schools in the state of Virginia.

2. BACKGROUND

Virginia has a very large spectrum of backgrounds, with some areas having larger population densities and urban environments such as Richmond and the Washington Metropolitan Area, while the southern and western parts are dotted with large rural areas with much less population density comparatively. Virginia is also home to many accredited universities such as the University of Virginia, Virginia Tech, and Virginia Commonwealth University.

3. RELATED WORKS

Similar works have been done, such at a state university in Turkey, where researchers analyzed 1,854 undergraduate students taking the Turkish Language 1 course using random forest regression to model and predict academic performance in the course. Their

model achieved a classification accuracy of 75% (Yağcı, 2022).

Another study published in India used similar methods to achieve a 66% model accuracy when attempting to make an early prediction as to whether a student will struggle or not. These students were sampled from computer science undergraduates at various Kolkata universities (Acharya & Sinha, 2014).

4. PROJECT DESIGN

The design of the data pipeline was primarily done in three separate sections: Cleaning out the data, running basic statistical analysis, and then constructing the predictive models themselves. The data itself was retrieved from Kaggle.com (Setash, 2024).

4.1 Data Cleaning

As all of the different categories of data were put into separate *.csv* files, all of the files had to be compiled into one combined dataframe. To do this, we made use of the *pandas* Python library. However, some of the files were given in a format that made it difficult to standardize, such as raw numbers of male and female students. We had to recalculate those demographics as percentages to avoid skewing the model in one direction or the other. This was done for disabled students, race demographics, military connection, homeless students, gender, students in foster care, English learners and economically disadvantaged students. Other qualitative variables such as the schools' Title I Code and official poverty level had to be one-hot encoded, as they were not quantitative variables.

4.2 Data Correlations

Once the data was collected into the one dataframe, the correlation coefficients were taken for each categorical variable. Afterwards, the data was collected into a heatmap to visualize which geographical areas might be struggling or performing. Once the basic correlations were taken, the data was fully cleaned and we could have some expectations of what the models might say.

4.3 Machine Learning Predictive Models

We focused on four main approaches to building predictive models—linear regression and other linear fitting models, random forest regressions and ensemble learning and K-Means clustering—to understand the geographical importance.

4.3.1 Linear Methods

We created a basic linear regression model, as well as a gradient descent model, using the out-of-the-box sci-kit learning Python packages. We split the 20% of the dataset into a testing set and imputed any missing values using a mean strategy for both methods. For the gradient descent, we used a stochastic gradient descent model with a five-fold cross validation across a multitude of parameters. The best parameters were scored using negative mean squared error. Afterwards, we initialized and evaluated the linear regression model and took the RMSE and R^2 value. We also measured which features were the most significant to the model's performance.

4.3.2. Ensemble Learning

We again used a 20/80% testing/training data split for a random forest regressor. We had a parameter grid which had a gridsearch run on it with five-fold cross validation scored with R^2. We then used the best model to calculate the RMSE and R^2 value when evaluated on the test set.

We then tried using XGBoost to squeeze more performance out of the gradient boost, training the model with early stopping if there was no improvement after ten rounds. Then predictions were made with RMSE, R^2, and measured weights for feature importance.

4.3.3. K-Means Clustering

Finally we tried K-means clustering to see if we could find a pattern of specific schools. We did so by creating an elbow plot to find the optimal number of K clusters. Afterwards, the schools categorized were graphed based on their latitude and longitude overlaid on a map of Virginia.

5. RESULTS

The results can be split between the correlations we found through basic statistical analysis and the performance of the predictive models.

5.1 Correlations

After analysis, we found that the features with the highest correlations were associated with both racial demographics and income level, while some of the least correlated features were gender distribution, the degree level of the features, and severely underrepresented racial demographics. See references for the raw Python Notebook used (Gu et al., 2024).

The full correlation chart can be seen in Figure 1 below:

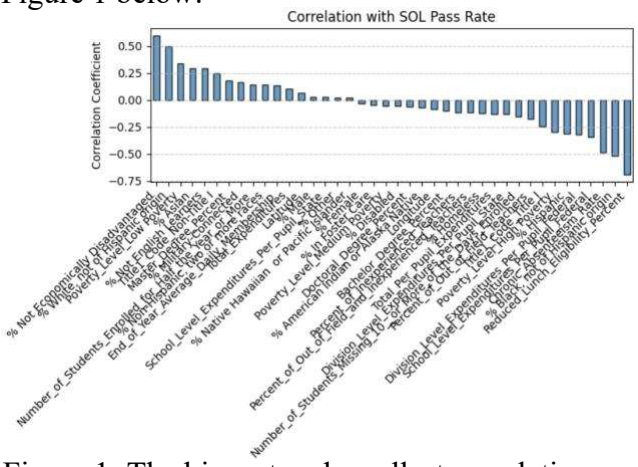

Figure 1: The biggest and smallest correlation coefficient belonged to % Not Economically Disadvantaged and Reduced Lunch Eligibility Percentage, respectively.

The heatmap for the scores and school attendance density were graphed on Figure 2 as shown below:

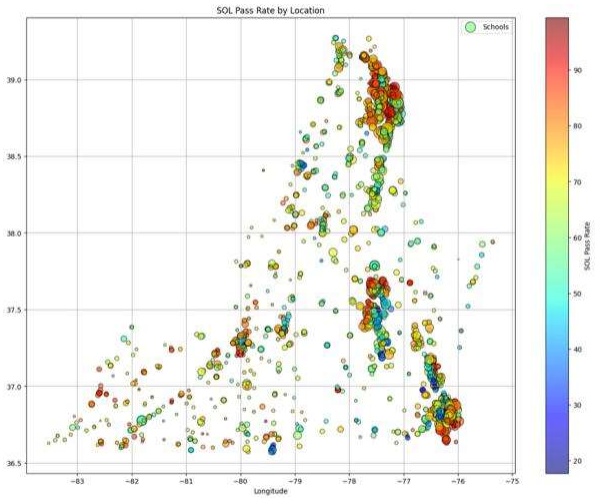

Figure 2: Larger circles indicate a higher concentration of daily membership. Notably, large urban centers seem to have a mix of extremely high and low scores.

5.2 Model Results

The quantitative predictive models' results are summarized in Figure 3 below:

Method	R^2	RMSE
Linear Regression	0.645	8.900
Gradient Descent	0.647	8.890
Random Forest Regressor	0.761	7.302
XGBoost	0.768	7.199

Figure 3: The linear methods we used to predict the SOL Scores.

Overall, the XGBoost model managed to capture the variance best and minimized the Root Mean Square Error, which is also not to the point where we would suspect overfitting.

5.3 K-Means

For the K-Means Graph, we chose five clusters, as shown in Figure 4 below.

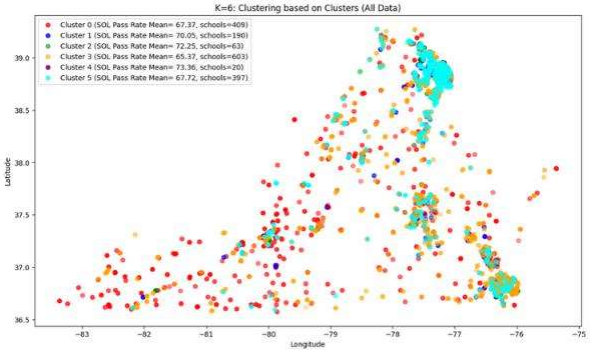

Figure 4: All the K-means color coded.

6. CONCLUSION

Schooling is an extremely complex problem, especially in the state of Virginia which has had a history of discrimination in school. While financial status and race are some of the best predicting factors, it is important not to see schooling as a "zero-sum" game. Rather, we all benefit from everyone having a quality education to facilitate and promote the spread of new and remarkable ideas. This project was created as course credit for CS 4774: Machine Learning at the University of Virginia and entered into the Machine Learning For Virginia (ML4VA) competition.

7. FUTURE WORK

While the K-Means was taken, little interpretation could be done due to time constraints. I believe that we can understand correlations and predict scores, but the true insight to solving the issue may lie in finding the common denominators within the clustered groups.

8. ACKNOWLEDGMENTS

Machine Learning for Virginia Group: Tara Morin, Rithwik Raman CS 4774 (Machine Learning) Professor and ML4VA coordinator: N. Rich Nguyen, Ph.D.

REFERENCES

Acharya, A., & Sinha, D. (2014, December). Early Prediction of Students Performance using Machine Learning Techniques. *International Journal of Computer Applications*, *107*(1). https://d1wqtxts1xzle7.cloudfront.net/54261624/pxc3899939-libre.pdf?1503890083=&response-content-disposition=inline%3B+filename%3DEarly_Prediction_of_Students_Performance.pdf&Expires=1737429308&Signature=fiPGuxNftLMQjA49dJfX1Ta7YD94PYbiE1Ru7qq0IuIK~Ao1Anar

Gu, J., Morin, T., & Raman, R. (2024). *Machine Learning for Virginia: Using Machine Learning to Predict Standardized Testing Scores* [Computer software]. GitHub. Retrieved February 9, 2025, from https://github.com/Jerry-Gu-SB/ML4VA

Memdova, S., & Pawlowski, E. (2019, July 2). Adult Literacy in the United States. *National Center for Education Statistics*. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2019179

Setash, Z. (2024, February 11). *Virginia Public Schools*. Kaggle.com. Retrieved February 9, 2025, from https://www.kaggle.com/datasets/zsetash/virginia-public-schools?resource=download

Yağcı, M. (2022, March 3). *Educational data mining: prediction of students' academic performance using machine learning algorithms - Smart Learning Environments*. Smart Learning Environments. Retrieved January 20, 2025, from https://doi.org/10.1186/s40561-022-00192-z