ADDRESSING SOL PASS RATES IN VA

ML4VA PROJECT F23

Oscar Lauth
University of Virginia
whd7zb@virginia.edu

Grant Sweeney
University of Virginia
gcx8yv@virginia.edu

Ethan Haller
University of Virginia
ttk4ey@virginia.edu

December 10, 2023

ABSTRACT

Standards of Learning (SOL) assessments are crucial in bench-marking academic performance in Virginia's public schools (1). Our project delves into the factors contributing to SOL pass rates across the state, aiming to enhance the prediction of educational outcomes using various features. We employed correlation analysis to identify initial feature importance and to understand the factors most strongly correlated with SOL performance. Additionally, we implemented clustering, specifically using K-means, to visualize and group our data, which revealed significant commonalities along geographical and regional lines. When comparing regression with clustering, we found that clustering provided a qualitative understanding of the problem. Conversely, regression, specifically a random forest regressor, proved more effective in concrete insights. The random forest model illustrated which specific features, such as percentage of students eligible for free or reduced lunch and percentage of students who are economically disadvantaged, contribute the most to influencing SOL pass rates. Now, it's in the hands of policymakers to use the similar school groupings discovered by clustering and the important underlying factors identified by regression to create policy and allocate funds to improve SOL performance across the state.

1 Introduction

Virginia, despite its high national ranking in public school education, faces persistent regional inequities (3). These disparities are particularly evident in the varying pass rates of the Standards of Learning (SOL) exams across different areas (2). SOL exams are a series of standardized tests covering various subjects, designed to set minimum knowledge and skill expectations for students at the end of each grade or course (5). For instance, the Science SOL for third grade specifies what students should ideally learn in that grade. This project was motivated by reports suggesting a significant decline in Virginia's standardized test scores since the pre-COVID era (1). Through a data-driven approach, employing machine learning techniques such as regression and clustering, we aim to identify factors that correlate with SOL pass rates. Uncovering these factors is crucial not only for providing insights into improving SOL scores but also for predicting the impact of economic and environmental changes on these rates. Previous research by the University of Virginia has explored the correlations between demographics and SOL pass rates using statistical analysis. Building on this foundation, our report delves into the application of regression and clustering techniques. We aim to compare and contrast these methods to provide both a current understanding and a predictive analysis of the key factors influencing SOL pass rates in Virginia's public schools.

2 Method

To begin, an extensive dataset about Virginia Public Schools was found on Kaggle (4). The dataset contains information such as demographics, economic factors, student behaviors, teacher information, and SOL pass rates for over 1600 public elementary, middle, and high schools in Virginia. However, this data is spread across about 20 different files,

so the next step was to collate the data from files such as SOL_Pass_Rate.csv, Economically_Disadvantaged.csv, and Funding.csv. Many datasets were pivoted to have one row per school. For example, the school "Albemarle High" might have two diffrent rows in Economically_Disadvantaged.csv: one with a value of "N" for the column "Economically Disadvantaged", and one with a value of "Y". Both of these rows had the column "Total Count". Pivoting the data added features such as "economically_disadvantaged_Y" and "economically_disadvantaged_N", where the value of these for each row is "Total Count". After pivoting many datasets, they could be joined together by their school and division features ("Sch_Div") to create a large "master" dataset.

The next step in the process was feature engineering. In order to achieve better performance in both clustering and in regression, it was helpful to have more relevant features. So, instead of having the total number of "Y" values and "N" values for features such as "Economically Disadvantaged", "English Learners", and "Foster Care", the numbers were converted to percentages. This was done for many different features that were given in this format. Examples include "percent_economically_disadvantaged", "percent_english_learners", "percent_foster_care", "percent_homeless", "percent_military", and "percent_disabled".

After the feature engineering, the data was cleaned and formatted through a Pipeline in order to prepare it to be used by machine learning algorithms. Since many of the NaN values in the master dataset are resultant of rows not existing in the smaller datasets, the Pipeline uses a SimpleImputer that replaces NaN values with 0. For instance, in the Economically_Disadvantaged.csv, if a school has a "N" for "Economically Disadvantaged" but does not have a "Y", it means that there are no students that are economically disadvantaged, so "economically_disadvantaged_Y" should be 0 instead of NaN. The Pipeline also uses a StandardScaler so that each feature is scaled to reduce outliers and have better impact on both clustering and regression tasks.

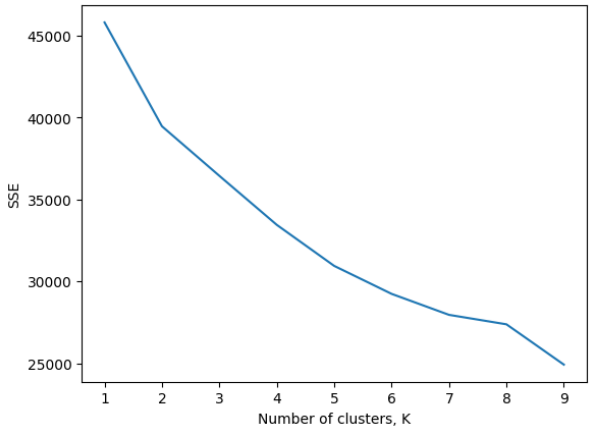
With this Pipeline in place, clustering was performed on the dataset through k-means clustering. The k-means algorithm was chosen not only because it guarantees convergence, but also because the centroids found through the algorithm would have meaning in the context of the problem. The final centroids from the algorithm can be compared to note key differences between them. These key differences may be significant factors relating to SOL pass rates. A more detailed description of the algorithm and the number of clusters chosen is described in the Experiments section.

From here, regression was done with a Random Forest Regressor. This was done for two reasons. First, using a Random Forest Regressor provides a model that can predict SOL pass rates given a number of factors. So, if government officials were considering implementing changes, they can see how it may affect SOL pass rates. Also, a Random Forest Regressor is built using decision trees, and decision trees can provide helpful insights into which features of the dataset are most important in predicting SOL pass rates. The Experiments section elaborates more on the different regression models tested.

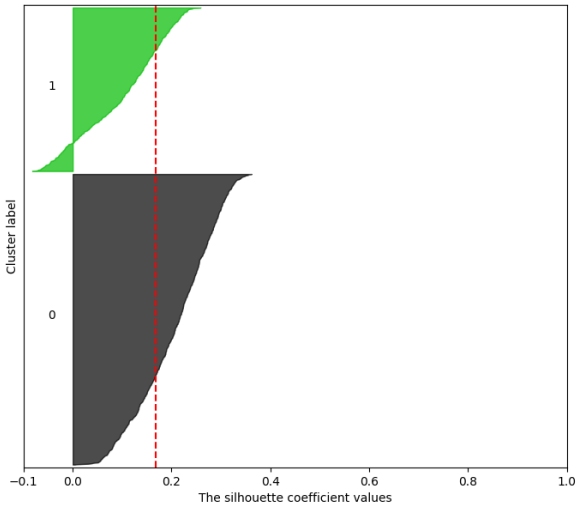
3 Experiments

First, basic correlation analysis was conducted on the data. From correlation with the label, SOL pass rate, we see strongest correlation (negative) with free or reduced lunch eligibility and cohort dropout rate. There is also strong positive correlation with white, non-Hispanic percentage. Correlation analysis helps gain early data insights for feature engineering and can later confirm the coefficients and feature importance scores of the regression model. K-Means clustering was also performed on the data. The elbow method was not entirely conclusive in terms of the optimal number of clusters (1a). So, an alternative technique, silhouette scoring was utilized to evaluate different numbers of clusters. The silhouette scores indicated that 2-3 clusters were optimal as they had the highest positive silhouette scores. Sample silhouette plots for 2 (1b), 3 (1c), and 4 (1d) clusters are shown.

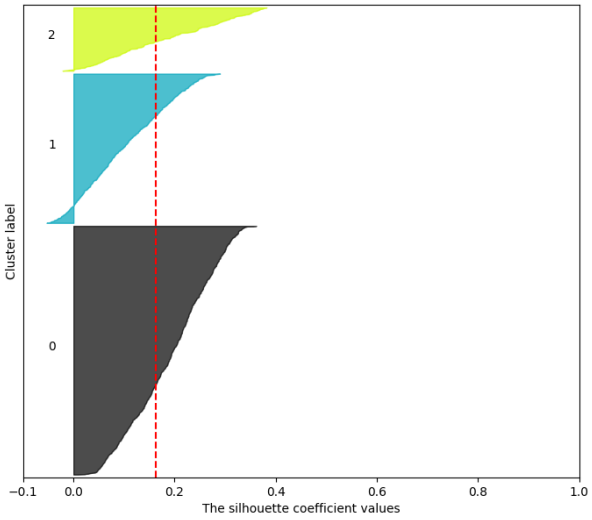
Lastly, regression was applied to this problem. A wide range of regression models were tested such as linear regression, ridge regression, decision tree, random forest, support vector, and multi-layer perceptron. Next, 5-fold cross validation with root-mean-squared-error (RMSE) was used to evaluate the efficacy of each model. From this model selection step, the random forest regressor had the lowest of all the models with a cross validation RMSE of 8.307. With the model now selected, next was model tuning and optimization. The random forest regressor was tuned using a randomized search with cross validation that tried different combinations of hyperparameters. The tuned hyperparamters for the random forest were 7 max features with 140 decision tree estimators. This tuned model yielded an improved 5-fold cross validation RMSE of 8.116 and test set RMSE of 8.688. This shows strong model generalization as the test and validation RMSE were very close.



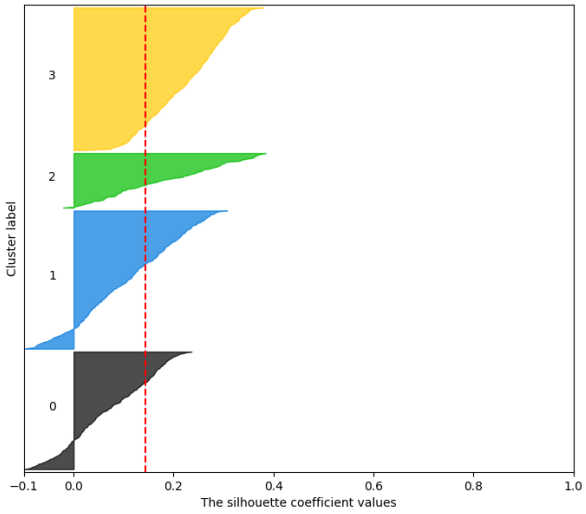
(a) Elbow Analysis of K-Means Clustering



(b) Sample K(2)-Means Clustering Silhouette Scoring



(c) Sample K(3)-Means Clustering Silhouette Scoring



(d) Sample K(4)-Means Clustering Silhouette Scoring

Figure 1: Subfigures of K-Means Clustering

4 Results

With the models tuned and optimized, they were now applied to this problem. K-Means with 2 clusters was plotted on a map of Virginia and the clusters seem to roughly align with densely populated areas (red cluster) and rural areas (blue cluster) (2). While silhouette scoring in the experimental phase indicated 2-3 clusters as optimal, it seems different numbers of clusters can be useful in different ways. For example, splitting the data into 5 clusters and plotting each on a map of Virginia (Fig. 3) produced interesting results. It seems clustering on all the features aligns pretty well with geography. The 5 clusters split the state into approximate regions such as Northern VA, Southwest VA, Coastal VA and more urban versus rural areas. These seemed to make sense and align with expectations as schools close to each other likely face the same challenges like poverty or share similar advantages such as higher funding or well-educated teachers.

Next up was applying the tuned Random Forest Regressor to this problem. Random forest regression works by an ensemble of decisions trees that split on features. So, from the model each feature has an importance value and can be ranked in order of importance. From these feature importance values, some of the least important features in predicting SOL pass rates for a school are percent homeless, percent other gender (non-binary), and cohort dropout rate. Some

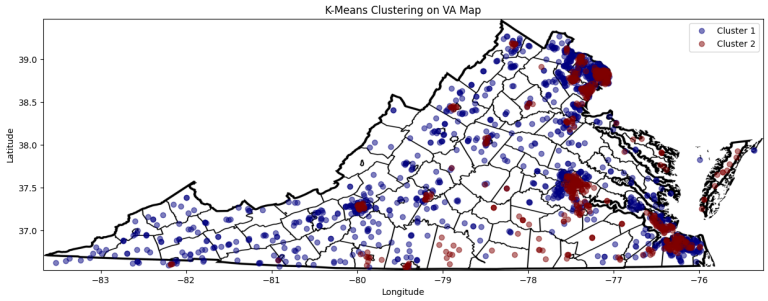


Figure 2: Sample K(2)-Means Clustering on Map of Virginia

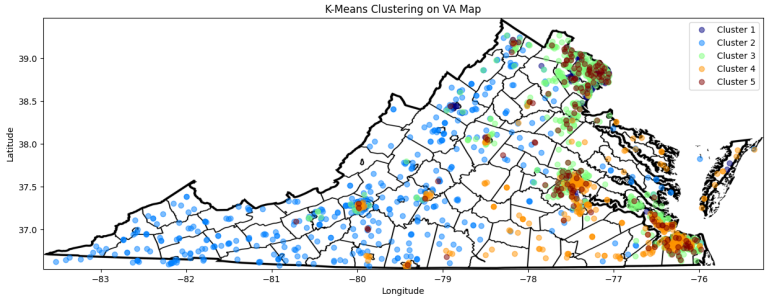


Figure 3: Sample K(5)-Means Clustering on Map of Virginia

of the most important features in predicting SOL pass rates are percent eligible for free or reduced lunch, percent White (non-Hispanic), and percent economically disadvantaged. This roughly mapped onto the expectations from the correlation analysis, however there are some notable differences. For one, cohort dropout rate had a correlation of -0.630 which is pretty strong, however the random forest's feature importance ranking had it at 5th to last among all 27 features. These feature importance rankings are super useful in discovering the underlying factors contributing to poor academics and test performance at schools. From the most important features, it seems economics plays a big role. Both economically disadvantaged and eligible for free or reduced lunch point pretty directly towards students in low socioeconomic standing. Further, White students may be generally wealthier and more established in the country, leading to strong contribution to higher SOL pass rates and another indication that the issue is one economical disparities.

5 Conclusion

This investigation into the Standards of Learning (SOL) pass rates across Virginia's public schools, using machine learning techniques such as regression and clustering, has unveiled significant insights. The correlation analysis highlighted key predictors of SOL performance and granted some starting intuition, while K-means clustering revealed distinct regional and geographical patterns in educational outcomes. In addition, clustering grouped similar areas together allowing policymakers and educators to make changes at a larger, cluster-wide scale as opposed to painstakingly tailoring changes school by school. Regression analysis proved instrumental in predicting future trends and highlighting the impact of specific features on SOL pass rates. These findings suggest that addressing educational disparities in Virginia might require region-specific strategies, informed by a deep understanding of local socioeconomic contexts. This is evident by the top contributing features to SOL pass rates being primarily economic such as percent economically disadvantaged and percent eligible for free or reduced lunch. As the project evolves, it aims to offer actionable insights for policymakers, potentially guiding efforts towards a more equitable educational landscape in Virginia.

6 Member Contributions

Oscar Lauth: Touched up pipeline and worked on clustering task. Evaluated different K cluster sizes and plotted K-means clustering on map of Virginia. Performed regression on data and did model selection + validation + testing of regression model.

Grant Sweeney: Refactored code gathering the data, dropping unnecessary columns, and merging the 18 datasets into one. Greatly cleaned up data preprocessing and loading. Collected references. Wrote conclusion.

Ethan Haller: Wrote the large initial skeleton of the data loading, collation, and processing, including how to join the 18 datasets. Worked on feature engineering and the Pipeline to prepare data to be used in both regression and clustering tasks. Developed the video in iMovie. Did model tuning and feature importance extraction.

References

[1] Dean Mirshahi. A look at SOL pass rates for Richmond area school districts. *COVID Effect on SOL*, 2023. URL: https://www.wric.com/news/virginia-news/a-look-at-sol-pass-rates-for-richmond-area-school-districts-chesterfield-henrico-petersburg-hanover/.

[2] Lisa Rowan. Few Virginia school divisions see bright spots in SOL scores. *COVID Effect on SOL*, 2023. URL: https://cardinalnews.org/2023/09/09/few-virginia-school-divisions-see-bright-spots-in-sol-scores.

[3] US News Education *Education Rankings by State*, 2023. URL: https://www.usnews.com/news/best-states/rankings/education.

[4] Virginia Public Schools. Compiled school level data for the 2021 - 2022 school year. *SOL Data*, 2022. URL: https://www.kaggle.com/datasets/zsetash/virginia-public-schools.

[5] Virginia Department of Education. K-12 Standards & Instruction. *K-12 Standards*, 2022. URL: https://www.doe.virginia.gov/teaching-learning-assessment/instruction.