Building a Deep Learning Model to Predict Academic Achievement

Jane Ahn, Haneul Shin, Max Guo

December 28, 2021

1 Introduction

In this paper, we develop two models to predict student academic achievement based on school spending, student race and gender data. We base our models off of the Kaggle dataset U.S. Education Datasets: Unification Project, which consists of K-12 financial, student, and academic achievement data in U.S. Education. This dataset provides detailed information on the different types of revenues and expenditures of each U.S. state from 1986 to 2019, as well as the average test scores of its students on math and reading assessments. We first perform an exploratory data analysis on the data. We then utilize two different machine learning models – decision trees and random forests – to predict student test scores based on financial and student data.

2 Background and Related Work

It is well-documented by many scholars that increased school spending is positively correlated with improved student outcomes. Wide-scale studies have found that schools with higher levels of funding tend to have higher student academic achievement as measured through standardized test scores and graduation rates [3]. Moreover, increases in school resources that require significant financial support, such as smaller class sizes and higher teacher salaries, have also been positively associated with improved academic achievement [3].

While many scholars agree on the blanket conclusion that school spending is a significant factor in determining student outcomes, there is no clear consensus on the relative impact of different types of spending or how they impact different student groups. For instance, some studies have found positive correlations for increasing certain types of expenditures (e.g. teacher benefits) and negative correlations for others (e.g. instructional services) [2], while others have found that additional instructional supports are essential to improve student outcomes [3]. Furthermore, there is no clear consensus on the extent to which different types of revenues and expenditures impact different student groups, e.g. racial groups.

3 Problem Specification

3.1 Data

We develop two deep learning models to predict the reading and math test scores of a given student group. Each data point, which is characterized by a student group, contains the following information. When pertinent, we provide the possible values of the inputs as well.

- Year the data was taken: 1986 - 2019

- State revenues (federal, state, local)
- State expenditures (instructional, supportive services, capital outlay, other)
- Number of enrolled students by race and gender: {American Indian/Alaska Native, Asian, Black, Hawaiian Native or Pacific Islander, Hispanic/Latino, Two or more races, White}, {Male, Female}
- Average grade 4 and 8 reading and math standardized test scores, by race and gender

Note that each average standardized test score corresponded to *either* a race group or gender group, not both.

3.2 Algorithms

Given the year, state revenues, and state expenditures information (as specified above) of a given student group (which may be further characterized by either race or gender), each of our models predicts the average grade 4 and grade 8 reading and math scores of the student group. In addition to predicting academic achievement, we will use these models to analyze the relative correlations of different types of revenues and expenditures with academic achievement. In particular, since our decision tree model iteratively splits on the feature that will yield the greatest information gain, the resulting tree will suggest which factors are most important to consider.

4 Approach

In order to intelligently design our machine learning models, we first performed data analyses on our given data set to better understand the relationships between the variables of gender, race, time, test scores, expenditure, and revenue. We detail the various analyses performed in the Experiments section below.

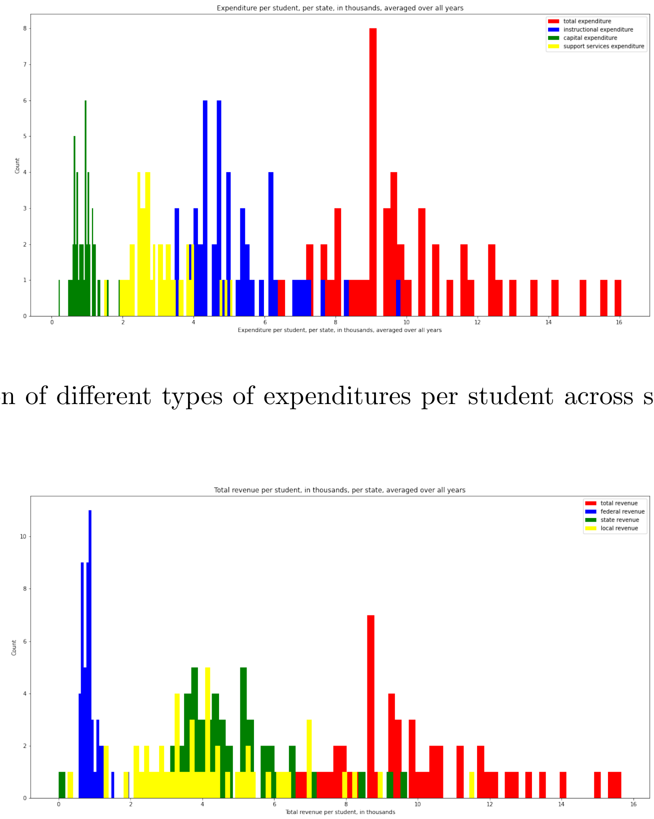
Next, we applied our decision tree and random forest models in order to determine how accurately we can predict test scores based on the features in the data set, as well as which features seemed to have the most significant impact on these predictions. We also describe the details of these algorithms at length in the Experiments section below.

5 Experiments

5.1 Data Analysis

5.1.1 Distributions of Revenues and Expenditures Across Students

We first analyzed the distribution of resources for each type of revenue and expenditure across different states. If a certain type of expenditure was consistently insignificant when compared to the total expenditure, we would expect it to be a less significant factor in predicting academic achievement in our model. As shown in Figure 1a, the relative ordering of expenses from least to greatest is capital expenditure, support services expenditure, and instructional expenditure. Finally, there is a relatively wide distribution of expenditure per student over all states, ranging from $\$6,000$ to $\$16,000$ dollars spent per student in a given year. Figure 1b demonstrates that there is a significant gap between types of revenue as well: federal revenue contributes the least, while state and local revenue contribute similar amounts.



(a) Distribution of different types of expenditures per student across states, averaged over time

(b) Distribution of different types of revenues per student across states, averaged over time

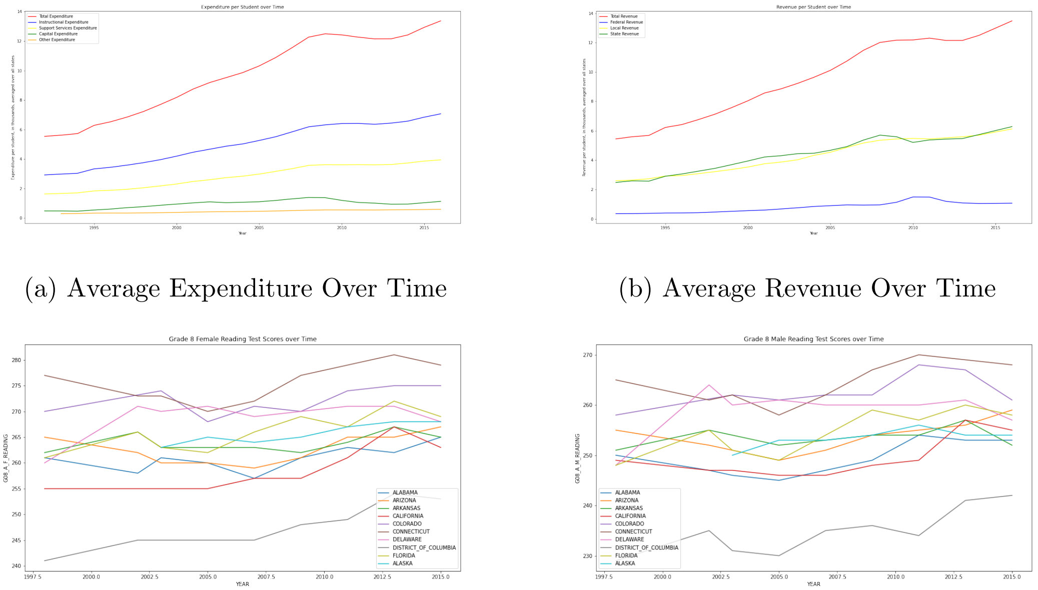
Figure 1: Distributions of Revenues and Expenditures Across Students

5.1.2 Expenditure, Revenue, and Test Scores Over Time

We then verified the conclusions from the aforementioned background research that school spending is positively correlated with academic achievement. Figures 2a and 2b show that expenditure and revenue have increased over recent years across all categories. Meanwhile, Figures 2c and 2d show moderate improvement in both Grade 8 female and male reading test scores since 1997 for several states. This observation suggests that the scope of this problem may be fitting for our proposed machine learning models, as higher spending and revenues seems to be associated with increased academic performance.

5.1.3 Race in Relation to Expenditure and Test Scores

To determine the significance of race in comparison to expenditure, we plotted reading and math scores with respect to per-student expenditure, grouping by race, as shown in Figures 3a and 3b. We observe that there may be a slight positive correlation between per-student expenditure and reading and math scores, especially within select racial groups. However, the figures also demonstrate that race itself is highly correlated with test scores. We thus expect our machine learning models to utilize race as a more important factor in determining academic achievement over expenditure on education.



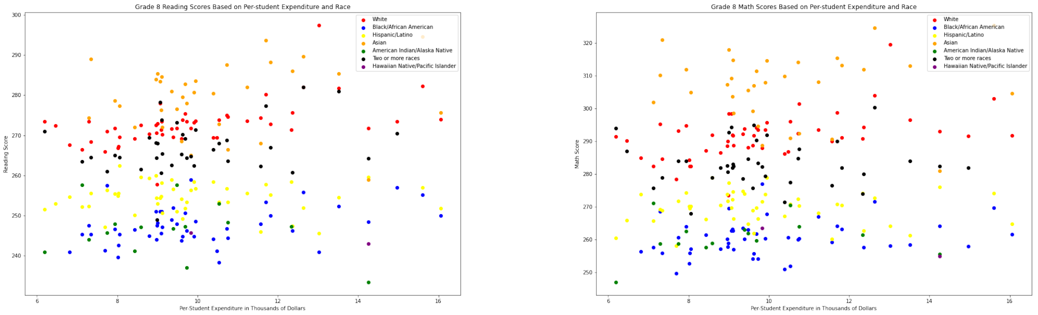
(a) Average Expenditure Over Time

(b) Average Revenue Over Time

(c) Grade 8 Female Reading Scores over time, across select states

(d) Grade 8 Male Reading Scores over time, across select states

Figure 2: Expenditure, Revenue, and Test Scores Over Time



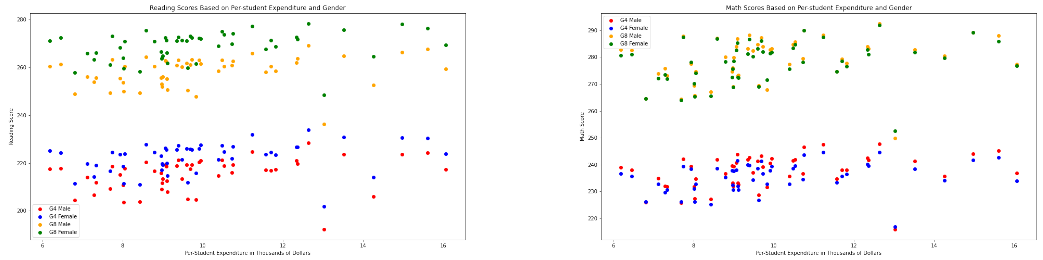
(a) Grade 8 Reading Scores Based on Per-student Expenditure and Race

(b) Grade 8 Math Scores Based on Per-student Expenditure and Race

Figure 3: Race in Relation to Expenditure and Test Scores

5.1.4 Gender in Relation to Expenditure and Test Scores

To determine the significance of gender in comparison to expenditure, we performed a similar analysis as in the previous section. We plotted reading and math scores with respect to per-student expenditure, grouping by gender, as shown in Figures 4a and 4b. As before, we observe a slight positive correlation between per-student expenditure and test score. There also seems to be a slight difference between male and female test scores for both math and reading, although the difference is larger for reading. However, compared to race, gender does not seem to correlate as strongly with test scores, so we would expect our machine learning models to weight expenditure relatively higher when grouping by gender than when grouping by race. Finally, we verify that



(a) Reading Scores Based on Per-student Expenditure and Gender

(b) Math Scores Based on Per-student Expenditure and Gender

Figure 4: Gender in Relation to Expenditure and Test Scores

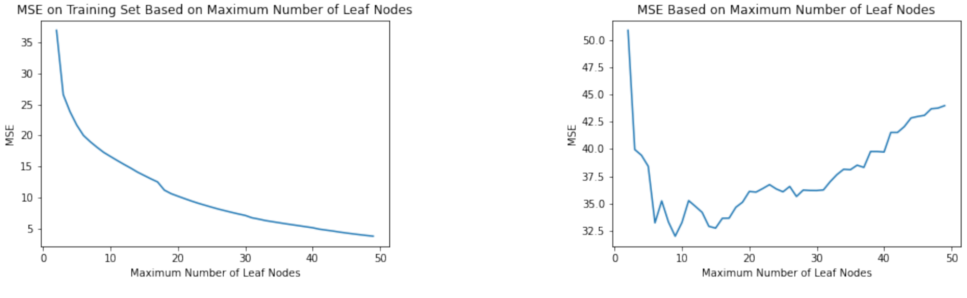
there is a definite increase in test scores between Grade 4 and Grade 8.

5.2 Machine Learning Models

5.2.1 Decision Tree

We began by running a decision tree on our data to determine the relative importance of the various revenues and expenditures of each state in determining average student performance. We divided our data into student groups so that each group would consist of students from the same state in a particular calendar year. The features we used in building the decision tree included the year, state's revenue per student (including federal, state, and local), and state's expenditure per student (including instruction, capital, support services, and other). Using these values, the decision tree would predict the average 4th/8th grade reading/math score of the group.

We used a random 75%-25% split to obtain our training/testing data. We tested the performance of a range of decision trees by varying the restriction on the maximum number of leaf nodes in each decision tree. We expected poor performance in decision trees with a small number of leaf nodes due to coarse branching. However, we also expected poor performance in decision trees with a very large number of leaf nodes due to overfitting on the training data.

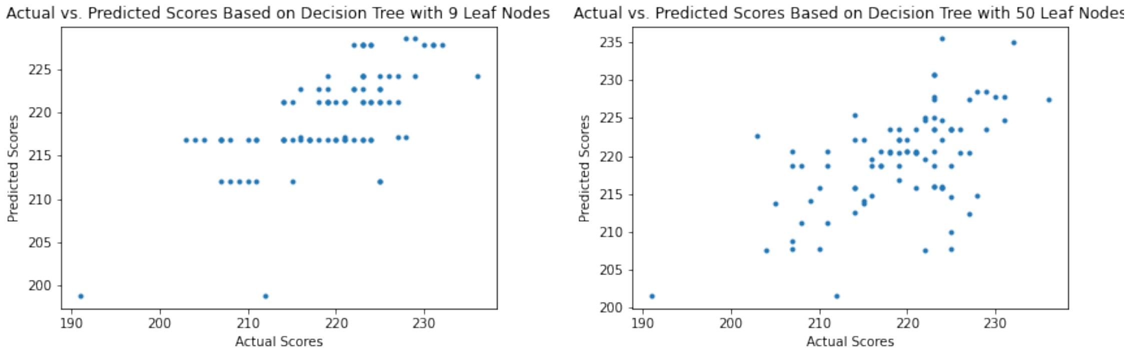


(a) MSE of Decision Tree in Training Set

(b) MSE of Decision Tree in Testing Set

Figure 5: MSE of Decision Tree against Maximum Number of Nodes

Graphs 5a and 5b plot the MSE for both the training and testing data of decision tree models against the maximum number of leaf nodes in the decision tree. While the MSE for the training set



(a) Decision Tree Prediction with 9 Leaf Nodes (b) Decision Tree Prediction with 50 Leaf Nodes

Figure 6: Decision Tree Prediction

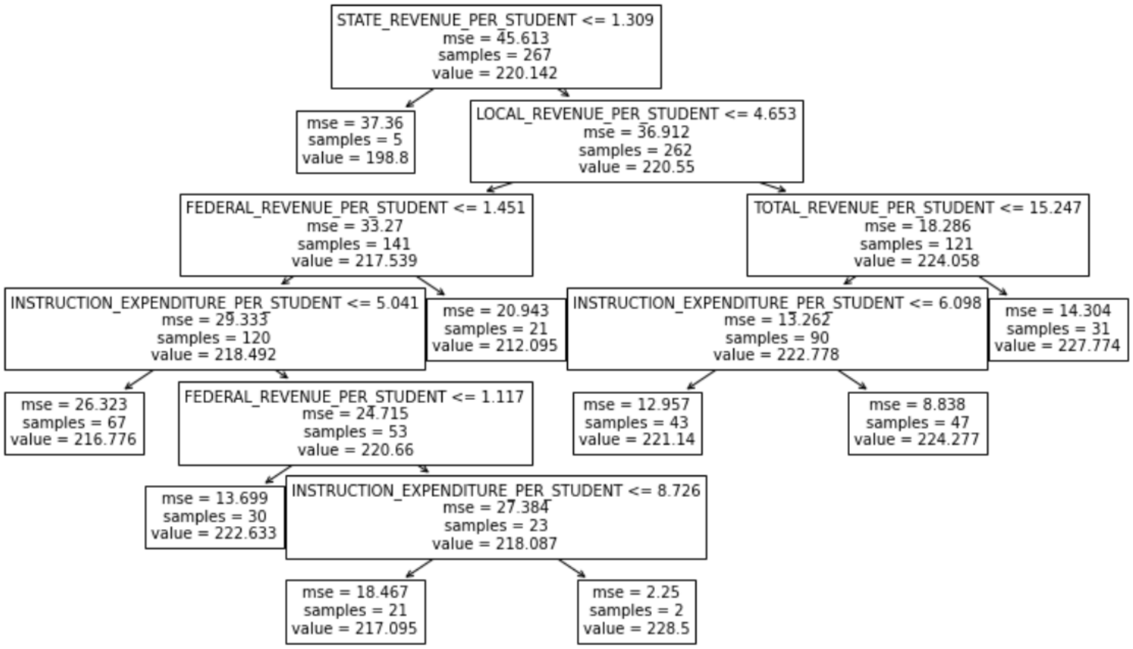
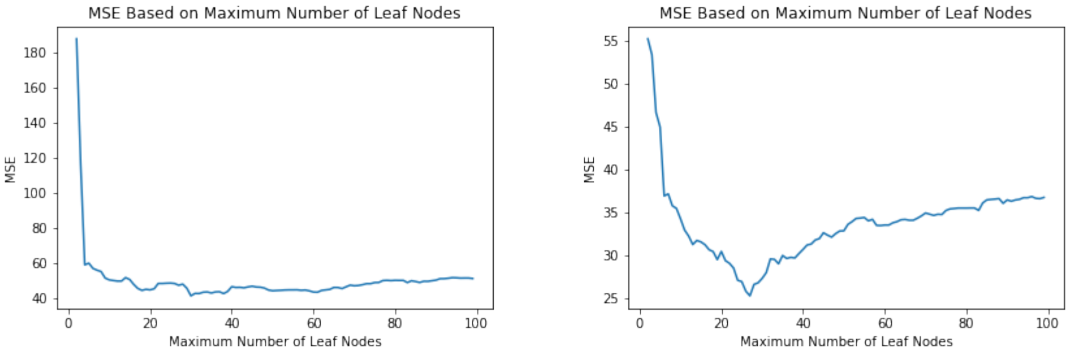


Figure 7: Prediction of Decision Tree with 9 Leaf Nodes

decreases as the maximum number of leaf nodes increases as expected, shown by 5a, we observe an initial sharp drop in the MSE of the testing set as our decision tree model becomes more precise, followed by a slow decline in performance as the decision tree model begins to overfit to the training set as depicted in 5b.

Furthermore, 6a shows the predicted scores of the test data using a decision tree with 9 leaf nodes, compared to a decision tree with 50 leaf nodes in 6b. As seen in 5b, the decision tree with 9 leaf nodes achieves a lower MSE than the decision tree with 50 nodes due to little overfitting.

Finally, we present the details of the decision tree with 9 leaf nodes in 7. The first couple of nodes



(a) Decision Tree MSE with Race Data

(b) Decision Tree MSE with Gender Data

Figure 8: Decision Tree Performance with Student Groups Further Partitioned by Race or Gender

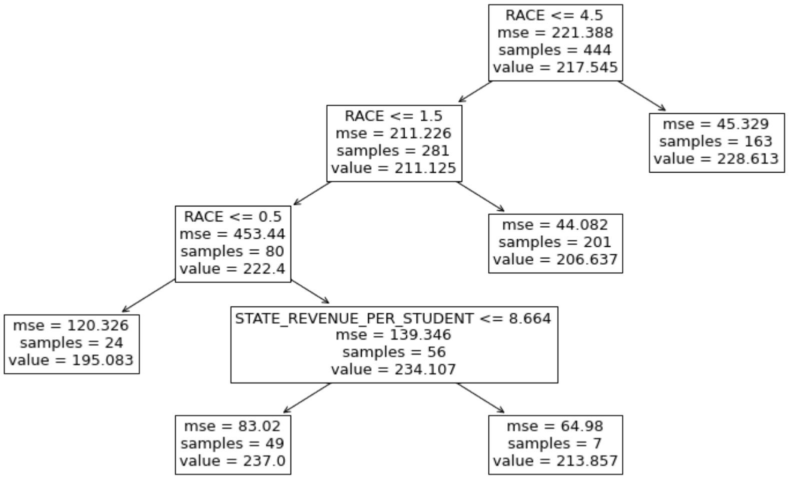


Figure 9: Decision Tree with Student Groups Further Partitioned by Race

all split based on revenue information, so it seems that the amount of revenue a state receives for educational purposes significantly affects students' average performance.

We also applied a decision model to predict the performance of student groups further partitioned by race or gender. We noticed the same trend in 8a and 8b where as the number of leaf nodes in the decision tree increased, the MSE on the testing set initially dropped sharply as the model became more precise, then gradually increased as the model began overfitting on the training data.

Looking at the first couple of nodes in the decision tree when we further partition student groups based on race, as shown in 9, race seems to be the most important factor in predicting student academic performance. Our decision tree split on nodes that yield the largest information gain, which suggests that race may be more significant than any state financial data, since it constitutes

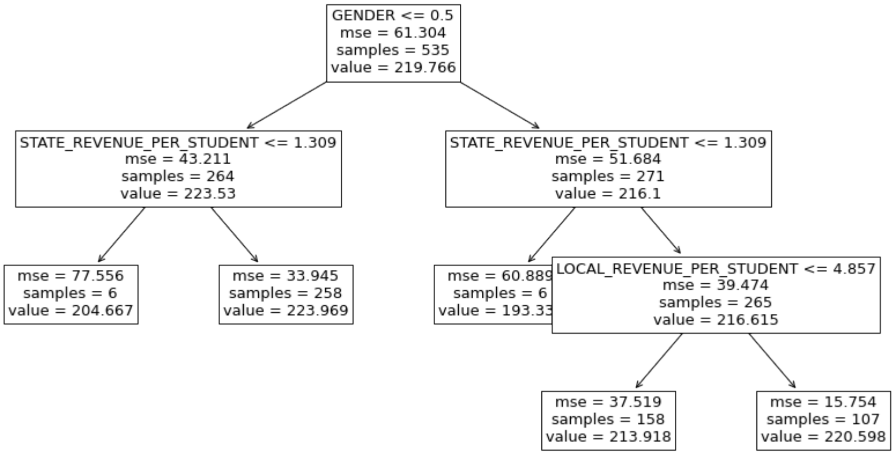


Figure 10: Decision Tree with Student Groups Further Partitioned by Gender

the first three of the four decision nodes. A similar but weaker trend appears in the decision tree where groups are further partitioned based on gender, shown in 10, since the first node splits on gender.

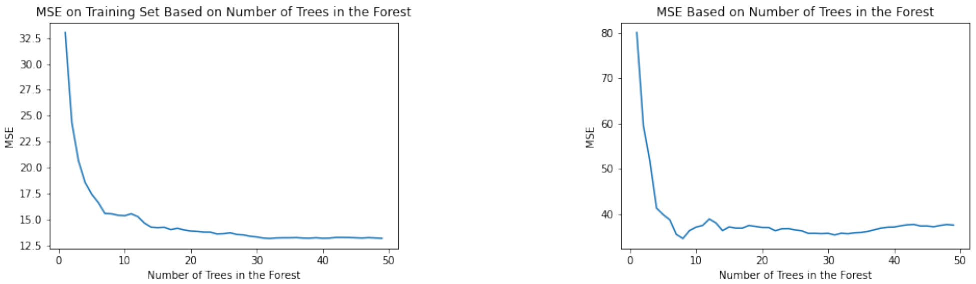
5.2.2 Random Forest

We next used a random forest model to predict academic achievement in order to account for the possible overfitting nature of decision trees. We again divided our data into student groups so that each group would consist of students from the same state in a particular calendar year. We used the same features as the decision tree, including the year, state's revenue per student (including federal, state, and local), and state's expenditure per student (including instruction, capital, support services, and other). Using these values, we wished to predict the average 4th/8th grade reading/math score of the group.

We again used a random 75%-25% split to obtain our training/testing data and observed the performance of our model based on the number of trees included in the random forest. We expected an initial increase in performance as the number of trees used increased until we reached a "sufficient" number of trees in the forest, at which point the MSE should stabilize.

Graphs 11a and 11b plot the MSE for both the training and testing data of decision tree models against the number of trees used in a random forest. The MSE on the training set mostly decreases as the number of trees in the forest increases. Regarding the MSE on the testing set, we observe an initial sharp drop in the MSE which then stabilizes at a relatively constant level of MSE.

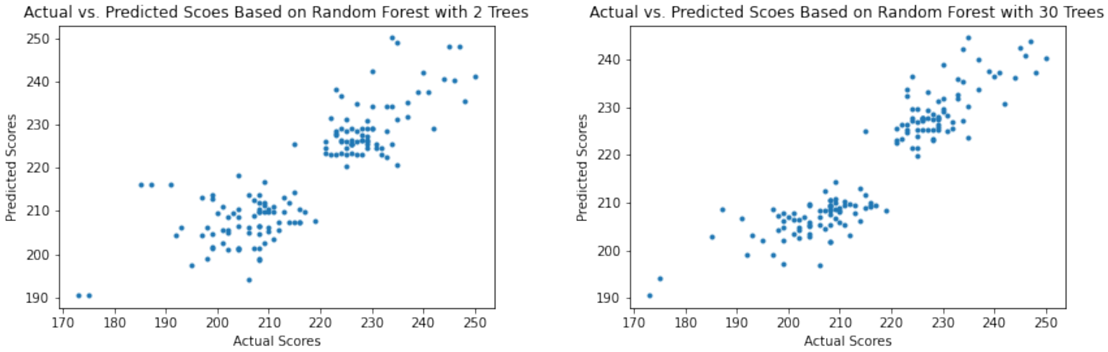
Next, 12a shows the predicted scores of the test data using a random forest with 2 trees, compared to a random forest with 30 trees in 12b. We know from 11b that the random forest with 2 trees



(a) MSE of Random Forest (Training Set)

(b) MSE of Random Forest (Testing Set)

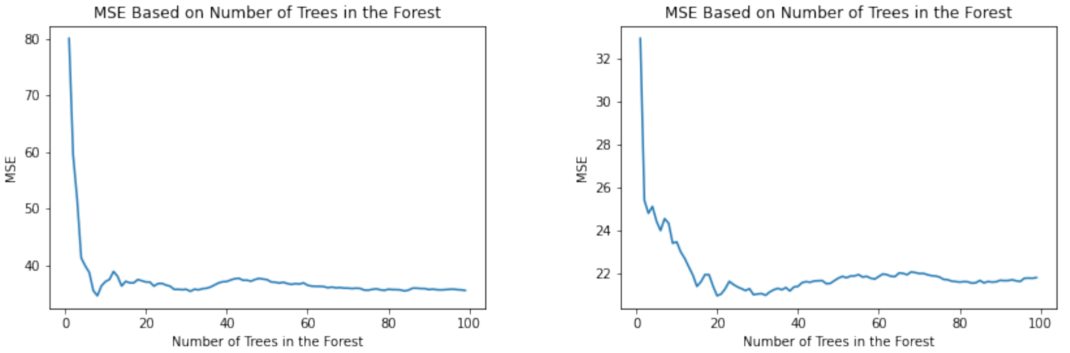
Figure 11: MSE of Random Forest against Number of Trees



(a) Random Forest Prediction with 2 Trees

(b) Random Forest Prediction with 30 Trees

Figure 12: Random Forest Prediction



(a) Random Forest MSE with Race Data

(b) Random Forest MSE with Gender Data

Figure 13: Random Forest Performance with Student Groups Further Partitioned by Race or Gender

results in higher MSE than the random forest with 30 trees due to more overfitting.

We also applied a random forest model to predict performance of student groups further partitioned by race or gender. We noticed the same trend in 13a and 13b where as the number of trees in the decision tree increased, the MSE on the testing set initially dropped sharply as the model became more precise, then stabilized at a low level of MSE.

6 Discussion

The focus of our project revolved around understanding the relationships among the variables of expenditure, revenue, race, gender, and year in predicting academic achievement, as well as attempting to develop models that allowed us to perform these predictions accurately. We performed data analyses and developed machine learning models in order to accomplish these goals.

Through various forms of data analysis, we observed a varied distribution of expenditure and revenue across different categories (1), increase in revenue, expenditure, and test scores over time (2), mild correlations between expenditure and test scores (3, 4), strong correlations between race and test scores (3), and moderate correlations between gender and test scores (4).

We then used two separate machine learning models – decision trees and random forests – to predict the scores of various groups of students, under three sets of features: {Revenues and Expenditures} (5a, 5b, 6a, 6b, 7, 11a, 11b, 12a, 12b), {Revenues, Expenditures and Race} (8a, 9, 13a), and {Revenues, Expenditures and Gender} (8b, 10, 13b). The decision tree model offered insights on which features provided the greatest information gain in predicting test scores in each scenario: State Revenue, Race, and Gender were the most important features in the three scenarios, respectively. We also anticipated and found that the decision tree model tended to perform better with more leaf nodes until a certain point, after which it tend to overfit the training data. The random forest, however, stabilized with more trees and performed as good as or better than the single decision tree in each scenario.

There are certain implications that arise from the results of this project. Based on the decision tree, we observe that higher revenue, regardless of the source, is a relatively strong predictor for higher academic scores. This implies that states should work toward acquiring more funding for educational purposes. The decision tree also suggests that the most effective way to spend any additional funding is on instructional expenditure, since this factor seems to be the most significant predictor of academic achievement among the different expenditure categories. Even with a limited budget, states should consider allocating more expenditure toward instruction in order to increase in students' academic performance. More in-depth data analysis and research would need to be performed in order to make more informed decisions, e.g. a state or a school with a certain demographic distribution can run our models for specific racial groups to determine a policy that would benefit it the most.

A Appendix 1 - System Description

The Python notebook that we used can be run on any system that can import popular libraries such as Numpy, Matplotlib, and Sklearn. We developed the code using a Deepnote notebook. To reproduce the results, run the cells in the notebook in order. Notebook is linked here:

https://deepnote.com/project/fe5f8678-0b92-4d57-abb5-4f6f6fd10d19

B Appendix 2 - Group Makeup

We (Jane Ahn, Haneul Shin, and Max Guo) each contributed to every part of the project as we intended in the beginning. This included the problem formulation, background research, data analysis, machine learning models, and the final write-up.

References

[1] Truckenmiller, A. J., Petscher, Y., Gaughan, L., & Dwyer, T. (2016). *Predicting math outcomes from a reading screening assessment in grades 3–8* (REL 2016–180). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs.

[2] James, L., Pate, J., Leech, D., Martin, E., Brockmeier, L. and Dees, E. (2011). ‘ Resource allocation patterns and student achievement’, *International Journal of Educational Leadership Preparation*, vol. 6, no. 4, pp. 1– 10.

[3] Baker, B. (2018). *How Money Matters for Schools.* Learning Policy Institute Research Brief.