Data Mining Final Project

Johnathan Bowman, Amal Kadri, and Alice Kemp

Spring 2022

Introduction

The subject of school performance has been heavily researched in the past with most studies coming to the conclusion that household income and racial/ethnic demographics are the most predictive factors of school performance. Students who come from households with predominantly high socioeconomic status tend to perform better than their peers who come from lower socioeconomic circumstances - this trend further aggregates to the school and district level with schools located in neighborhoods of higher socioeconomic status typically outperforming those in poorer areas, as measured by metrics such as standardized test scores, graduation rates, and college acceptance rates. However, in a state as racially diverse as Texas, how well do these trends explain over versus under-performance at the district level? In this report, we will analyze district-level data gathered from the Texas Education Agency during the 2019 to 2020 school year covering student and faculty demographics, SAT/ACT test scores, median household income, enrollment, and graduation rates. From this data, we will identify districts that over or under perform their predicted outcome score and use machine learning techniques to analyze the correlated variables responsible. By doing so, we hope to uncover the key factors that make a district out or under perform other districts with similar demographic makeup. By doing so, we will hopefully deepen our understanding of school performance and use our findings to narrow the achievement gap between districts in Texas and beyond.

Methods

The subject of school performance has been heavily researched in the past with most studies coming to the conclusion that household income and racial/ethnic demographics are the most predictive factors of school performance. Students who come from households with predominantly high socioeconomic status tend to perform better than their peers who come from lower socioeconomic circumstances - this trend further aggregates to the school and district level with schools located in neighborhoods of higher socioeconomic status typically outperforming those in poorer areas, as measured by metrics such as standardized test scores, graduation rates, and college acceptance rates. However, in a state as racially diverse as Texas, how well do these trends explain over versus under-performance at the district level?

Data

The Data we used for our analysis was gathered primarily from the TEA, USDA, and the Census. We gathered TEA data on educational outcomes (Graduation Rates, Standardized Test Scores, Attendance, etc.), and school-district-level covariates (Student-Teacher Ratio, Teacher Pay, Disciplinary Activity, School Meals, etc.). We merged these variables onto socioeconomic indicators such as poverty rate, median income, and education levels. This aggregation had to be done at the county level, which means some researcher bias had to be introduced when deciding how to most appropriately aggregate outcomes data gathered at the district level up to the county level. All said, we had 87 covariates for analysis on 8 outcome variables: * `ERW`: Average SAT evidence-based reading and writing score * `MATH`: Average SAT mathematics score * `TOTAL`: Average SAT total score * `ann_grad_count_1819`: The number of students who graduated during the 2018-19 school year, including the summer of 2019. This count includes 12th grade graduates, as well as graduates from other grades. * `avg_sat_1819`: The average of SAT total scores (a sum of evidence-based reading and writing and mathematics) for 2018-19 graduates who took the SAT divided by the number of

2018-19 graduating SAT examinees. Total scores for the SAT range from 400 to 1600 for evidence-based reading and writing and mathematics combined. Total score for each examinee is calculated based on the best section scores from all SAT tests taken by the examinee anytime during their high school years. * `avg_act_1819`: The average of ACT composite scores (an average of English, mathematics, reading, and science), created by summing the composite scores for 2018-19 graduates who took the ACT divided by the number of 2018-19 graduating ACT examinees. Scores on each of the ACT sections range from 1 to 36. * `Above_Crit_Rate`: Percent of graduating examinees receiving SAT total scores of 1180 or higher * `Above_TSI_Both_Rate`: Percent of graduating examinees meeting the college-ready graduates TSI criteria for the SAT on both ELA and mathematics An unfortunate issue when it comes to using Education data from a data analytics perspective is the issue of "Masking". Because of privacy considerations, schools must take care to not release any data that could be potentially used to identify specific students. As an example, if there are only a handful of Hispanic students in a given school, the school might have to mask any statistics on the racial/ethnic breakdown of educational outcomes in order to prevent the possibility that the data can be easily used to find the scores, economic, or disciplinary status of specific students. In aggregate, this means that there are a significant number of N/As and masked codes that had to be dealt with in order to proceed with the analysis. As a result, our data is biased slightly in favor of being more accurate for schools with larger, more diverse school populations, and may not capture all the useful variation for smaller school districts. However, given that this is a limitation with all publicly available, and our goal was to identify which patterns/abnormalities in the data we *could* see, rather than a more rigorous causal analysis, this seemed to us an acceptable constraint.

Analysis

To start our analysis, we first merged our data and then created an aggregated "outcome" variable to measure district performance across a variety of metrics including SAT ERW and Math scores, previous year graduation rate, previous year SAT and ACT scores, and percentages of graduating students meeting the college-ready measures for SAT scores. To create this outcome variable, Principle Component Analysis (PCA) of rank 1 was used to reduce the dimensionality of our outcome variables and create one PC of weights that maximizes the variance found in the original outcome data. The resulting PC1 in Table 1 shows large weights from every performance metrics, with the minimal exception of previous year graduation rate.

Table 1: PCA of performance metrics

	PC1
Above_TSI_Both_Rate	0.2617
Above_Crit_Rate	0.1985
avg_act_1819	0.2186
avg_sat_1819	0.2978
ann_grad_count_1819	0.0931
Total	0.5001
Math	0.5006
ERW	0.4961

OLS Model

Next, we fit the loadings of PC1 on to our original district-level data resulting in a singular outcome performance variable. The next step in our analysis was to create a simple linear OLS regression of our new weighted outcome variable on a selection of covariates believed to be most indicative of performance based on previous literature. These features included median household income measured as a percent of the statewide median, the child poverty rate in each county, average faculty salary, along with percentages for student population by race (Black, Hispanic, White, Asian, Native American, Pacific Islander, and multi-racial). Child poverty percent effects has both the resources a child has access to and is associated with other deviant

behavior like truancy. Percent state median income defines the percent of the mean income for the county of the states median income. For instance, a county with 1.5 is a county with a mean income 50% higher than the state median. Average school salary describes the amount of resources that are going to schools. As observed in the model results in Table 2, we find that percent of state median household income and average school salary are highly statistically significant in predicting district performance. This result supports other studies on the subject, however, this does not explain why some school districts who would be predicted to perform at a certain level do not. These districts are represented in the residuals of our linear model, which we will next use as the outcome variable in a series of machine learning models.

Table 2: Linear Model - PC1

	Dependent variable:
	PC_outcome
child_poverty_percent	0.023
	(0.017)
pct_state_median_HH_income	0.018***
	(0.005)
avg_salary_school	0.00002***
	(0.00000)
st_pct_black	−0.177
	(0.743)
st_pct_hisp	−0.165
	(0.743)
st_pct_white	−0.157
	(0.743)
st_pct_asian	−0.065
	(0.743)
st_pct_native	−0.149
	(0.741)
st_pct_pac	−0.079
	(0.753)
st_pct_mult	−0.084
	(0.745)
Constant	12.422
	(74.349)
Observations	1,060
R^2	0.137
Adjusted R^2	0.129
Residual Std. Error	1.762 (df = 1049)
F Statistic	16.672*** (df = 10; 1049)
Note:	*p<0.1; **p<0.05; ***p<0.01

Decision Tree

For our decision tree and random forest models, we first scaled our variables that used counts as to normalize based on student enrollment in each district. Then, a decision tree was fitted with our residuals from the OLS model as the outcome variable of interest. As seen in Figure 1, the most important split was on previous year's total expenditure per student, with districts spending more per student having higher residuals, i.e. over performing their predicted performance. Interestingly, the model did not pick up current year's expenditure per student as a significant split, although a high "fund balance", or the remaining district funds at the end of the school year, was the next significant predictor for over-performing schools. Next, we observe average salary for central faculty (i.e. Principals, Vice Principals, and other administrative positions), with a lower salary indicating negative residuals, or under-performance. Racial diversity was another significant factor with

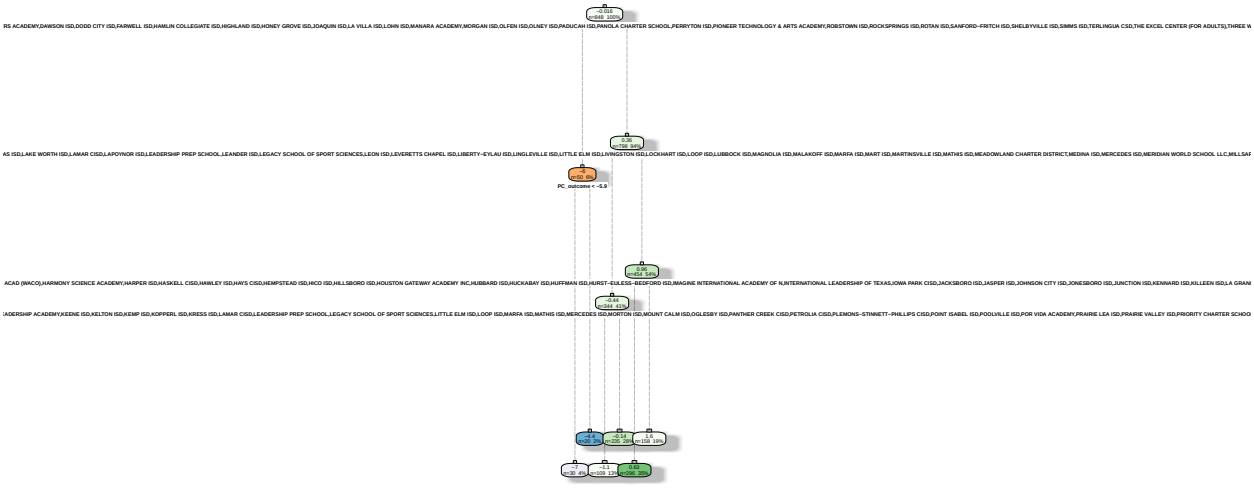


Table 3: CART: Variable Importance

feature	importance
DistName	2492.92718
PC_outcome	2160.39204
FreeElig_mean	181.14928
st_pct_careertech	112.92529
st_pct_ecodis	69.02964
exp_pct_na	58.15815
st_pct_native	54.80287
exp_pct_basiced	37.64176

Random Forest

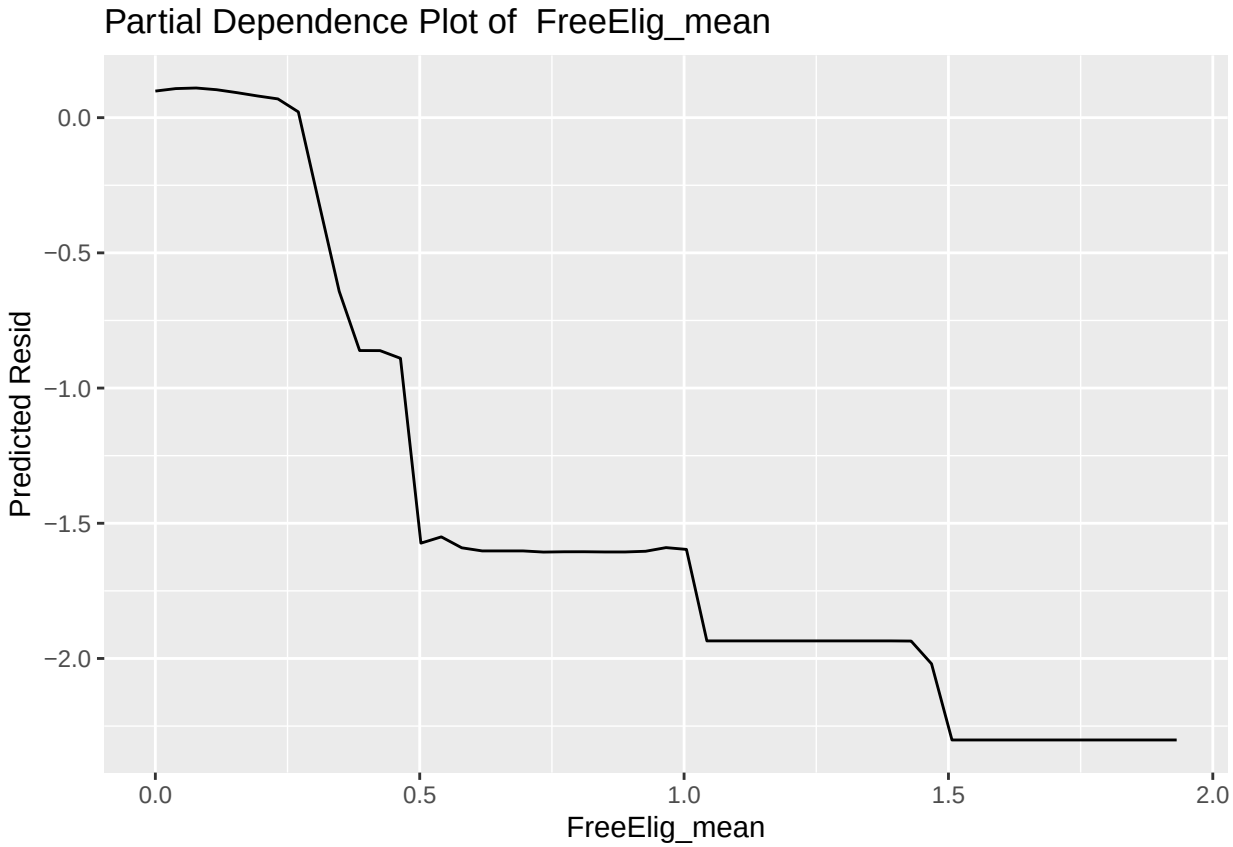
Our next step in modeling over and under performing school districts was to use a random forest on the residuals found in the linear OLS regression. Using a random forest with all features as possible splits, we created a model with a root mean squared error of 1.17, lower than the error from the CART regression above. The most important variables selected in the new random forest model were the average students eligible for the free meal program, the percent of students in career/tech programs, the rural index of the district,

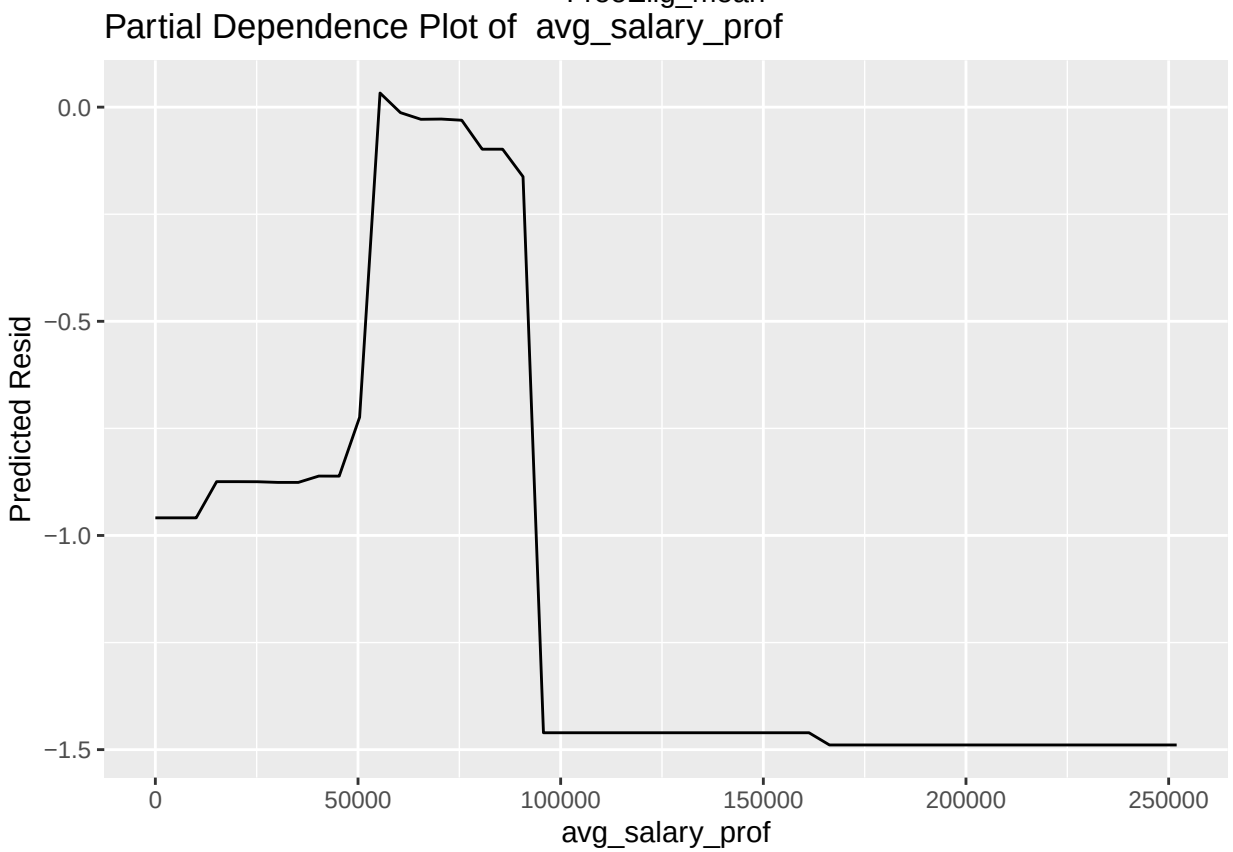
the average salary for district employees working in professional services, and the district's teacher turnover rate. Looking at partial dependence plots of these top five features, we can interpret the marginal effects of increasing these feature values on the residuals found in the OLS model, thus giving an easily interpretable estimate of how certain district characteristics lead to over and under performance.

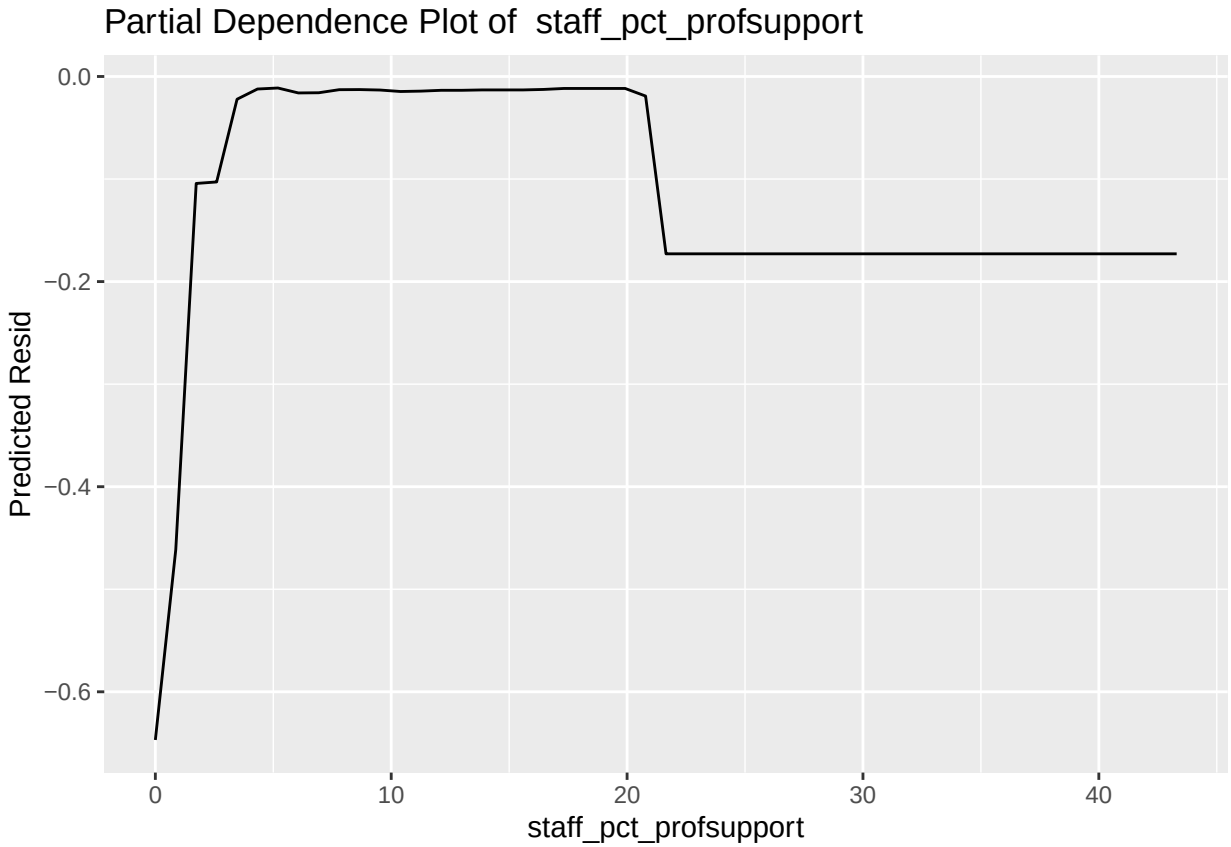
The first partial dependence plot demonstrates the negative correlation between the number of students eligible for free meals, based on their household income, and the predicted residuals. The graph implies that districts with relatively low (0 to 0.25) shares of students eligible for free meals tend to have positive residuals, or over performance. However, as this share increases, the residuals become more negative, implying that districts with larger shares of students eligible for free meals tend to under-perform their predicted performance. Moving on to the share of students in career/tech programs, we observe that districts with larger shares actually tend to under-perform their predicted outcomes thus implying a negative correlation. Next, we look at the partial dependence of RUC code, which measures how rural a district is with 1 being the most rural and 9 being the most urban. We find varying results for this variable, with all codes correlated with negative predicted residuals - further analysis in later sections of this report will attempt to uncover the true correlation. Moving on to the average salary for professional services, we find results for negative predicted residuals only and find that there is an interesting dependence with the marginal effect of increasing salary first moving predicted residuals towards zero before dropping steeply. Overall, average professional services salaries in a district above $75,000 appear to be correlated with under-performing districts, although this trend bottoms out quickly. Lastly, we observe the partial dependence of teacher turnover rate with lower rates appearing to be correlate with over-performing districts, while higher rates above approximately 18% being increasingly negatively correlated with under-performance.

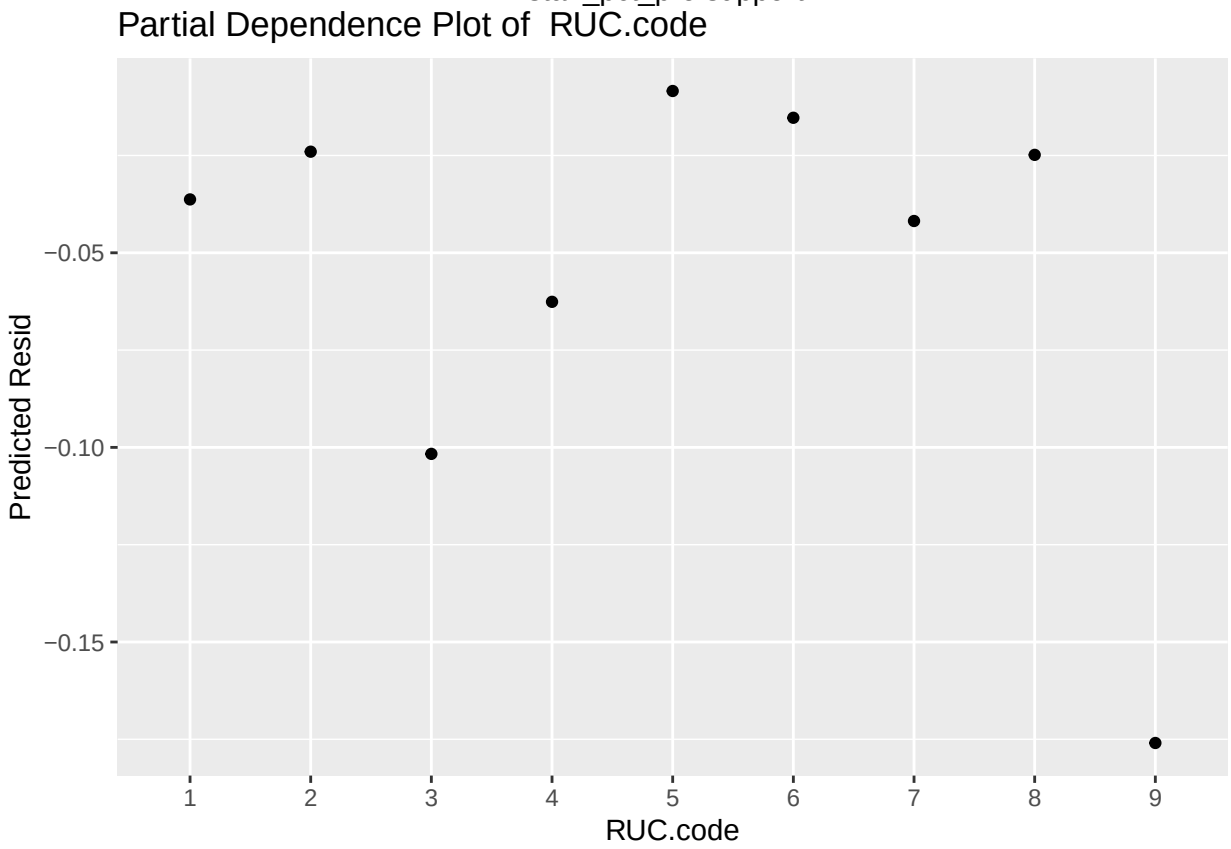
Table 4: Random Forest: Variable Importance for All Districts

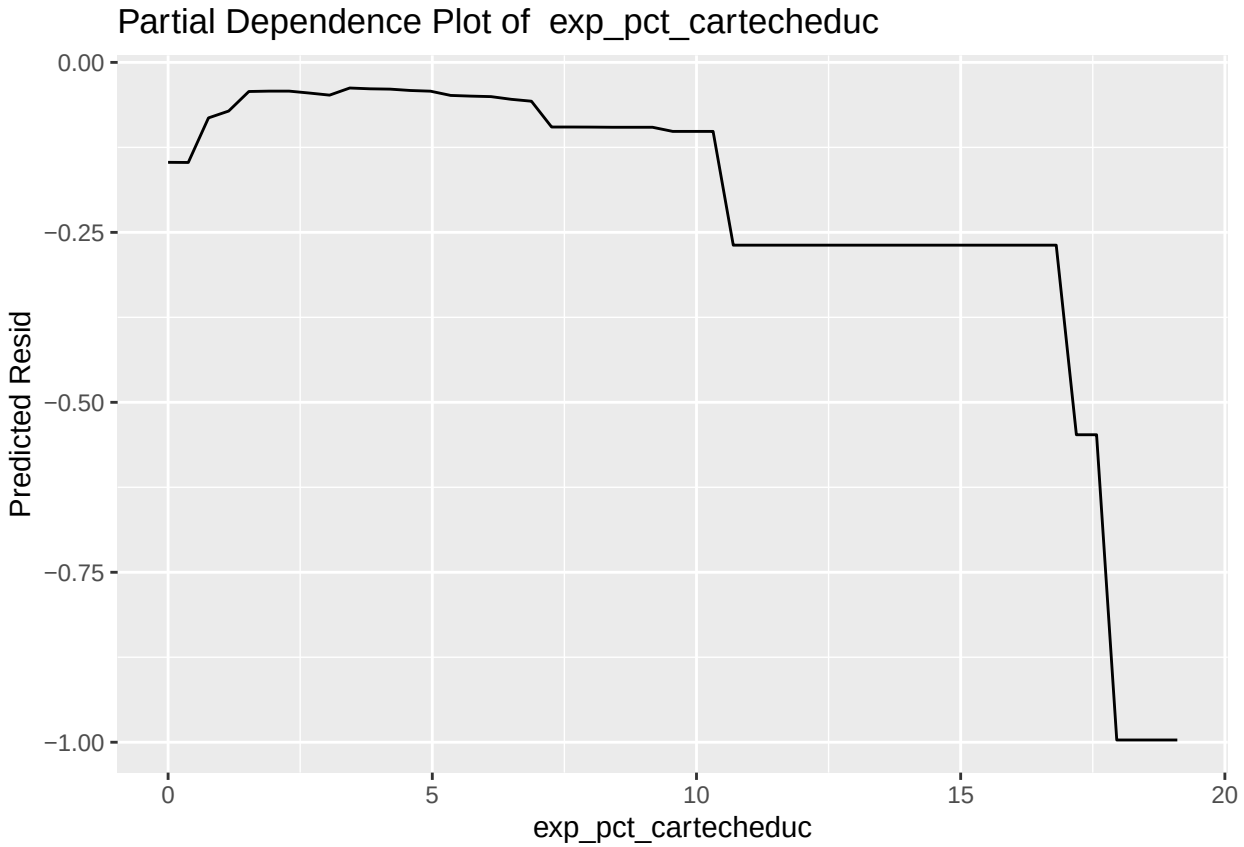
feature	importance
FreeElig_mean	303.90963
avg_salary_prof	157.42754
staff_pct_profsupport	106.44108
RUC.code	98.44220
exp_pct_cartecheduc	77.82229


Partial Dependence Plot of FreeElig_mean
Predicted Resid
0.0
−0.5
−1.0
−1.5
−2.0
0.0
0.5
1.0
1.5
2.0
FreeElig_mean


Partial Dependence Plot of avg_salary_prof
Predicted Resid
0.0
−0.5
−1.0
−1.5
0
50000
100000
150000
200000
250000
avg_salary_prof


Partial Dependence Plot of staff_pct_profsupport
Predicted Resid
0.0
−0.2
−0.4
−0.6
0
10
20
30
40
staff_pct_profsupport


Partial Dependence Plot of RUC.code
Predicted Resid
−0.05
−0.10
−0.15
1
2
3
4
5
6
7
8
9
RUC.code

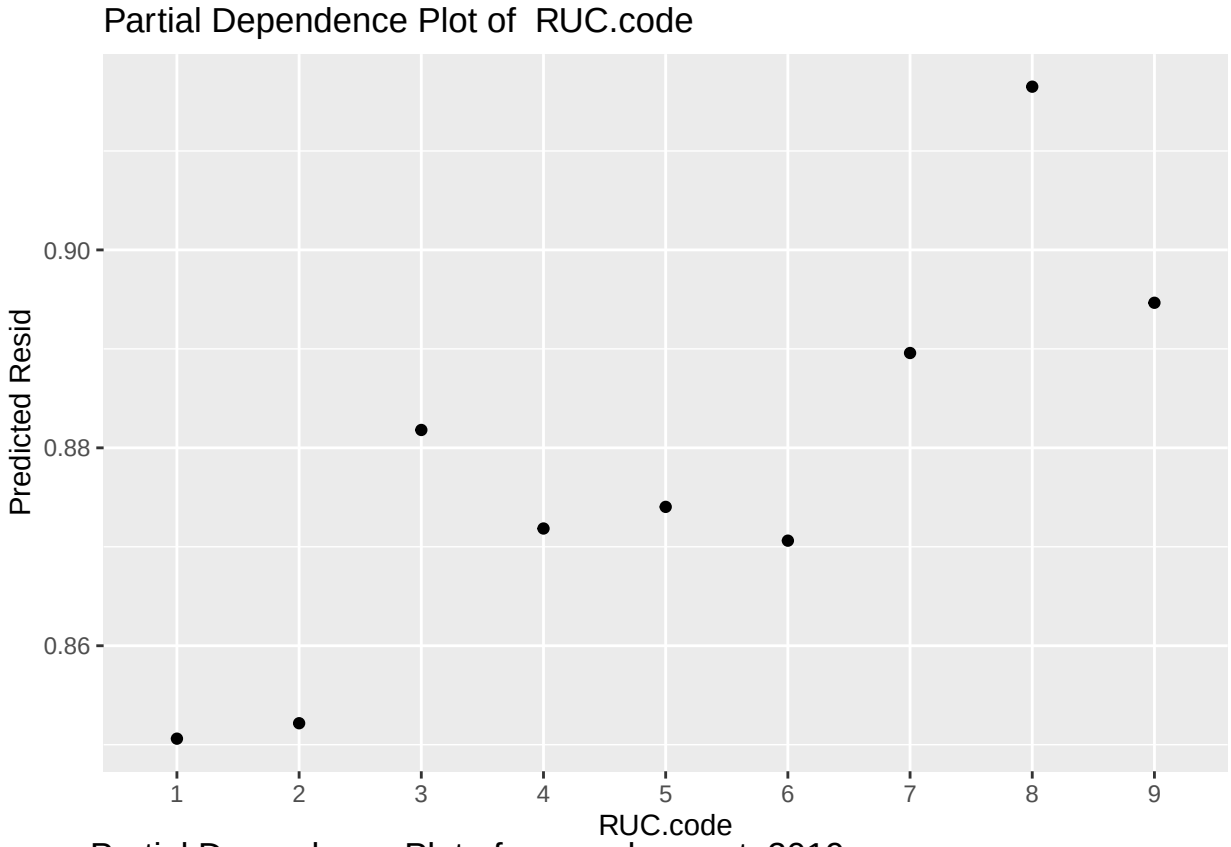


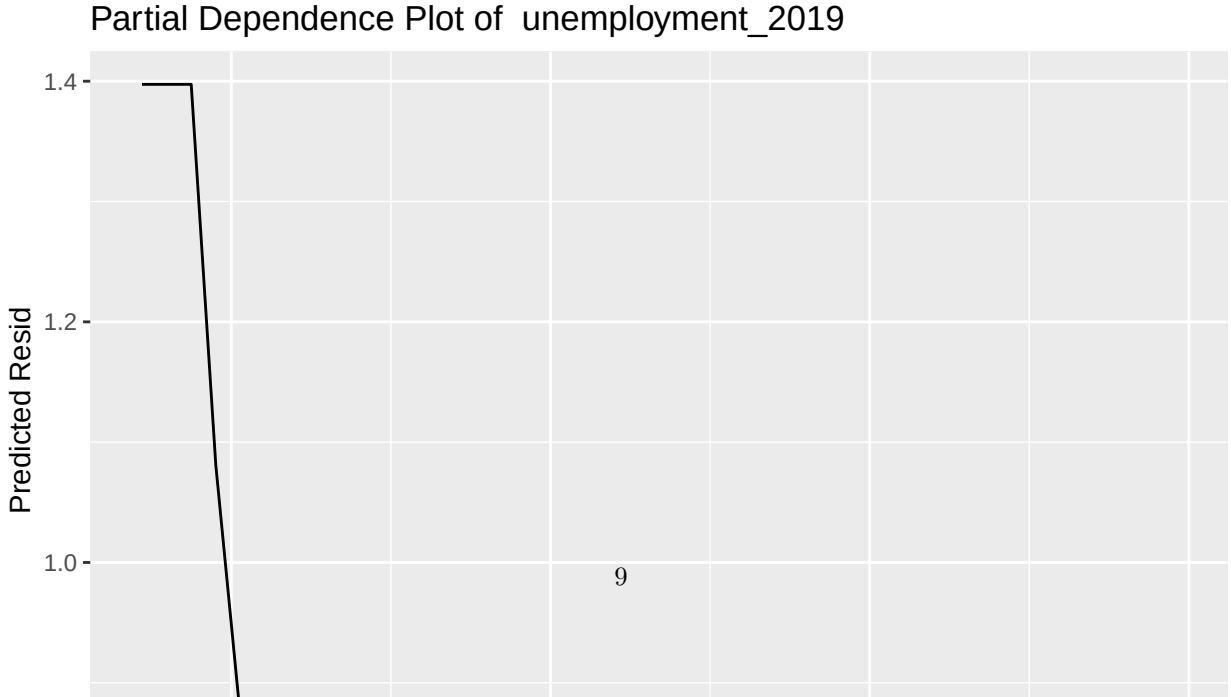
To further deepen the model, we next split our data into districts with positive and negative residuals, representing districts that over performed relative to their predicted performance and those who under performed relative to their predicted performance. We used random forests of 50 trees to identify the features most predictive of an over or under performing district and created variable importance plots to visualize the partial dependence of the most important variables for over and under performing districts. For the over performing school districts with positive residuals, we found the most important feature to be unemployment in 2020, with lower levels of county unemployment generally yielding higher performance than predicted. Similarly, we found that the next most important feature was the ten-year county population percent change, with districts in counties with decreasing populations actually yielding higher over-performance than districts in counties with increasing populations. Next, we found that the percentage of students who were economically disadvantaged had a negative effect on over-performance with over-performing districts with the highest percentages of economically disadvantaged students seeing the lowest gain over predicted performance. This seems to align with findings in other studies that state schools with higher populations of students who come from lower socio-economic households having lower relative performance. On the other hand, the percentage of students who are deemed "gifted and talented" appears to have a positive effect on over-performance, with districts with higher percentages of gifted students having higher than predicted performance scores. However, this feature is likely prone to sample bias, with certain districts that are likely already high-performing testing their students at a higher level than lower performing districts. The final feature in the top five is RUC code, a measure of how rural versus urban a school district is on a scale from 1 to 9. As seen in the figure below, we observe that there is overall in positive and increasing effect of RUC on over-performance, with more urban districts having higher than predicted performance than more rural districts.

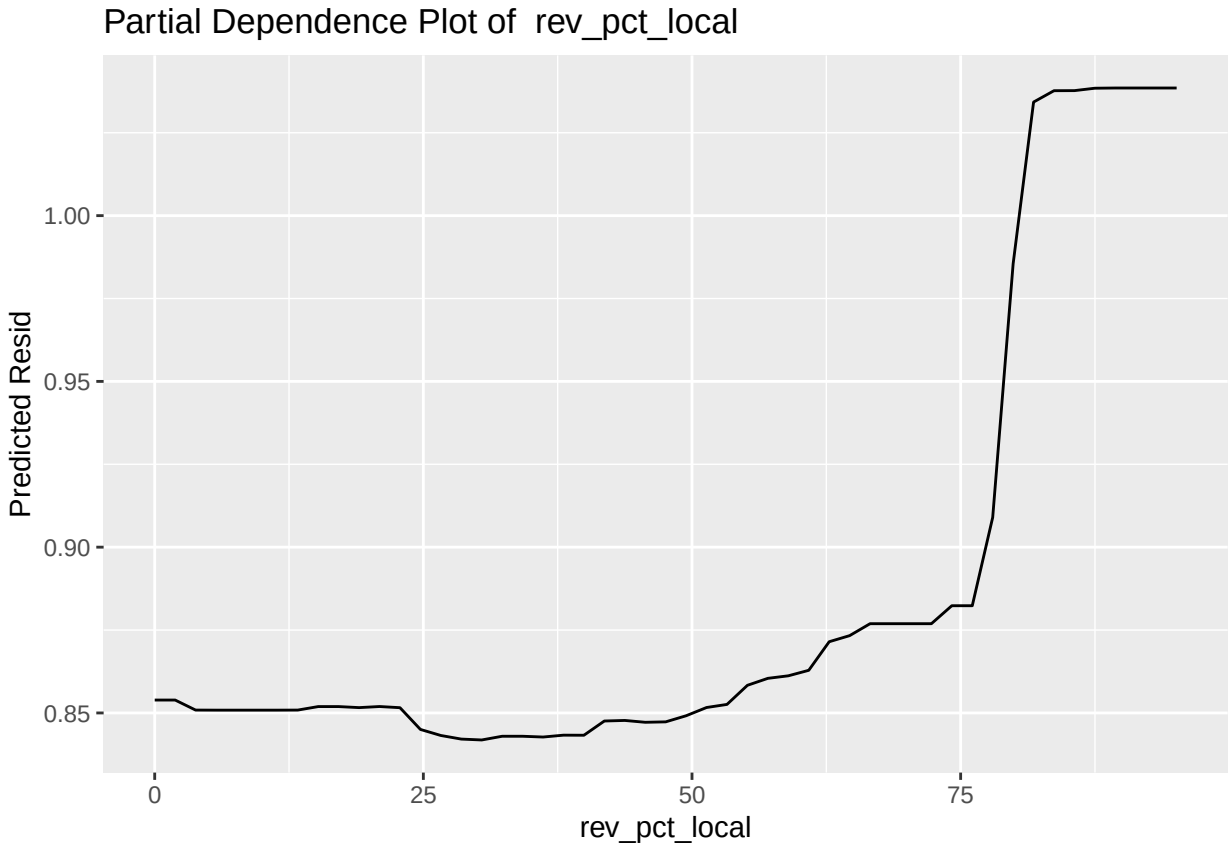
Table 5: Random Forest: Variable Importance for Overperforming Districts

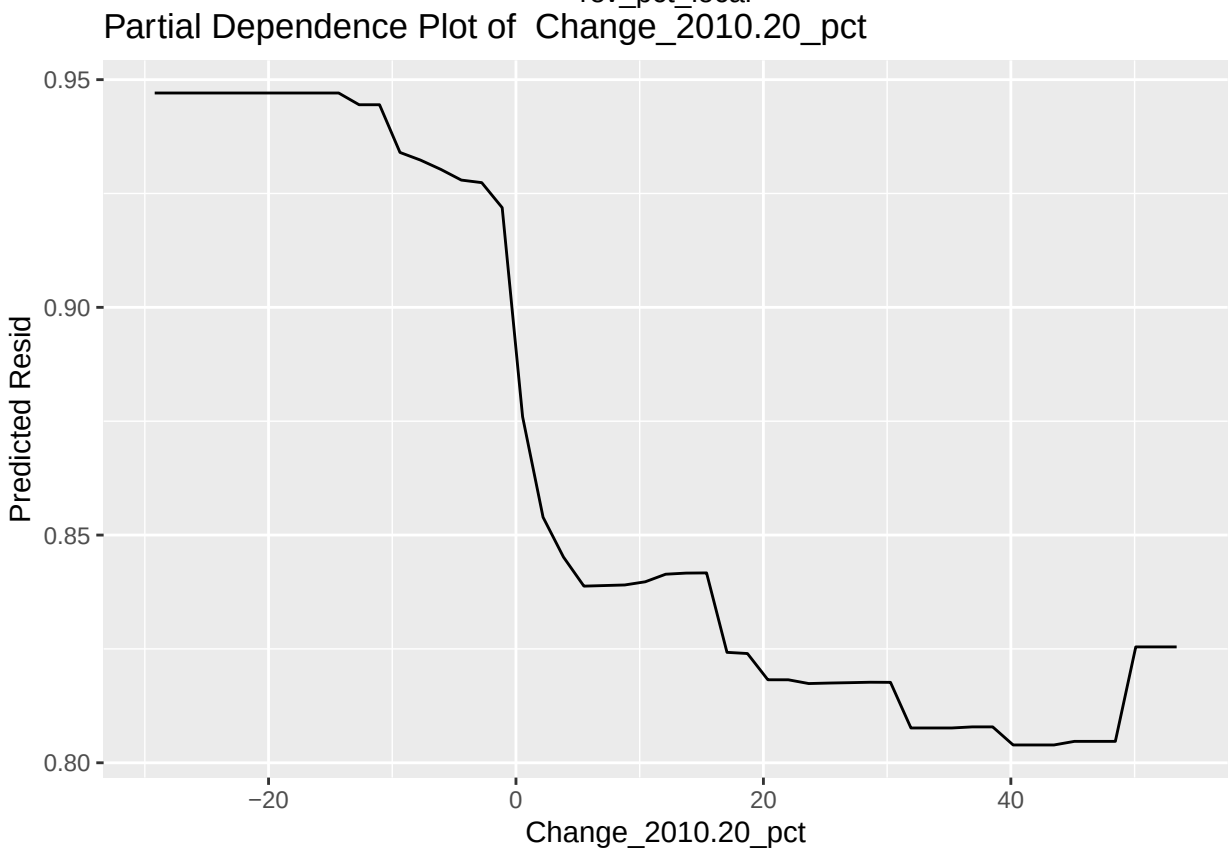
feature	importance
RUC.code	7.358867
unemployment_2019	6.917333
rev_pct_local	6.632933
Change_2010.20_pct	5.704039
st_pct_speced	5.127102

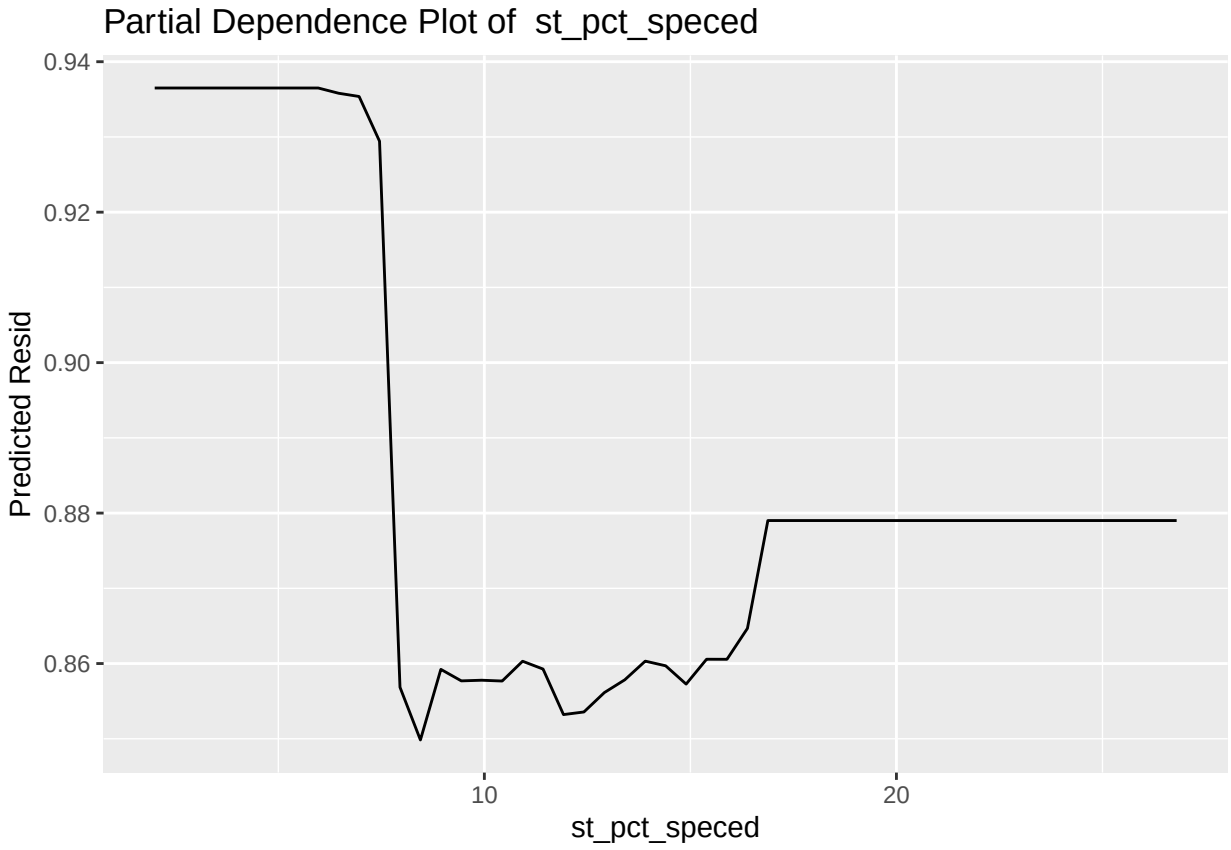
Overperforming Districts






Partial Dependence Plot of rev_pct_local
Predicted Resid
1.00
0.95
0.90
0.85
0
25
50
75
rev_pct_local

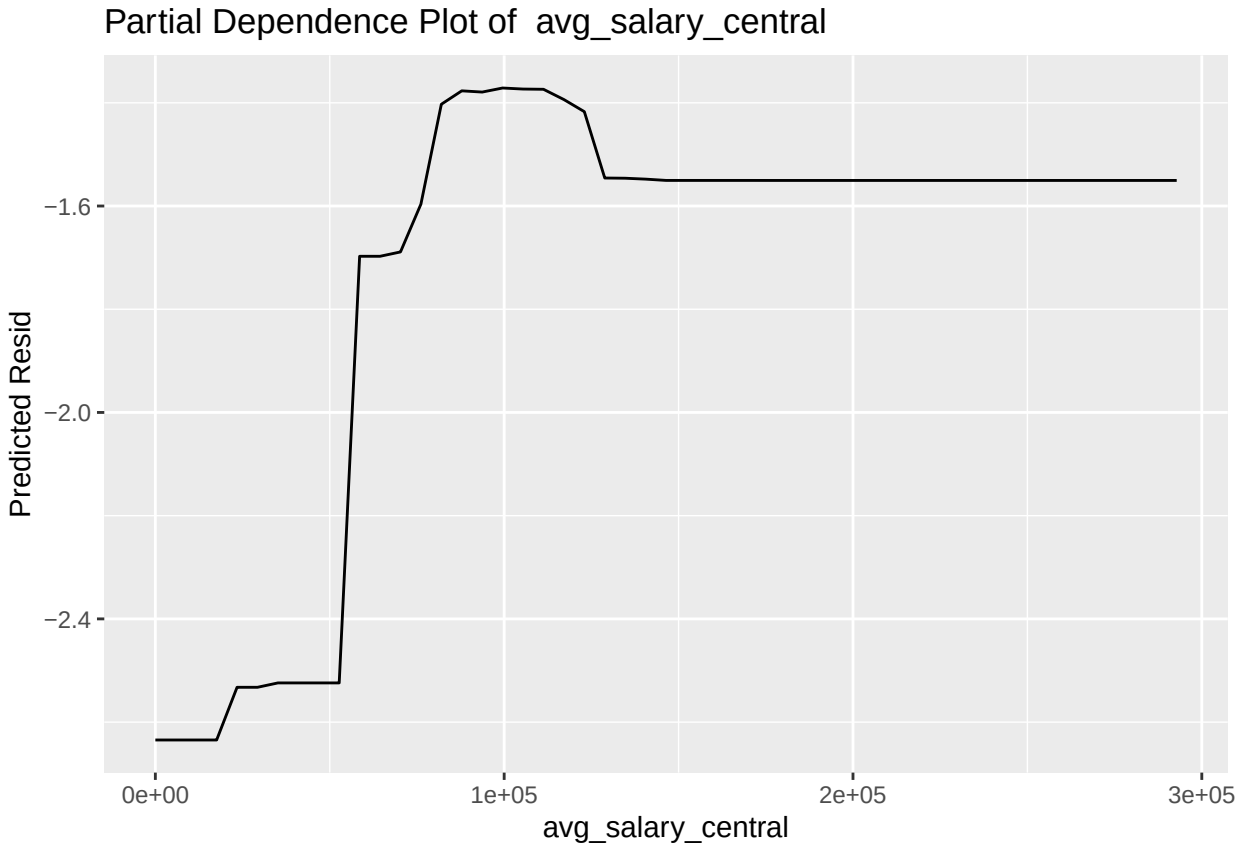

Partial Dependence Plot of Change_2010.20_pct
Predicted Resid
0.95
0.90
0.85
0.80
−20
0
20
40
Change_2010.20_pct

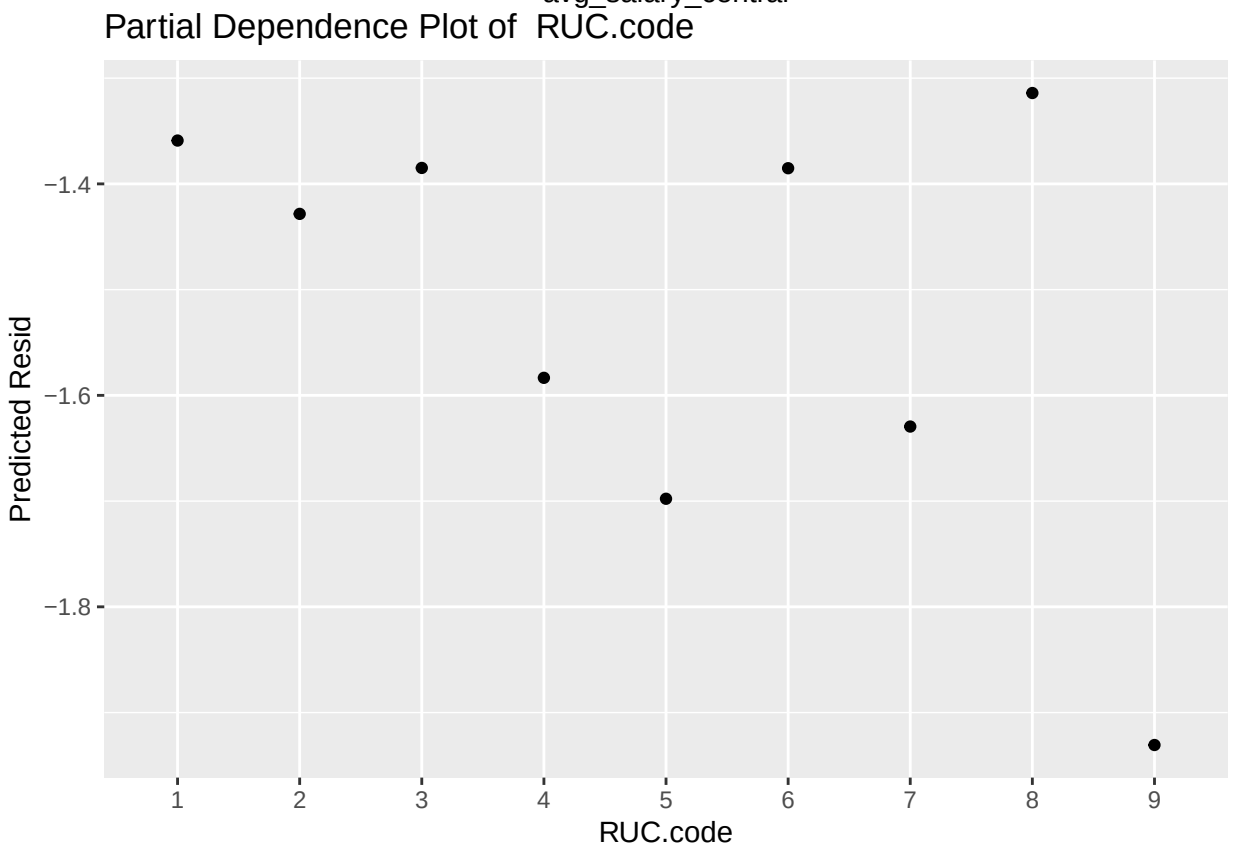


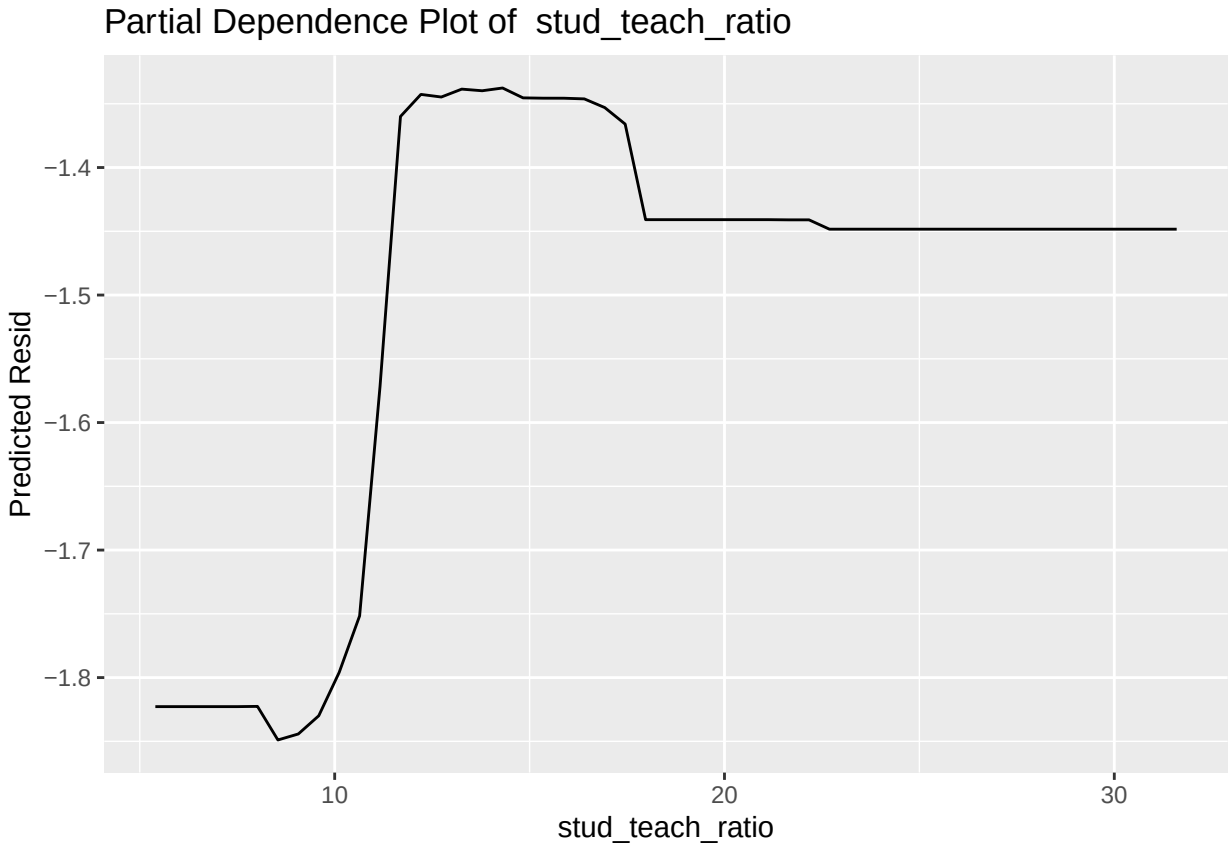
Underperforming Districts

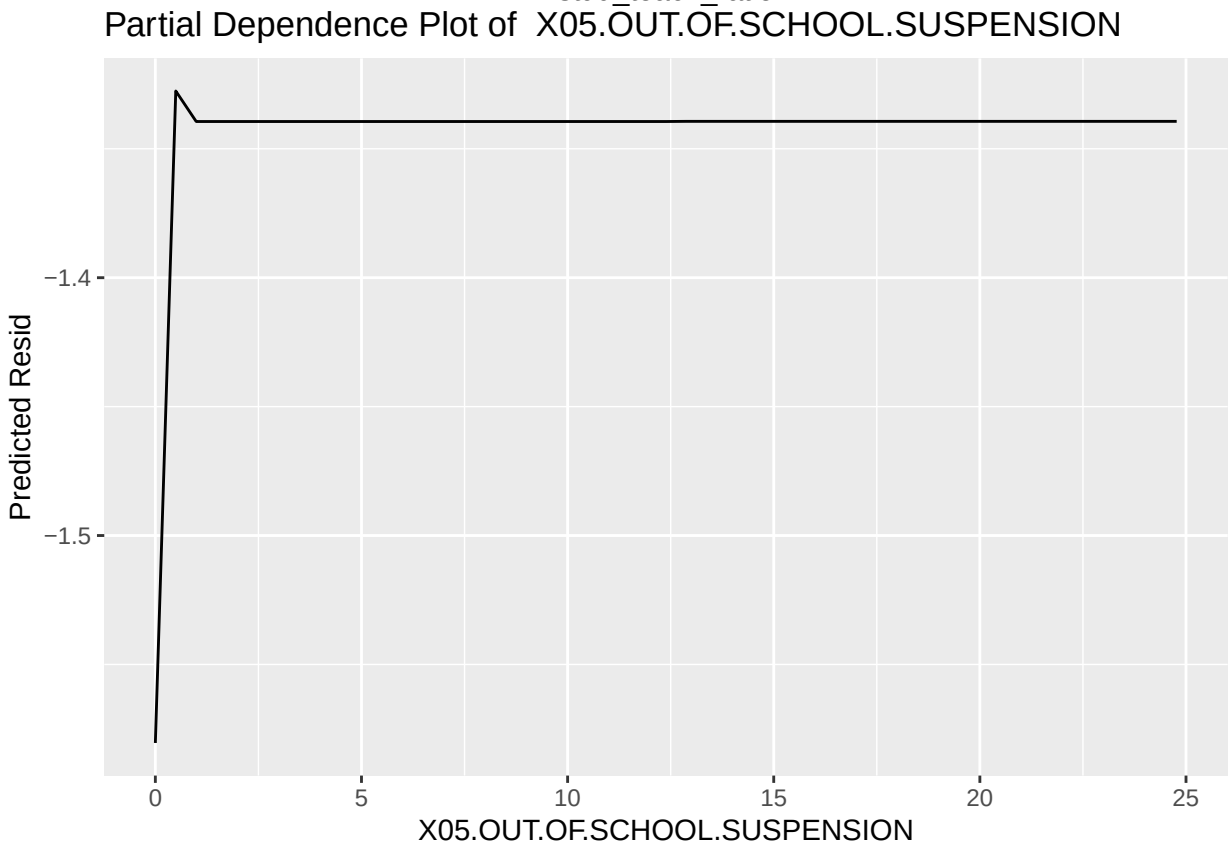
Table 6: Random Forest: Variable Importance for Underperforming Districts

feature	importance
avg_salary_central	62.94977
RUC.code	54.58913
stud_teach_ratio	53.45356
X05.OUT.OF.SCHOOL.SUSPENSION	50.13959
avg_salary_prof	49.35879


Partial Dependence Plot of avg_salary_central
Predicted Resid
−1.6
−2.0
−2.4
0e+00
1e+05
2e+05
3e+05
avg_salary_central


Partial Dependence Plot of RUC.code
Predicted Resid
−1.4
−1.6
−1.8
1
2
3
4
5
6
7
8
9
RUC.code


Partial Dependence Plot of stud_teach_ratio
Predicted Resid
−1.4
−1.5
−1.6
−1.7
−1.8
10
20
30
stud_teach_ratio


Partial Dependence Plot of X05.OUT.OF.SCHOOL.SUSPENSION
Predicted Resid
−1.4
−1.5
0
5
10
15
20
25
X05.OUT.OF.SCHOOL.SUSPENSION

Findings

Conclusion

A severe limitation to the data was the masking of many of the student fields. The reason the data was masked is because it was available to the public. If certain values were made available it would give anyone the means to impute the identities of students who are in these groups. An unfortunate side effect is the elimination of many fields which could have been descriptive for controlling and predicting education outcome. This included but is not limited to fields like immigration or migrant status, dyslexia, foster care status, homelessness, or military connected. These were among the fields which had to be removed in the cleaning process due to NA values so the variation in these fields was not able to be accounted for. A massive improvement to this study would be to conduct it with access to this masked data, which may need to occur within a state regulatory institution which has clearance to view the information.