Data Mining Final Project

Johnathan Bowman, Amal Kadri, and Alice Kemp

Spring 2022

Abstract

In an effort to identify patterns in School Performance and student educational attainment, we chose to look at the ways that schools defy prevailing trends. In order to do this, we used OLS to regress classical economic and demographic variables on a School Performance Index that we defined through a Principal Components Analysis of different educational outcomes. We then used Random Forest Regression to predict the *Residuals* generated by our original linear model. Our analysis is broken down into "overperforming" and "underperforming" schools, schools with positive and negative residuals respectively. We find that Urbanization is a strong but inconsistent contributor educational outcomes for both overperforming and underperforming schools, and that teacher and staff pay are both correlated with higher performance. Overall our findings lead us to the somewhat unsurprising conclusion that teacher pay tends to be correlated with higher performing students; and that expenditures that take away from teacher pay, either by deliberate reallocation or mismanagement, tend to reduce outcomes. Specific student aid, such as Special Ed Teachers, or Gifted-and-Talented programs, are also correlated with higher attainment for students. Our findings our limited by the masking of student data for privacy considerations, and the relative dearth of data on schools and counties with lower populations (predominantly more rural ones), and would benefit from extension to a national level and more granular data.

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

For the first stage of this study, we took various education outcomes like graduation rate and standardized test scores and use principal component analysis to compress these outcomes into a single principal component education score. Then we defined a linear model with three critical socioeconomic variables: median household income measured as a percent of the statewide median, the child poverty rate in each county, average faculty salary with additional controls for demographic makeup. We regressed this model against the education score

and collected the residuals. We then ran random forest models containing 82 different features against the positive residuals, negative residuals, the entire population of residuals, and on final random forest against the original education outcome principal component. We recorded and analyzed the results and discussed future directions for this line of inquiry.

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 relatively 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

First, we fit the loadings of PC1 on to our original district-level data resulting in a singular outcome performance variable. We then 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 a district. After fitting the OLS model, we then use the residuals, or the difference between predicted performance versus actual performance, as the outcome variable for a random forest model. This machine learning method uses the aggregate results of "n" individual, uncorrelated decision trees to minimize overfitting of any singular tree. We then further our analysis by running separate random forest models on the positive and negative residuals found in our original OLS model, with positive residuals representing districts that over perform relative to their predicted outcome, and negative residuals representing districts that under perform. In doing so, we hope to isolate any variables that affect over performing versus under performing districts. For all random forest models, root mean squared error was calculated using 80% train-test splits, and results are visualized using partial dependence plots of the top five most important variables to each model.

Findings

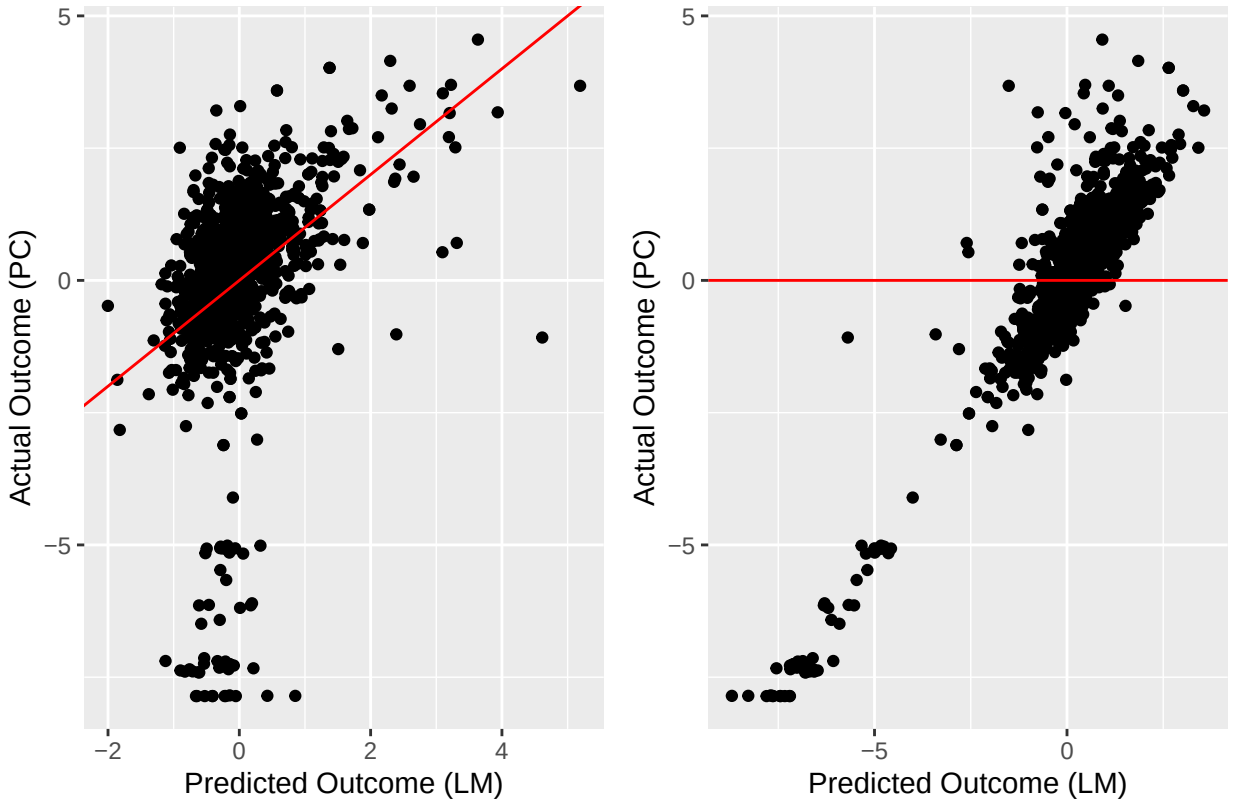
OLS Model

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

	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

Fig 1: Linear Model Predictions and Residuals



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 the scaled share of students in a district eligible for free meals, which is based on a student's household income level. For over-performing districts with positive residuals, we see a share less than 31%, indicating that these districts tend to have higher average household income. Over-performing districts with a predicted average residual of 0.38 also have lower un-allocated expenditure levels, scaled per student, and higher percentages of bilingual students. For districts with negative residuals, we find somewhat contradictory results, with RUC codes of both 4 and 9 leading to some of the most under-performing districts. RUC codes measure how rural a district is, with 1 being the most urban and 9 being the most rural. Thus, according to our decision tree, the most rural districts and those in the middle tend to be the most under-performing. We also see that lower expenditure on high school programs leads to higher levels of underperformance, which seems to fit classic theories relating to funding's effect on school performance. Finally, smaller shares of discipline records per student appears to predict better performing districts, however the predicted residuals for both sides of this node are negative. Since we are concerned with overfitting our data, we will next use random forest models to minimize this risk.

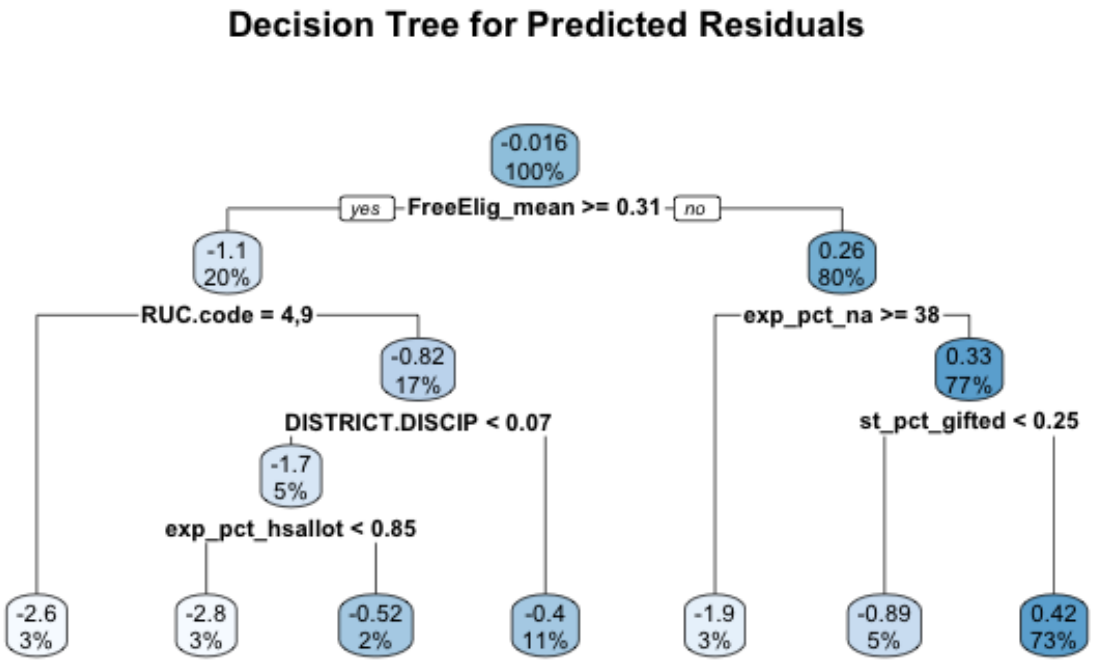


Table 3: CART: Variable Importance

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 1.53, lower than the error from the CART regression above of 1.66. In addition, when we plot our predicted residuals from the random forest model versus the actual residuals calculated from the linear OLS model, we see a relatively close fit along the 45 degree line with a few observations being largely underpredicted. However, most of our data appears to fit well, indicating that our random forest model does a sufficient job at predicting over and underperformance relative to our original linear residual findings. 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 urban and 9 being the most rural. 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 unallocated expenditure per student - actual expenditures that districts spent, but did not distribute into other programs. We find an interesting trend for this variable, with over-performance linked to districts spending less than 28% of their expenditures on un-allocated spending and an increasingly negative correlation with under-performance after this threshold.

Table 4: Random Forest: Variable Importance for All Districts

feature	importance
FreeElig_mean	273.40828
st_pct_careertech	133.35813
RUC.code	103.75358
avg_salary_prof	96.65566
exp_pct_na	79.31202

Fig 3: Predicted Residual vs. Actual Residuals

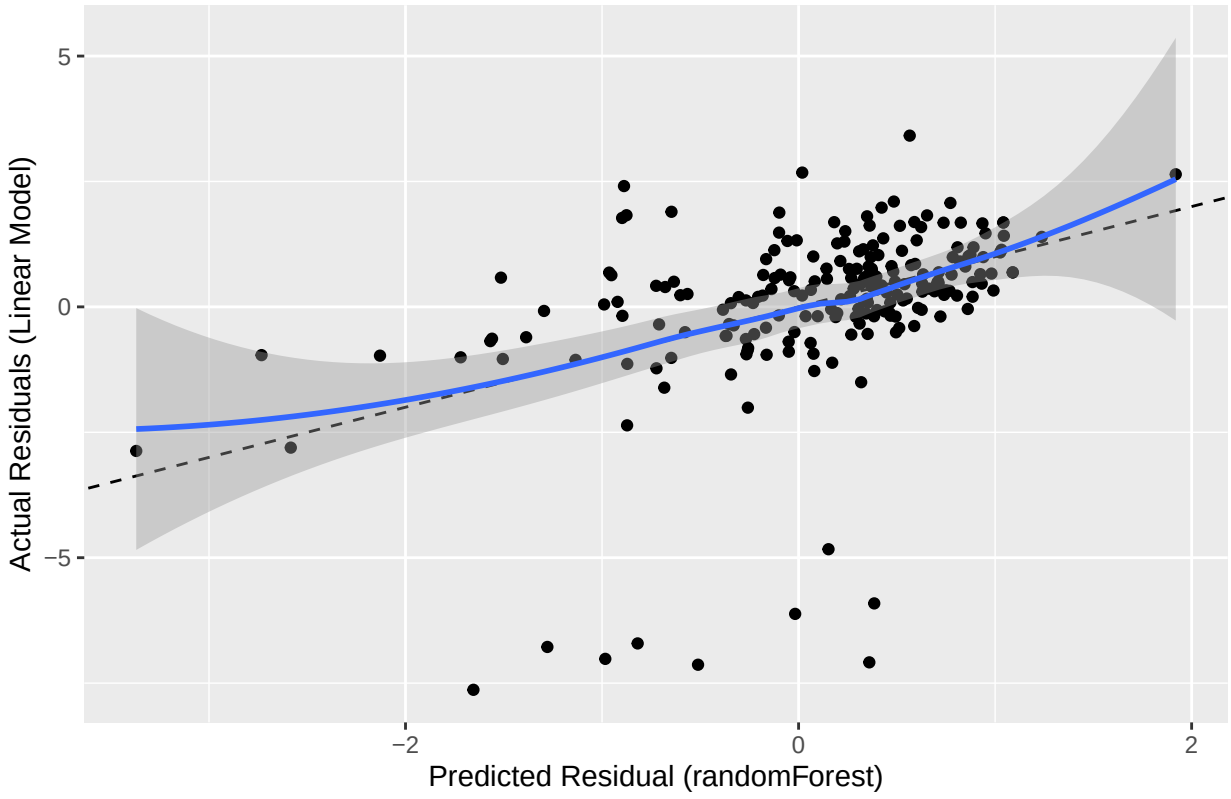
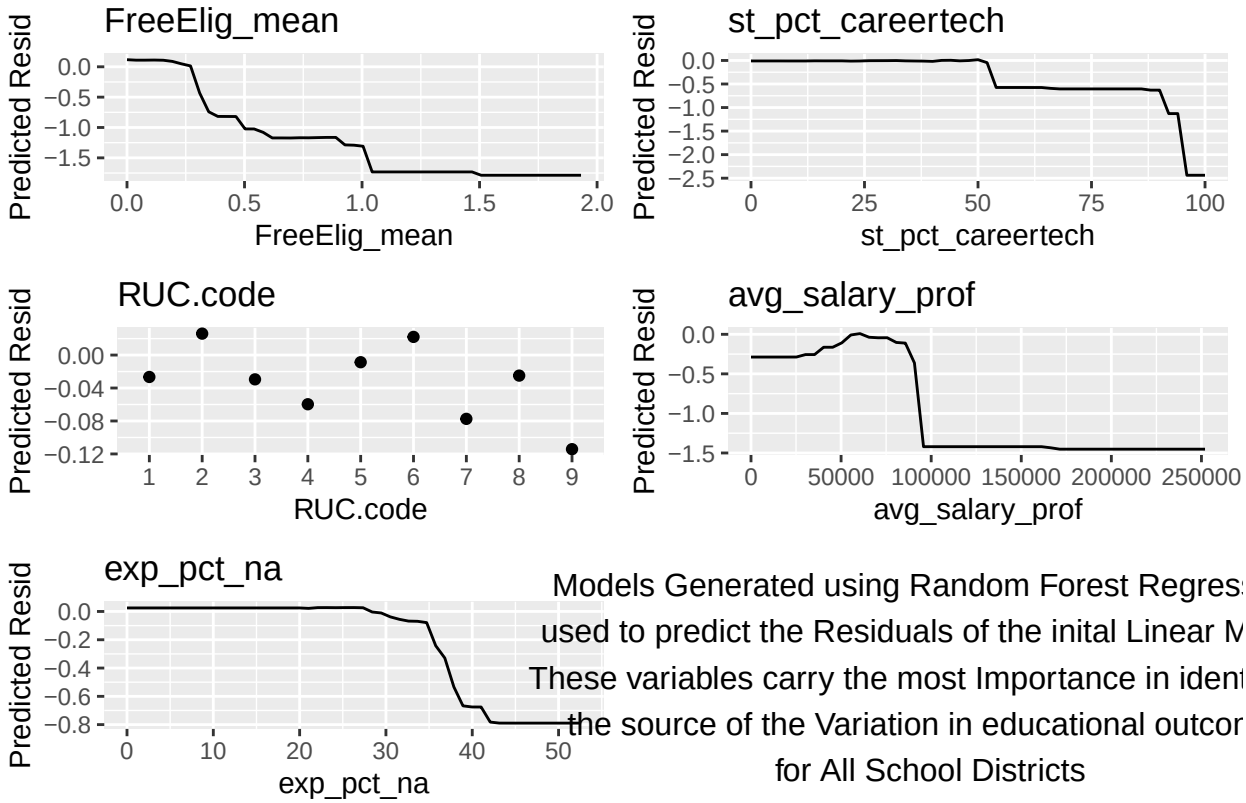


Fig 4: Partial Dependence Plots for All Schools



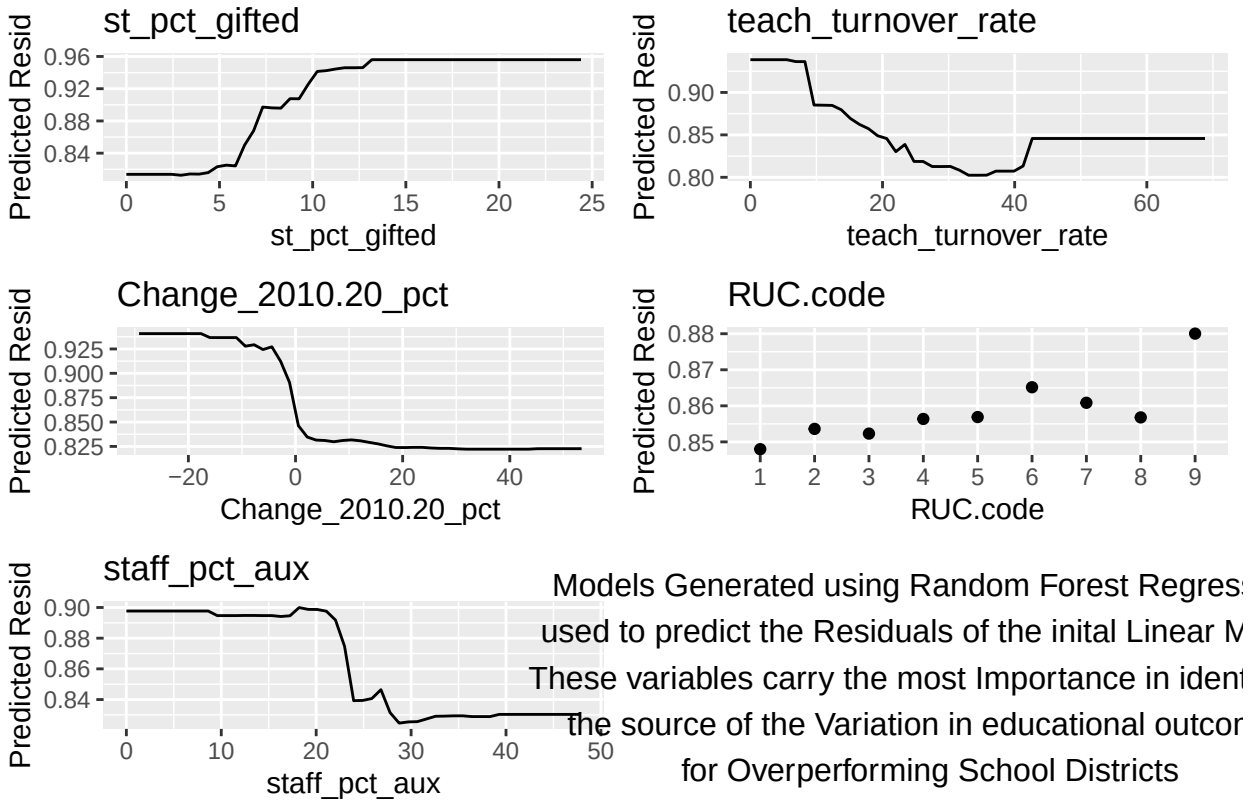
To further our analysis, we next split our data into districts with positive and negative residuals, representing districts that over performed relative to their predicted performance and those who under performed relative to their predicted performance. We used random forests of 50 trees to identify 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 the percentage of students characterized as "gifted and talented" in a district. We find a positive overall correlation with overperformance, with increasing shares of gifted students leading to over-performance of a school district. The next plot shows the partial dependence of the change in county population from 2010 to 2020, which utilized county-level data from the U.S. Census. We observe a surprisingly negative trend between population change and over-performance, with districts in shrinking counties having higher positive residuals than districts in growing counties. One possible explanation for this trend is that the counties that shrunk the most tended to be less heavily populated than the counties that grew the most between 2010 and 2020. Thus, we have far fewer observations in for the school districts in shrinking counties, which may be unfairly coloring our results. The next dependence plot covers the marginal effect of RUC code, a measure of how rural versus urban a county is on a scale from 1 to 9, with 1 being the most urban and 9 being the most rural. As seen in the figure below, we observe that there is overall a positive and increasing effect of RUC on over-performance, with more rural districts over-performing by a higher margin than more urban districts. This seems to contradict what our predictions would be regarding higher performing urban districts in major metropolitan districts. However, since there a nadmany more schools within an urban district versus an urban district, our results are likely heavily biased towards overperformance in rural districts with very few schools aggregating to the district level. The final plot shows the partial dependence of the percent of district staff that work as auxiliary faculty, including food service workers, bus drivers, secretaries, and custodial staff. We observe an overall negative trend, with a steep dropoff in overperformance at staff percentages over 20%. This trend indicates that as the share of auxiliary staff increases, we likely see a corresponding decrease in the share of teachers, which would likely lead to higher student-to-teacher ratios, a classic indicator of lower student performance.

Overperforming Districts

Table 5: Random Forest: Variable Importance for Overperforming Districts

feature	importance
st_pct_gifted	7.123796
teach_turnover_rate	6.036541
Change_2010.20_pct	5.935104
RUC.code	5.168892
staff_pct_aux	5.129073

Fig 5: Partial Dependence Plots for Overperforming Schools



For under performers, the most consequential feature was avg_salary_central, which showed higher rates of underperformance as the average salary of a central administrators decreased. It plateaus around 100,000 then decreases slightly before plateauing from 125,000 through salaries as high as 300,000. This indicates that while central administration compensation is important, districts which averaged beyond 100,000 it tends to have no additional positive effects on education outcomes, and can even cause slight decreases in education outcomes The next most important feature was rural indicator codes, which displayed a heterogeneous effect from varying levels of urban development. What is striking is the massive gap between the two most rural codes, 8 and 9, which represent counties with fewer than 2,500 residents either adjacent to a metropolitan area for 8 or non-adjacent to a metropolitan area for 9. This is likely due to small sample sizes of each group in the underperforming districts, 8 and 9 being containing 13 and 11 observations respectively, and being the fourth and second smallest respectively. What is apparent is higher performance, or less underperformance, from counties with the highest urban development category 1, which represent populations of over 1 million people. While this category has the most observations in the underperformer subset, it is likely because it the largest group in the original data set and yet still has fewer than a third of its observations in the underperformer subset, in addition to being the second lowest underperformer. Urban metropolitan

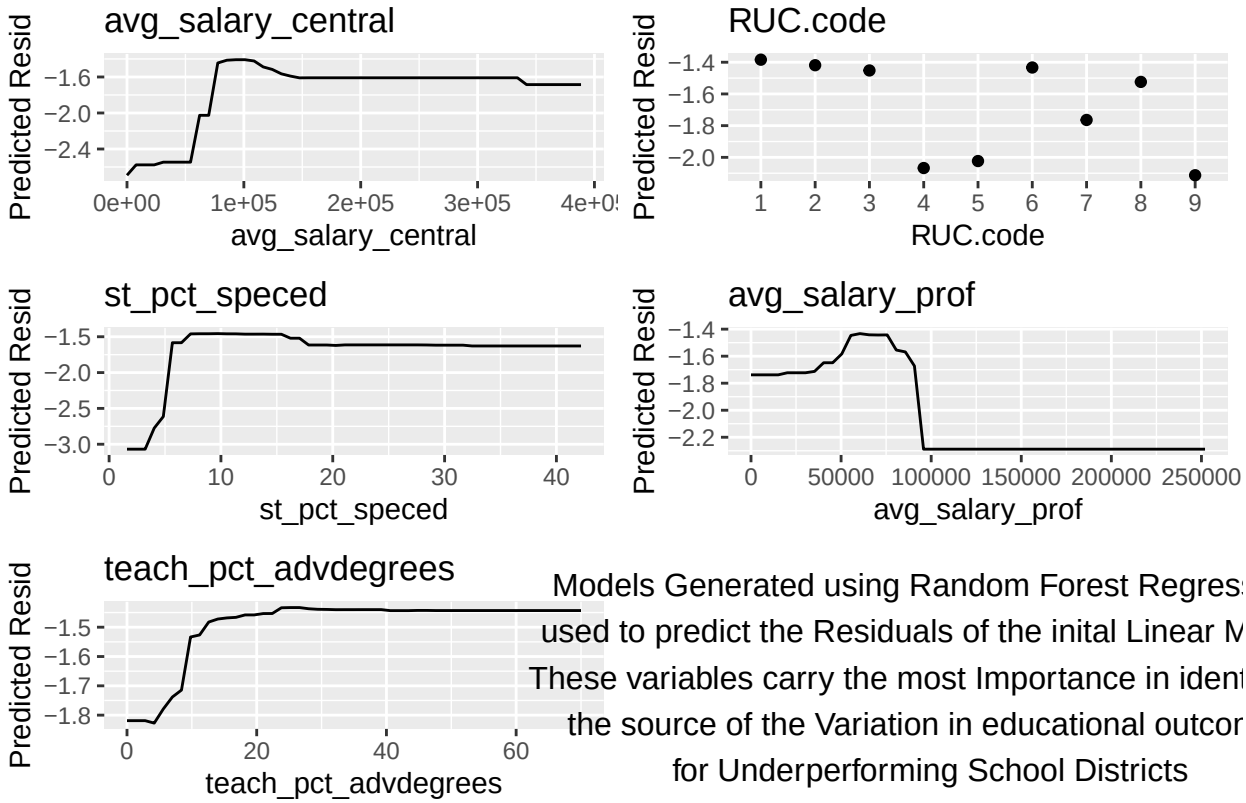
areas with comparable underperformance were in category 6, which represented counties with populations between 2,500 and 19,999 which were adjacent to metropolitan areas. Ultimately these results of gross under performance were obscured by small sample sizes, as counties sizes which were exceptional tended to have very few observations like 5, urban counties with greater than 20,000 not adjacent to metropolitan areas, which had only 4 observation in the underperformer subset. The next most important variable for underperformers is st_pct_speced, which refers to the percentage of students in special education. We see a massive rise if even over 5% are in special ed, and mostly a plateau afterwards. This variable is likely a proxy for resources available to the school in general. Special education has a perception as a luxury and has been the subject of cuts in the state of Texas in the past. If a school has enough funding to has special education at all, it also has funding for numerous other features which boost education outcomes. The next most important variable for underperformers is avg_salary_prof, or the average salary of teachers in a district. We see a steady rise till around 55,000 where the residuals plateau. At 75,000 the residuals plumet and remain regardless of increases to teacher salary. This is an unexpected result as generally a higher paid teacher would have more certification, and therefore be more qualified. Considering there are only 2 observations above 100,000 the higher salaries can be taken as outliers, but there are more observation greater than 75,000 that follow the trend as well, warranting further study into the circumstances of these well paid underperforming teachers. The final underperforming partial dependence plot somewhat refutes the former, as it shows the precent of teachers with advanced degrees. There is a sharp decrease in underperformance if the percentage of teacher with advanced degrees rises above even 3% and sharply increases until roughly 25%, after which it plateaus. This indicates that higher salary does not necessarily mean more qualifications, since the two variable have different associations with outcome variance. It also shows that while having some advanced degree holders absolutely is crucial in mitigating underperformance, there is a very clear limit to how much improve performance, after which it is mostly ineffective.

Underperforming Districts

Table 6: Random Forest: Variable Importance for Underperforming Districts

feature	importance
avg_salary_central	100.30370
RUC.code	96.00668
st_pct_speced	57.90635
avg_salary_prof	48.30812
teach_pct_advdegrees	45.38318

Fig 6: Partial Dependence Plots for Underperforming Schools

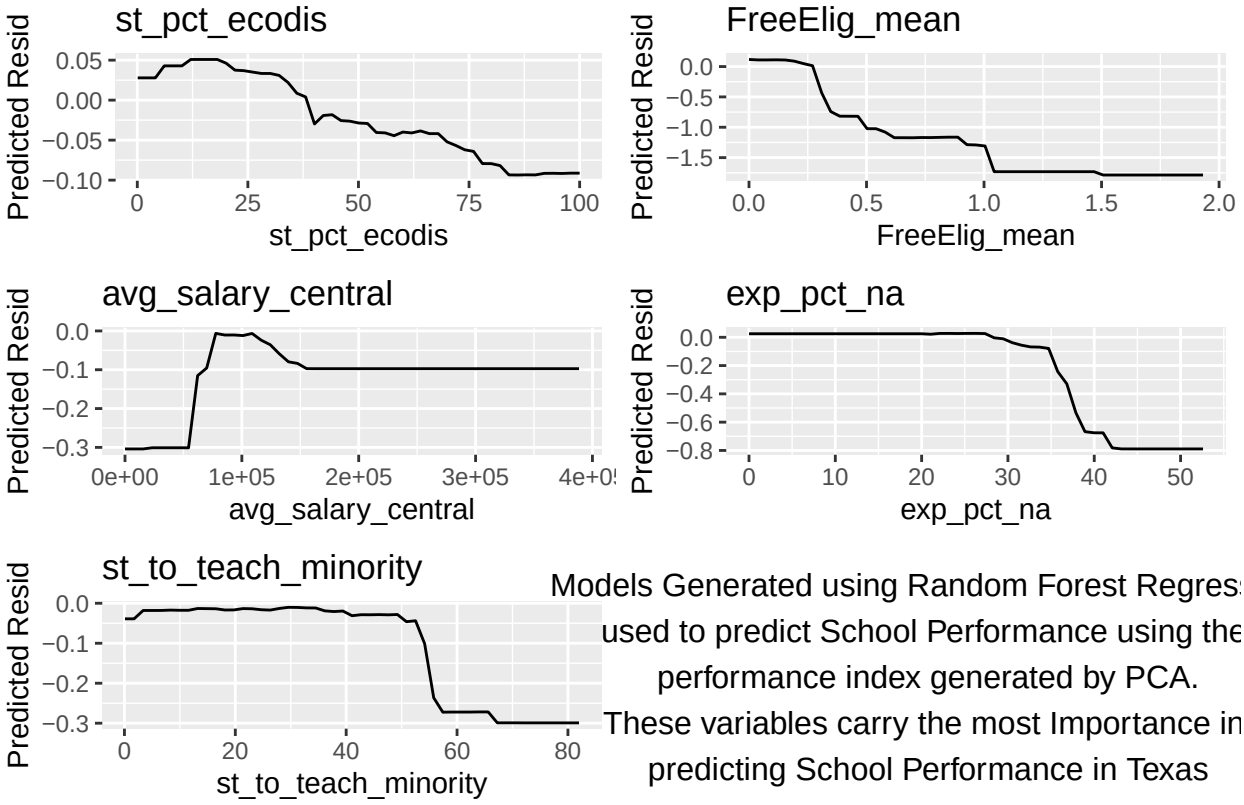


The most explanatory feature for determining education outcomes is st_pct_ecodis, which is the percentage of students who are economically disadvantaged. While there is some variance, the line steadily decreases as percent increases. There are likely several reasons for this. First, education outcomes often are effected by the home life of the child and the resources they have access to. This can effect their ability to buy school supplies, get access to food, and also be a general source of stress in their life, all of which can effect education outcomes. The second reason is that high percentages of economically disadvantaged people means a poorer overall population. Because schools are funded by property taxes, if all the children in the school are economically disadvantaged, it is likely that the school district is not bringing as much money in to pay for teachers, programs, and amenities. The second most important variable for education outcomes is FreeElig_mean, which is the average number of students who are eligible for free school meals. This variable tells a similar story to the last, which is that the more people who are in need of resources, the lower the expected outcomes of educational achievement. The next most important variable for education outcomes is avg_salary_central. Much like the underperformers, this variable rises steeply then plateaus, though a bit earlier around 75,000. This highlights the importance of a well funded administration as well as the pit falls of over funding, with virtually no increases after the 75,000 threshold. The fourth most important outcome variable is exp_pct_na, which is the precent of funding per student which is not allocated. This breaks from the previous narrative to refocus on mismanagement. As funding per student decreases, there is a lower and lower expected outcome, with a small exception around the 25% mark. This indicates that as the proportion of funding per student increases which has not been allocated to a specific program, a greater and greater negative outcome should be expected. This could be seen as a companion feature to the school administration salary, as a poorly run school will likely not allocate funds properly and will also have a meager administrating budget. The last feature for determining education outcomes is st_to_teach_minority, which is the ratio of students to teachers who are racial minorities. The disparities in outcomes between different racial and ethnic groups is well documented across the US and Texas education outcomes are no exception. There are likely confounding variables as individuals from traditionally disenfranchised racial groups often intersect with other disadvantageous socioeconomic factors like lack of healthcare and housing.

Table 7: Random Forest: Variable Importance for Education Outcomes

feature	importance
avg_salary_central	100.30370
RUC.code	96.00668
st_pct_speced	57.90635
avg_salary_prof	48.30812
teach_pct_advdegrees	45.38318

Fig 7: Partial Dependence Plots for All School Performance



Conclusion

School performance has been heavily researched with many results pointing to household income, racial makeup, and student-to-teacher ratios being the most impactful determinants to high student performance. However, few studies have analyzed the existing pool of over-performing and under-performing schools to determine what differentiates a school that should be performing well, but doesn't. The state of Texas provides a sufficiently large sample of school districts for such analysis and also features comparatively high racial, urban, and attainment diversity across districts and counties.

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.

A recurring highly explanatory feature was rural coding, particularly when looking at the under and overachieving districts separately. It is likely part of this explanatory power is due to reduced sample sizes for each of these random forests. This is another reason why this study could benefit from a regional or nationwide comparison, so the rural coding can reduce the impact of outliers by increasing observations from those group. If this feature is explanatory, then having more data to drown out the noise could help better illustrate how urban vs rural environment s effect education outcomes. If the feature is not explanatory then it will drop out of the top 5 for a better fitting parameter. This questions warrants further study but is beyond the scope of our analysis, due to the massive amount of data aggregation which would need to be performed to get the same feature list for all 50 states or even the sunbelt states.

In our study, we find that the share of students eligible for free meals acts as a quasi measure of average household income and an significant indicator for underperformance. In addition, we find that higher average salaries for professional support staff including therapists, counselors, nurses, librarians, and department heads, tends to be correlated with underperformance, but suffers from a deficiency in data at lower and upper quantiles. We also find a negative correlation with unallocated expenditure per student, with higher shares of a district's budget spent on items not distributed to specific programs being correlated with underperformance at shares above 28%. This variable may serve as an indicator of fiscal responsibility, with districts spending large portions of their budgets on unallocated items seeing higher levels of underperformance. We next break down our results further by zeroing in on the most important features that predict over versus under performance. For overperforming districts, we find that the share of students deemed "gifted and talented" is the most important feature and is positively correlated with higher shares indicating higher levels of over-performance. We also find that county-level population change from 2010 to 2020 is a significant predictor, however we find a negative trend in predicting overperformance. We believe that this is another result of flaws in data collection, with the smallest counties with the least number of observations heavily biasing our data. Finally, we observe the percent of district workers that work as auxiliary staff on overperformance and find an overall negative trend, indicating that as the share of auxiliary staff increases, we likely see a corresponding decrease in the share of teachers, which would likely lead to higher student-to-teacher ratios, a classic indicator of lower student performance.