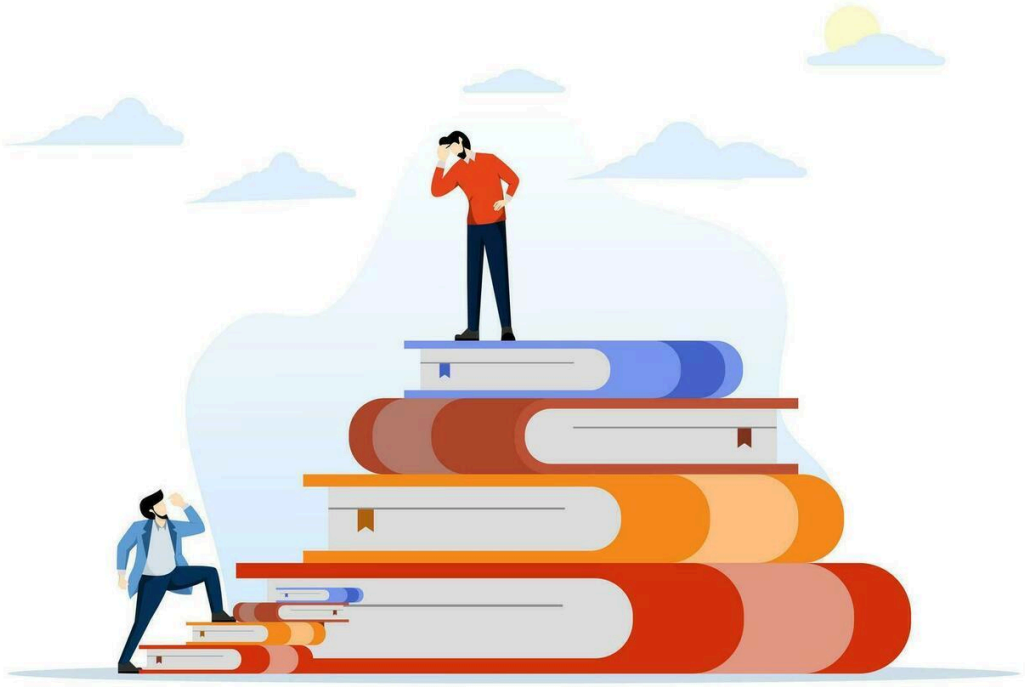
Capstone Final Report

Student Success Prediction



Contents:

Introduction:

Problem:

Educational inequities persist across various school districts in Texas, potentially impacting student academic performance and long-term success. Despite the state's efforts to allocate funding based on Average Daily Attendance (ADA), disparities in funding raise concerns about the adequacy of resources available to each student. The core issue is whether the differences in funding correlate with academic outcomes and whether this could be a contributing factor to the educational inequities lurking beneath the surface of Texas' complex system.

Project:

My aim was to examine the relationship between various educational variables -- school district funding in particular -- and student academic performance in Texas as measured by standardized test scores. By leveraging detailed financial, attendance, and academic data from the Texas Education Agency (TEA) and other sources, the project seeks to identify patterns and correlations that could inform policy decisions and interventions aimed at reducing educational disparities.

Clients:

The findings of this study will be of interest to a broad range of stakeholders, including educational policymakers, college admissions offices, school district administrators, teachers, parents, and students. It will particularly benefit decision-makers within the TEA and local educational authorities who are responsible for allocating resources and designing strategies to improve educational equity and outcomes across Texas.

Methodology:

Data:

The datasets for this project are drawn from a few sources such as the U.S. Census Bureau and WalletHub, but they're primarily sourced from the TEA. The TEA provides comprehensive records on school district funding, student attendance rates, and various academic performance indicators. I gathered data from 2018-2022 and analyzed on a per-school district basis. Several features were incorporated, including but not limited to, funding per student, attendance rates, socioeconomic levels, teacher expertise, and student-to-teacher ratios. The primary goal was to develop a predictive regression model that accurately reflects the complex dynamics between educational variables and student success while highlighting any significant disparities that warrant attention.

Data Preprocessing:

I conducted extensive data cleaning of ten individual datasets before joining them together for further analysis. I leveraged visualizations to identify potential outliers and feature correlations, generated statistical summaries, and uncovered the nature of distributions for each variable. I also engineered some critical features, removed missing values, and imputed the missing median income values using the year-to-year

percentage changes for the state median income. I then split the data into training and testing subsets, and scaled the data for model preparation. I decided to drop PCA in my analysis because there weren't enough features to warrant reducing them, and the performance metrics for PCA transformed features were worse than with the full 33 feature set.

Insights from Exploratory Data Analysis:

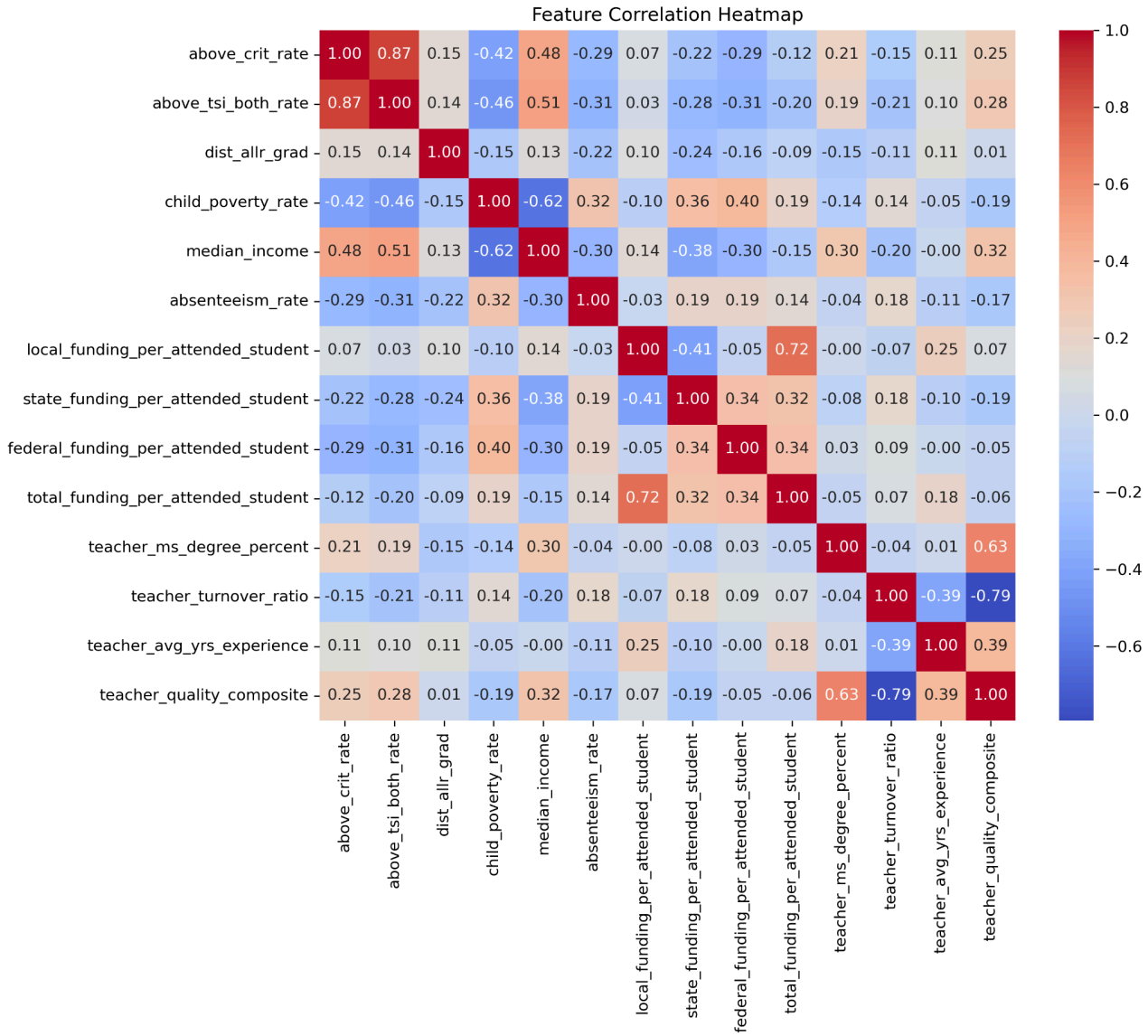


Figure 1: Correlation Matrix

- Due to HS Graduation Rate having zero variables with any positive or meaningful negative correlation to it, we dropped this value as a target variable and focused instead on SAT-ACT Scores as our sole target (Above TSI Both Rate).
- Overall, the correlation matrix highlights complex relationships between educational funding, socioeconomic factors, and student success metrics (see Figure 1 above). These insights could be useful for policymakers and educators in understanding the multifaceted influences on educational outcomes and in designing interventions to improve student success.
    - **Both state and federal funding per attended student is positively correlated to child poverty rate** (0.33 and 0.37), indicating that the state and federal governments are more likely to increase funding to districts with higher child poverty, which is a good thing.
    - The state claims that they generally allocate more state funding to districts with lower local funding from property taxes, and this appears to be true from the data: the correlation between state funding per attended and local funding per attended is -0.41.
    - **Imputed median income has the strongest absolute correlation to test scores** along with child poverty rate, indicating that **economic inequality and educational inequality are inextricably linked**.
    - Interestingly, **NONE of our funding specific variables show positive correlations with our target variable** of SAT-ACT Score (above_tsi_both_rate). <u>This debunks our initial hypothesis that higher funding is correlated with higher student success, indicating other factors are more important</u>.

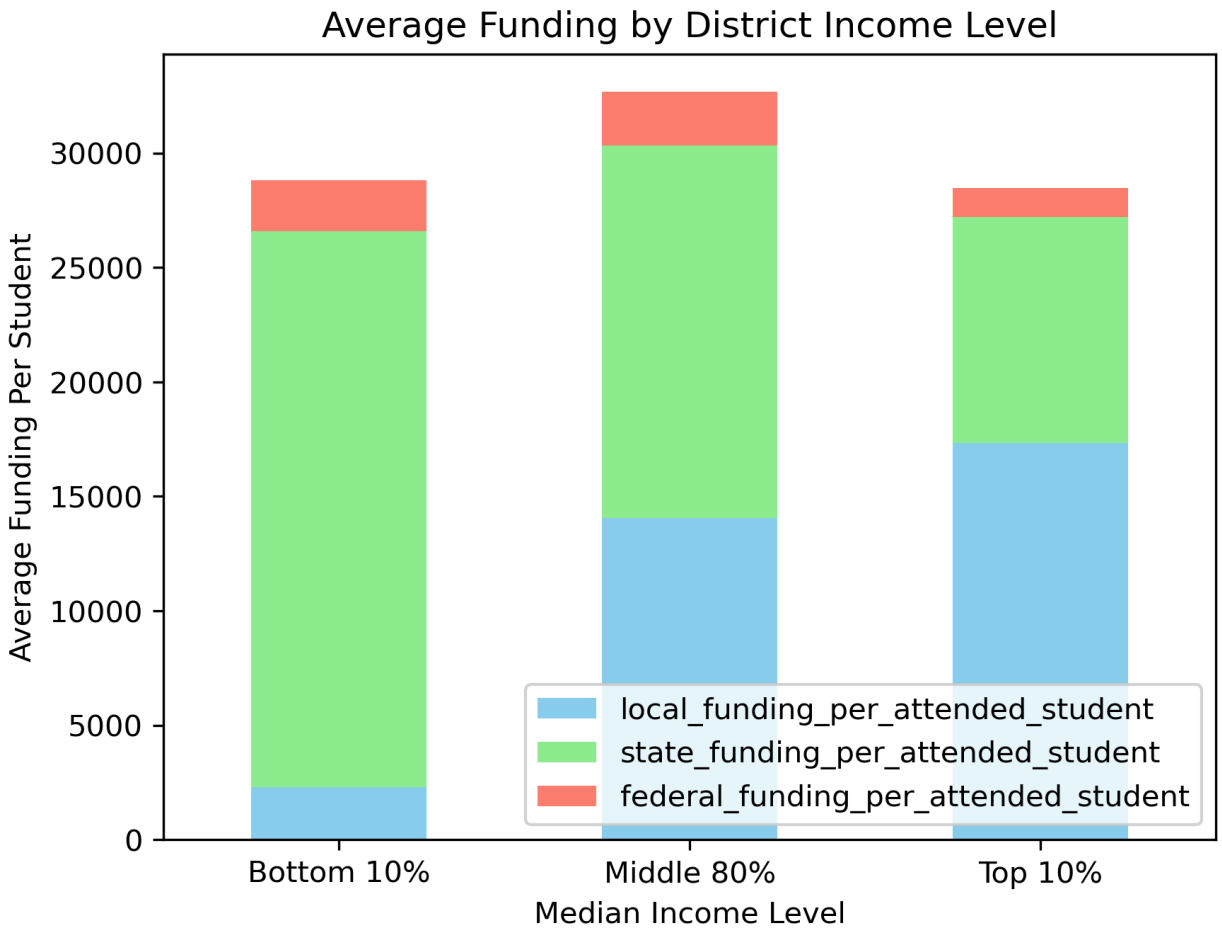


Figure 2: Average Funding by District Income Level

- There are disparities in district funding by median income levels, particularly with local funding amounts (see Figure 2 above). This could have implications for policy and resource allocation to address these disparities.
    - **The districts with the bottom 10% median income receive significantly less local funding on average compared to both the middle 80% and the top 10%, as well as a much lower level of total funding compared to the middle 80%**. Although the bottom 10% receives total funding per student that is roughly equal to that of the top 10%—thanks to higher state funding—this amount is still insufficient compared to the middle 80%. This situation highlights substantial inequalities among districts, where poorer neighborhoods face disadvantages by receiving significantly less total funding per student—thousands of dollars less than those in higher income brackets.

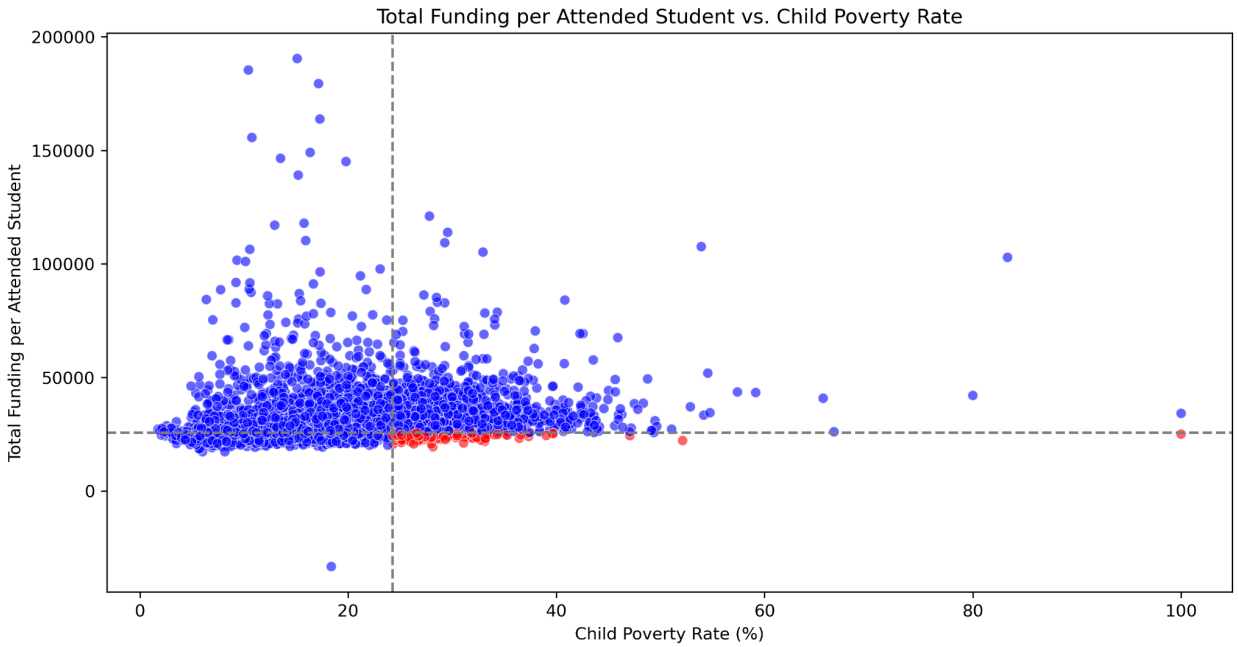


Figure 3: Funding vs Child Poverty

- The scatterplot illustrates the relationship between total funding per attended student and child poverty rate across districts, with districts falling into the low funding and high poverty category marked in red (see Figure 3 above). The grey dashed lines indicate the thresholds for low funding (below the 25th percentile of total funding) and high child poverty (above the 75th percentile).
- Districts marked in red are of particular concern, as they represent **situations where low funding coincides with high poverty rates**, potentially exacerbating educational inequities.

Results and Analysis:

Model Description:

I compared the performance of four different models to determine which one provided superior results, looking at simpler models first and creating more complex models as we progressed. I also performed cross-validation to evaluate how well a model is likely to perform on unseen data by using different subsets of the training data for both training and validation.

Here are all the models tested with their descriptions and parameters:

- *Linear Regression Model:* predict a response using a linear combination of input features.
    - Default Parameters: fit_intercept=True, normalize=False, copy_X=True, n_jobs=None
- *Random Forest Model:* ensemble learning method that builds multiple decision trees during training and outputs the class that is the mode of the mean prediction (regression) of the individual trees.
    - Parameters: n_estimators=100, random_state=42
- *Gradient Boosting Model:* a powerful ensemble technique that builds models sequentially, each new model correcting errors made by the previous ones.
    - Parameters: n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42
- *Decision Trees:* non-linear predictive modeling tool. They divide the dataset into branches to form a tree with decision nodes and leaf nodes.
    - Parameters: max_depth=5, random_state=42

Performance Metrics:

- *R-Squared (R^2):* Higher values are better. Indicates a model's ability to explain the variability of the target variable.
- *Mean Absolute Error (MAE):* Lower values are better. Average of the absolute differences between predictions and actual observations. It gives an idea of how wrong the predictions were, without considering direction.
- *Mean Squared Error (MSE):* Lower values are better. Measures the average squared difference between the estimated values and the actual value.
- *Root Mean Squared Error (RMSE):* Lower values are better. The square root of the MSE. Measures the average magnitude of the error. Especially useful when large errors are particularly undesirable. The square root allows for interpreting the errors in the same units as the response variable.

These metrics help determine not only how close the predictions are to the actual data in terms of error magnitude, but also in terms of the proportion of variance explained by the model.

Model Performance:

Looking at all the models, *Random Forest and Gradient Boosting were the top performers based on the metrics*. Both models have similar performance metrics, with Gradient Boosting slightly edging out in terms of a lower Test MSE and RMSE, and a

slightly higher R^2. Their MAEs are also comparable, and both models demonstrate better generalization ability (lower CV Mean MSE) compared to Linear Regression and Decision Tree.

While Gradient Boosting slightly outperforms Random Forest in some metrics, the differences are not substantial enough to clearly favor one over the other based on performance alone. Compared to Gradient Boosting, Random Forest is:

- More interpretable: feature importances are more easily gleaned
- Less sensitive to overfitting: this is from the averaging of multiple decision trees
- Easier to tune: this is due to having fewer hyperparameters that critically affect performance

To balance the need for a high-performing model with the practical considerations of explainability, ***I opted for the Random Forest model as the best model to use.***

Conclusion and Recommendations:

Model Findings:

After tuning the Random Forest model hyperparameters, I experimented with some reduced feature amounts and found that reducing the original 33 feature set to 10 features yielded results similar to the full set, indicating the reduced amount had a minimal impact on model predictive accuracy. I then re-tuned the hyperparameters on the reduced feature set using GridSearchCV.

Our best parameters that lowered our MSE were:

- max_depth: 30
- min_samples_leaf: 1
- min_samples_split: 2
- n_estimators: 300

After hyperparameter tuning on the reduced feature set, the model achieved an MSE of 116.9, an MAE of 8.5, an R-Squared 0.53, and an RMSE of 10.8.

The feature importance analysis revealed that **variables such as Child Poverty Rate, Average Daily Attendance (ADA), and Participation Rate** (rate of participating in taking standardized tests) **are crucial in predicting test scores**, emphasizing the impact of socioeconomic factors and attendance on educational outcomes. **Teacher**

Quality and Median Income also showed among the most important predictive variables (see Figure 4 below).

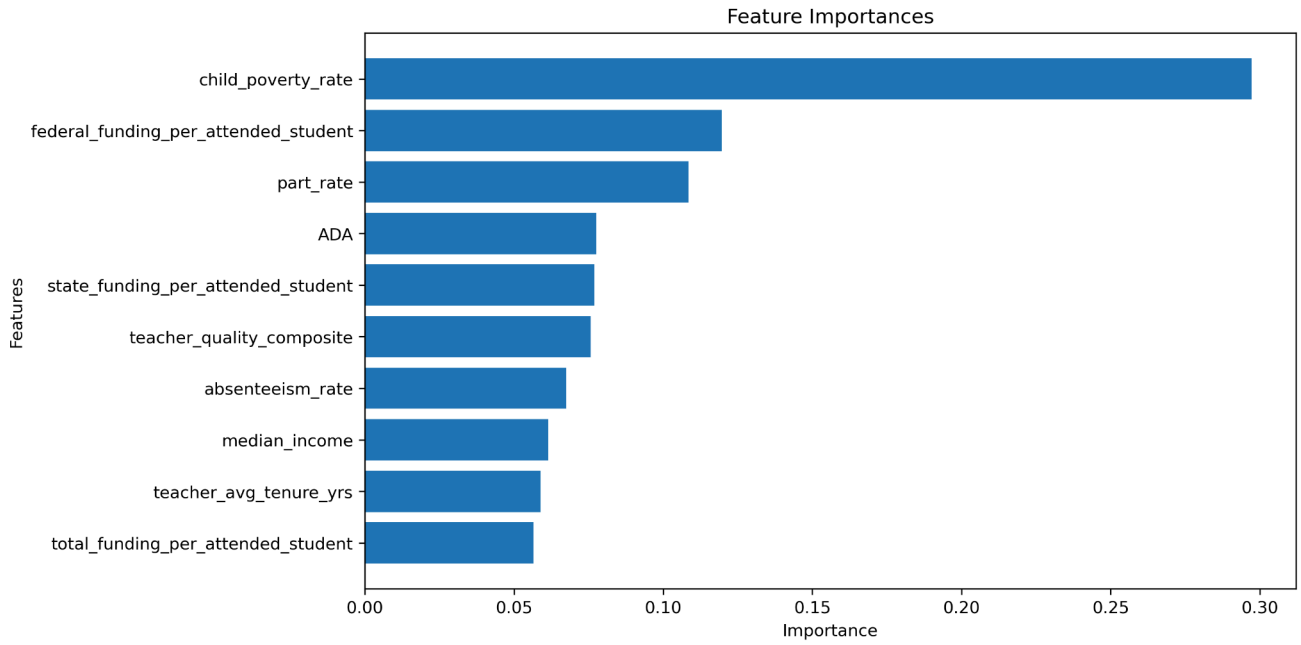


Figure 4: Feature Importances

Here are some additional insights gleaned:

- **Child poverty and attendance are among the most important features**, as higher levels of poverty and absenteeism tend to predict lower test scores among school districts (see Figure 4 above).
- **Teacher quality is among the top 4 most important features** for predicting student test scores (excluding state and federal funding since these are lagging indicators, not leading indicators).
    - *NOTE on teacher quality composite: this was a metric I created during the EDA stage. Three primary variables indicate teacher quality: teacher avg years of experience, percent of teachers with masters degree, and teacher turnover ratio. Each of these variables have small or minimal correlations to test scores in isolation, but I tried combining all three through creating a composite variable. The resulting composite correlation was 0.28 – higher than any of the three feature correlations in isolation – so we included this metric in our modeling. See the EDA notebook for calculation details.*
- Federal funding and state funding are likely lagging indicators rather than leading indicators (because, as mentioned, low test score districts receive higher state

and federal funding, rather than state and federal funding leading to lower test scores).

Recommendations:

Based on the modeling results and overall analysis, I'll present some action steps for educational policymakers, school district administrators, and the various levels of government.

Here are three concrete recommendations:

1. **Provide more targeted funding to high absenteeism districts:** Knowing attendance is one of the most important factors in predicting student outcomes (see Figure 1 and Figure 4), policymakers should identify districts with higher absenteeism and provide more funding towards:
    a. Expanded transportation options to ensure all students have reliable access to school, particularly in rural or underserved areas.
    b. Health and wellness programs including mental health support to address health-related absenteeism. These can include routine health checks, counseling, emergency health services, and more.
    c. Engagement and enrichment programs that increase student engagement through clubs, sports, arts, and other extracurricular activities.
2. **Increase equity of qualified teachers among districts**: We know that teacher quality is another important factor in predicting student outcomes (see Figure 4), so it's crucial to invest in our teachers. School district administrators can increase the equity of teacher quality through:
    a. State-funded incentive programs that attract and retain highly qualified teachers in underserved or lower-performing districts. This could include higher pay scales, signing bonuses, housing allowances, or student loan forgiveness.
    b. Professional development and continuing education opportunities for teachers to pursue advanced degrees, certifications, and specialized training.
    c. Equitable funding models that ensure that funding is specifically earmarked for initiatives that directly enhance teacher quality, such as bonuses for teachers who achieve certain professional milestones or who effectively implement innovative teaching practices.
3. **Enact targeted interventions for high inequality districts (low funding + high poverty):** Despite the state's efforts to level the playing field, substantial disparities in local funding continue to drive funding inequities across districts (see Figure 2), and there are still districts in Texas with both low total funding

AND high child poverty rates (see Figure 3). These districts require targeted interventions to address disparities and support students from these socioeconomically challenged backgrounds. While we know that funding does not correlate to student outcomes in and of itself (see Figure 1), and we may not be able to eliminate poverty altogether, the state and federal government can still grant more funding that focuses on reducing absenteeism and enhancing teacher quality for these poorer districts.

Ideas for Further Research:

Further analysis could investigate the deeper reasons behind the disparities in funding across different districts, as well as its impact on educational opportunities, and potential policy implications. If possible, it would be interesting to expand the scope and analyze similar data for the entire country to see if similar trends are present.

References:

Here are all the data sources used:

Texas Education Agency (TEA):

- 4-Yr HS Graduation Rates: https://tea.texas.gov/reports-and-data/school-performance/accountability-research/completion-graduation-and-dropout/four-year-graduation-and-dropout-data-class-of-2022
- SAT-ACT Data: https://tea.texas.gov/reports-and-data/school-performance/accountability-research/satact/sat-and-act-data-class-of-2022
- District Funding Totals: https://tea.texas.gov/finance-and-grants/state-funding/state-funding-reports-and-data/peims-financial-data-downloads
- District Property Values: https://tea.texas.gov/finance-and-grants/state-funding/state-funding-reports-and-data/peims-financial-standard-reports

- AP-IB Scores & Participation: https://tea.texas.gov/reports-and-data/school-performance/accountability-research/apib/advanced-placement-and-international-baccalaureate-data-2021-22
- Teacher Count and Avg Salary: https://rptsvr1.tea.texas.gov/adhocrpt/adpeb.html
- Average Daily Attendance (ADA): https://rptsvr1.tea.texas.gov/adhocrpt/adpeb.html

U.S. Census Bureau:

- Kids in Poverty by District: https://www.census.gov/data/datasets/2022/demo/saipe/2022-school-districts.html

WalletHub:

- Median Incomes per District: https://wallethub.com/edu/e/most-least-equitable-school-districts-in-texas/77134