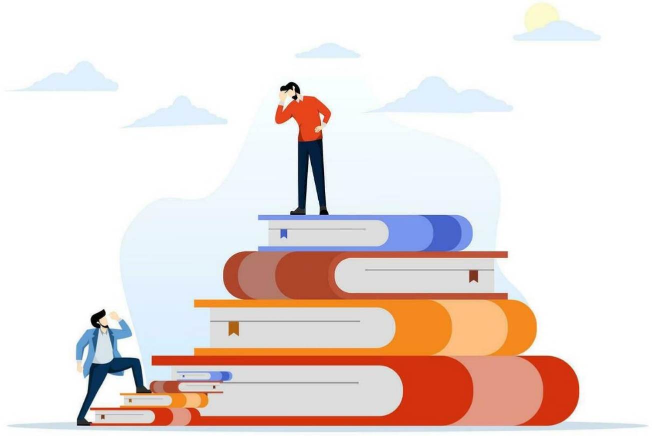
Texas Student Success Prediction

By Joshua Beasley

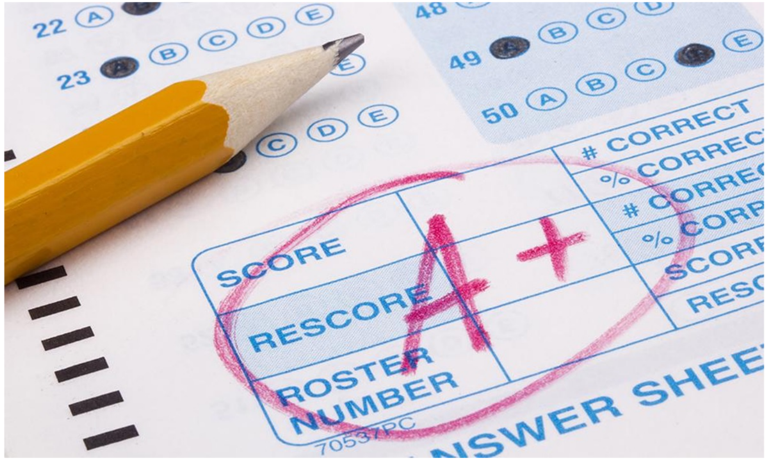
Problem Statement

- What educational inequities exist within the Texas public school system?
- How do these inequities impact student academic performance and long-term success?
- Which educational factors correlate with academic outcomes?



Predicting Test Scores

- How can machine learning help predict student standardized test scores?
- Identify patterns and correlations that could inform policy decisions and interventions aimed at reducing inequality
- Datasets: Texas Education Agency (TEA), U.S. Census Bureau, WalletHub



Data Wrangling

- Years analyzed: 2018-2022
- Cleaned 10 individual datasets and joined
- ~6,400 observations with 57 features initially
- Dropped redundant and unhelpful features, and rows with missing values
- Engineered several helpful features
- Reduced to ~4,500 observations and 33 features after cleaning
- **Target Variable: Standardized Test Scores** (Above TSI Both Rate)

Imputations for Median Income

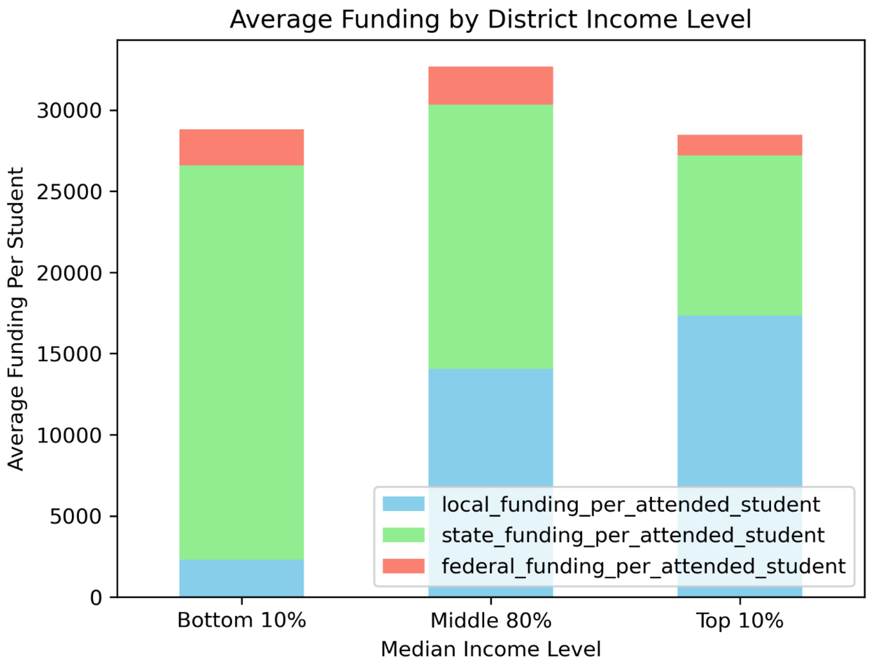
- Imputed median income values for 2018-2022
- Used year-to-year % changes for state median income
- Allowed estimation of district-level incomes for missing years
- Assumes the economic trends affecting income are relatively consistent statewide
- Used as a proxy, since real world data is messy/imperfect

Correlation Insights

- Complex relationships between educational and socioeconomic variables
- Imputed median income had strongest correlation to test scores, along w/ child poverty rate
- District funding was NOT correlated with the target

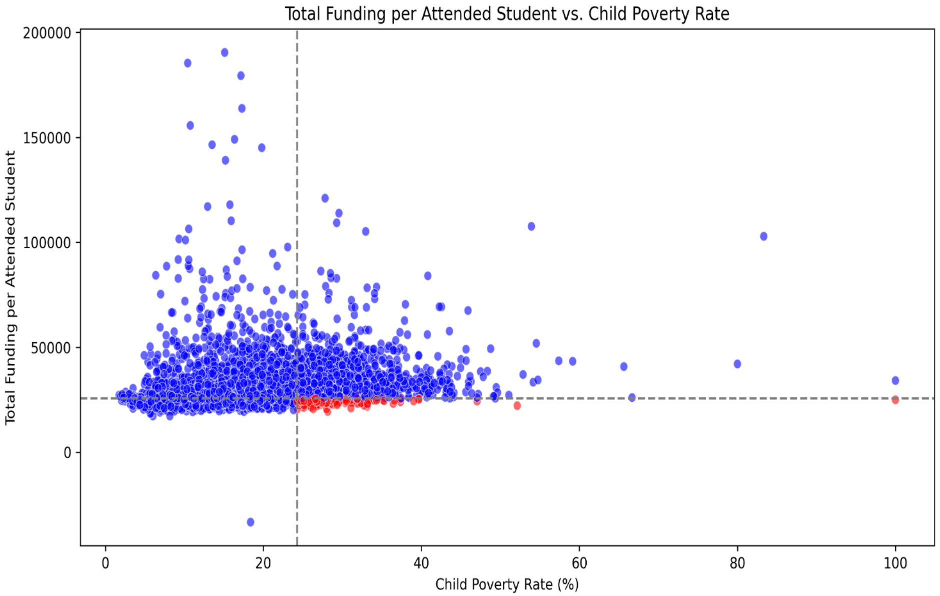
Funding by Income Level

- Disparities in district funding by income levels
- Bottom 10% districts receive much less funding compared to middle 80%
- Inequalities in funding exacerbate economically disadvantaged districts

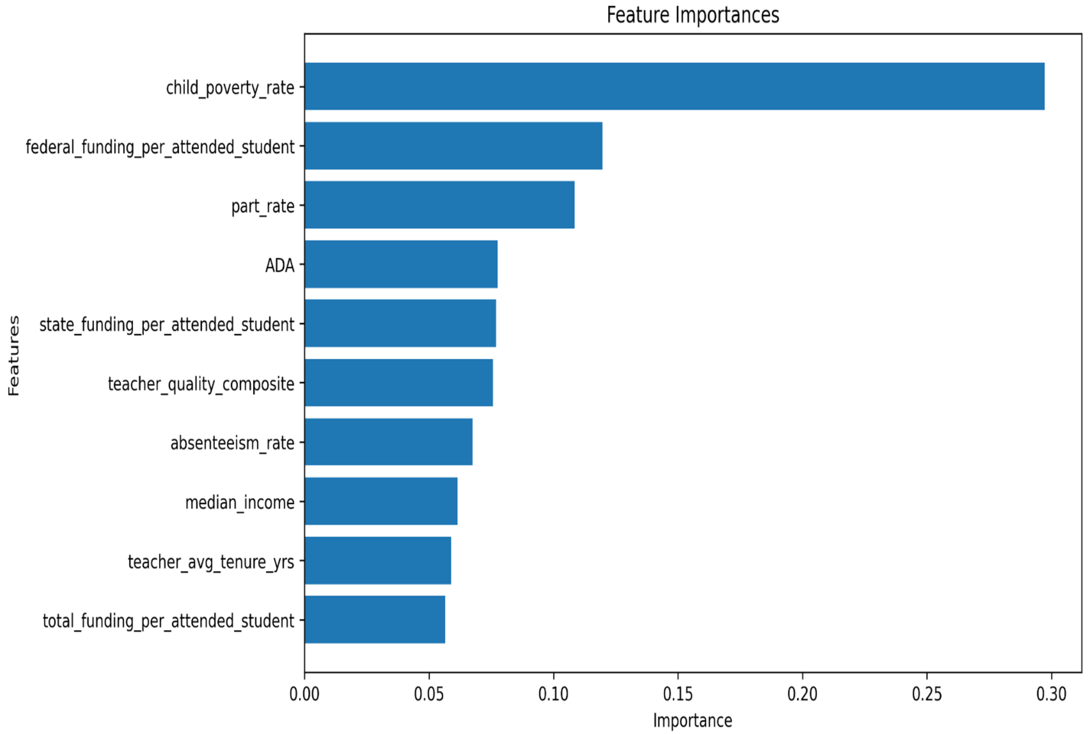


Districts with Ultimate Inequality

- Grey dashed lines = thresholds for low funding (below 25th percentile) and high child poverty (above 75th percentile)
- Red districts = low funding AND high poverty
- Red districts show situations of ultimate inequality



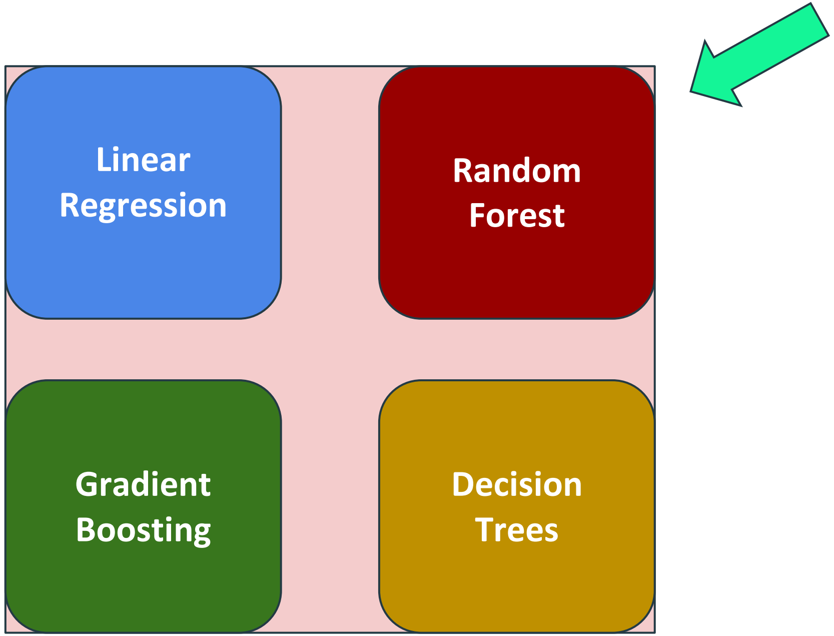
Most Impactful Features



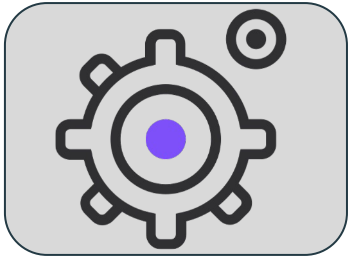
Metrics for Evaluation

- **R-Squared (R^2)**
  - Model's ability to explain the variability of the target variable
- **Mean Absolute Error (MAE)**
  - Avg of the absolute differences between predictions and actual observations
- **Mean Squared Error (MSE)**
  - Avg squared difference between the estimated values and the actual value
- **Root Mean Squared Error (RMSE)**
  - The square root of the MSE; measures the average magnitude of the error

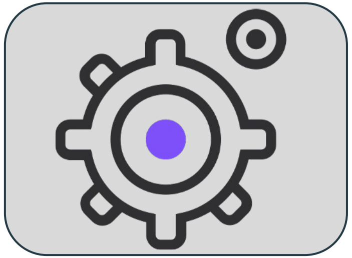
Model Selection



Hyper-Parameter Tuning



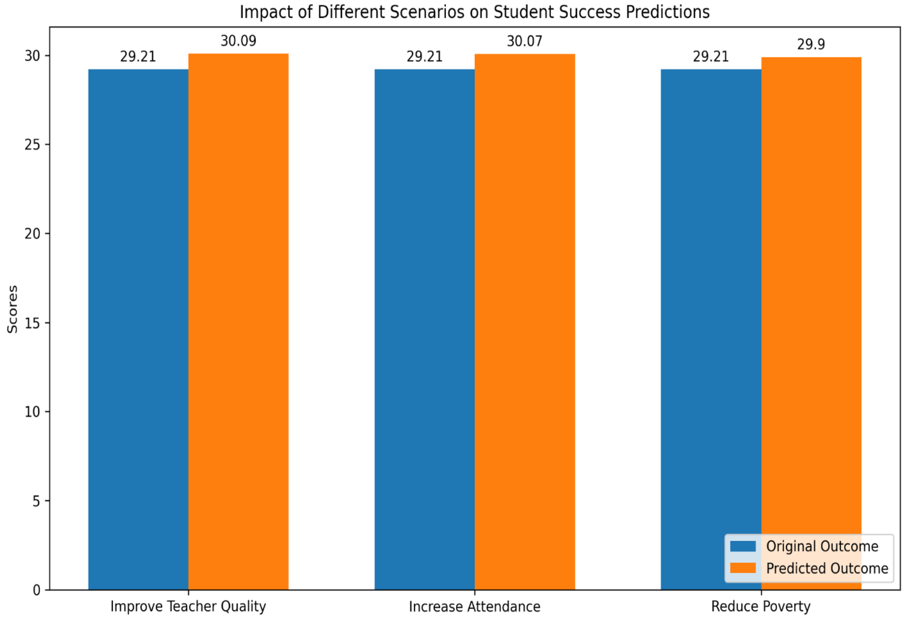
- Reduced features from full 33 to 10 (similar results)



- Discovered optimal parameters that lowered MSE (max_depth, min_samples, min_samples_split, n_estimators)

Predictions

- Showcases impact of interventions and/or policy changes
- Scenarios Tested: Reduce poverty rate 10%, increase attendance 5%, increase teacher quality score 5%
    - Results: ~1% increase in test scores
- Significant impact when scaled up across districts



Takeaways

- **Economic inequality and educational inequality are inextricably linked**
- **Subset of poorer districts receive inadequate total funding compared to most better-resourced districts**
- **Factors such as attendance, teacher quality, and socioeconomic status are crucial in academic performance**

Recommendations

1. **Provide more targeted funding to high absenteeism districts**
   a. Expanded transportation
   b. Health and wellness programs
   c. Engagement and enrichment programs
2. **Increase equity of qualified teachers among districts**
   a. State-funded incentive programs
   b. Professional development and continuing education
   c. Equitable funding models
3. **Enact targeted interventions for high inequality districts**