


Can surrounding socioeconomic factors predict ACT scores?

Genny Sheara
DATA 3320

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

Questions and overview

This project seeks to address discrepancies in socioeconomic status (measured in aspects such as income, percent of students receiving free lunch, unemployment rates of surrounding areas) and discrepancies in college readiness as measured by ACT scoring.

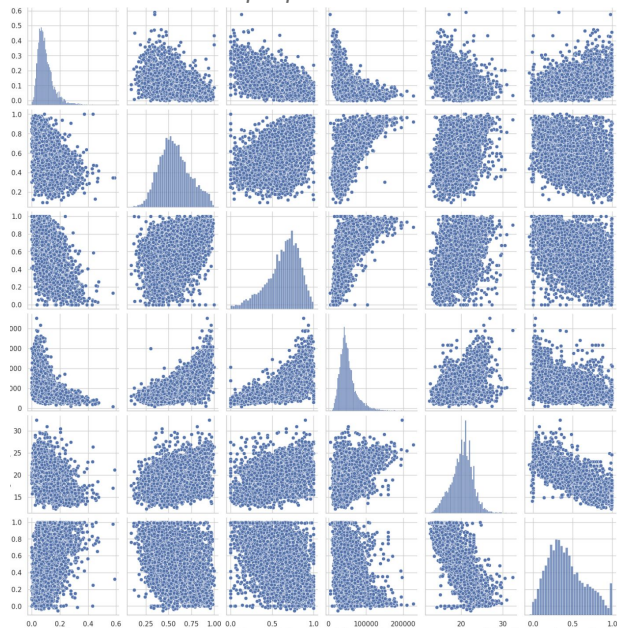
To explore the overall question of if and how ACT scores can be predicted by socioeconomic data, I have isolated the three following questions:

1. Can numerical socioeconomic data be used as a predictor of ACT scores?
2. If so, which aspect of this data can be used as the most accurate predictor?
3. Can school type (as an example of categorical data) be used as a predictor of ACT scores?



Data preparation

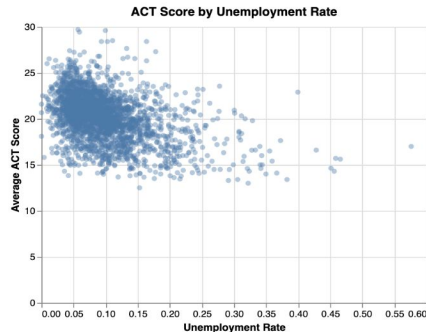
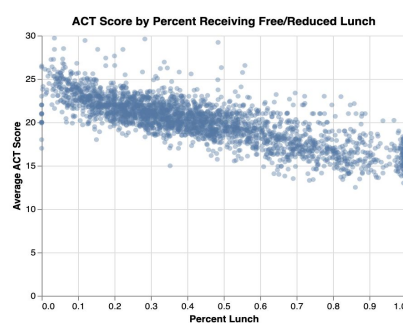
Chart 1: A basic pairplot of numerical variables.



While there are too many numerical variables visualized here to really dig into, we can observe that ACT scores seem to generally respond to socioeconomic factors.

- Focus on numerical data, with additional step of exploring how to work with school type (regular versus non-regular) as categorical data.
- Filtered out extraneous data such as location
- Imputed missing values using training set
- Divided data into numerical and non-numerical, and then did basic train/test split for each group

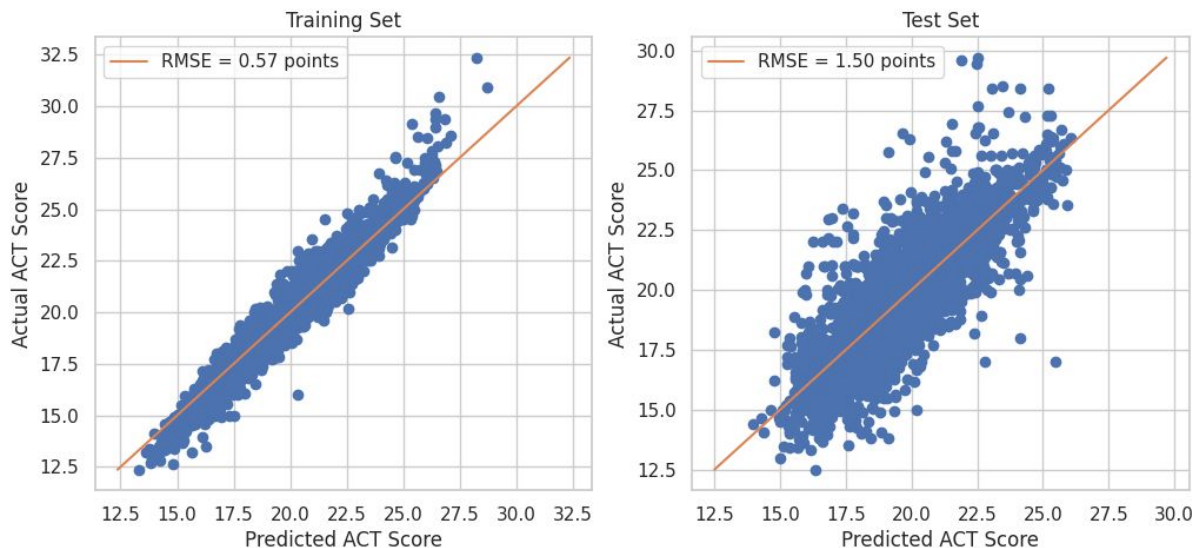
Charts 2 and 3: Quick visualizations to observe basic patterns in data.



Analysis (Questions 1 and 2)

Q1: It seems that numerical socioeconomic data can be used as a pretty viable predictor of ACT score, within a point and a half. One thing of note in this chart is the fact that there seems to be a cutoff in the Test Set where the predicted ACT score doesn't go much past 25 or 26, despite actual ACT scores going up to 30. A similar but more subtle pattern of under-predicting high ACT scores is apparent in the Training Set as well.

Chart 4: Training and test sets of numerical variables using random forest regressor.

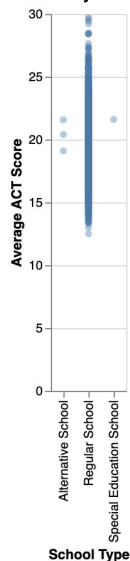


Q2: Using linear regression to find the coefficients for each column, unemployment rate was found to have the highest value and may act as a strong predictor of ACT score.

Analysis (Question 3)

Chart 5: Distribution of school type

ACT Score by School Type



Q3: Neither a random forest nor linear regression produced an accurate predictor of ACT score using school type. This could possibly be due to the overwhelming number of 'Regular' schools as seen in Chart 5, and their high range of score. Alternative, special education, and career/technical schools were encoded as a binary 'non-regular' school category to simplify the one-hot encoding process. The models' predicted ACT score seems to align with the overall mean of the data.

Chart 6: Random forest regression

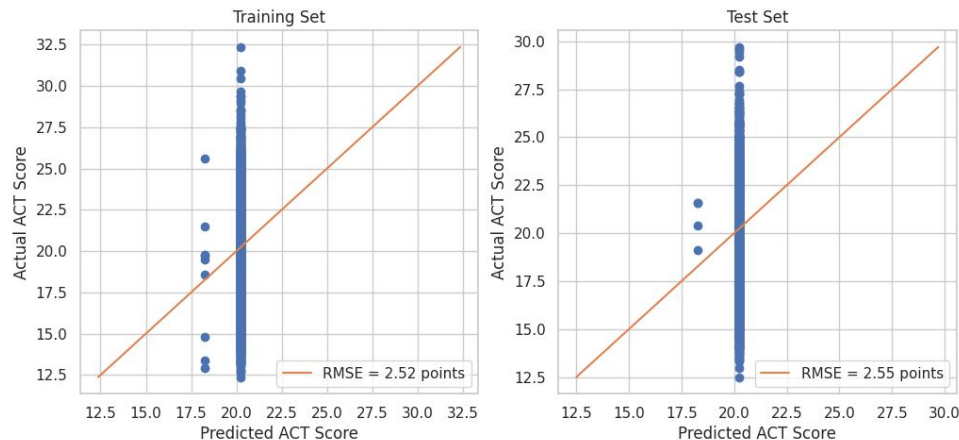
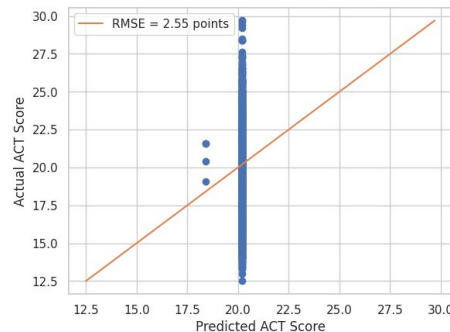


Chart 7: linear regression



Conclusion and further work

In conclusion, it seems that the numerical data collected on socioeconomic factors in surrounding areas of different high schools can be used as the best predictor of ACT scores. There was about a point and a half of variance overall, and some difficulty in accurately predicting the higher scores of the datasets. The categorical data used, school type, is possibly more difficult to compare because an overwhelming number of schools are regular as opposed to career and technical, alternative, and special education, and this regular school group also saw the biggest range. Neither a random forest or linear regression model produced a good prediction of the data.

Further work is also needed to truly address the discrepancies in college readiness and test scores and their relationship to socioeconomically disadvantaged and underserved communities.

Further analysis on other demographic factors, particularly those that are more easily quantified such as population percentages on race and gender, may also provide additional insights into the relationship between socioeconomic status and ACT testing.

Sources

The data utilized for this project comes from EdGap.org, an online tool for visualizing factors contributing to standardized testing results across the United States, and the National Center for Education Statistics (school info dataset).

Education Gap dataset link:

https://github.com/gsheara/Education-Inequality/blob/373577c9a9887a1c77aee25b1a6cdb50109bc72/EdGap_data.xlsx

School Info dataset link:

<https://drive.google.com/file/d/1HvW2w-o2XZzCm4KTvnb1Bb3BvoAa14BP/view>

GitHub repository link containing all steps and relevant files:

<https://github.com/gsheara/Education-Inequality>