Education Regression Analysis

THE IMPACT OF SOCIOECONOMIC FACTORS AND INVESTMENT IN SCHOOL RESOURCES ON ACT SCORES.

Joel Uy

The problem:

Education inequality exists in the US which has unfair effects on the ACT/SAT performance of students depending on where they go to school. This inequality can be signaled by socio-economic statistics. To better understand how inequality is affecting student performance, regression analysis can be used to quantify how various socio-economic factors relate to each school’s average ACT/SAT scores.

DATA SOURCES

EdGap.org

EdGap is an online resource that visualizes average ACT/SAT on a US map. Various socio-economic statistics for each area can also be viewed.

Data from this source is each school area's:

1. Average ACT/SAT scores
2. Median household income
3. Unemployment rate
4. Proportion of adults holding a college degree
5. Proportion of students in a married parents family
6. Proportion of students that receive free or discounted lunch

NCES (School Info)

The National Center for Education Statistics aims to provide basic information and descriptive statistics for every public elementary/secondary school in the US .

Data from this source is each school's:

1. School year of data recording
2. State by location
3. Zip code by location
4. School type (regular, alternative, technical/career, etc)
5. School level (primary, middle, high, etc)

Census Data

Data is pulled from the US Census *Annual Survey of School System Finances Tables*. The 2019 regional data is used to match the EdGap Data. Per pupil metrics are used to minimize the bias of population on spending.

Data from this source is each district's:

1. Per pupil average instructor salary
2. Per pupil student support spending
3. Per pupil staff support spending
4. Per pupil average admin salary

ANALYSIS STEPS

- Prepare the Data
  - Merge data into one data frame
  - Perform quality control
  - Split data into training and testing set
  - Normalize quantitative data
  - Impute missing quantitative data
- Identify best regression features
- Model selection and parameter sweep
- Model results and performance

DATA PREPARATION

Merge Data
1

EdGap and school info merged using NCESSH id. Then, census data is merged using district NCES id, a substring of NCESSH id.

Quality Control
2

Impossible values are removed (eg percentages outside 0% and 100%, negative monetary values, etc). Rows with missing data for qualitative variables are dropped.

Train/Test Split
3

Data is split into a train and test set with a 80:20 ratio. Individual rows are randomly distributed into each split.

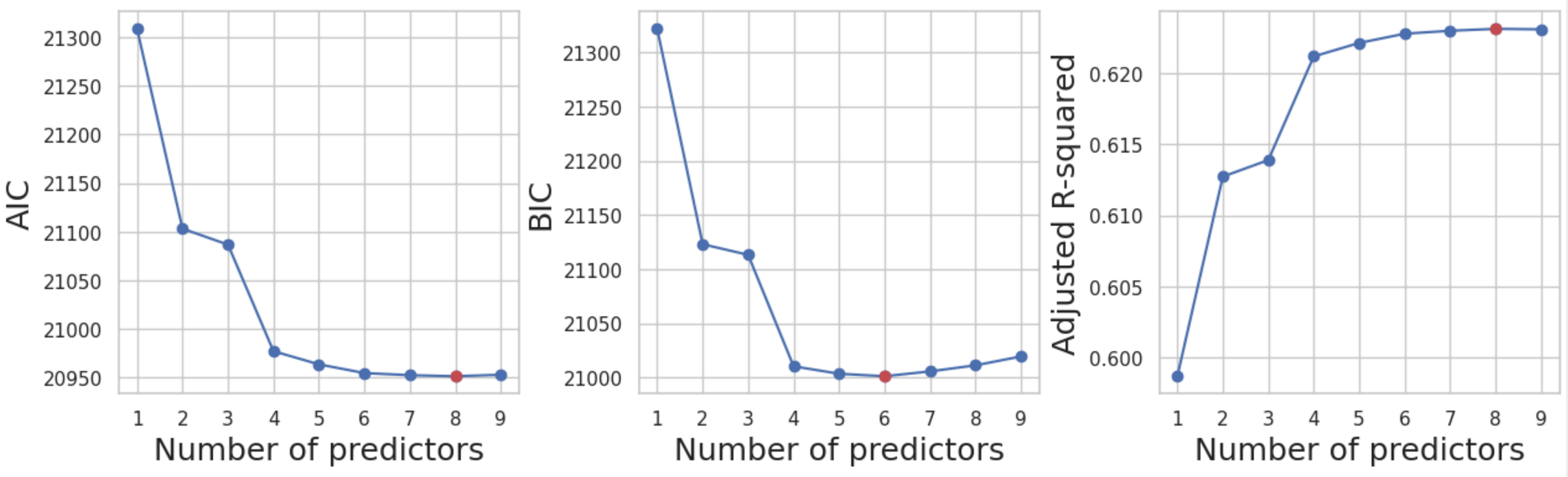
Normalization
4

Quantitative variables are normalized against the training set means and standard deviations as to make regression coefficients more interpretable.

Imputation
5

Missing data for quantitative variables are imputed using MICE imputation.

FEATURE SELECTION



Optimal features subset size seems to be 6, 7, or 8

- **minimizes AIC, BIC**
- **maximizes adjusted R-squared**

FEATURE SELECTION

Potential best features:

- Socio-economic features
  - Unemployment rate
  - Proportion of adults holding college degrees
  - Proportion of students receiving free/reduced lunch
  - Proportion of students in married families

- Capital allocation features
  - Per pupil instructor salary
  - Per pupil support services
  - Per pupil staff support services
  - Per pupil admin salary

```
subset_6 = ['rate_unemployment',
  'percent_college',
  'percent_lunch',
  'teacher_salary',
  'student_support',
  'staff_support']

subset_7 = ['rate_unemployment',
  'percent_college',
  'percent_married',
  'percent_lunch',
  'teacher_salary',
  'student_support',
  'staff_support']

subset_8 = ['rate_unemployment',
  'percent_college',
  'percent_married',
  'percent_lunch',
  'teacher_salary',
  'student_support',
  'staff_support',
  'admin_salary']
```

MODEL SELECTION & PARAMETER SWEEP

Generate tuning parameter

Generate 10,000 evenly spaced values between 0.01 and 1, and select 1000 of them randomly.

Apply elastic net regression

Loop through all 1000 potential parameters by applying elastic net regression. Start with the subset of 6 features.

Elastic net has the upside of favoring small regression coefficients like Lasso, but also preventing high volatility due to correlated features like Ridge.

Store best parameter

With each loop, take note of mean squared error on the training set. Take note of the parameter that generates the lowest mean squared error, and store this error value as well.

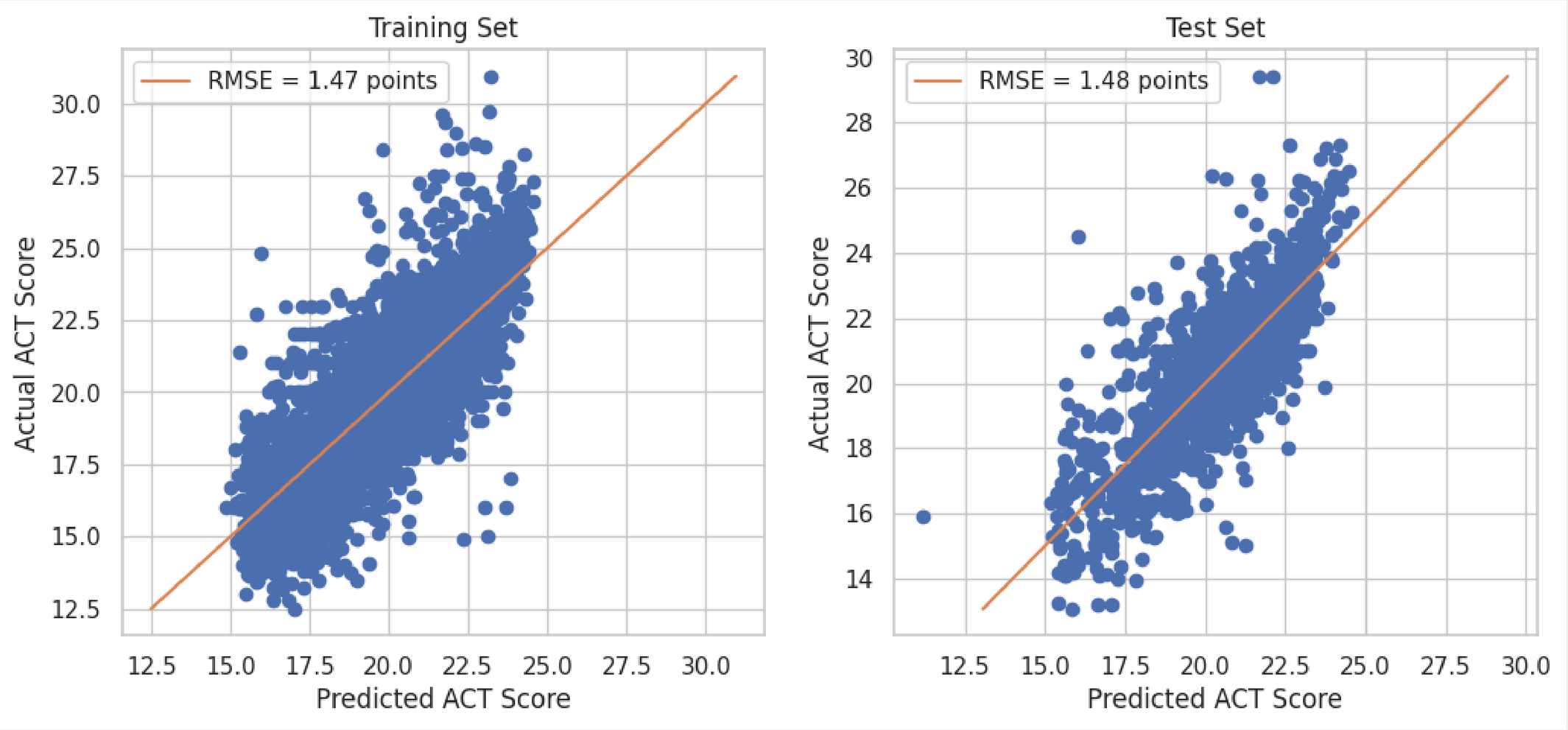
Repeat on subsets 7 and 8

Repeat the previous 3 steps on subsets with 7 and 8 features. The subset that generates the least error is best.

After comparing...

- **All 3 subsets generate similar error values. Smallest error is determined by randomly generated tuning parameters.**
- **Proceed with subset of 6 features as the least computationally intensive set**

MODEL RESULTS



MODEL RESULTS

Model can consistently predict ACT scores within 1.5 points using 6 features. These 6 features and their coefficients are:

```
rate_unemployment: -0.0931
percent_college: 0.2708
percent_lunch: -1.6275
teacher_salary: 0.3007
student_support: -0.3164
staff_support: 0.0634
intercept: 20.397941045233843
```

If a school is average in all these metrics, they are forecasted to have ACT scores of 20-21.

Proportion of students recieving free/reduced lunch has the biggest impact among all socio-economic variables. Every 1 standard deviation increase in this metric translates to 1.6 points lost on average for that school's ACT scores.

Unemployment rate has a minimal negative relationship, and proportion of college holders has a middling positive relationship to ACT scores.

For capital investment, it seems like higher instructor salaries boost scores, higher student support actually reduces scores, and more staff support has a minimal positive effect.

CONCLUSIONS

Social class matters

All socio-economic factors deemed relevant by the model suggest students attending schools in higher class areas (low free lunch availing, low unemployment rate, high concentration of degree holders) are able to score higher ACT scores.

Instructor support helps

Capital allocation features suggest it is best to invest in the school's instructors (primarily in salary, secondarily in support services) in order to improve ACT scores.

The best instructors will flock to the schools that support them the most, and this translates into better education and thus better ACT results.

The impact of high quality instructors is so significant that investing money elsewhere (eg student support services) can hurt test results more due to opportunity cost. There is also the potential issue of students not having the initiative to take advantage of resources.