March 24, 2022

Predicting U.S. Public School Reading Performance

A Classification Modeling Project

The COVID Learning Crisis

News

Massachusetts' virtual schools have 'poor academic outcomes' but state's education commissioner still recommends renewing certificates

Updated: Mar. 23, 2022, 3:42 p.m. | Published: Mar. 23, 2022, 1:28 p.m.

STUDENT & SCHOOL PERFORMANCE ASSESSMENT AND TESTING TEACHING & CLASSROOM

High school grades are up, but test scores aren't. W

By Matt

K-12 EDUCATION

Ector County Schools Fight Learning Loss With Virtual Tutors

A school district in Odessa, Texas, has adopted a highly structured virtual tutoring program that connects students with instructional support from other parts of the country and pays the contractor based on results.

March 23, 2022 • Giovanni Albanese

Content from Verizon

1 Education & Sustainability

2 Higher Ed & Blended Security

3 Higher Education Under Attack

BLOG STANDARDS, TESTING, AND ACCOUNTABILITY

Declining NAEP Scores Are Flashing Red Lights for the Covid Generation

My claim that the Great Recession was to blame for lower test scores was met with mockery. Here's the case.

Michael J. Petrilli

THE HILL

Student test scores fall for first time in national test's history

BY MONIQUE BEALS - 10/14/21 07:29 PM EDT

1,548 COMMENTS

3,100 SHARES

SHARE TWEET

Just In...

Research Problem

Does virtual schooling have as big of a detrimental effect on learning (and more specifically, reading performance) for students as many people fear?

If not, can we determine the other predictive features of schools that do well when it comes to teaching students how to read?

About the Data

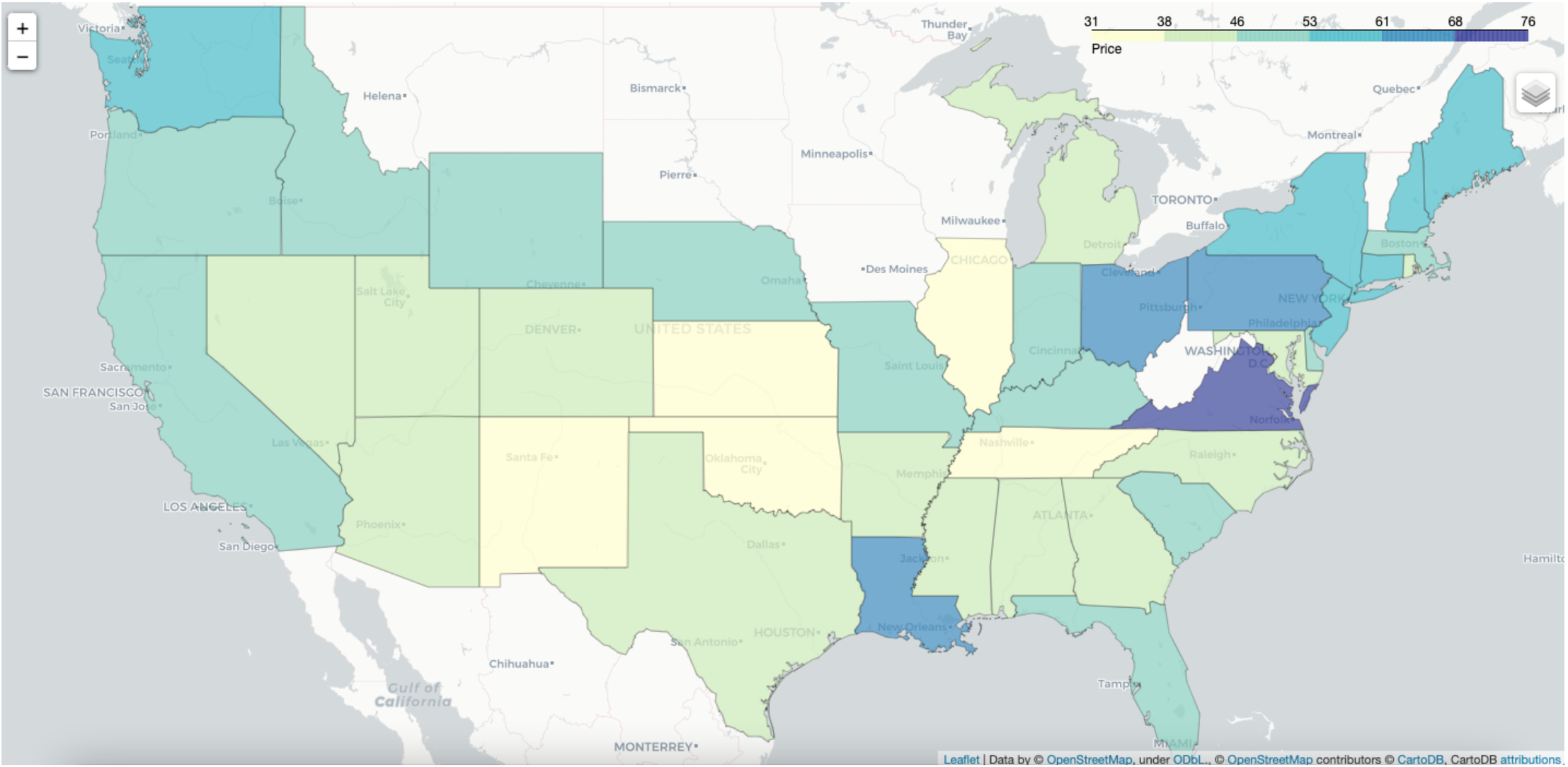
Contained numeric and categorical data from over 72,000 K-12 public schools from 43 states. The final model was trained on over 14,000 schools.

Test score, financial and characteristic data came from 2020-2021 test scores from the U.S. Department of Education's Common Core of Data. Income data came from the U.S. Census.

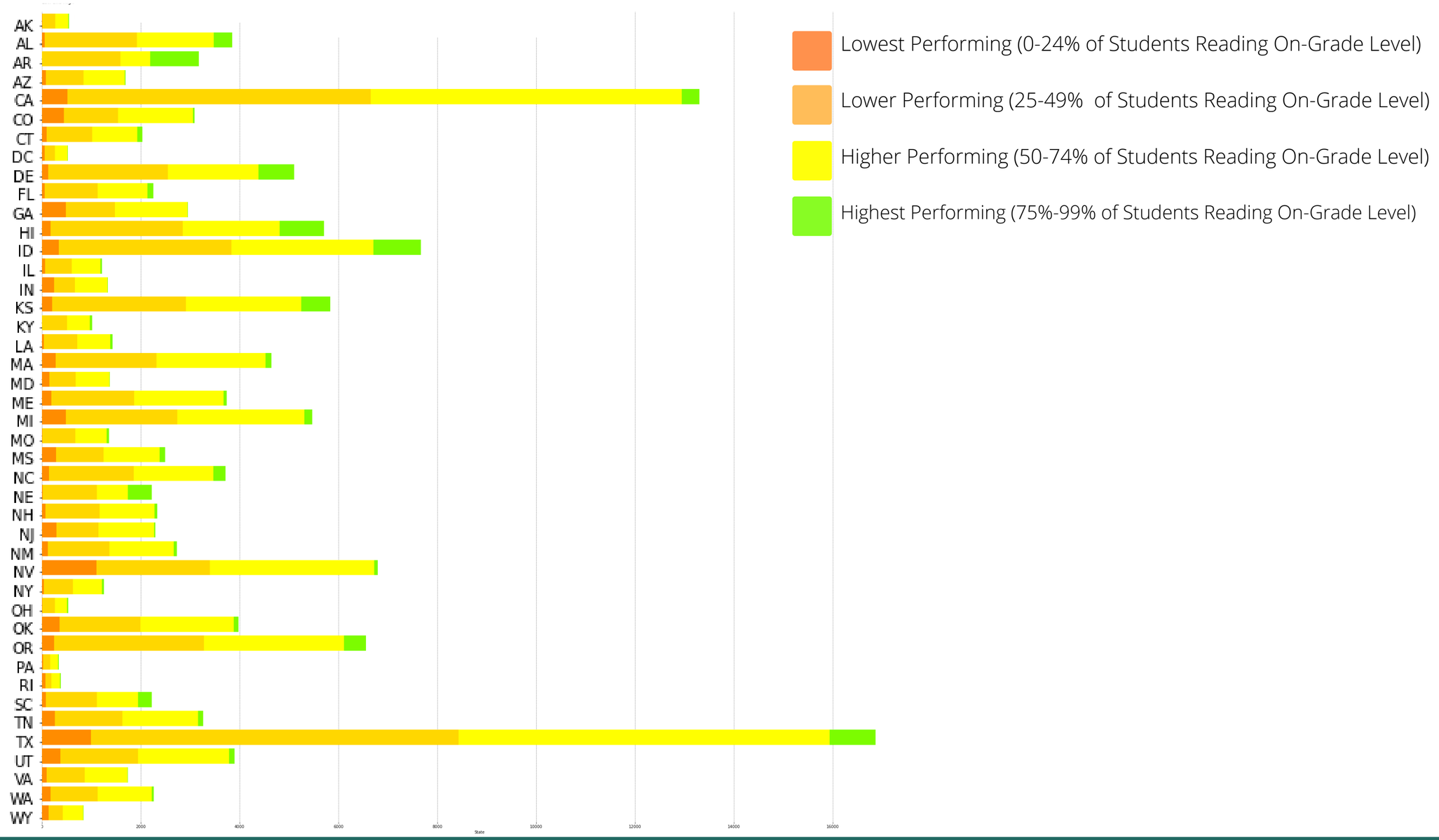
School-level descriptive data included Title I status, the presence/absence of virtual instruction, and the per-pupil amount spent on teacher salaries.

The target variable was a school's reading proficiency rate, defined as the percentage of students who scored "proficient" or above on state reading assessments.

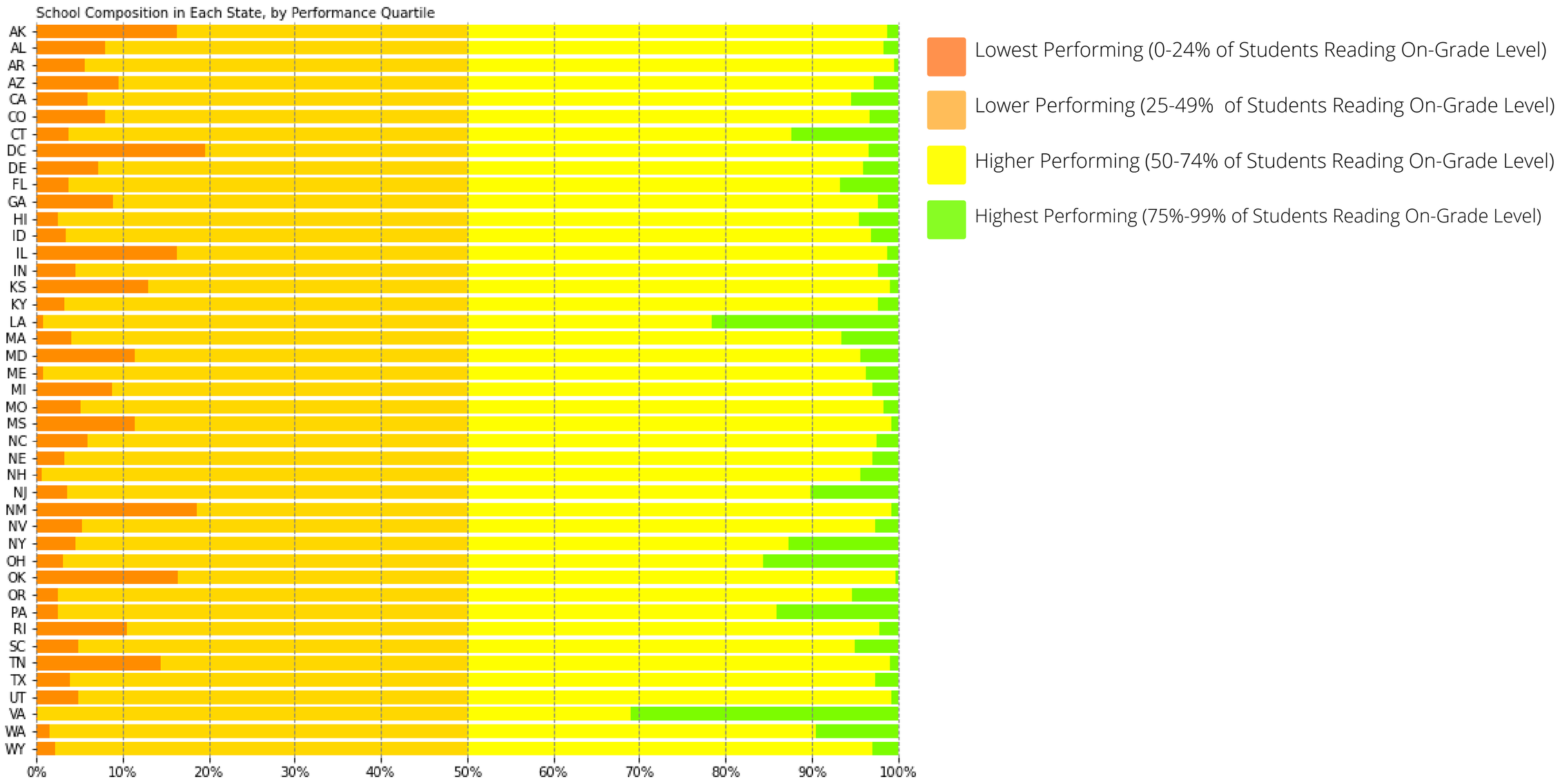
Percentage of K-12 Students Testing as Proficient in Reading, Averaged by State



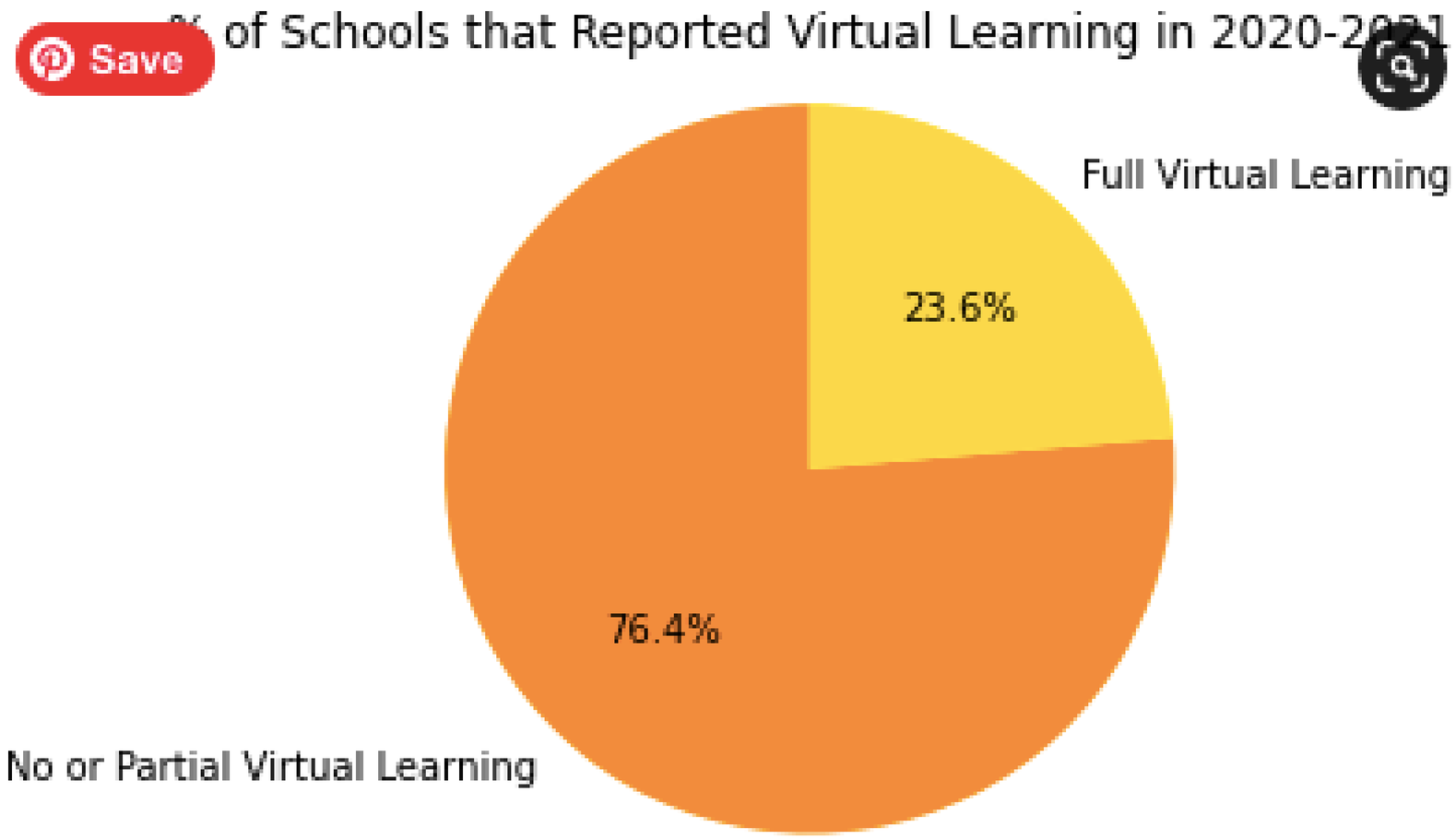
Number of High- and Low-Performing Schools by State



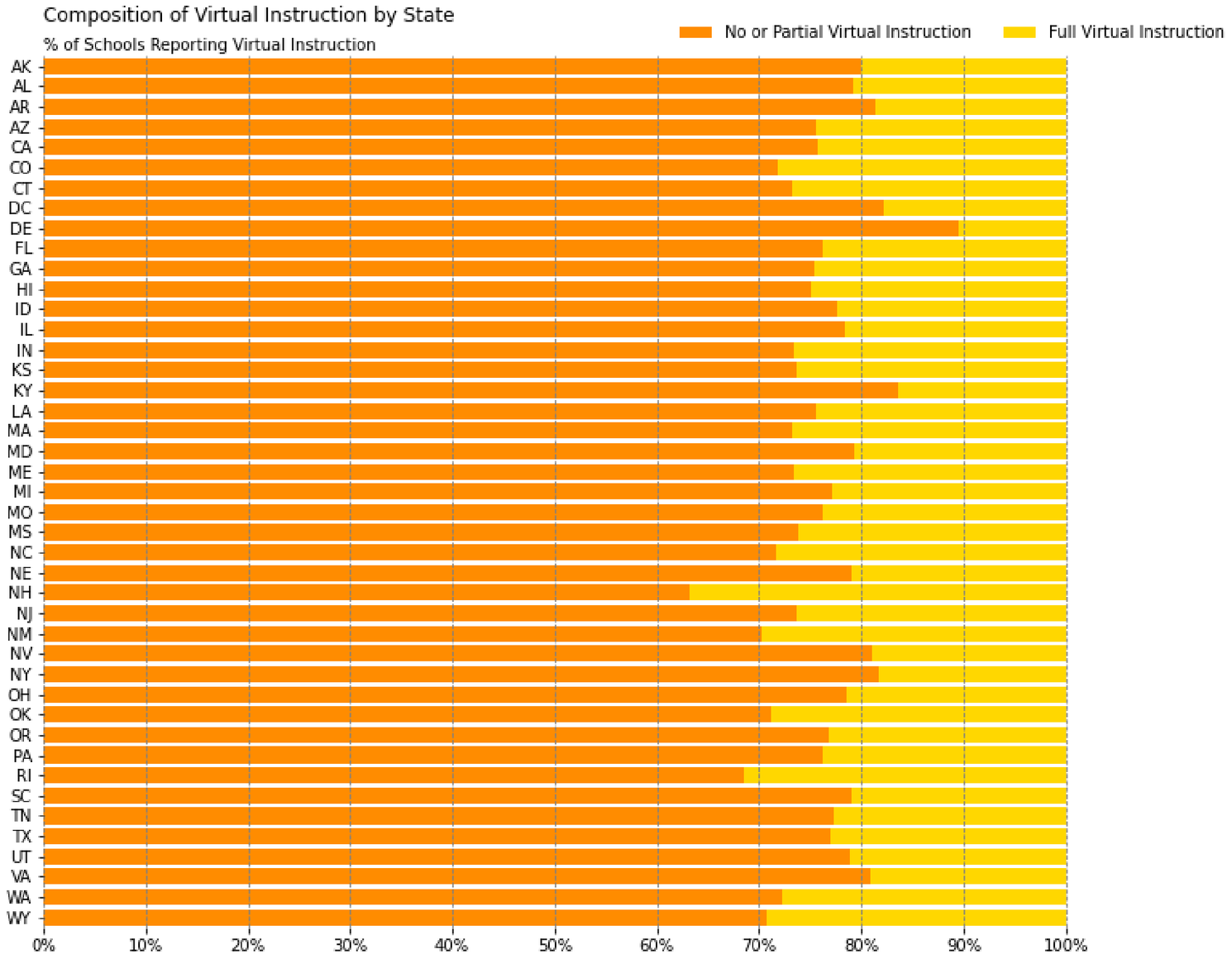
Share of High vs. Low-Performing K-12 Schools, by State



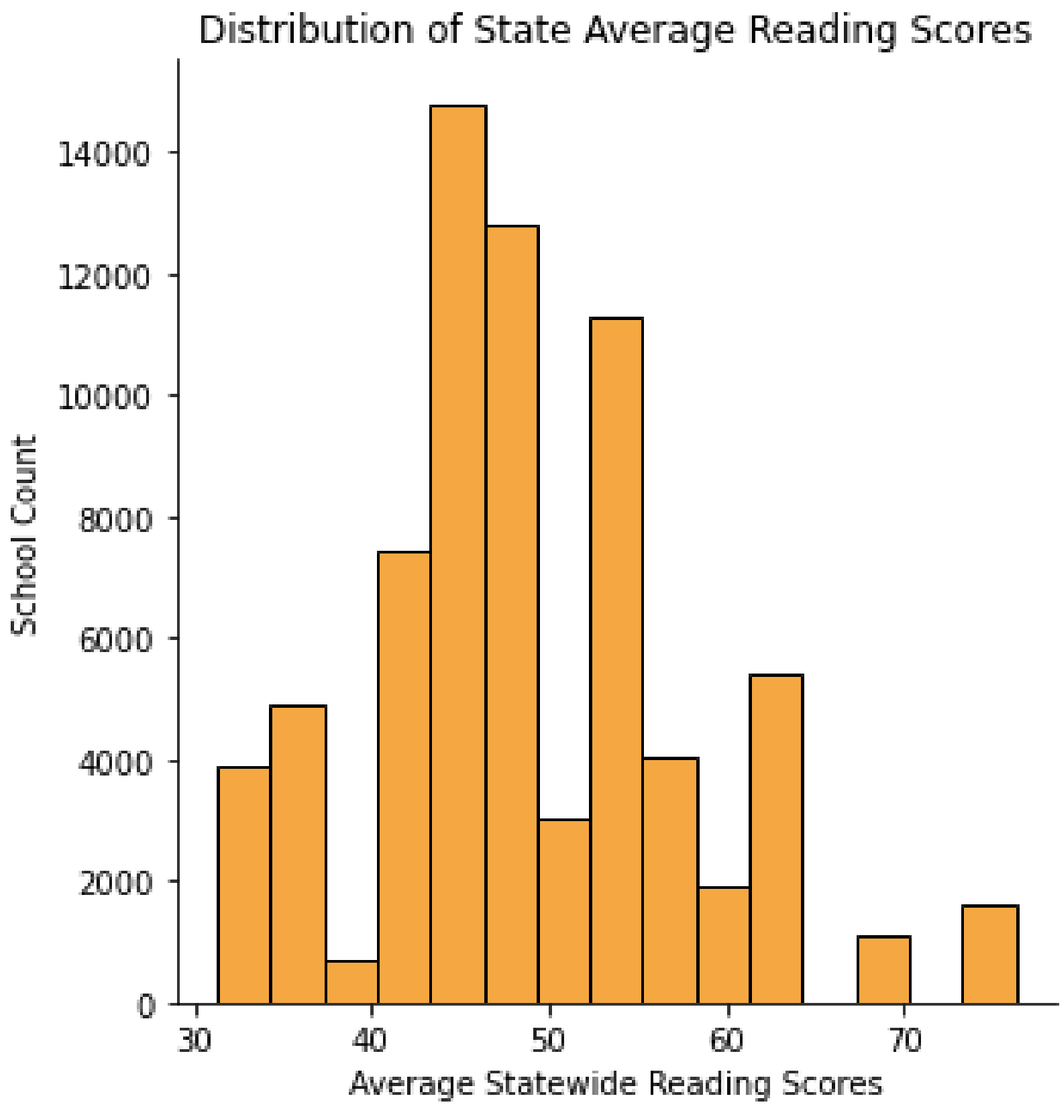
Share of U.S. Schools Who Reported Virtual Learning in 2020-2021



Share of U.S. Schools Who Reported Virtual Learning in 2020-2021, by State



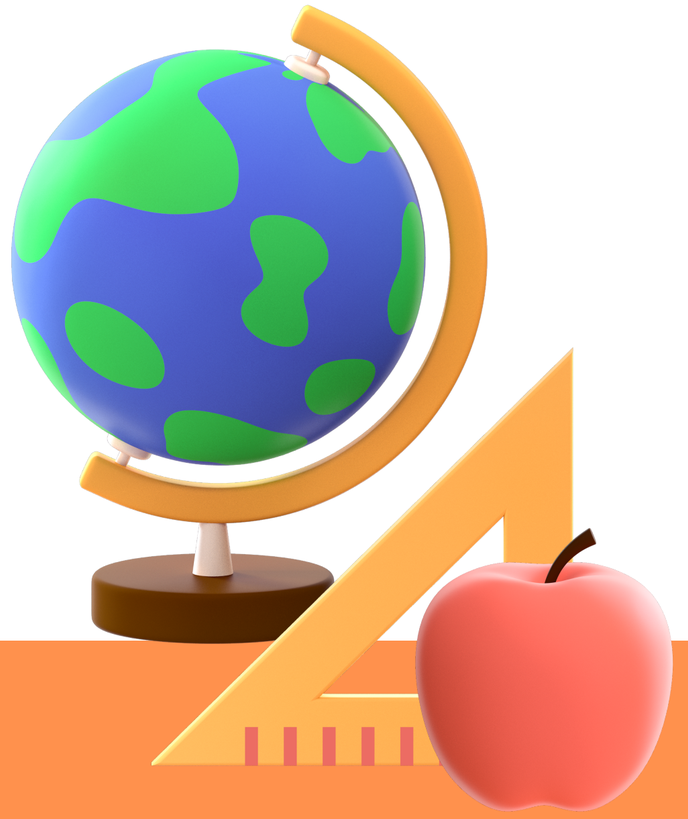
Distribution of Target Variable



Modeling Approach

- Classification Modeling
  - Characterized each school as either **"higher performing"** (reading proficiency rate of >50%) or **"lower performing"** (reading proficiency rate of <50%).

- Model attempted to **use the characteristics of each school** to predict whether the school was higher- or lower-performing.

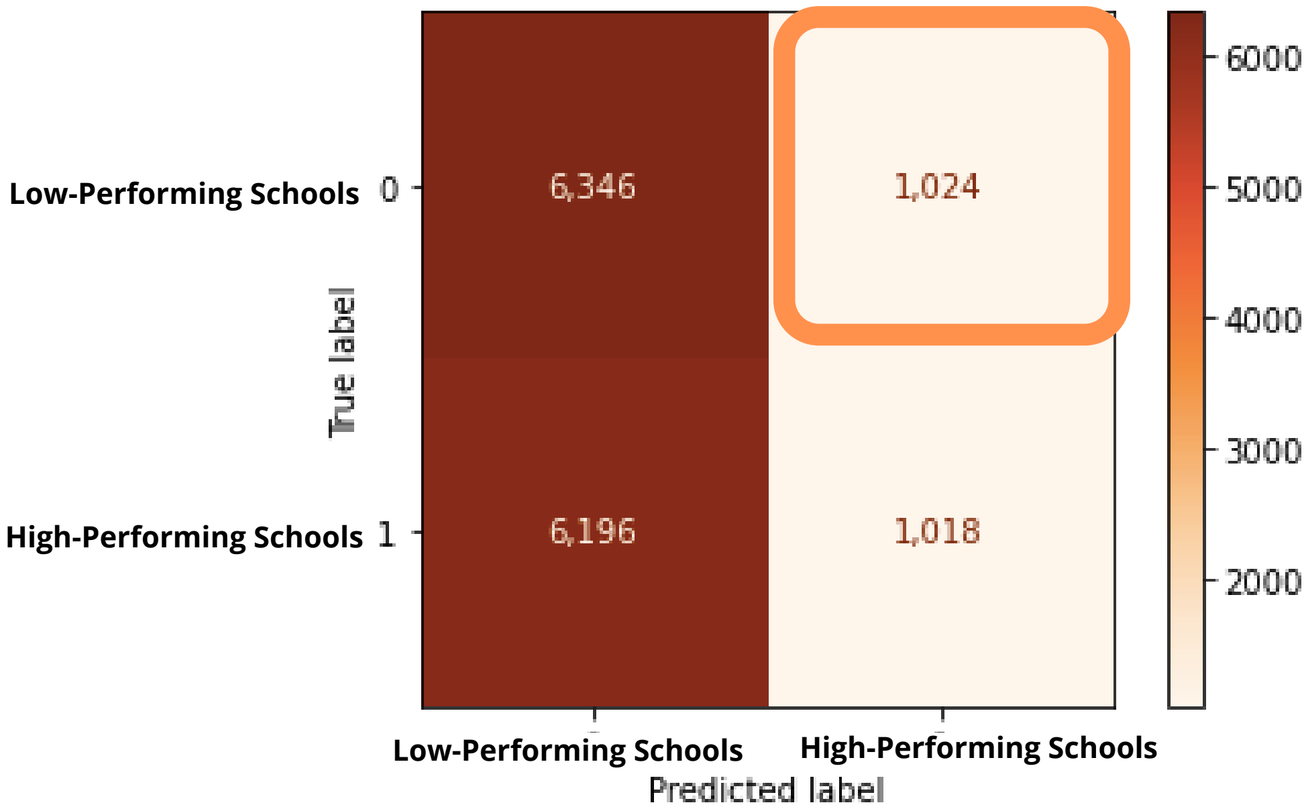


Baseline Modeling

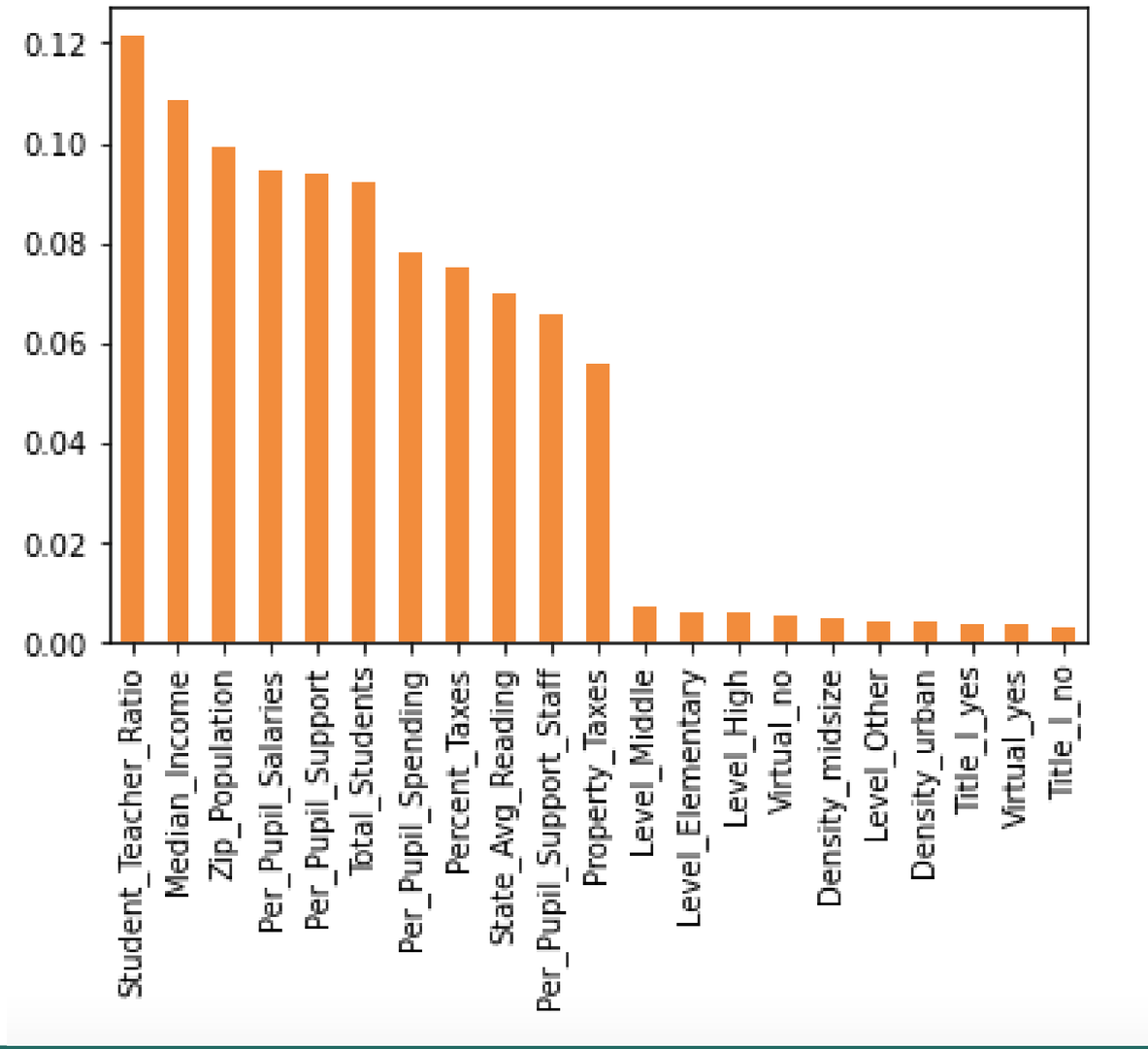
- Initial baseline models included:
  - Trying to predict a school's performance based on its **state average reading** proficiency
    - **42% accuracy**
  - Logistic Regression
    - **49-50% accuracy**

Early Random Forest Model

- Optimizing for **precision**
  - The cost of not acting is higher
  - Wanted to minimize the number of low-performing districts mistakenly labeled as high-performing
- Initial random forest model outperformed the other algorithms, but still was only correctly identifying about **50%** of high-performing schools correctly.



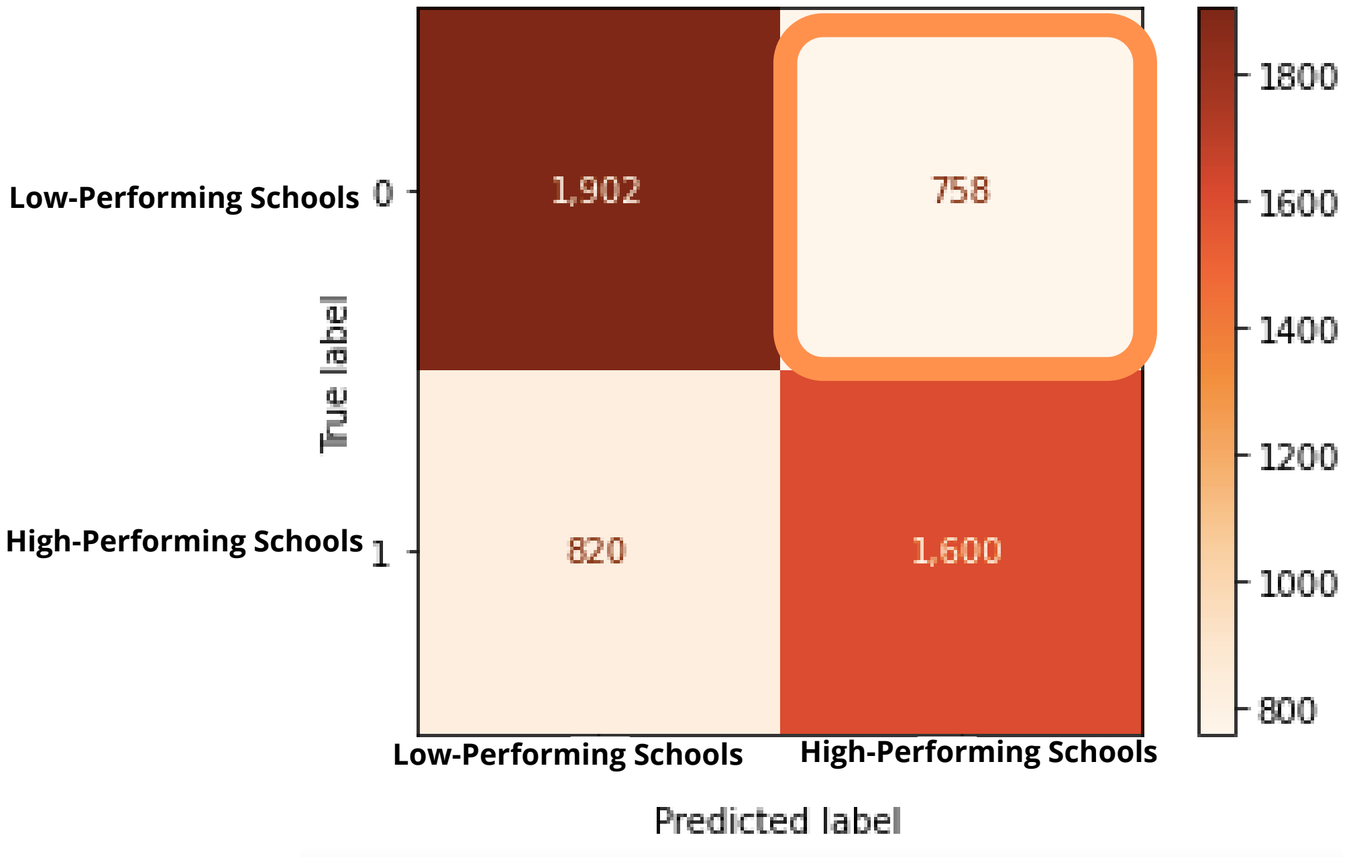
Initial Feature Importance Analysis



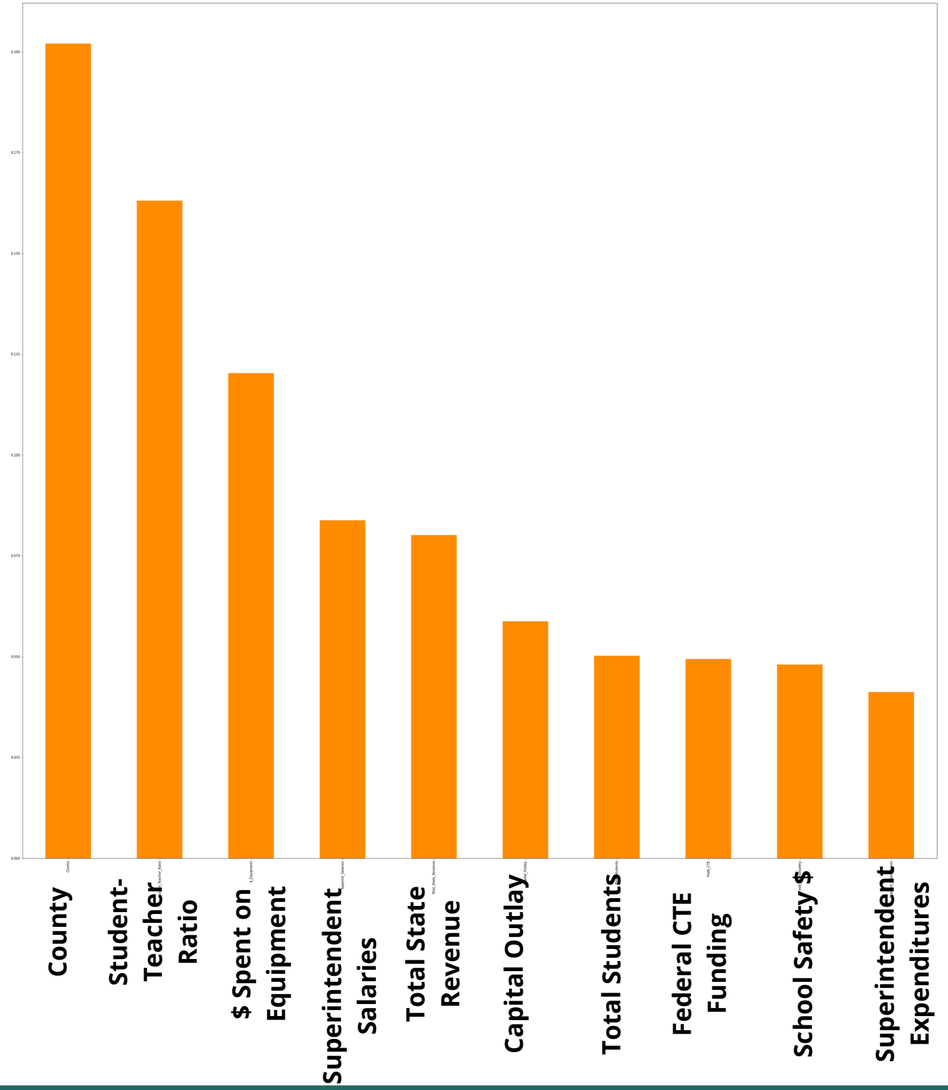
- Categorical data (such as presence of virtual learning) turned out to be not nearly as helpful to my model as I had thought.
- Student-teacher ratio was suggested to be one of the stronger predictors.
- I selected the significant features from this initial model and eventually added in some additional financial and funding data to (hopefully) improve precision.

Final Random Forest Model

- My final random forest model was faster, less prone to overfitting, and more **precise** than the other algorithms I tried
- My final random forest model performed with **70% precision** on the training set and **69.6% precision** on the testing set.



Final Feature Importance Analysis



- Adding in non-instructional, financial- and budget-related data greatly helped feed my model some valuable information about a school's performance.
- Two predictors (the **county** in which a school Is located, and **student-teacher ratio**), seem to be especially powerful.
- Student-teacher ratio, in particular, is at least twice as significant as many of the financial inputs (such as money spent on capital outlay).

Thoughts & Future Work

- **Money and location matter** --at least on a macro level.
  - Many features that ended up being the most useful for my model (e.g., county, superintendent salaries, state revenue) are indicative of a certain community's level of wealth--not so much the resources dedicated to students.

- However, a school's **student-teacher ratio** may carry a lot of leverage when it comes to how well its students learn to read.
  - A possible next step would be adapting this into a linear regression problem, so that promising variables like these could be examined using additional methods (e.g. p-values).

Thank you!

Alexis Kedo

GitHub: github.com/alexiskedo