




Is a School's Location Predictive of it Closing?

School Closure Trends in California

Presented by: Emily Lopez-Lemus
Last updated: May 23, 2024





Through my experience with AmeriCorps Reading Partners Silicon Valley, I learned about the daily reality of teachers, students and staff at a California public school post-pandemic. I was happy to see similarities to my elementary school years, but I was also very surprised to see the dramatic effect that COVID-19 had on school processes, staff retention, and student outcomes.

I served as the onsite Program Coordinator for our literacy program so I was able to work closely with volunteers, teachers, and staff. I learned that many students were far behind the reading level for their grade, there were serious shortages of teachers, and students frequently enrolled and transferred schools due to school closures or family circumstances. I also learned that some teachers must teach 2 different grade-levels in the same class due to lack of instructors.

Reading Partners specifically functions at Title 1 elementary schools where most students' families experience financial hardship. Based on the definition of a Title 1 school, most Title 1 schools are in districts that are located in neighborhoods with reported low incomes. Since I first learned about school closures while working for this program, I was curious to see if there's a relationship between school closures and districts that are in neighborhoods with historically low-reported incomes.

LOCAL NEWS - News

South Whittier School District to close Monte Vista Elementary School

Monte Vista students will go to Los Altos Elementary School, only a hop, skip and a jump away.



Monte Vista Elementary School, 12000 Loma Dr, Whittier, will close in the 2021-22 school year. (Staff photo by Mike Sprague)



By **MIKE SPRAGUE** | msprague@scng.com | Whittier Daily News
PUBLISHED: April 21, 2021 at 7:43 p.m. | UPDATED: April 22, 2021 at 5:09 p.m.

I've attended title 1 schools K-12 in Whittier, CA, so I felt a strong desire to understand the current circumstances regarding school closures and students' experiences.

Opened 1968. And closed 2021

Closing Monte Vista, will save more than \$500,000, much of that savings is coming from a loss of one principal, Gonzales said.

<https://www.whittierdailynews.com/2021/04/21/south-whittier-school-district-to-close-monte-vista-elementary-school/>

Table of Contents

School Closure Trends in California

1. Purpose Statement
2. Data Analysis
3. Predictive Modeling
4. Next Steps
5. Potential Applications

Objective

Identify if a **school's location** is predictive of **school closure** by observing trends at the district level in California.

Through predictive modeling, I would like to understand if a school's location is predictive of it being closed or open. These datasets from the National Center for Education Statistics (NCES) on California Local Educational Agencies (LEA) could help me move towards this goal by observing school closure trends at a state level and/or district level.

Data Analysis

Definitions

Local Educational Agency (LEA): An LEA is a local entity involved in education including but not limited to **school districts**, county offices of education, direct-funded charter schools, and special education local plan area.

Biased model: The presence of systematic errors in a model that can cause it to consistently make incorrect predictions, such as being biased towards the **majority class** in the dataset

Local Educational Agency (LEA): An LEA is a local entity involved in education including but not limited to school districts, county offices of education, direct-funded charter schools, and special education local plan area (SELPA).

- For purpose of this study, LEA refers to a school district

Biased model

- In our study it refers to being biased towards majority class

National Center for Education Statistics (NCES) Data

Dataset 1: Common Core of Data, Local Education Agency (LEA)

Dataset 2: Public School District Geocode File

- Dataset 1: Common Core of Data (CCD), Local Education Agency (School District) Survey Data for nonfiscal year 2022-2023
 - Gives us the main information, like location, name and status.
 - Will be our primary dataset.
- Dataset 2: GEO - Public School District Geocode File for 2022-2023
 - Can tell us if LEA is located in suburban, town, rural, or city areas.
 - Will be our secondary dataset.

Data Cleaning & Transformations

Scope:

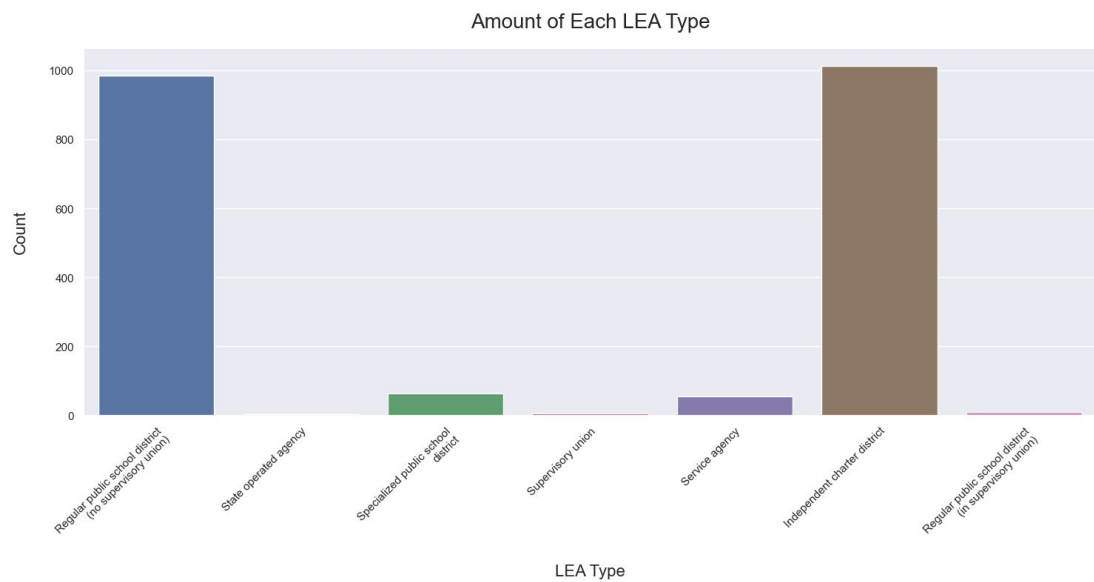
- California (CA)
- 2022-2023 school year

leaid	name	city	zip	state	locale	lea_type	lea_type_text
1232 600001	Acton-Agua Dulce Unified	Acton	93510	CA	41	1	Regular public school district that is not a c...
1233 600002	California School for the Blind (State Special...	Fremont	94536	CA	21	5	State operated agency
1234 600003	California School for the Deaf-Fremont (State ...	Fremont	94538	CA	21	5	State operated agency
1235 600006	Ross Valley Elementary	San Anselmo	94960	CA	21	1	Regular public school district that is not a c...
1236 600007	CA Sch for the Deaf-Riverside (State Special S...	Riverside	92506	CA	11	5	State operated agency

start_status	start_status_text	updated_status	updated_status_text	effective_date	operational_schools	lowest_grade_offered	highest_grade_offered	lea_level
1	Open	1	Open	8/21/23	3	KG	12	Other
1	Open	1	Open	8/21/23	1	KG	12	Other
1	Open	1	Open	8/21/23	1	KG	12	Other
1	Open	1	Open	8/21/23	5	KG	8	Elementary
1	Open	1	Open	8/21/23	1	KG	12	Other

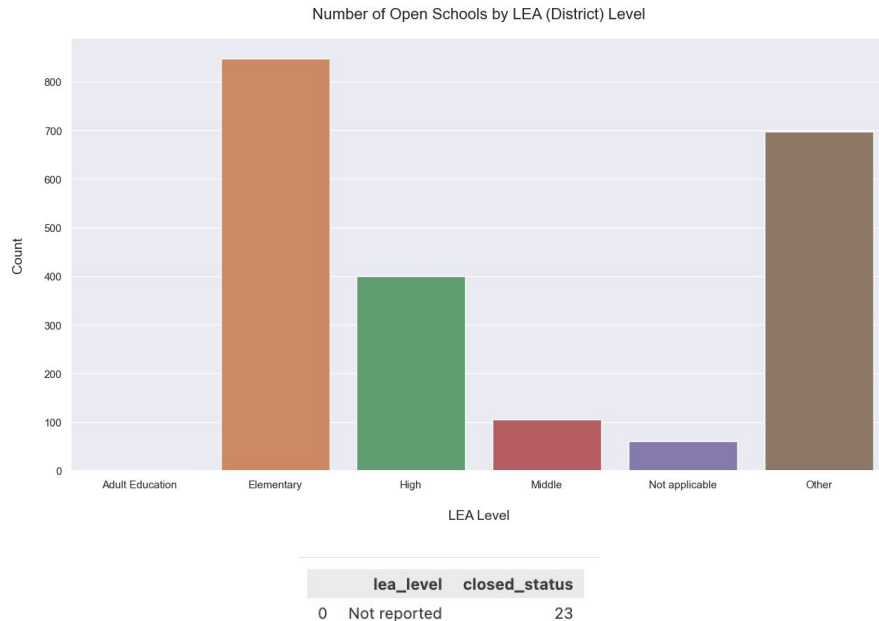
- Scope: Focus only on LEAs in California
 - I reduced the scope to CA LEAs to understand the relationship between closures of LEAs and locations in CA.
- 2022-2023 school year

- The datasets looked like I expected them to based on the documentation.
- The state name, id, and FIPST were all consistent (location values that should be consistent with each other),
- we only have data for 2022-2023
- we don't have missing values for our state of interest.
- The GEO dataframe had the same number of values/rows as the LEA set,
 - which makes sense and implies we are not missing any district/LEA data if there are no missing values.
- I also checked the primary key LEAID to make sure it made sense to join both tables on it.
 - My analysis showed that each LEAID value in LEA and GEO were unique
 - and there are no null or missing values;
 - therefore, it was safe to use LEAID as the primary key.



The top 3 LEA Types are

- 'Independent charter district' with 1025 values,
- 'Regular public school district that is not a component of a supervisory union' with 983 values,
- 'Specialized public school district' with 64 values.

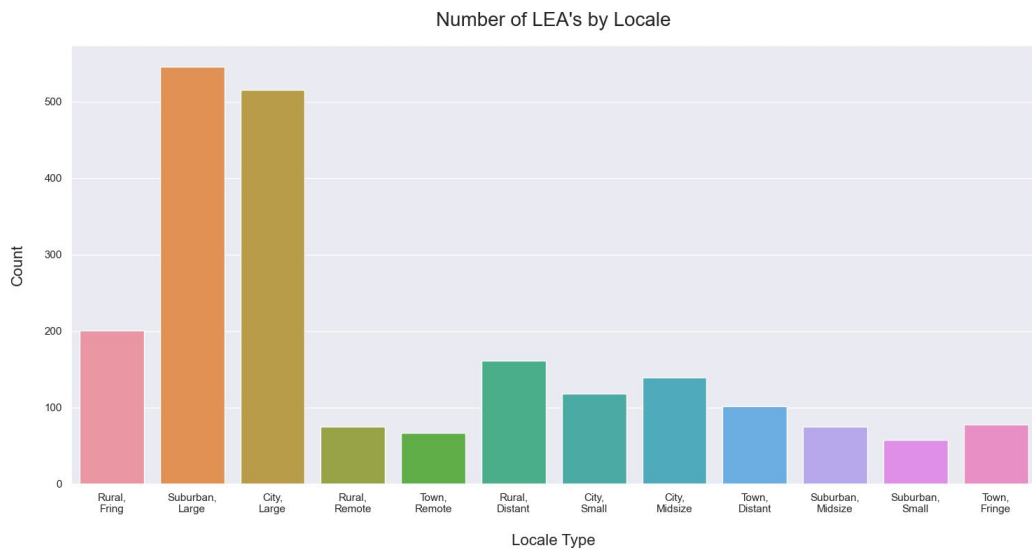


We see the top 3 levels are

- elementary with 850
- , "other" with 699
- and high with 401.
- Middle school which follows high only has 105.

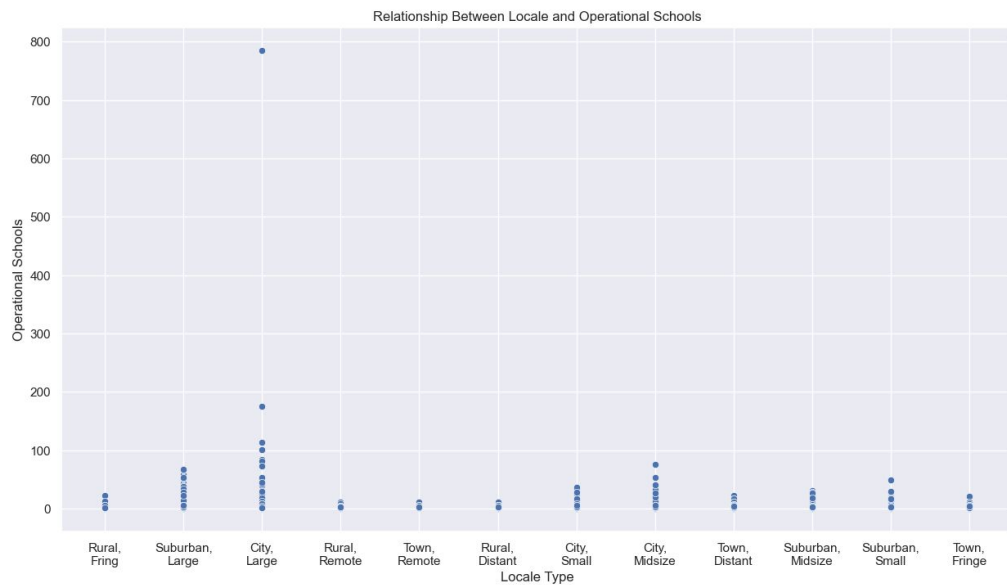
It was interesting to see that one of the most popular levels in California was "other"

- I've done some research, but it is unclear what "other" may stand for.
- Generally, it is a LEA or School level that does not fall under the categories listed.
- My best understanding based on documentation:
 - may stand for "Other High School Completers" which they define as "a certificate of attendance or other certificate of completion awarded in lieu of a diploma. Not included are equivalency diplomas, such as the GED or the HSED. Recipients of equivalency diplomas, such as a GED or HSED, are excluded from this survey."



Most CA LEAs are located :

- within Large Suburban with 548 values,
- Large City with 519, and
- Rural Fringe with 203 values.



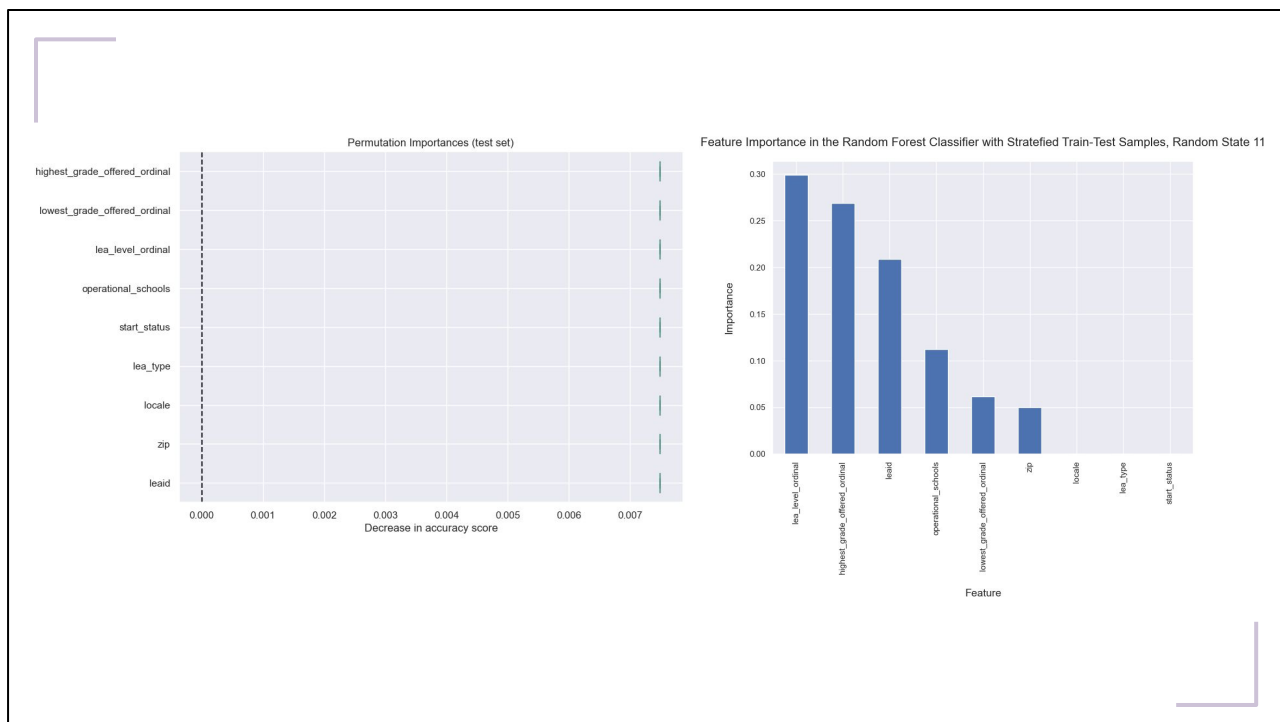
After looking at a scatter plot of these locales,:

- we see it closely resembles the LEA count x locale bar graph;
- upon closer look, we see an outlier under city,large locale for operational school counts that belongs to Los Angeles Unified.
- This did not appear to skew our data since suburban large and city large were far above the other locales.

```
updated_status
1    2112
2      23
Name: count, dtype: int64
-----
Class='Open' : 2112/2135 ( 0.989)
Class='Closed' : 23/2135 ( 0.011)
```

Within the CA subset, the most common status by far was 'open' at 2112 followed by 'closed' at 23.

Predictive Modeling



After discovering the significant imbalance among open vs closed LEAs in our final dataset for CA sites (0.99 open vs 0.01 closed):

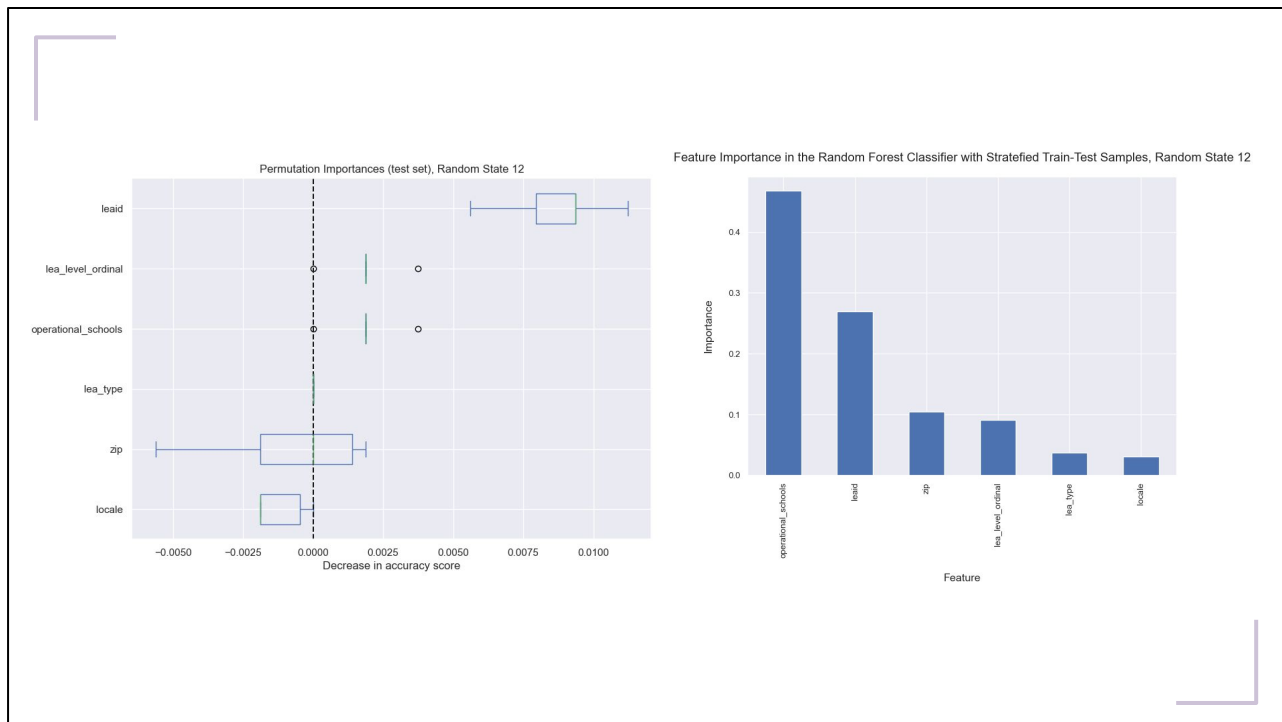
- I had to switch from my initial plan of using a logistic regression to an ensemble method for the predictive model.
- Having an imbalanced dataset when doing our predictions could result in:
 - a biased model towards the majority class ('Open' status LEAs),
 - poor generalization to new data especially for the minority class ('Closed' status LEAs)
 - and misleading evaluation metrics, like accuracy and feature importance since the model might predict the majority class most of the time.

After seeing the performance of the bagging approach and random forest classifier:

- I discovered that although both had misleading model accuracy results, the random forest seemed to have the most potential for fine-tuning
- since bagging models had an Accuracy score of 1.0 for training and test data whereas the random forest classifier had an accuracy score under 1.0.
- may be because the random forest classifier samples from a subset of all the training features in addition to the bootstrap samples of the training dataset and fits a decision tree to each
 - . I think having random features helps us see how certain features have greater importance than others for predicting the outcome (open/closed) and prevents overfitting.

I looked for ways to reduce bias towards the majority class through stratified sampling for train-test sets which

- appeared to have a less biased approach when used in conjunction with a random forest classifier.
- With stratified sampling, we maintain the same class distribution in each subset of the randomly split dataset which is necessary for our imbalanced dataset (imbalanced closed/open status outcomes).



After looking at feature importance through two different evaluators:

- I was surprised to see that zip and/or locale did not have as big of an importance in determining if a site would be closed or open.
- In the impurity-based feature importance evaluation, operational schools feature was consistently most important,
 - which I believe may be because in the dataset if a site is closed they have 0 operational schools.
- The permutation feature importance evaluation showed that LEAID is consistently the most important feature of a closed or open LEA.
 - I believe this may have some connection with the LEA location but further analysis would be required.
- LEA level also appeared to have an influence on whether a site would be closed or open as well,
 - but upon a closer look at the data, all closed LEAs had an LEA level of "not reported."
 - This may have the same effect on our model as operational schools x closed LEAs.

Next Steps

How to improve the model

1. Use a larger scope: analysis on more than one state
2. Improve district level analysis to observe wealth gaps
3. Additional predictive modeling approaches
4. Additional NCES dataset on School Neighborhood Poverty Estimates

****Scope****

One potential idea was to approach it by looking:

- at closed vs open LEA sites on a bigger scale,
- Looking at more than one state and not just CA.
- However, when looking at the percent of closed and open schools in our entire available dataset,
 - we see that across 56 US States and territories, there is a similar ratio of 0.99 open schools to 0.01 closed schools.
 - we would likely run into the same issue of having a highly imbalanced dataset regarding our desired outcome variable, updated status (closed/open).
- district level could also help us see the wealth gap more clearly
 - since situations district to district could be very different.
 - For example, the resources and student outcomes for students in Piedmont Unified School District vs Oakland Unified School District may be very different due to income inequality between these two districts, despite being in the same County.

****Predictive model****

For future next steps, it would be good to adjust our model

- to account for the operational_schools feature when sites are closed since it is intuitive that there would be 0 operational schools

- we do want to keep the other operational_schools values.
- account for LEA Level, since all LEAs with "Closed" status are listed as "not reported"
- may be contributing to overfitting as opposed to helping the prediction.

beneficial to determine which features may be correlated, > cluster these features, > only keep one feature from each cluster

- get a better understanding of what features are actually important.
- For example, location features may be highly correlated and in effect skew our evaluation of most important features when using the permutation approach.

continue to explore other predictive models that are suited for imbalanced datasets
continue with feature selection to refine our random forest classifier.

However, another solution may be finding an alternative outcome/target/dependent variable to analyze that may give us similar insight.

****Additional datasets****

- incorporate the (NCES) dataset on School Neighborhood Poverty Estimates
 - better understand the historical economic landscapes of school district's surrounding neighborhoods, particularly LEAs who face more closures.
 - Currently, our model and analysis needs more fine-tuning before introducing this supplemental information.
 - Additionally, the most recent data available is for 2020-2021, so it would be better to include current data once it becomes available.
 - Lastly, this dataset is at a school level, not district so further brainstorming will be needed to see how to best integrate it.

Potential Applications

EQUITY & DIVERSITY

Race Is a Big Factor in School Closures. What You Need to Know

By Eric Lipton & Seara Najarro — November 28, 2023 5 min read



— Stock/Getty Images Plus

SCHOOL & DISTRICT MANAGEMENT

Pressure to Close Schools Is Ramping Up. What Districts Need to Know

By Mark Lieberman — January 24, 2024 8 min read



— Illustration by Liz Tapp/Education Week (Images: Stock/Getty)

These findings could **inform strategic planning and interventions** to facilitate closures and **minimize negative implications** of these transitions on students, educators and neighborhoods.

Once this analysis and modeling is complete, the insights could:

- what districts need additional funding, resources and support,
 - like teaching staff,
- help decision makers determine best course of action for distributing funds and resources
- help institutional analysts understand how neighboring LEAs within the same city compare to one another
- compare to:
 - student assessment data/ performance metrics
 - student and teacher satisfaction metrics to see what implications the school closures may have in the locations we identify

Overall, these findings could:

- help education decision makers and administrators determine which areas may experience more closures in advance.
- inform strategic planning and interventions to facilitate closures
 - could minimize negative implications of these transitions on students, educators and neighborhoods.

Thank you!