

Predicting Lead Contamination in NY Schools

Erdos Data Science Bootcamp

November 2025

Cami Goray, Hana Lang, Ranadeep Roy

Invisible Hazard: Lead Exposure Through School Water in New York

Goal: Analyze a school's risk for contaminated lead pipes based on geographic, infrastructural and demographic factors

Motivation:

- Provide insights to make informed policy decisions, awareness to NY families
- Identify key risk predictors
- Understand which populations are most at risk

Motivation

Lead is a toxic metal that can cause health problems, especially in children:

- According to the CDC, “there is no safe blood lead level that has been identified for young children.”¹
- The EPA has “set the maximum contaminant level goal for lead in drinking water at zero.”¹
- Even small amounts can **affect memory, behavior and learning ability.**

A 2024 study from Cornell University found that **47% of schools** had at least one outlet that tested above 15 ppb.²

¹<https://www.cdc.gov/lead-prevention/prevention/drinking-water.html>

²New York State Water Resources Institute

Any amount of lead exposure is unsafe to children.

Therefore, we ask the question:

Any amount of lead exposure is unsafe to children.

Therefore, we ask the question:

Can we predict whether a school has contaminated drinking water?

Data Info & Data Cleaning

Main Dataset: Lead Testing in School Drinking Water, from the NY State Department of Health.

Size: 3028 rows and 24 columns

Removed:

- Rows with negative values
- Features which were redundant, unnecessary, or obviously correlated

Target Variable:

- Binary target variable: whether or not there is a drinking outlet in the school with lead contamination > 5 ppb
- This was engineered from the original “Number of Outlets Sampled Above 5 ppb” column

Feature Engineering

Combined and cleaned data from multiple sources, utilizing web scraping to obtain the majority of features.

NYC counties: data sourced from
NYC Environment and Health Data
Portal

Non-NYC counties: data sourced
from NY state Department of Health
Childhood Lead Exposure dataset

Socioeconomic:

- Student/Teacher Ratio
- % students eligible for Free & Reduced Lunch

Demographic:

Enrollment by race for each school

Infrastructure:

% of houses per county built before 1950

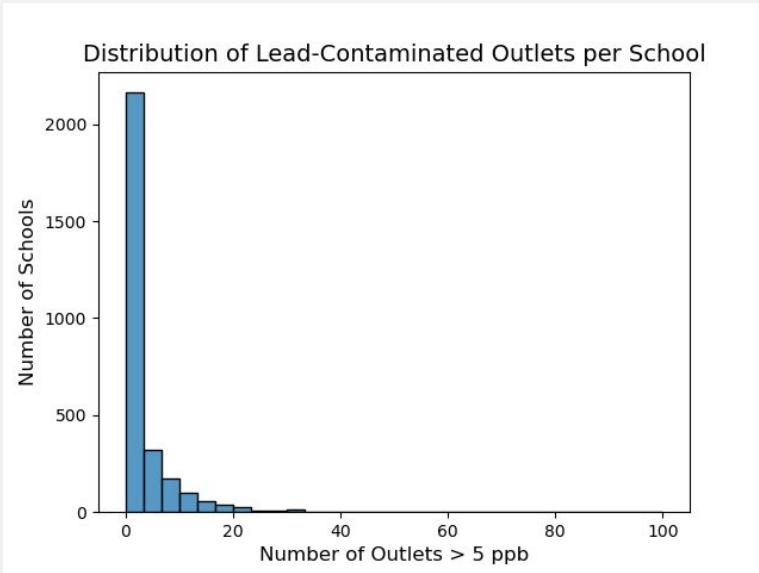
Geographic:

- City location
- School district

Scraped from National Center for Education Statistics website

From main dataset

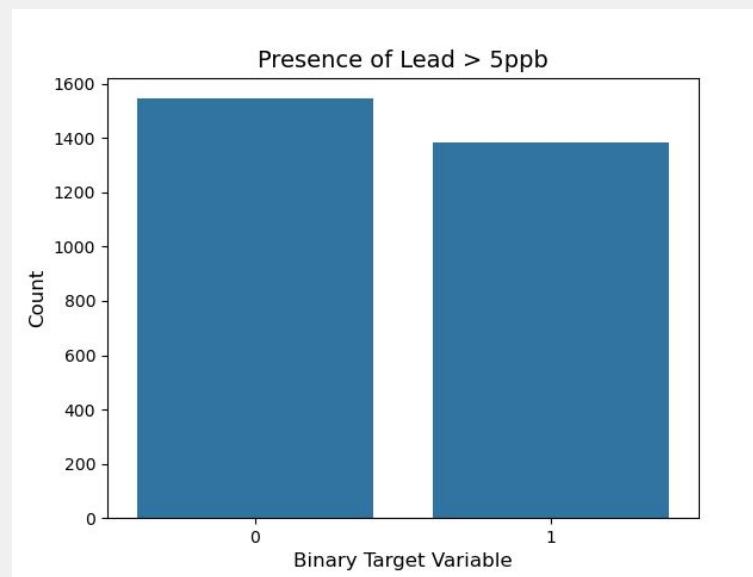
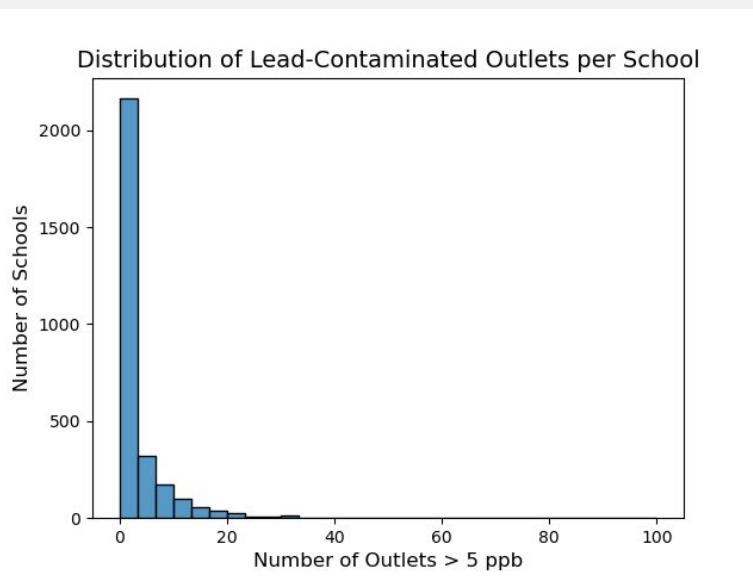
Exploratory Data Analysis



Skewed distribution:

The majority of schools in the state have relatively few outlets with lead > 5 ppb.

Exploratory Data Analysis



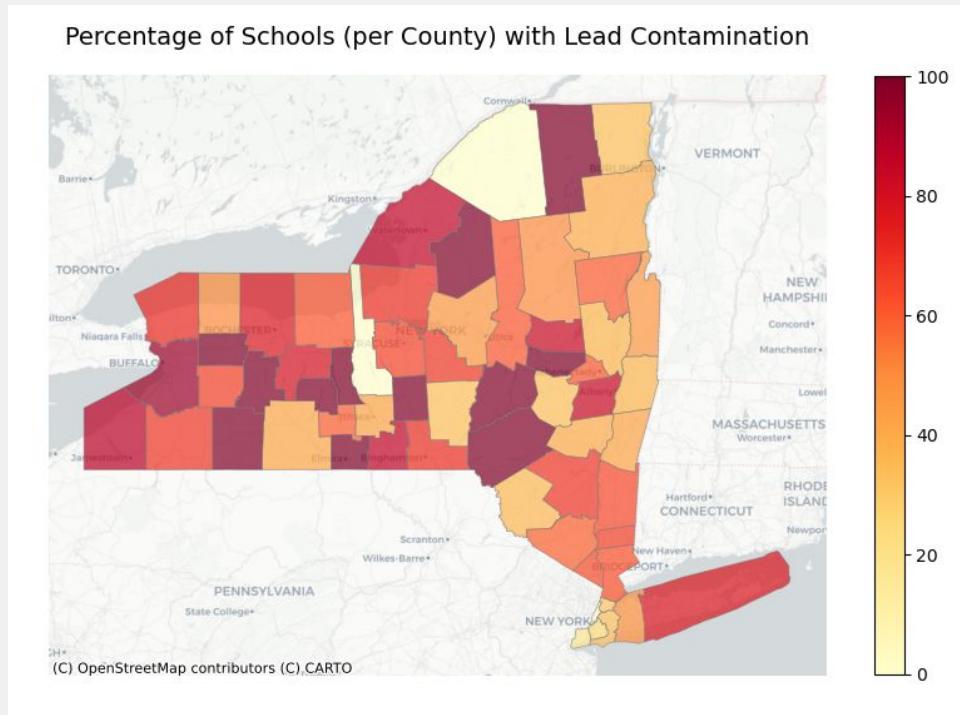
Skewed distribution:

The majority of schools in the state have relatively few outlets with lead $> 5 \text{ ppb}$.

Balanced distribution:

Creating a binary target variable leads to relatively balanced data.

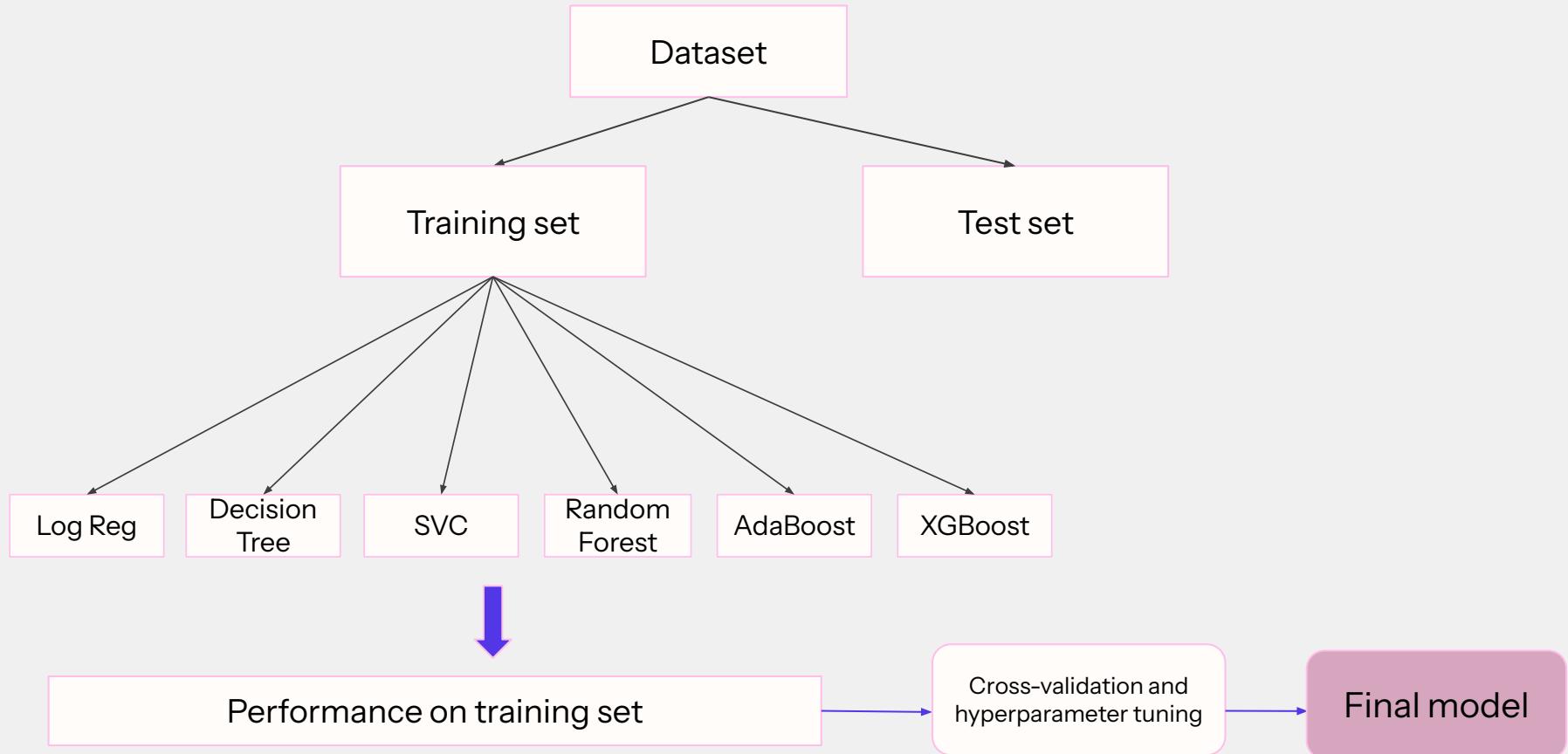
EDA: Geographical



Geographical correlation:

Certain counties have a higher proportion of schools with lead contamination in drinking water than others.

Modeling Approach



Model Comparison

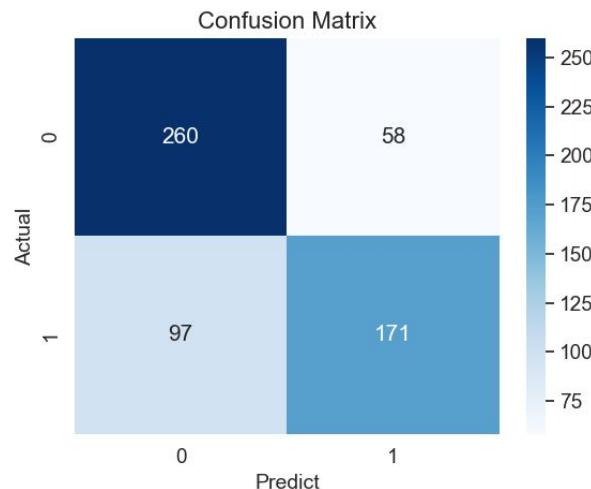
Model	ROC-AUC score (mean)
1. Logistic Regression	0.7621
2. Decision Tree	0.7118
3. SVC	0.7383
4. Random Forest	0.7259
5. AdaBoost	0.7358
6. XGBoost	0.7314



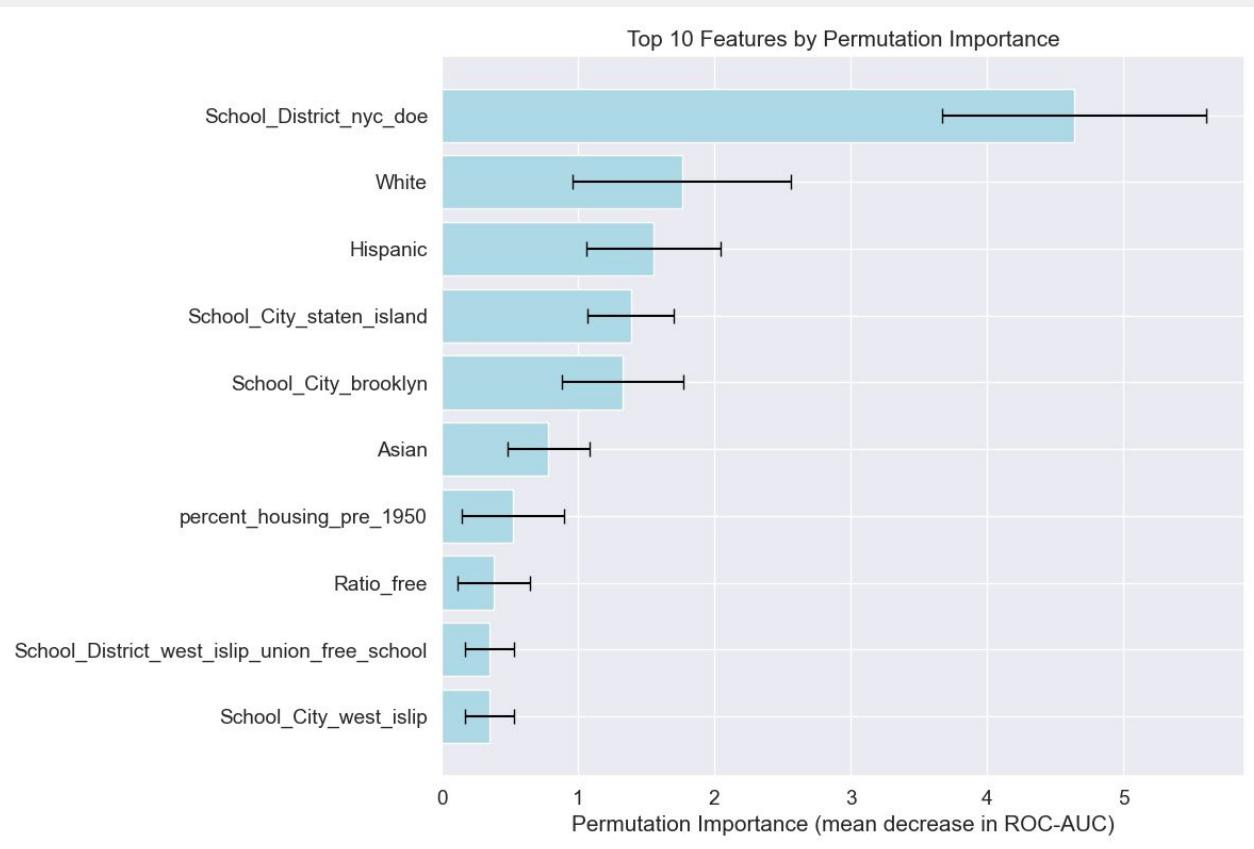
On Test Set

Best model: Logistic Regression with tuned hyperparameters

ROC-AUC Score: 0.7278



Feature Importance for Final Model



Observations:

- **Geographical features** (school district, city location of school) are strong predictors
- Features **White**, **Ratio of Free and Reduced Lunch** **Eligible students** also appear (as suspected in EDA)

Conclusion & Next Steps

Spatial and
Environmental
Analysis



Since geography plays a critical role in lead detection, incorporate **spatial modeling** to identify “hotspots” of elevated lead levels. Link **water system maps, soil contamination data, or infrastructure age** to reveal patterns across neighborhoods and help explain the presence of these hotspots.

Conclusion & Next Steps

Spatial and Environmental Analysis



Since geography plays a critical role in lead detection, incorporate **spatial modeling** to identify “hotspots” of elevated lead levels. Link **water system maps, soil contamination data, or infrastructure age** to reveal patterns across neighborhoods and help explain the presence of these hotspots.

Geographic Scope



Lead contamination is not limited to New York. Expand modeling to **incorporate data from other states**.

Acknowledgements

Thank you to the Erdos Institute and to our project mentor, Hannah Lloyd.