Predicting Lead Contamination in NY Schools

Github: https://github.com/Erdos-Projects/fall-2025-healthcare-costs-and-preventative-care

Datasets:
https://health.data.ny.gov/Health/Lead-Testing-in-School-Drinking-Water-Sampling-and/rygk-rhum/about_data
https://nces.ed.gov/ccd/schoolsearch/
https://www.health.ny.gov/statistics/environmental/public_health_tracking/about_pages/childhood_lead/export
https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/lead/?id=16#display=summary

Background and Project Overview
We developed a comprehensive analysis to predict the presence of lead contamination in New York school drinking water using a rich dataset of demographic, socioeconomic, infrastructural, and geographic features. We obtained our main dataset from New York's Department of Health.[1] Our target variable is a binary: 1 if there is a drinking outlet with above five parts per billion (>5 ppb) lead contamination, and 0 if there are no such drinking outlets. This level is recommended by the New York Public Health Law, which governs school potable water testing standards.[2] Our predictive features include school name, school district, percentage of students on free or reduced lunch, student-teacher ratio, demographic enrollment percentages (White, Black, Hispanic, Native American, Multiple Races) per school, and percentage of buildings in each county built before 1950, since building materials pre-1950 were more likely to contain lead in plumbing and paint materials.

Stakeholders
Our primary stakeholders are education administrators in the State of New York, as well as the NY State Department of Health, New York Schools, and NY parents and families of schoolchildren, as this project aims to predict lead contamination and therefore improve safety of school facilities.

Data Cleaning
We removed features which were redundant, unnecessary, obviously uncorrelated, or had ambiguous interpretation, including:

- Lead testing data from 2023 and 2024 (it was not clear whether outlets were resampled year-to-year)
- Extraneous geographical features such as zip code, county, county location, state, and school street name.

We created:

- A binary target column for lead presence > 5 ppb: 0 or 1. This was engineered from the Number of Outlets Sampled Above 5 ppb column.

---

[1] https://health.data.ny.gov/Health/Lead-Testing-in-School-Drinking-Water-Sampling-and/rygk-rhum/about_data

[2] https://www.nysed.gov/new-york-state-school-deaf/lead-testing-drinking-water

We transformed:

- Categorical Data: applied One Hot Encoding
- Numerical Data: applied StandardScaler
- Imputed data using a median strategy, since some features had skewness

Modeling Approach

That dataset was split into training and testing sets, with 20% of the data set aside as the final test set. For each model, we performed hyperparameter tuning with nested cross validation using random search. Models were considered in increasing order of complexity: logistic regression (with and without penalty), decision trees, support vector machines, random forest, AdaBoost and XGBoost. All the models were evaluated on the basis of mean ROC-AUC scores. This metric was chosen because our dataset was nearly balanced and we wanted to ensure that our model can distinguish between the two classes well.

Results

Our findings revealed that the hyperparameter-tuned logistic regression model produced the best mean AUC score in the training stage, with an average ROC-AUC score of 0.7621 across the five outer folds of the nested cross-validation. On the final validation set, we obtained an ROC-AUC score of 0.7278 on the validation set using the Logistic Regression with penalty and the best hyperparameters found in the training stage.

We also performed a feature importance analysis using permutation importance methods. The most important features of a school were: (1) a school district of NYC Department of Education, (2) the proportion of white students, (3) the proportion of Hispanic students, (4) a city location of Staten Island, and (5) a city location of Brooklyn. Based on EDA, we had expected geographical features to appear as important predictors (see the county-wide heat map displaying the proportion of schools with target variable 1). We discuss the implications of the importance of geographical features in the next section.

Limitations & Next Steps

Since geographical features (school district and city) were strong predictors of whether or not a school has lead-contaminated water, in further work, we would explore other features capturing why location plays a critical role in lead detection. For example, this might include incorporating spatial modeling to identify “hotspots” of elevated lead levels, including water system maps, pollution metrics, and so on. We would also hope to obtain institution-specific building age for future modeling, rather than using a county-wide metric.

While student/teacher ratio and the proportion of students eligible for free and reduced lunch acted as proxies for how well-funded a given school is and students’ family income, we might also look to obtain more socioeconomic data, for example school board funding, with the hypothesis that better funded schools serving wealthier students might be more likely to have already addressed elevated lead levels in drinking water. Finally, lead contamination in school drinking water is not a problem limited to New York State. In future work, we would look to expand modeling to include other states.