Executive Summary: Lead Exposure Project

Katherine Laliotis, Alex Schimmoller, Brock Grafstrom

July 1, 2025

1 Introduction

Lead can damage nearly every system in the human body, and has harmful effects on both adults and children. It is a serious environmental public health threat to people across the United States, including those living in Central Ohio. Oftentimes, lead exposure takes place within one's own home through ingestion of water carried through lead piping. In an effort to raise awareness of citizens' possible exposure to lead, the City of Columbus makes publicly available a Service Line Material Inventory detailing the type of line used for each home. However, this data is oftentimes self reported, and there exist many homes not registered with the inventory. In this project, we will use the inventory and data on homes available through the Franklin County Auditor's Website to train a model to determine a home's probability of containing lead piping. This model could then be used to assign unregistered homes with a relative risk of containing lead pipe.

2 Stakeholders and KPIs

The following list includes potential stakeholders who might benefit from the results of this project:

- Columbus, Ohio residents, in particular those whose homes are not reported in the Water System Service Line Inventory
- The City of Columbus, whose current initiatives include the Lead Education and Poisoning Prevention Program and the Lead Service Line Replacement Program.

Key performance indicators (KPIs) we will use for this project include:

1. Accuracy of the model in predicting homes that we know do contain lead using cross-validation and forward selection
2. Assuming the model proves to be accurate, results of this project could be used by the City of Columbus to create a list of all homes with the highest risk of lead exposure from water service lines and motivate the highest priority projects within the Lead Service Line Replacement Program. As there is no known "safe" blood lead level above zero, successful implementation of this program can lead to the minimization of lead poisoning cases in Central Ohio.
3. Knowledge of the number of homes whose pipes need to be replaced would also help the City of Columbus to design an accurate and cost-effective plan to meet the EPA's deadline of 10 years to replace lead pipes. Developing a plan and a cost estimate takes time, and throughout that time inflation and increasing materials costs would cause the replacement to only get more expensive for the city as time goes on. Using our findings, the city could minimize the impact of inflation on the pipe replacement project.

3 Methods and Models

We constructed our dataset by combining the Franklin County Parcel data and Lead Service Line Inventory using matching addresses. We then proceeded with cleaning the data of any null entries.

We achieved a final dataset describing 224,190 properties in Columbus, which make up 79% of all properties listed in the Service Line Inventory.

Our first concern within this dataset was biasing our model due to an imbalanced data set; only about 10% of our data contains "positive" detections of lead. This imbalance in data could cause the model to think it was achieving high accuracy through only detecting negatives and false negatives (see Fig. 1).

Because of this, we used methods of data analysis that are more robust to imbalanced data. Those include:

- Analysis of build year and lead pipes to get a general sense of the data
- Logistic regression using stratified K-folds for the train-test-split. Ensuring that each K-fold contains the same positive-negative detection split makes the results more reliable given the imbalance in our data. The logistic regression itself assigns balanced class weights to try and re-balance the regression results.
- Nearest-Neighbors algorithm: We chose to try this model for two reasons. First, it is generally more robust to imbalanced data. Second, we expect one of the main predictors of positive detections to be nearby homes with positive detections, so this method is a very interpretable way of analyzing this problem.

4 Results

A central complication of this data science problem is the different impact of false negatives and false positives on our stakeholders. A false negative is, in the case of lead detection, much more harmful than a false positive. To address this, in each of our models we fine-tune certain parameters to prioritize *recall* rather than precision.

The main variable that impacts the precision-recall split is the probability threshold for classification as a positive detection. In order to decide what threshold to use, we employed the F2 Score: a weighted mean of precision and recall that weights recall more heavily than precision.

4.1 Regression

For regression, we found that a detection threshold of 0.6 maximized the F2 score. Using that threshold, our model achieved Accuracy: 0.9154(+0.0720), Precision: 0.5035(+0.4179), Recall: 0.8840(+0.7984).

4.2 K-Nearest-Neighbors

Using the stratified K fold method for training and maximizing the threshold for recall using the F2 test (as with the regression) we found that a threshold of 0.128 would be ideal for this analysis. Using KNN with the latitude and longitude and year built as model parameters, we were able to obtain Accuracy : 0.930, Precision: 0.553, Recall : 0.926, and F2 Score : 0.816.

5 Future Goals

We would like to extend this project to provide useful information to our stakeholders including KPIs 2 and 3. To do this, we could run our model over the remaining homes from the Franklin County Auditor's Website that were *not* included in the Service Line Material Inventory, and come up with a list of those homes most likely to contain lead in their pipes. The list of homes and their respective lead-contamination probabilities could be used by the city of Columbus to do risk assessment and cost minimization for the imminent pipe replacement projects.