Lead Poisoning: Executive Summary

Anna Pham and Matthew Smylie

April 21, 2025

1 Introduction

There is no safe level of lead in the blood. Childhood exposure to lead remains a vital public health issue, as it has been linked to chronic health problems later in life. However, there are many places in the United States and around the world in which there is a dearth of actual statistics concerning blood lead levels (BLLs) in children. The goal of this project is to examine the data that does exist and to compare it to other social and economic variables. Ideally, we will be able to find good predictors that are better measured, and we may be able to identify regions that are most at risk.

2 The Data

The data used for this project comes from multiple sources. Our primary reference for the target variable (fraction of BLLs above the reference value) is the data compiled by the Centers for Disease Control (CDC) as part of their Childhood Lead Poisoning Prevention Program. Testing and surveillance were performed by individual states and reported to the CDC. However, the available state-level tables from the CDC only contain statistics from 22 states. We supplement this with other data from state health departments (specifically AZ, CA, IN, MA, MN, ND, NM, OH, and TX) for a total of 31 states. For the remaining states, there is not enough publicly available county-level data to include in the fit. In 2021, the CDC changed their BLL reference value from 5.0 μg/dL to 3.5 μg/dL. Since we are missing more recent data from many counties, we elect to use only 2021 data for consistency. We proceed with the assumption that the county-level rates have not varied significantly over the past four years.

The features on which we train the model include housing, economic, and occupational factors obtained from the U.S. Census Bureau. In addition, we include as a feature the rate of diagnosis of Attention Deficit Hyperactivity Disorder (ADHD) in children. There have been studies that show associations between lead exposure and ADHD, and further research is needed in this area. We use state-level statistics on ADHD compiled also from the CDC.

3 Data Challenges

As mentioned previously, there are many counties in the U.S. that do not have good BLL data available. There are other counties for which the data exists but has high margins of error, presumably due to issues of small sample sizes. For example, a reported value of 0 could mean that there are no cases of elevated BLL, or it could mean that the testing is inadequate. For the training features, census data is collected far more often and is far more complete on higher population counties than on the less populous counties. This coincides with the previous issue to make it difficult to get an accurate picture of rural counties.

4 Model Selection

Given the large number of independent features, we focus on tree-based regressors. Specifically, we consider the random forest, extra trees, and XGBoost regressors. Using 80% of our data for training, we implement 5-fold cross validation for model selection and hyperparameter tuning.

5 Results and Discussion

We find that the extra trees regressor performs the best in cross validation. However, when applied to the testing data, this model seems to overestimate low values of the target variable and underestimate high values.

A benefit to the tree-based methods we apply here is that feature importances are tracked. We find consistently that the most important feature for the cuts is the age of the houses. This may make intuitive sense, as older houses may have used lead-based paint.

There are many possible directions for future work. More data and different features could conceivably improve performance. In particular, our model only has access to state-level averages for ADHD, and we expect that county-level statistics would be very relevant.