Predicting Blood Lead Levels

Matthew Smylie and Anna Pham

April 21, 2025

The Erdős Institute
Data Science Boot Camp

Motivation

- There is no safe level of lead in the blood.
- Lead exposure contributes to millions of deaths worldwide.
- Particularly harmful to young children and pregnant women, as it can damage the nervous system and delay development.
- Public health agencies need to be able to respond to crises (e.g. Flint, Milwaukee).

Project Goals

- There are many areas in the United States with very limited data on blood lead levels (BLLs) in children.
- Funding and political issues may cause even more difficulties in the near future.
- Can we use other data to identify areas most likely to be at risk?

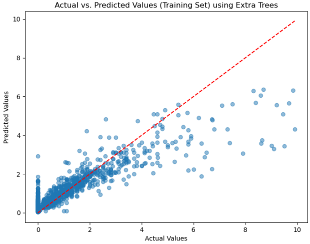
Data Gathering

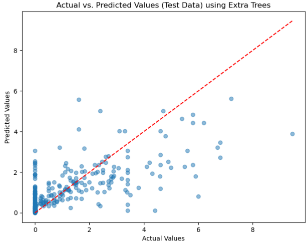
- Target variable: percentage of elevated BLLs in tested children by county.
  - CDC, state and local health departments.
  - 31 states and over 1300 counties.
- Features from the Census Bureau (American Community Survey):
  - Housing data: year built, price, plumbing.
  - Poverty rates.
- Geographic location.
- Rates of ADHD diagnosis.

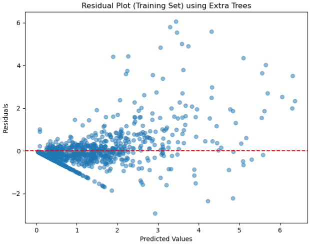
Model Selection

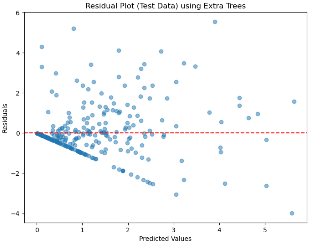
- 20% of the data was set aside as a testing set.
- Focused on tree-based methods:
    - Random forest.
    - Extra trees.
    - XGBoost.
- 5-fold cross validation was used for model selection and hyperparameter tuning.
- Minimize mean squared error (MSE).
- Selected an extra trees model with CV MSE of 1.78.

Extra Trees Plots

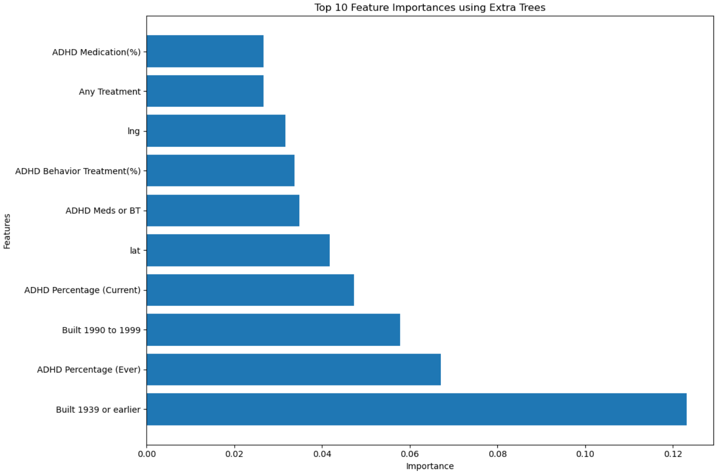

Actual vs. Predicted Values (Training Set) using Extra Trees


Actual vs. Predicted Values (Test Data) using Extra Trees


Residual Plot (Training Set) using Extra Trees


Residual Plot (Test Data) using Extra Trees

Extra Trees Plots



Results and Discussion

- When applied to the testing set, the model has a RMSE of 1.35 and an R^2 of 0.42.
- Model seems to have issues with low and high BLL values.
- Top feature importances are consistent.
- Counties with small lead sample sizes may under- or overestimate value for the target.