Does Your Home Contain Lead?

Predicting the prevalence of Lead pipes for homes in Columbus Ohio

Katherine Laliotis, Alex Schimmoller, and Brock Grafstrom

The Team

Katherine Laliotis

Alex Schimmoller

Brock Grafstrom

Final Year Ph.D. candidates at The Ohio State University in Physics

Primary Objective

For a given address in Columbus, Ohio…

What is the likelihood that building receives its water from lead pipes?

Current city database indicates that roughly **1 in 10** homes are serviced by lead pipes

Primary Objective

https://www.epa.gov/ground-water-and-drinking-water/infographic-lead-drinking-water

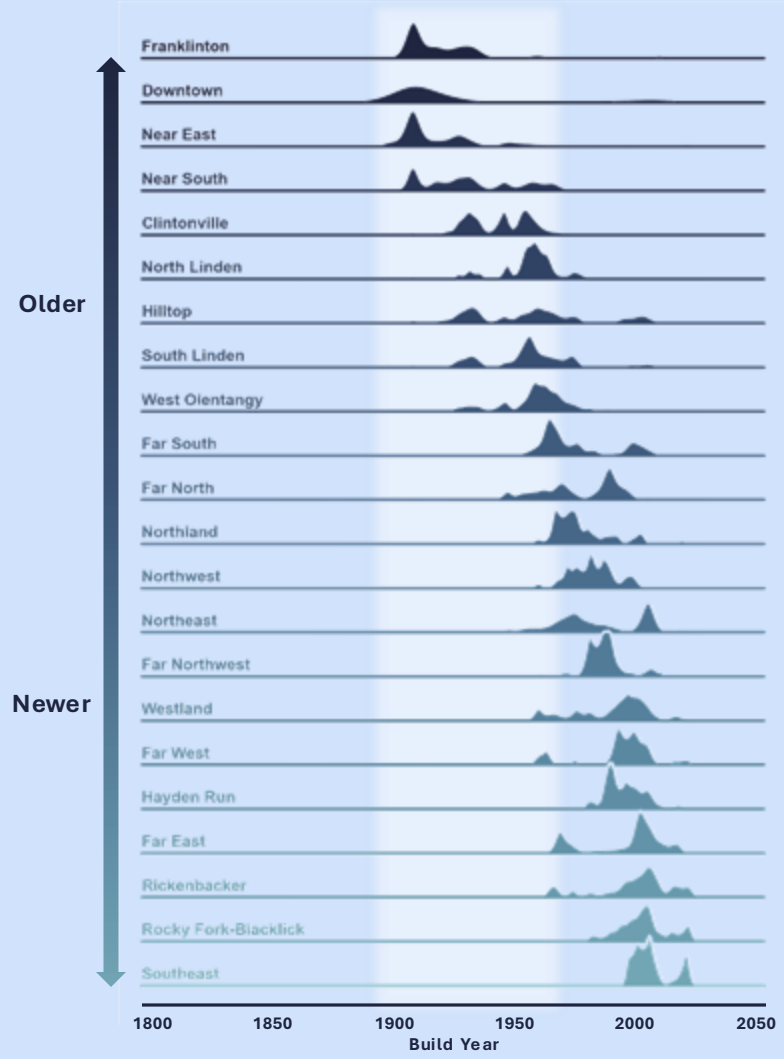
Raw Datasets:

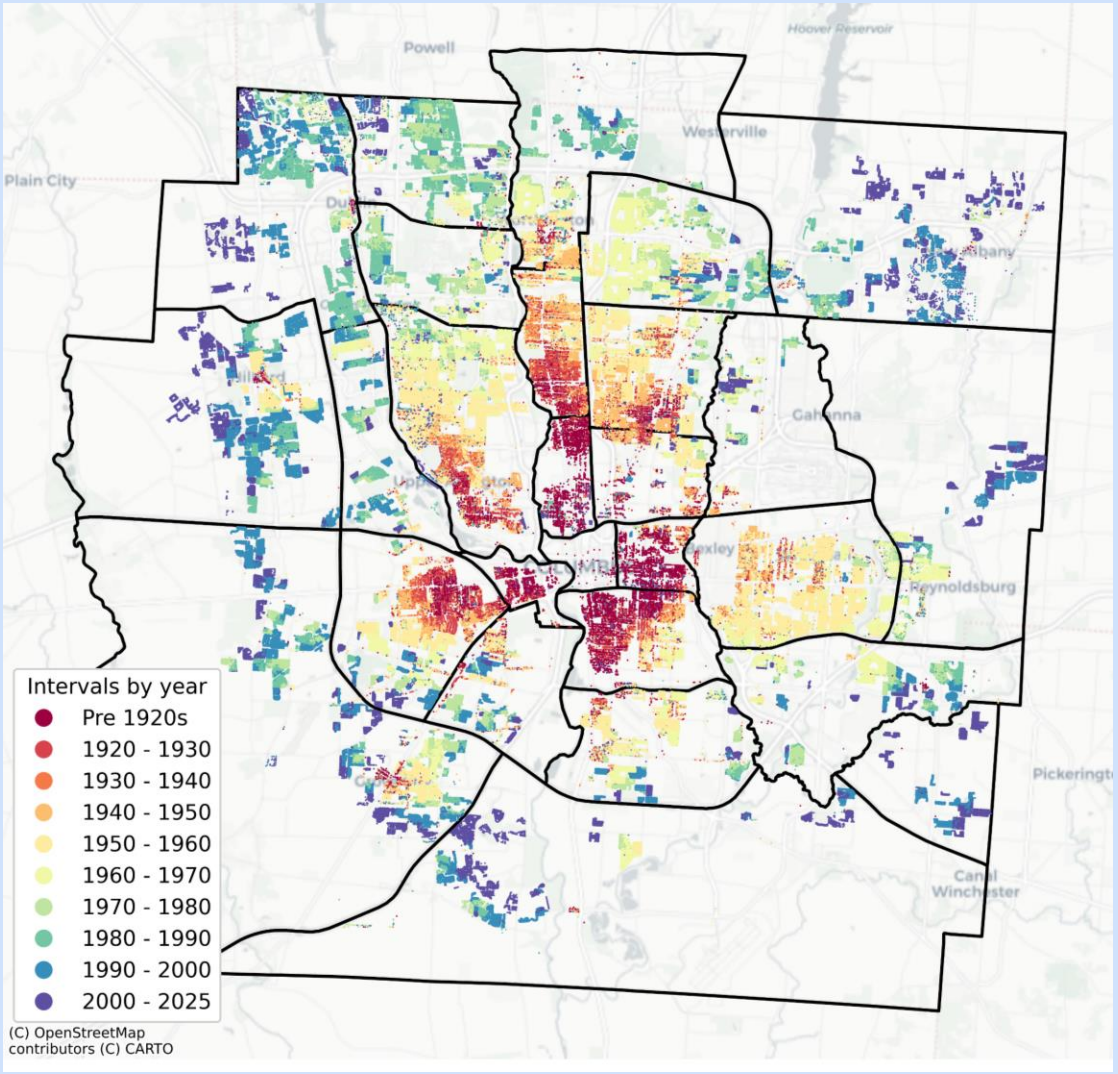
Columbus Public Water System Service Line Inventory

Franklin County Auditor's Office Parcel Inventory

Age of Homes in Columbus

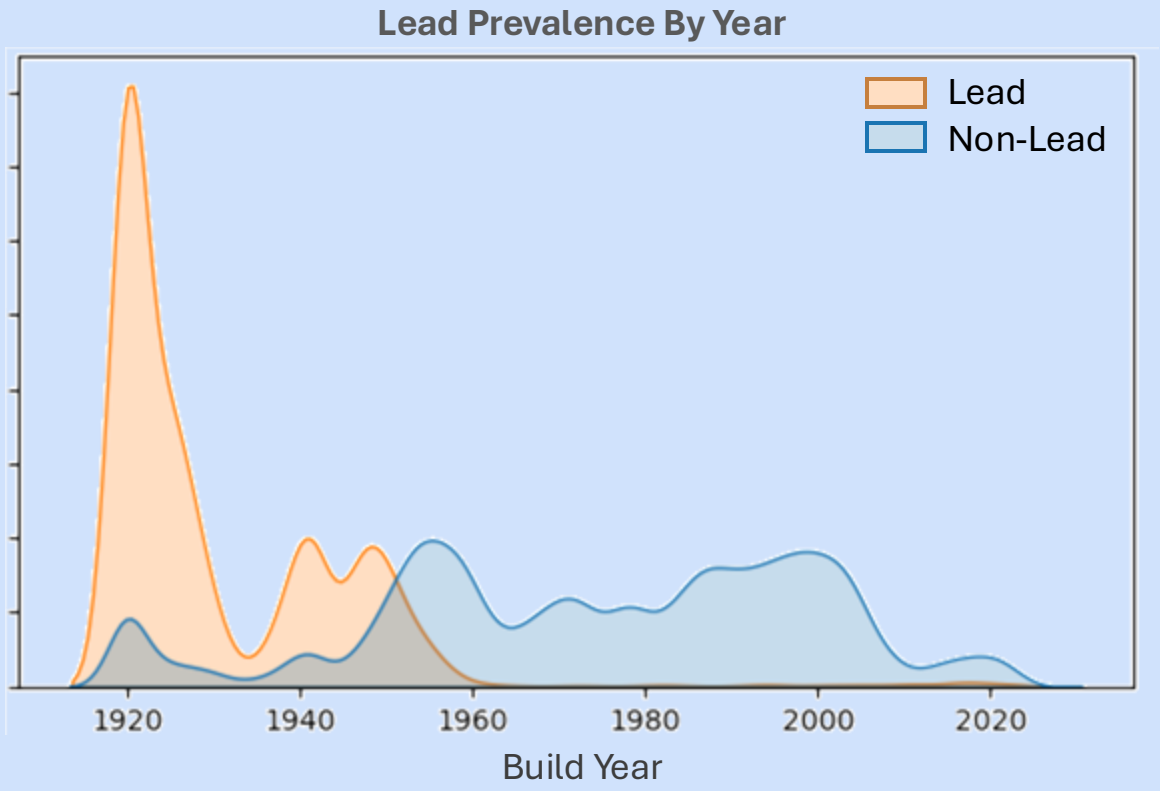
Started with an analysis of build year and neighborhoods

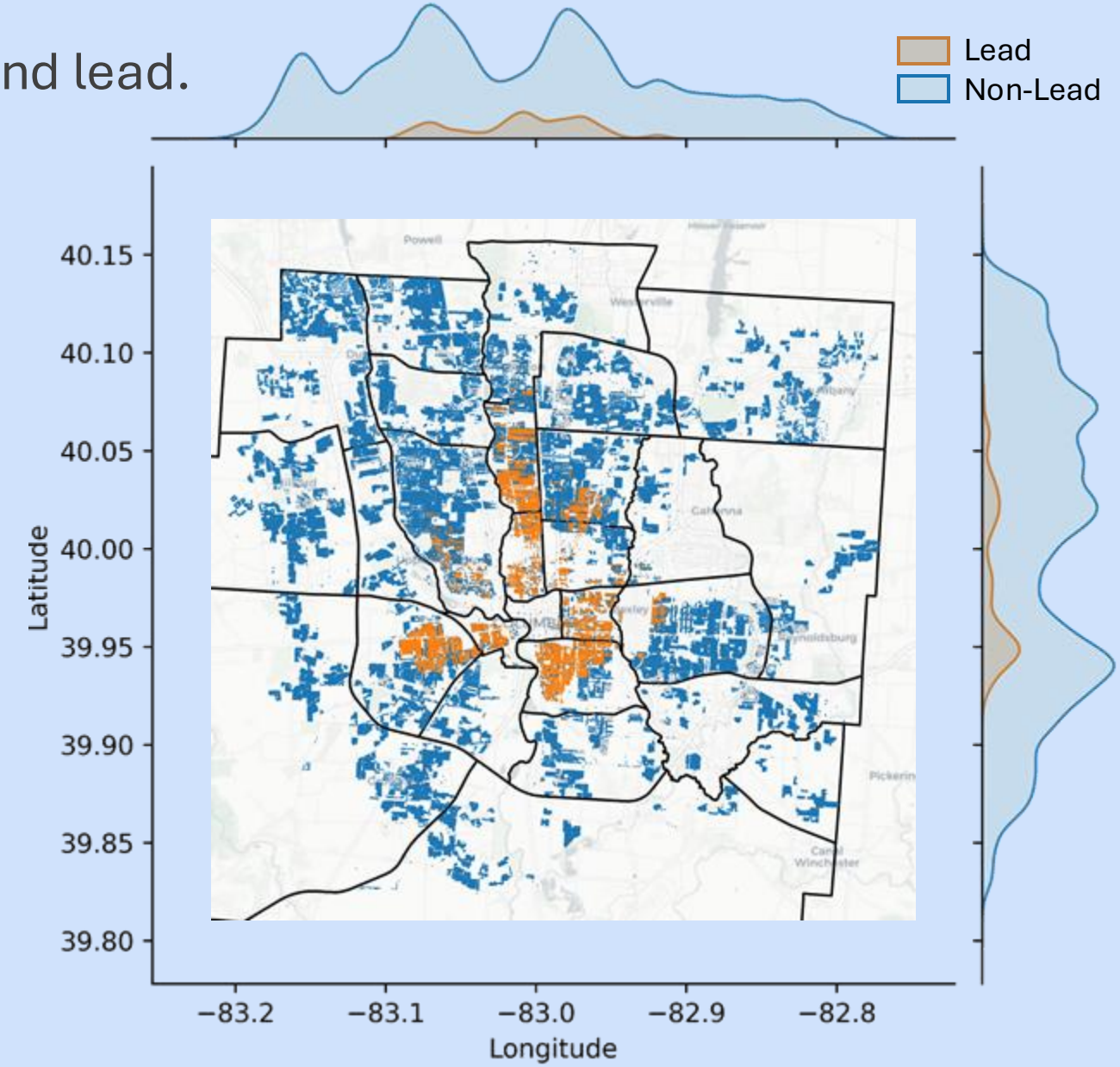




Lead vs Non-Lead in Columbus

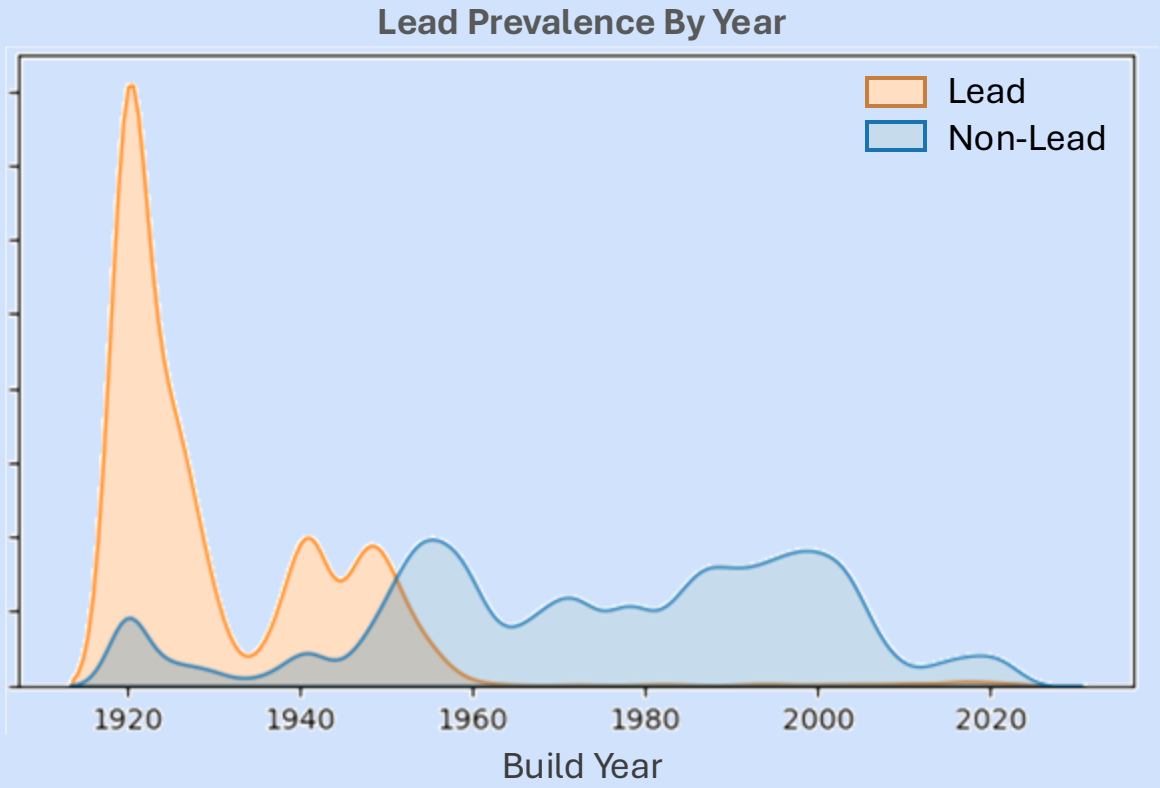
- Strong connection between home age and lead.
- Also reflected geographically.
- Not adequate for a full prediction.

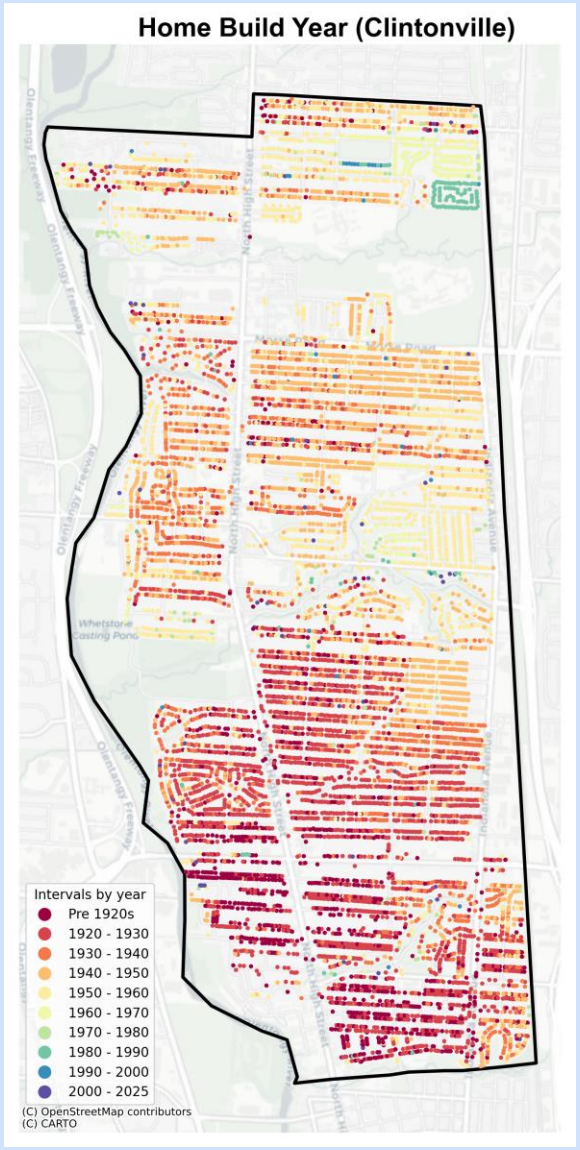


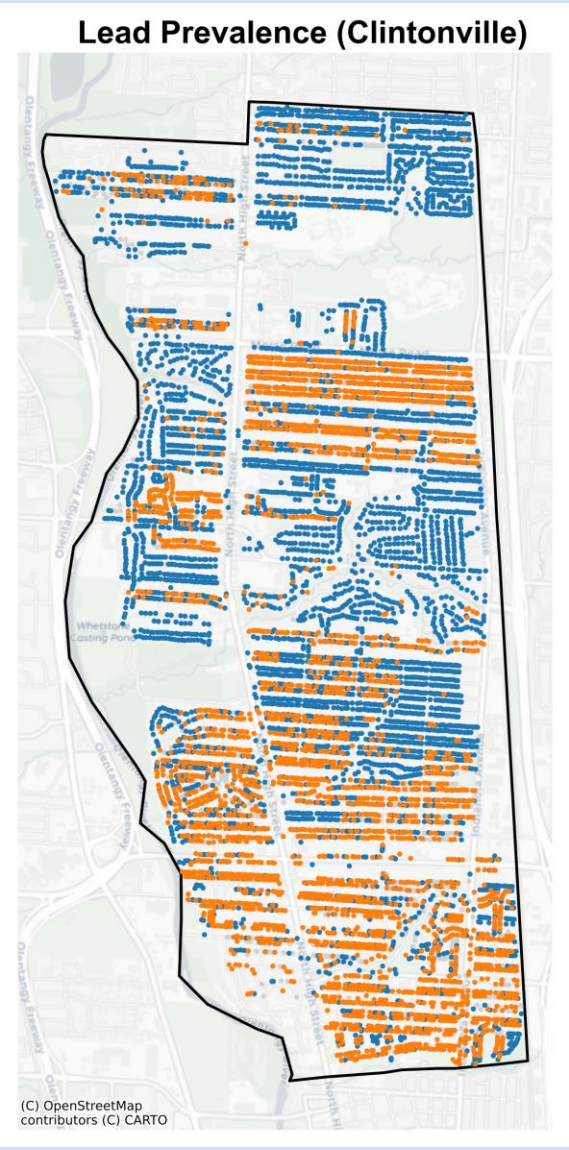


Lead vs Non-Lead in Columbus

- Strong connection between home age and lead.
- Not adequate for a full prediction.
- New construction in older districts leads to outliers.







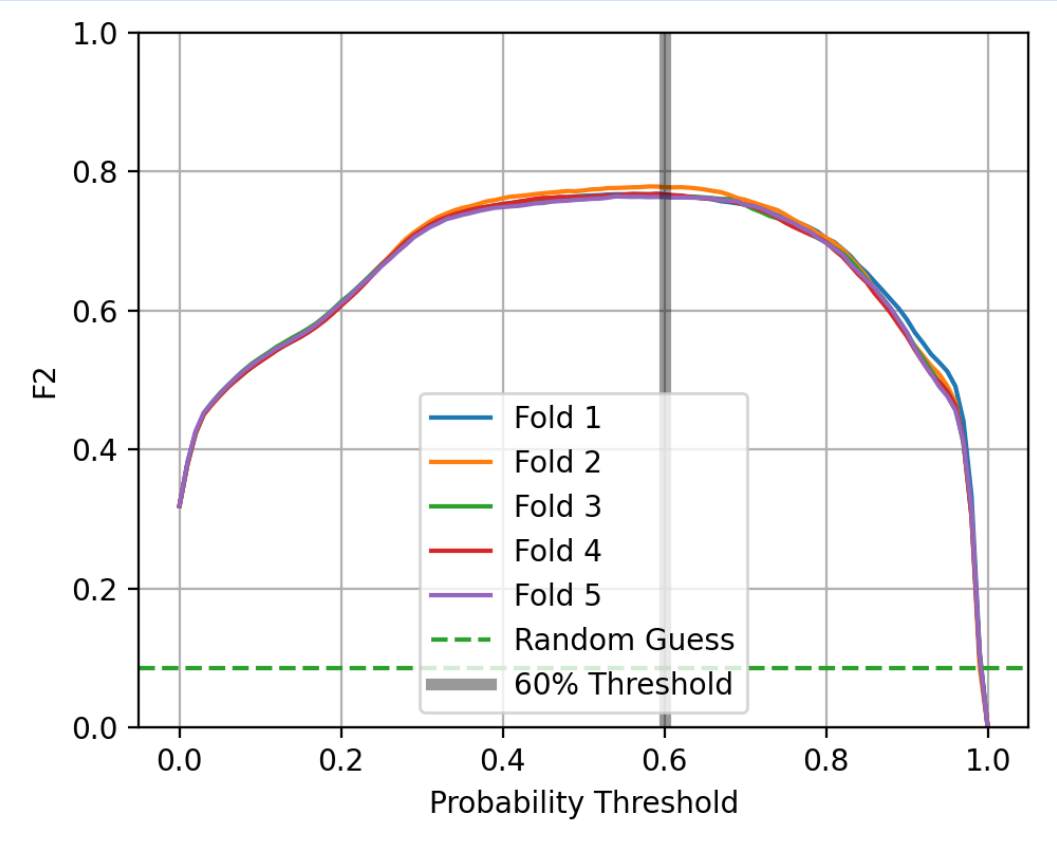
Logistic Regression: Setup

- Testing was performed for five stratified K-folds consisting of 20% of the total homes each, where ~9% of each fold were “lead positive” homes.
- Features include home build year and weighted nearest-neighbor lead value

Target:	is_lead (binary)
Feature:	YEARBLT (integer)
Feature:	nn_is_lead_weighted $= \frac{\pm 1}{d_{nn}}$

$+1 \rightarrow$ nearest neighbor has lead
$-1 \rightarrow$ nearest neighbor does not have lead
$d_{nn} \rightarrow$ distance from nearest neighbor

Logistic Regression: Results



Metrics	Log-Reg	Rand Guess
Accuracy =	0.915	0.844
Precision =	0.504	0.086
Recall =	0.884	0.086
F2 Score =	0.768	0.086

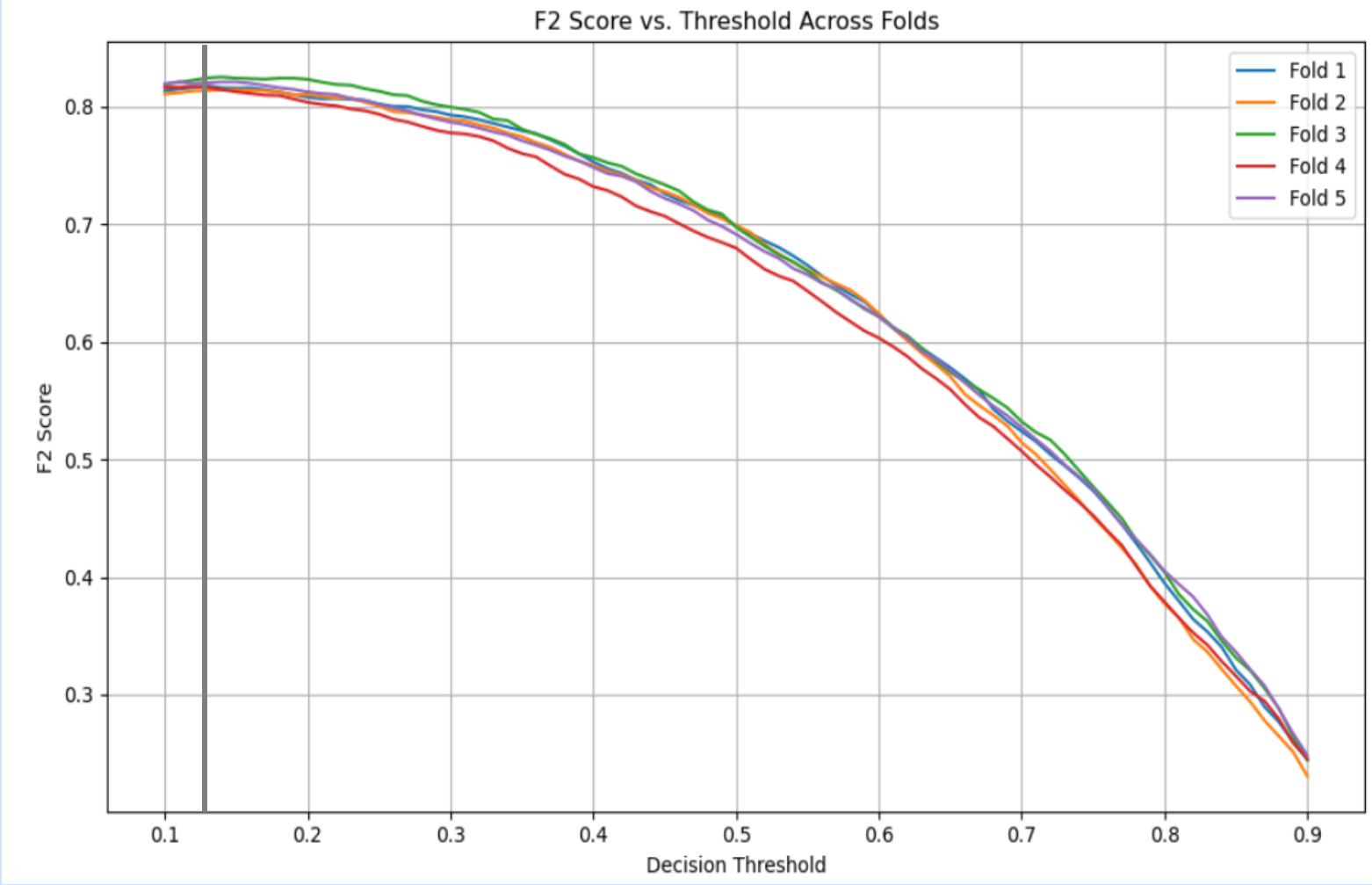
- Maximized F2 to find optimal decision threshold (aim to reduce false negative rate)
- **Performs better than random guess** and has a simple interpretation (newer homes have a decreased likelihood for lead, hence the negative linear slope, while nearest-neighbor terms are positively correlated).
- Accuracy is only marginally better than random guessing.

Model Coefficient Averages

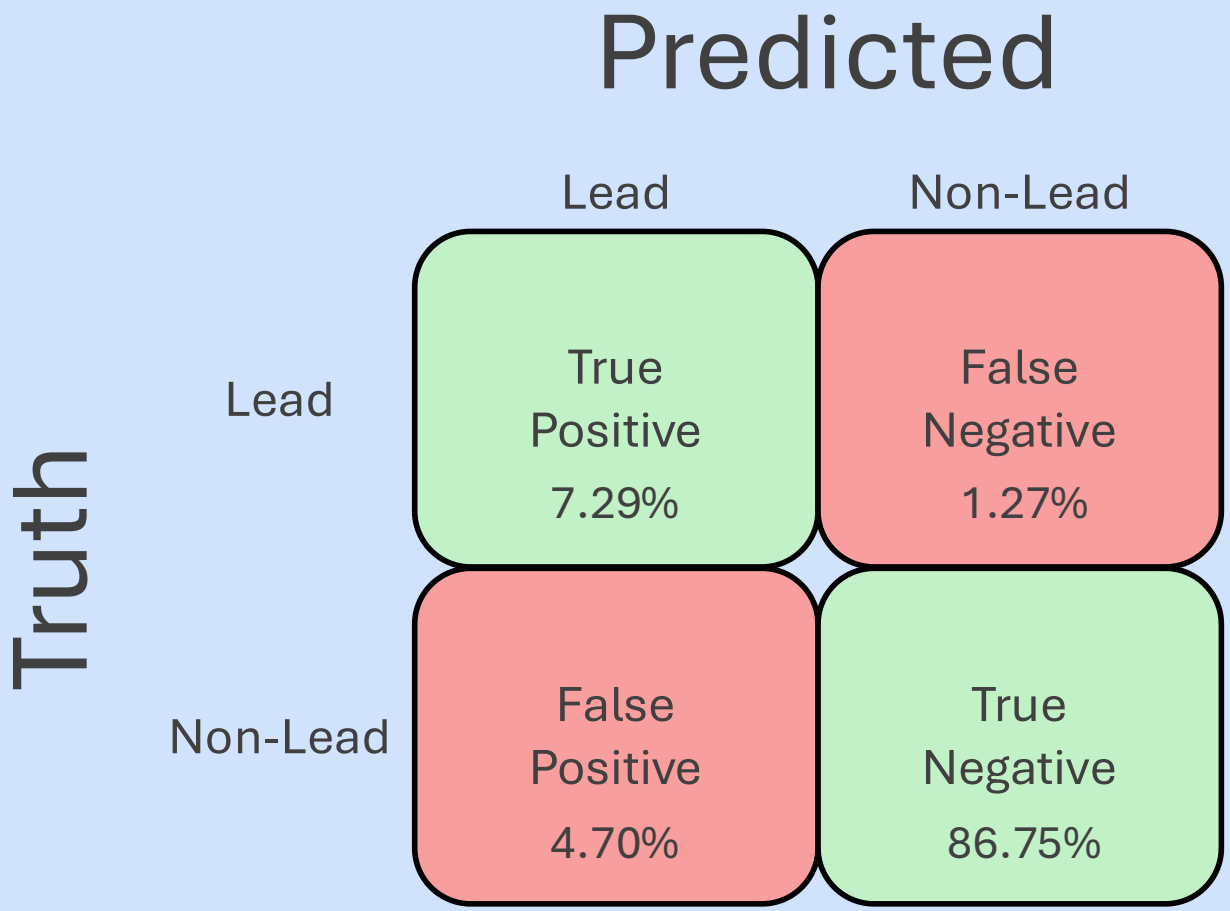
Intercept (β0)	= -2.3845
YEARBLT (β1)	= -2.3035
NN_is_lead_weighted (β2)	= 1.1628

Nearest Neighbor (KNN): Optimizing Across Folds

- Nearest neighbor classification
- Wrapped in stratified K-fold Cross-Validation
- Maximized F2 score to find the optimal decision threshold
- Mean Optimal Decision threshold was determined to be **0.130**



Nearest Neighbor (KNN): Results

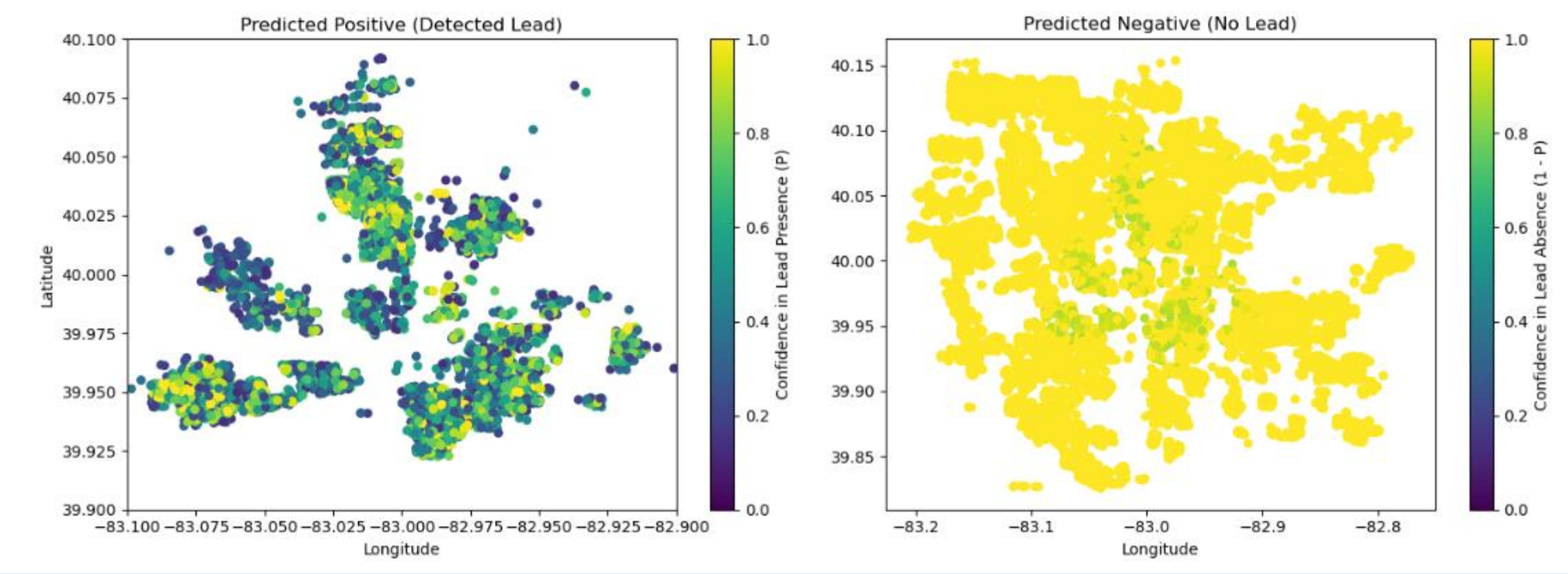


Metrics	KNN	Log-Reg	Rand Guess
Accuracy =	0.930	0.915	0.844
Precision =	0.553	0.504	0.086
Recall =	0.926	0.884	0.086
F2 Score =	0.816	0.768	0.086

- Used the optimal threshold on a final test set for the final analysis result
- KNN consistently outperforms logistic regression across all performance metrics.
- On the same testing set, KNN does better at minimizing false negative rate.

Confidence of Prediction

- KNN can predict suburban neighborhoods with nearly 100% certainty, since a majority of homes were built after 1960 in those regions.
- For downtown neighborhoods, both predicted positive and predicted negative rates increase, however the confidence surrounding predicted negative homes is constrained between 60% - 100%.



Future Extensions

- Can utilize the sharp correlation between home age and lead to extend the number of homes that are represented in the city's Service Line Data Inventory.
- The accuracy of these predictions will vary based on location, but for neighborhoods on the periphery, KNN is almost 100% accurate. Conversely, predictions for downtown regions are less certain.
- Only considered the primary connections between home age, nearest neighbors, and lead prevalence. Therefore, other factors such as home price (or rating) would likely aid in refining the predictive capabilities for future models. Need to be careful of overfitting.
- Could retrain the models on different cities. This could be used to determine "universal indicators" for an increased risk of lead pipes in homes.
- Could also bin home data by school district to see if there is a direct correlation with student test scores on statewide assessments.

Thank You