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

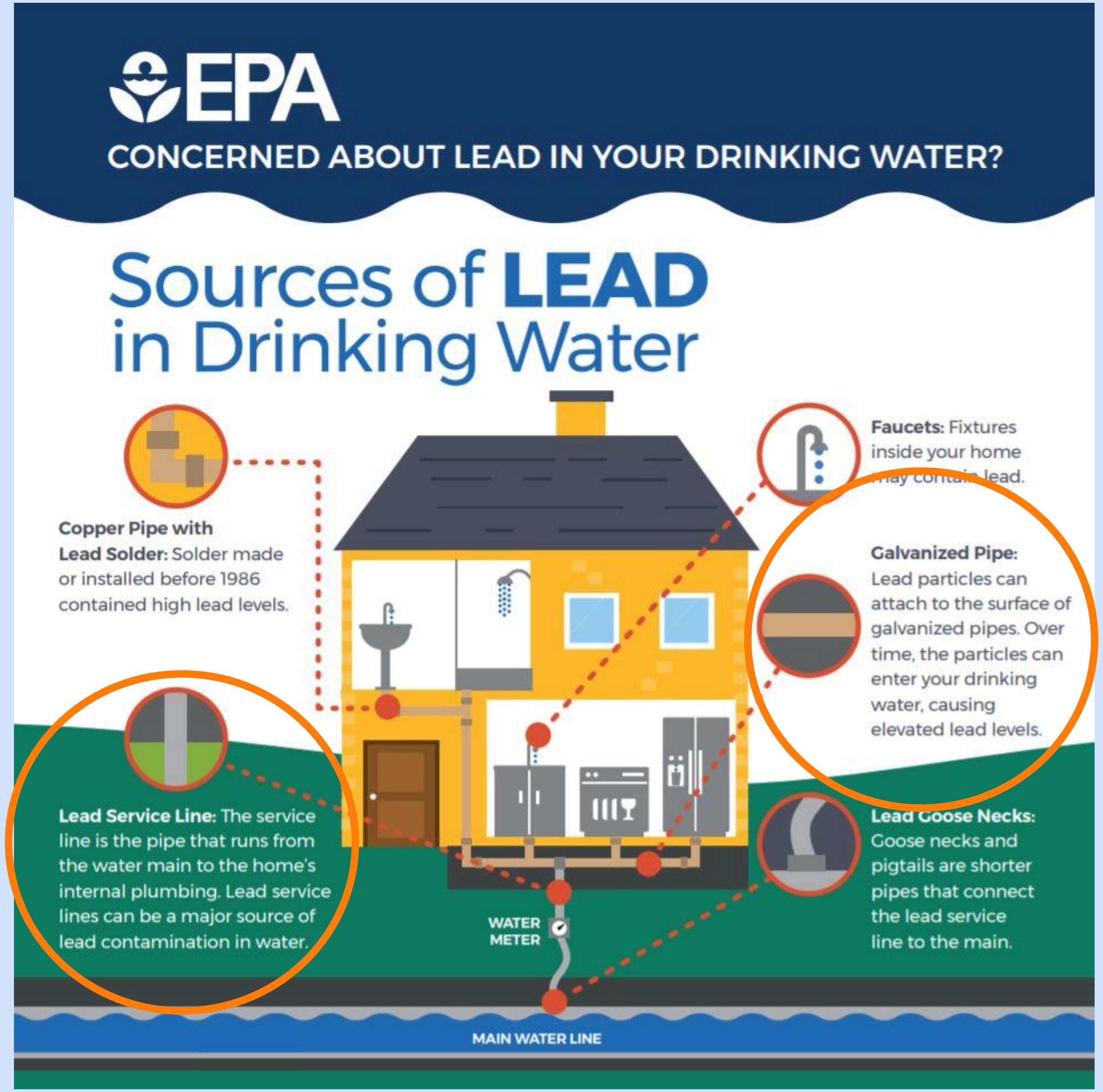
Primary Objective

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 exposed to lead

Primary Objective



EPA

CONCERNED ABOUT LEAD IN YOUR DRINKING WATER?

Sources of **LEAD** in Drinking Water

Copper Pipe with Lead Solder: Solder made or installed before 1986 contained high lead levels.

Faucets: Fixtures inside your home may contain lead.

Galvanized Pipe: Lead particles can attach to the surface of galvanized pipes. Over time, the particles can enter your drinking water, causing elevated lead levels.

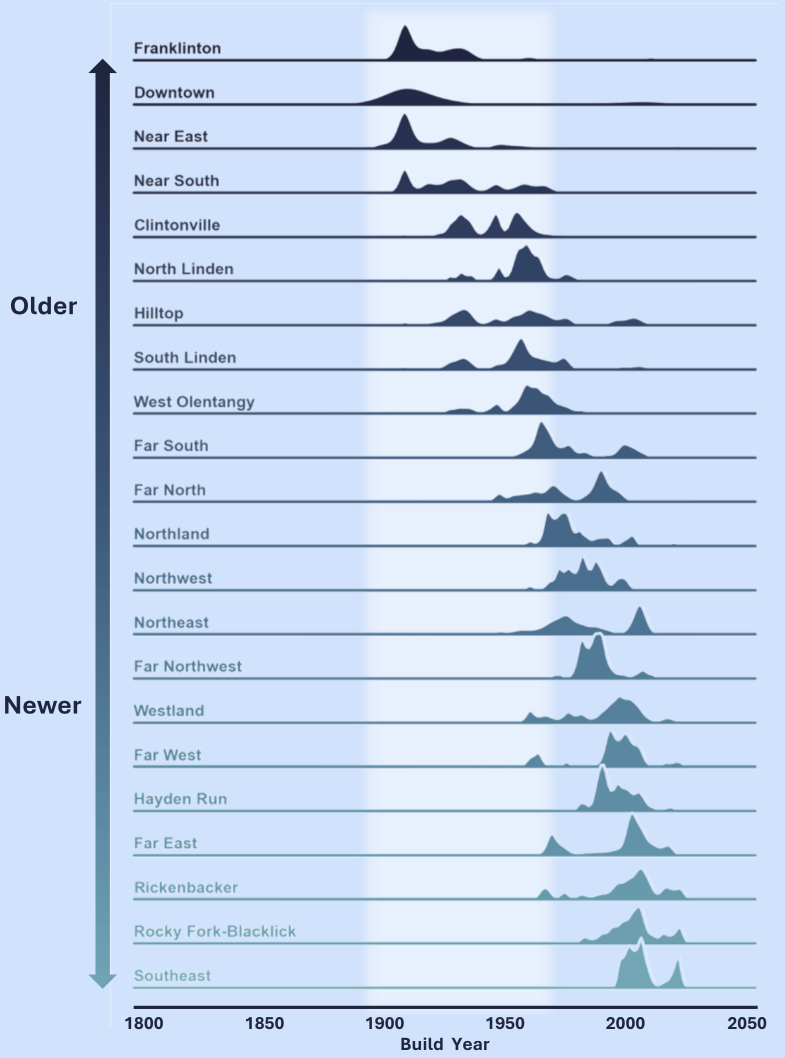
Lead Service Line: The service line is the pipe that runs from the water main to the home's internal plumbing. Lead service lines can be a major source of lead contamination in water.

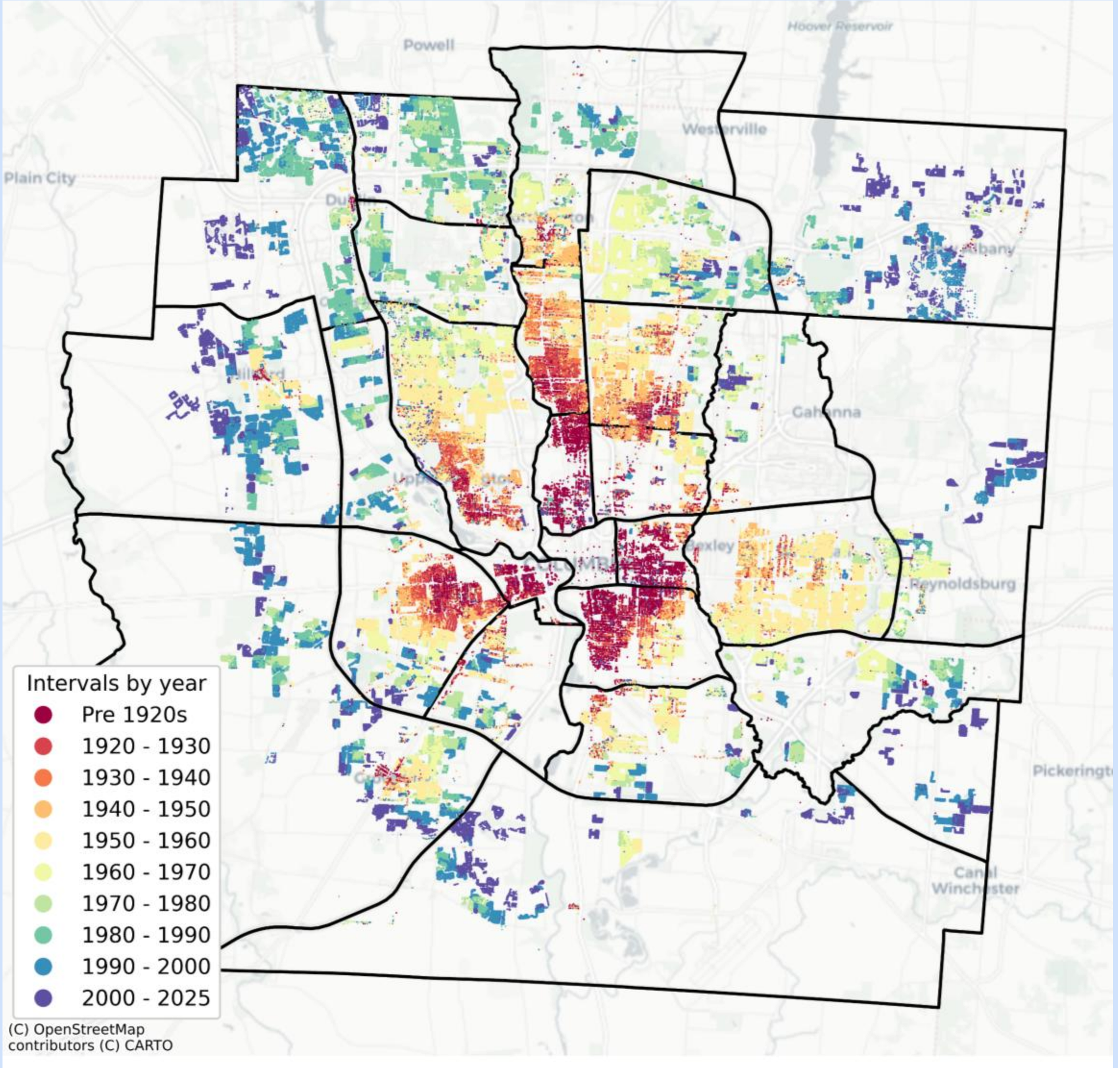
Lead Goose Necks: Goose necks and pigtails are shorter pipes that connect the lead service line to the main.

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

Age of Homes in Columbus

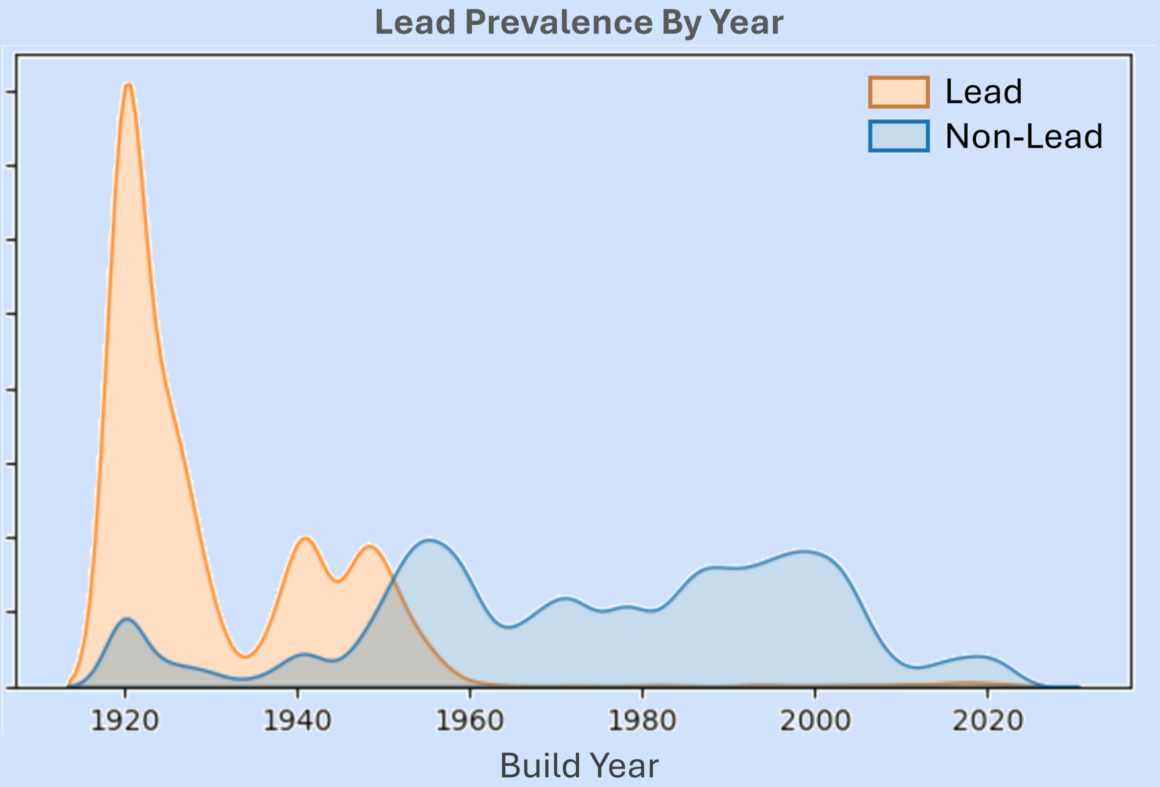
Started with a time series analysis by build year and neighborhood

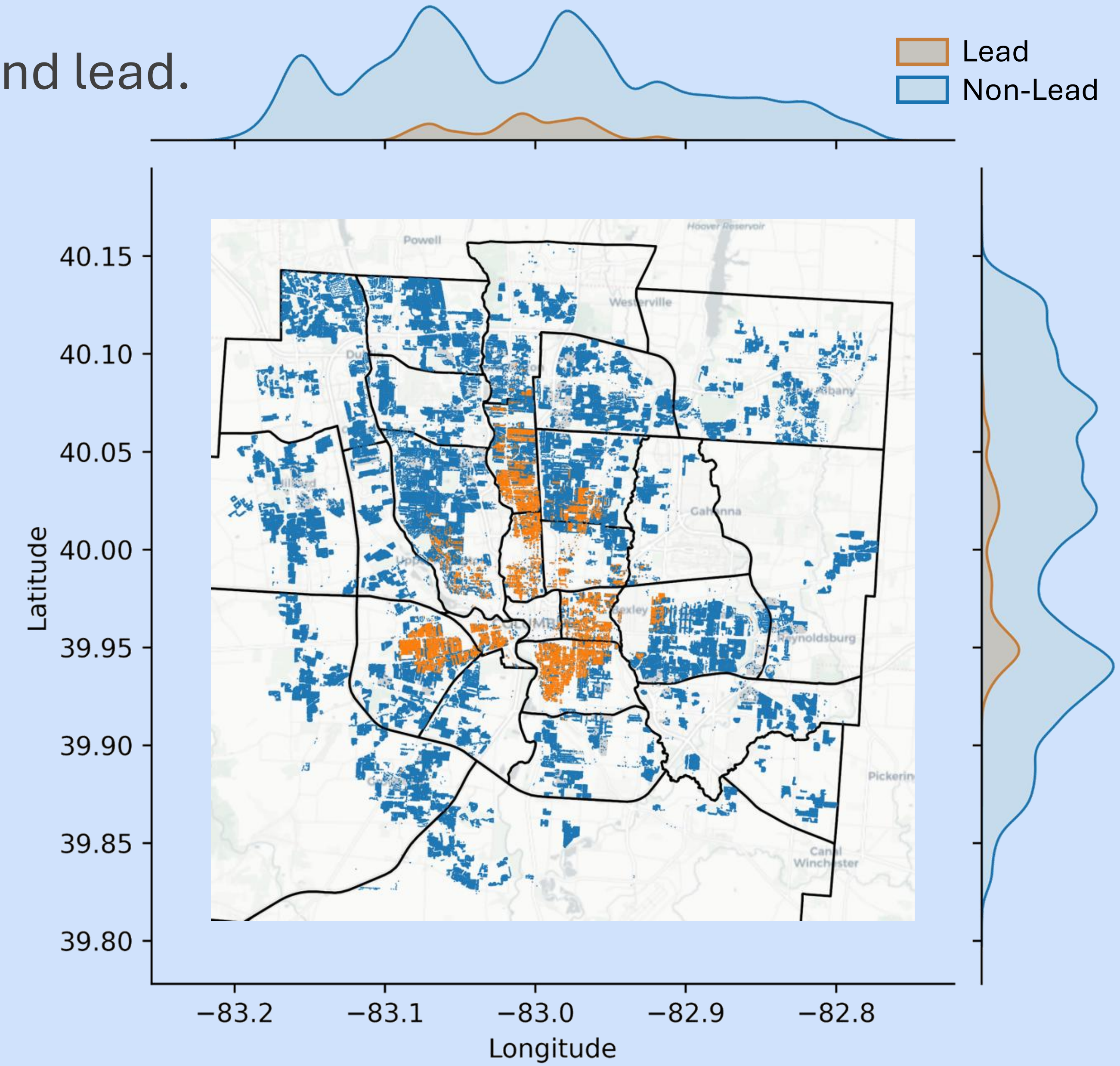




Lead vs Non-Lead in Columbus

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

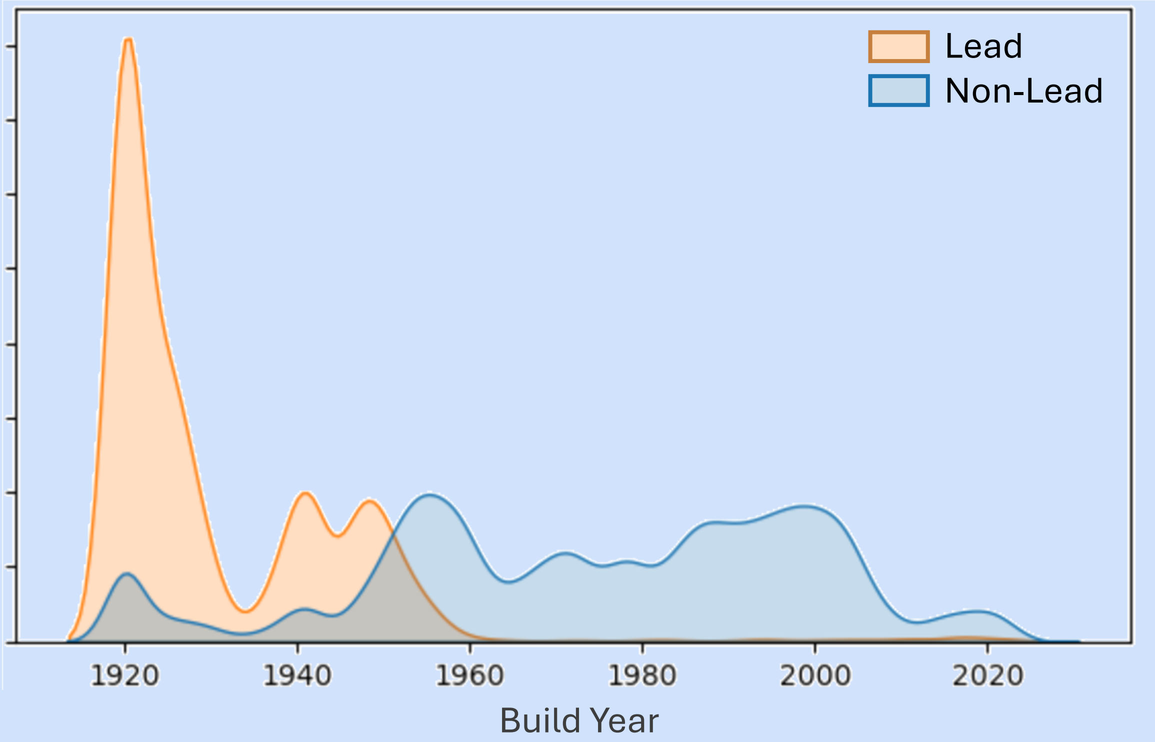


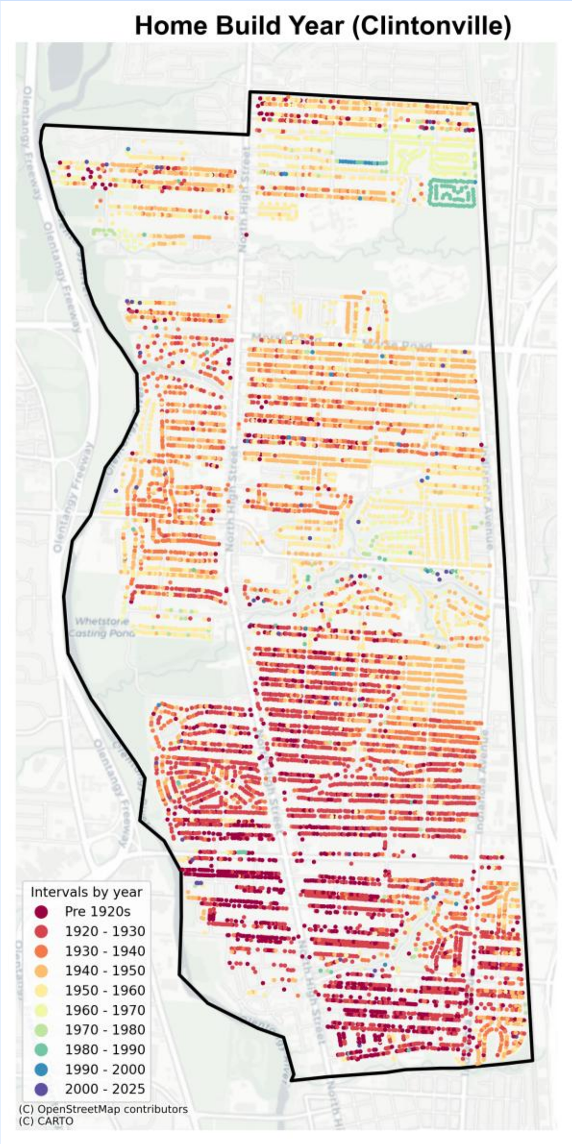


Lead vs Non-Lead in Columbus

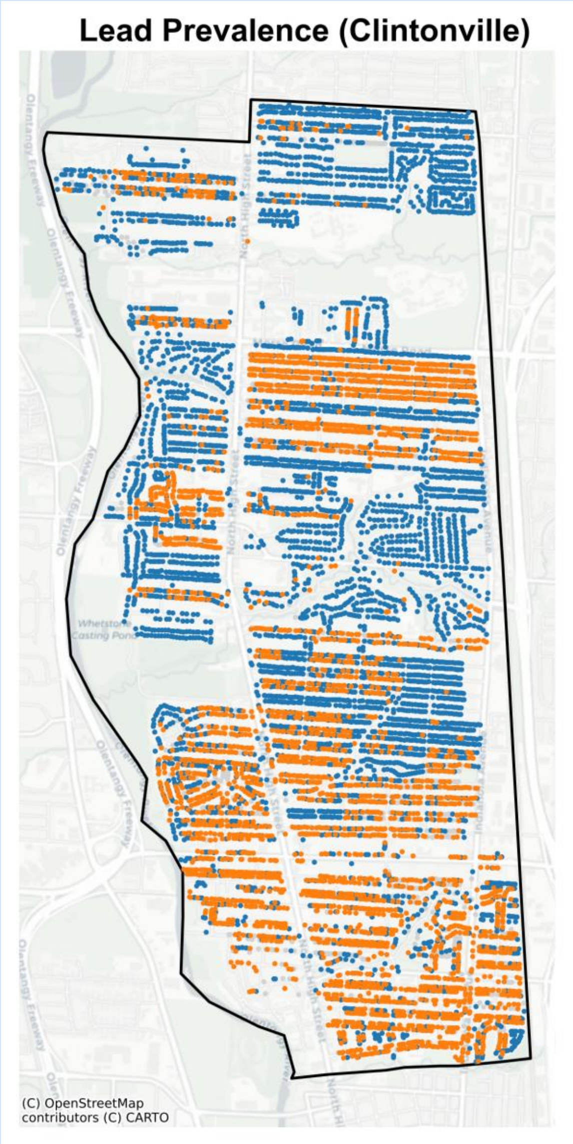
- Strong connection between home age and lead.
- Not adequate for a full prediction.
- New construction in older districts leads to outliers

Lead Prevalence By Year



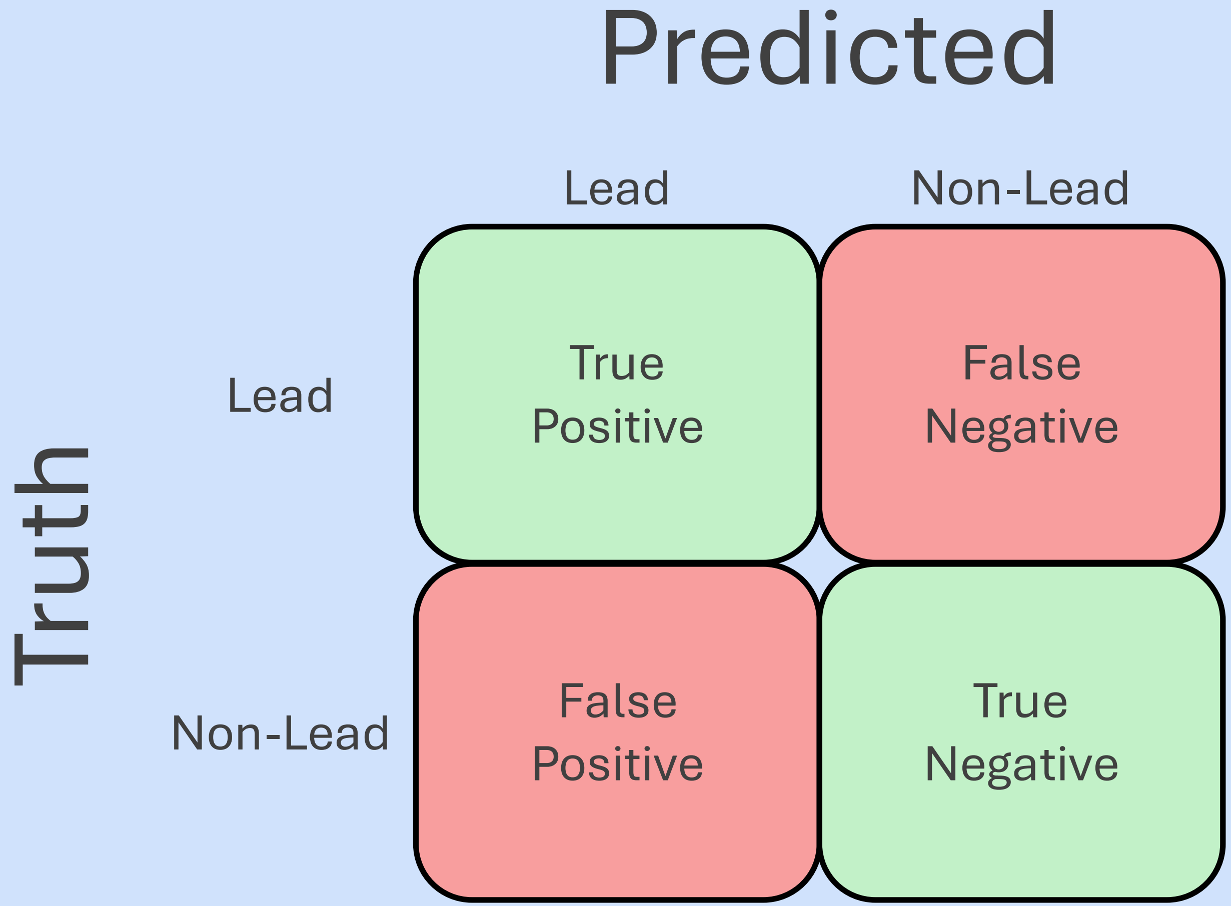

Home Build Year (Clintonville)

(C) OpenStreetMap contributors
(C) CARTO


Lead Prevalence (Clintonville)

(C) OpenStreetMap contributors (C) CARTO

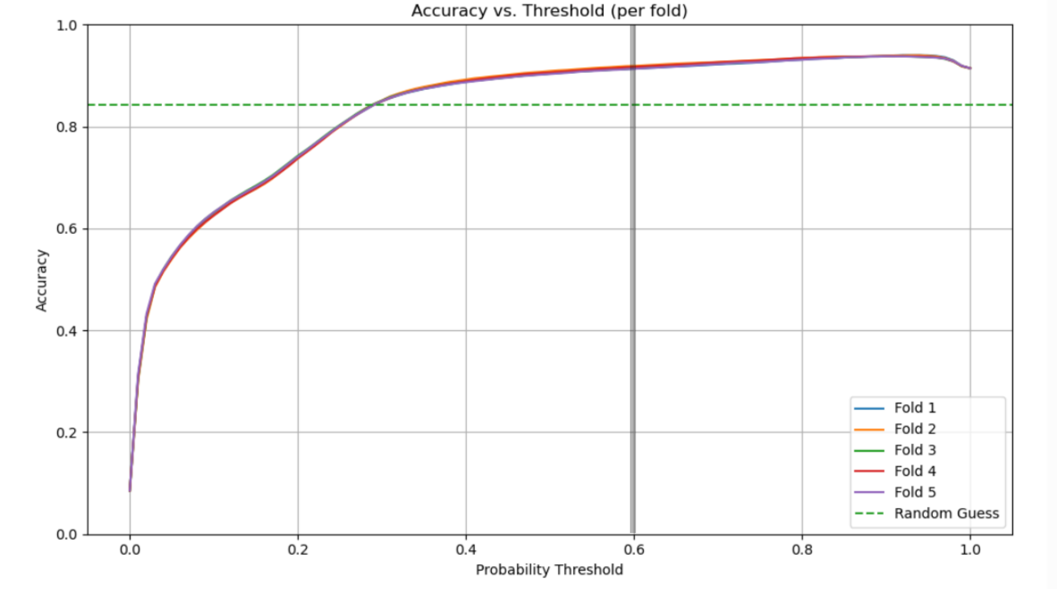
Naïve Approach (Random Guess)

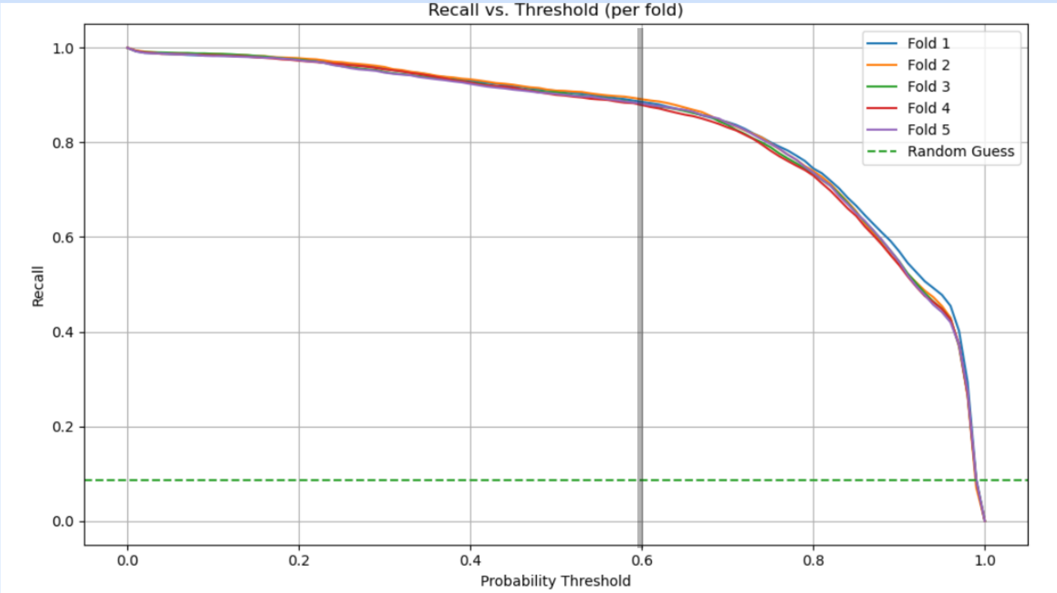


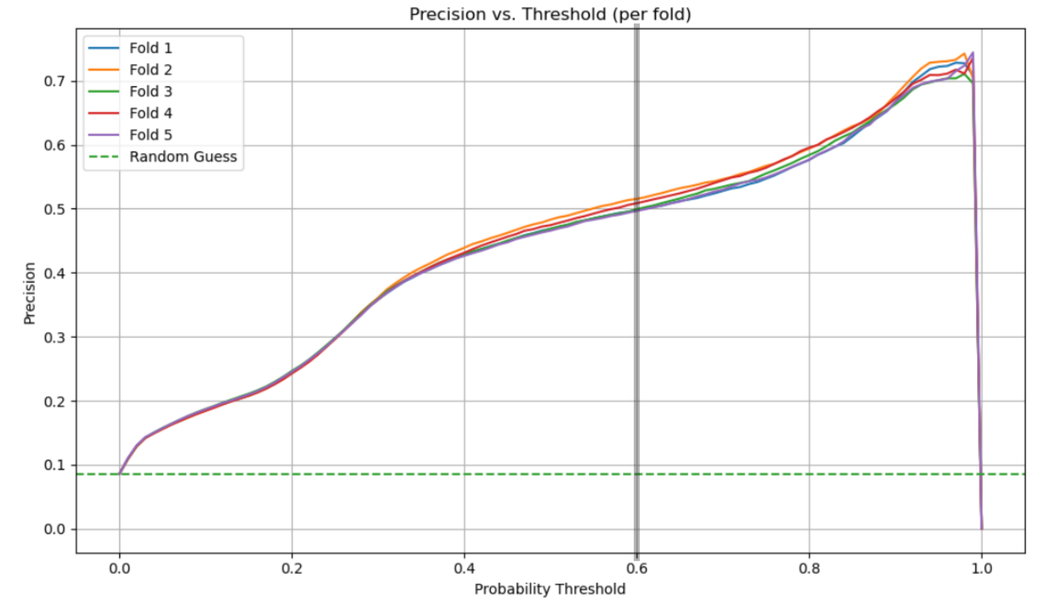
Predicted

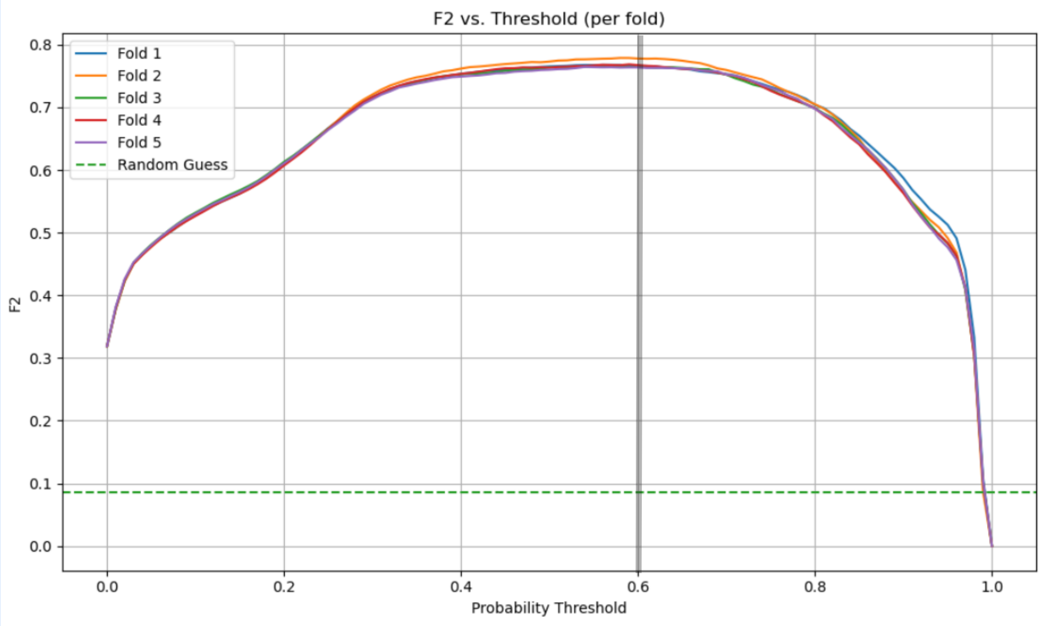
Truth	Lead	Non-Lead
Lead	True Positive	False Negative
Non-Lead	False Positive	True Negative

Logistic Regression


Accuracy vs. Threshold (per fold)


Recall vs. Threshold (per fold)


Precision vs. Threshold (per fold)


F2 vs. Threshold (per fold)

Logistic Regression

- Testing was performed for five stratified K-folds consisting of 20% of the total homes, where 10% of each fold were “lead positive” homes.

- This was done to address the unbalanced nature of our overall dataset (~10% lead, ~90% non-lead).

- Performs better than random guess and has a simple interpretation (newer homes have a decreased likelihood for lead, head the negative linear slope, while nearest-neighbor terms are positively correlated).

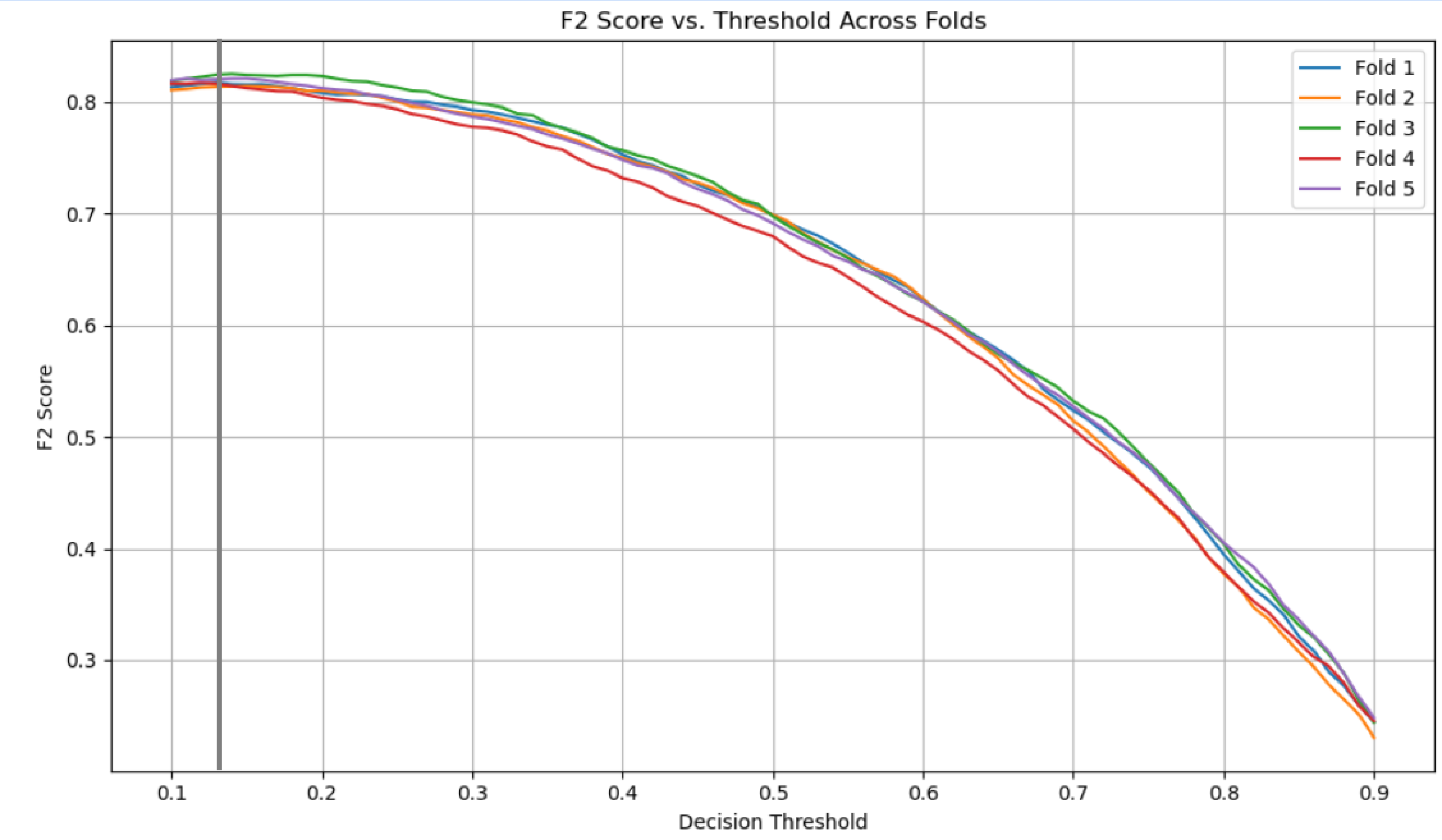
- Accuracy is only marginally better than guessing.

```
Model Coefficient Averages
 Intercept (β0) = -2.3845
   YEARBLT (β1) = -2.3035
NN_is_lead (β2) =  1.1628
```

```
Metrics for optimal threshold (~0.60)
 Accuracy = 0.915 (+0.0720)
Precision = 0.504 (+0.4179)
   Recall = 0.884 (+0.7984)
 F2 Score = 0.768 (+0.6823)
```

Optimizing Across Folds

Mean Optimal Decision threshold was determined to be 0.130

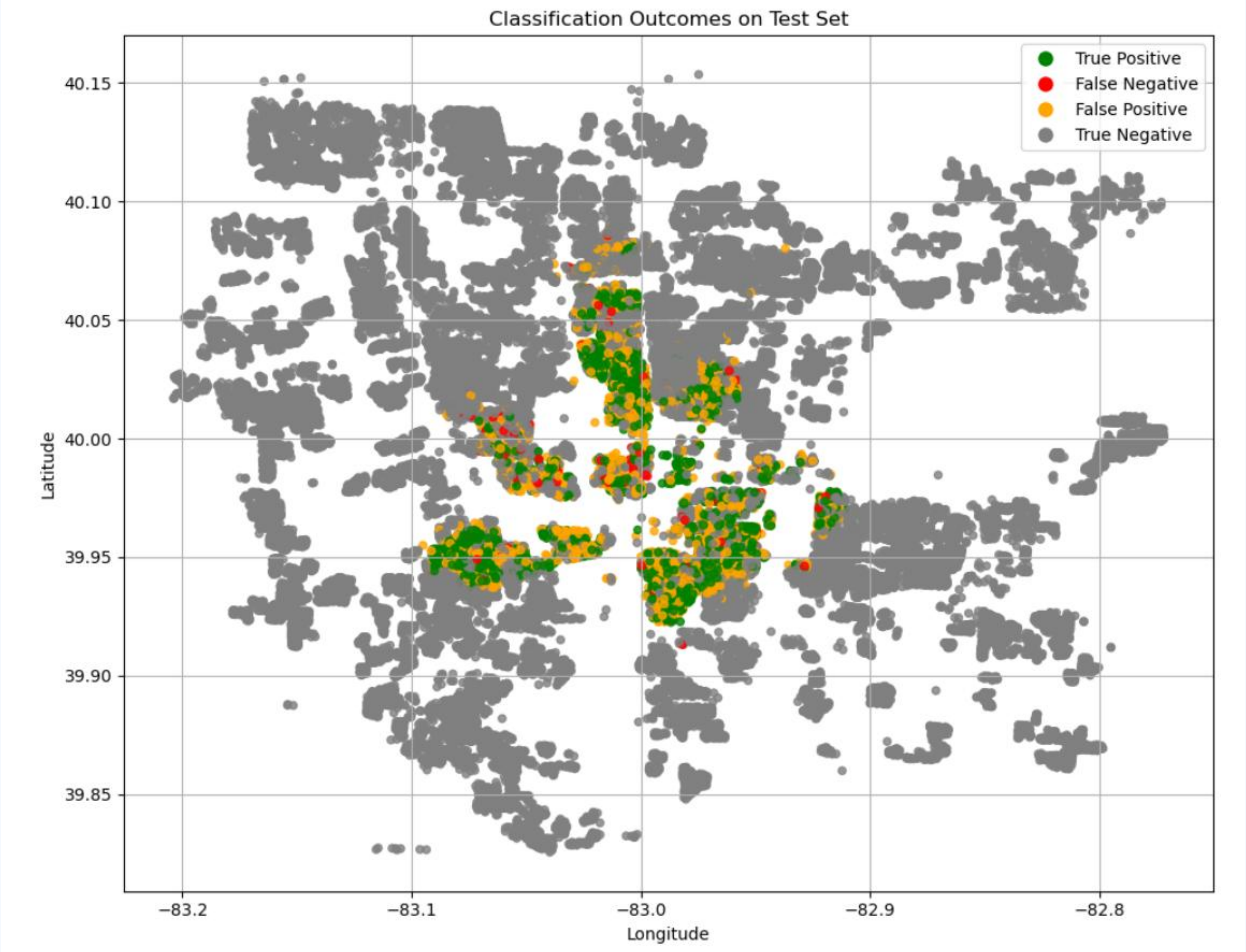


Nearest Neighbor (KNN)

- Effectively perfect prediction rates for regions located further from downtown Columbus.

- Can still predict reasonably well in the Downtown region.

- No clear trend behind the locations of false positives and false negatives.



Nearest Neighbor (KNN)

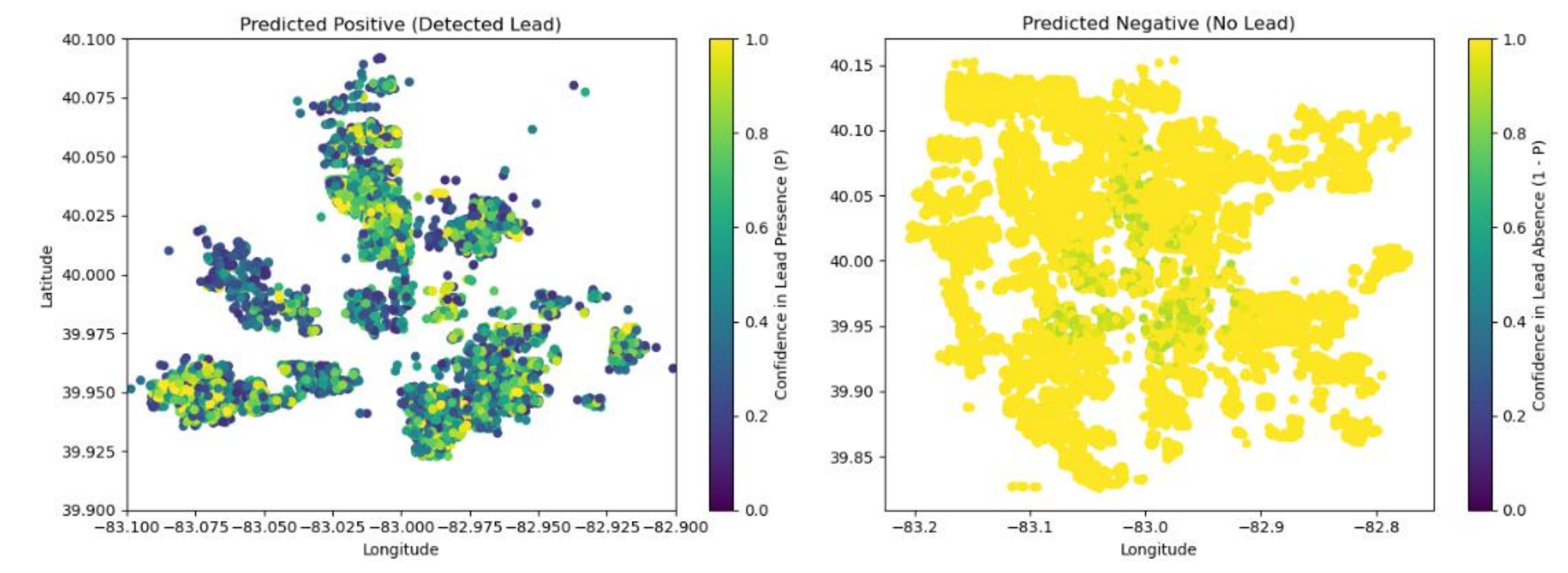
		Predicted	
		Lead	Non-Lead
Truth	Lead	True Positive 7.29%	False Negative 1.27%
	Non-Lead	False Positive 4.70%	True Negative 86.75%

```
Metrics          KNN    Log-Reg
 Accuracy = 0.930   0.915
Precision = 0.553   0.504
   Recall = 0.926   0.884
 F2 Score = 0.816   0.768
```

- 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% percent accurate. Conversely, predictions for downtown regions will be 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