Capstone Milestone 2: BlueConduit

Machine Learning to Detect Lead Service Lines

Lead Pipes: What is at stake?

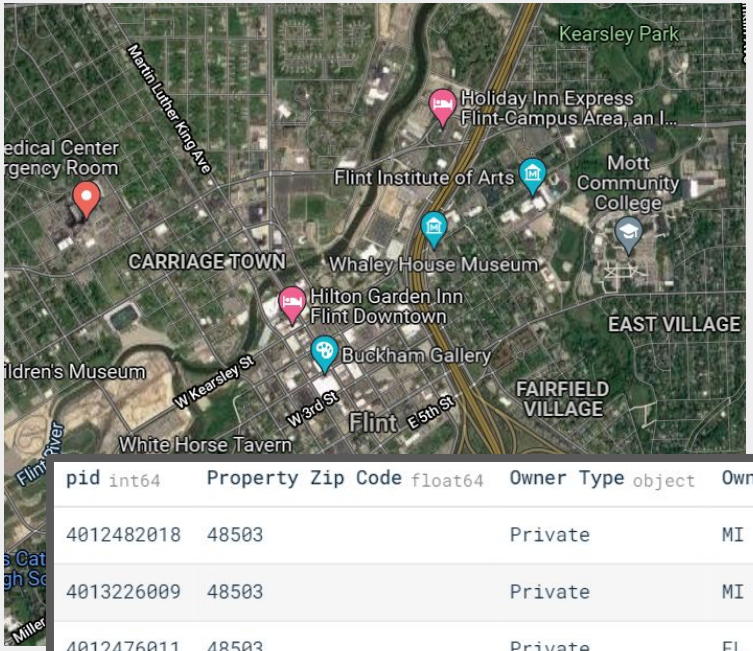
- When ingested, lead is highly poisonous to humans
  - Young children are particularly vulnerable

- Commonly used metal due to its malleability
  - Banned from inclusion in paint in 1978 & all pipes in 1986

Why is lead hard to find?

- Municipal records are scarce
- Digging pipes to confirm material is expensive

BlueConduit's Innovation



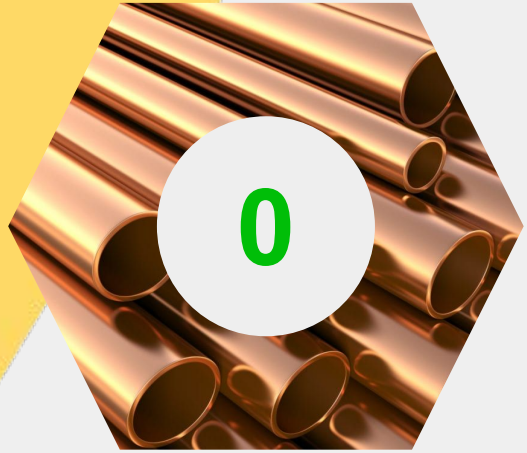
- Collected detailed data on homes in Flint, working with city and residents

pid int64	Property Zip Code float64	Owner Type object	Owner State object	Homestead object	Homestead Percent float64	HomeSEV int64	Land Value int64
4012482018	48503	Private	MI	Yes	100	18400	932
4013226009	48503	Private	MI	Yes	100	11800	420
4012476011	48503	Private	FL	No	0	0	602
4012481022	48503	Private	MI	Yes	50	4550	781
4013226025	48503	Private	MI	Yes	100	12800	510

BlueConduit's Innovation

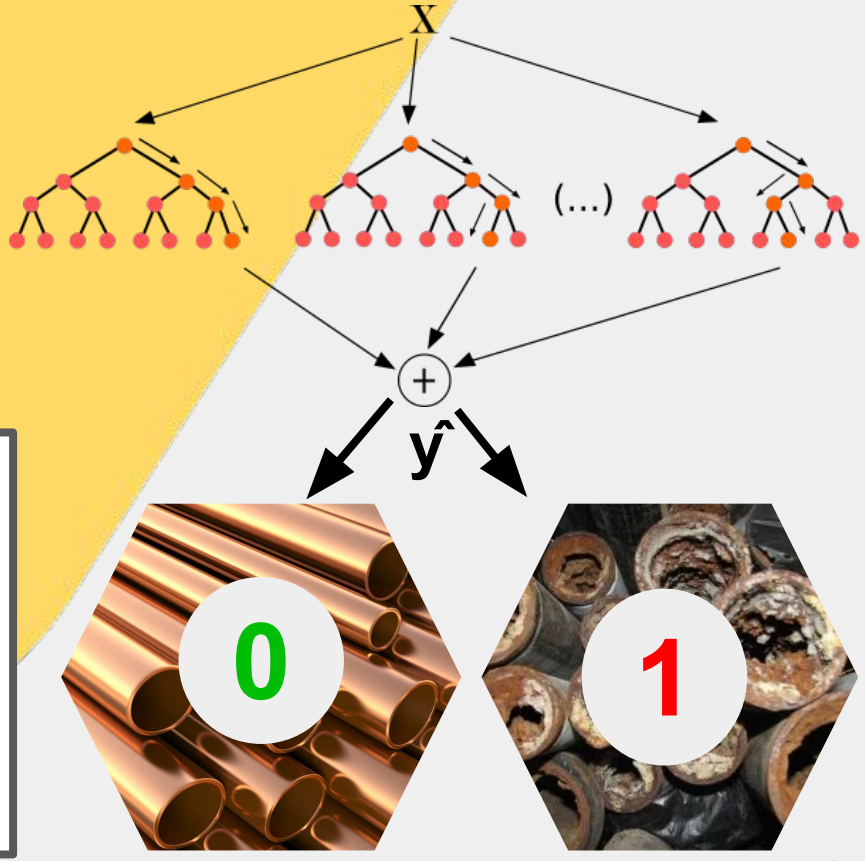
Used machine learning to predict copper/lead.

- City's initial digging:
  - 15% Hit Rate
- BlueConduit's digging:
  - 81% Hit Rate

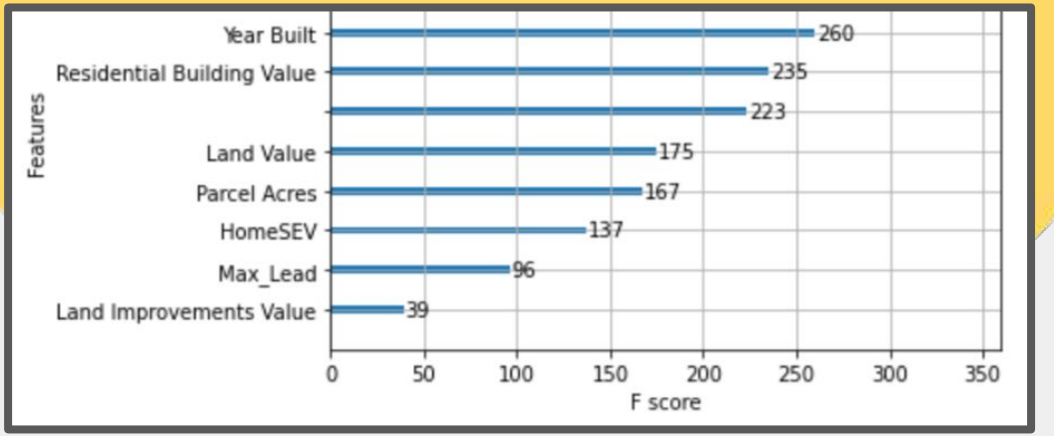


BlueConduit's Model

XGBoost

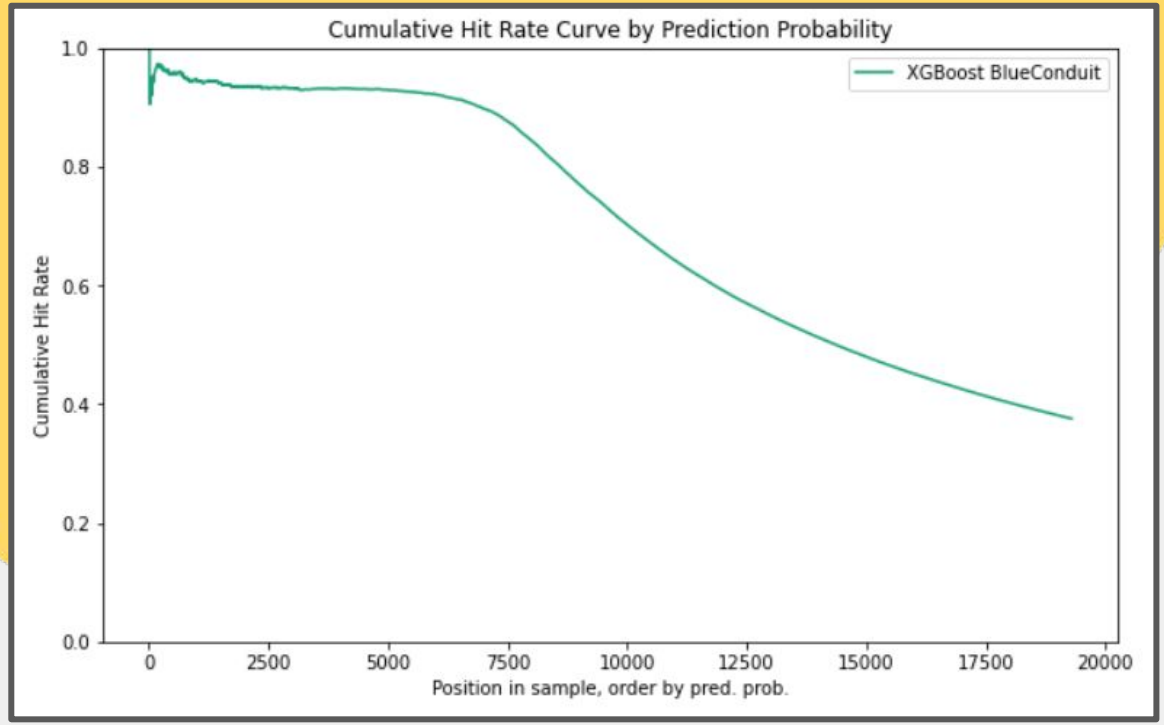


Feature Importance



BlueConduit's Model

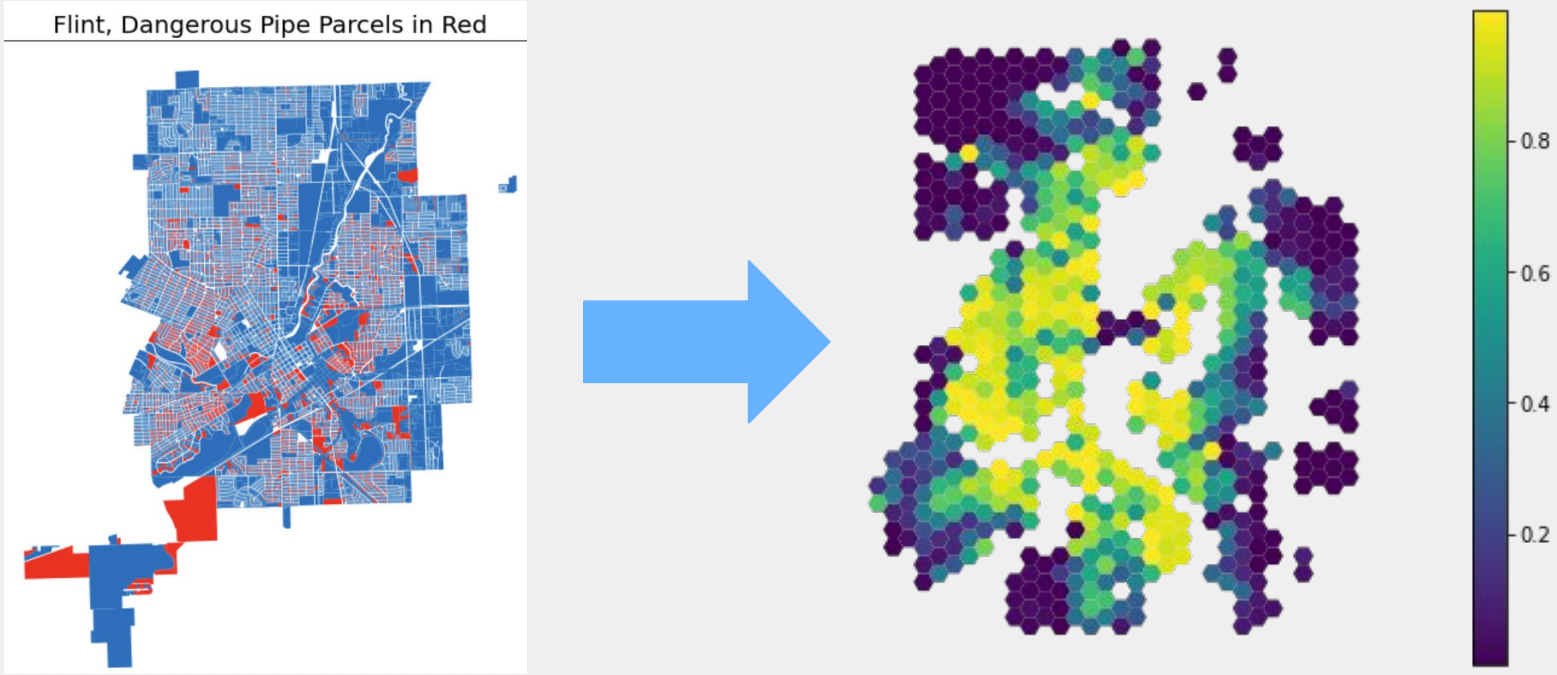
XGBoost



Performance: Hit Rate Curve

Motivation

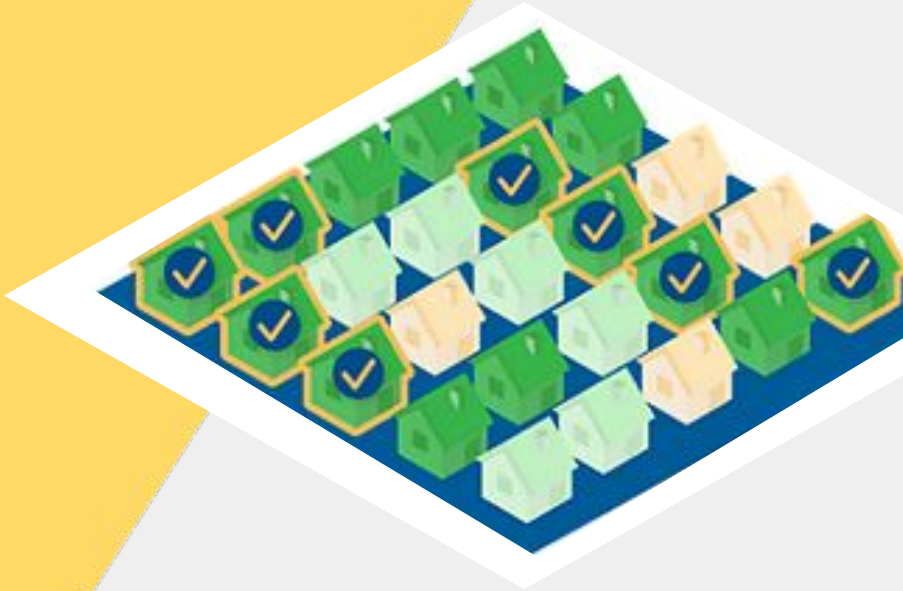
Can neighbours inform lead probability?



Scope of Work

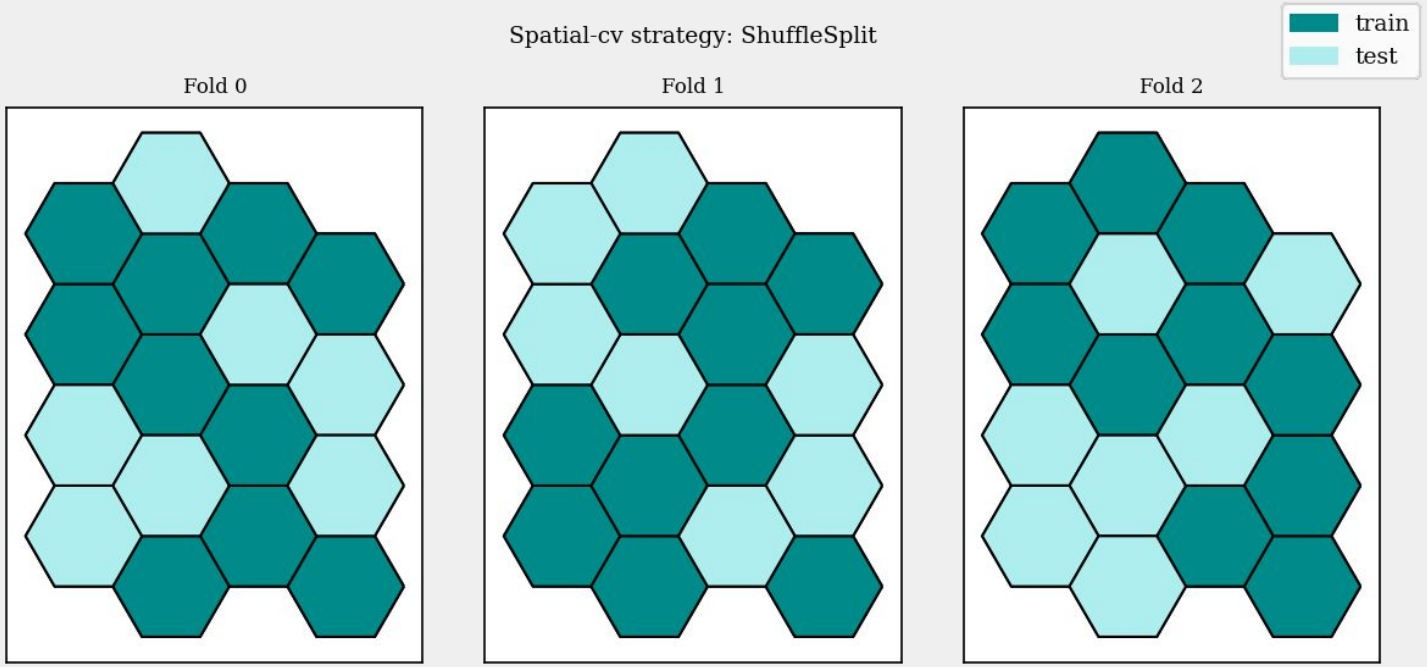
BlueConduit's model currently **does not use** spatial information.

Our task: Investigate whether using spatial information can help BlueConduit's model.



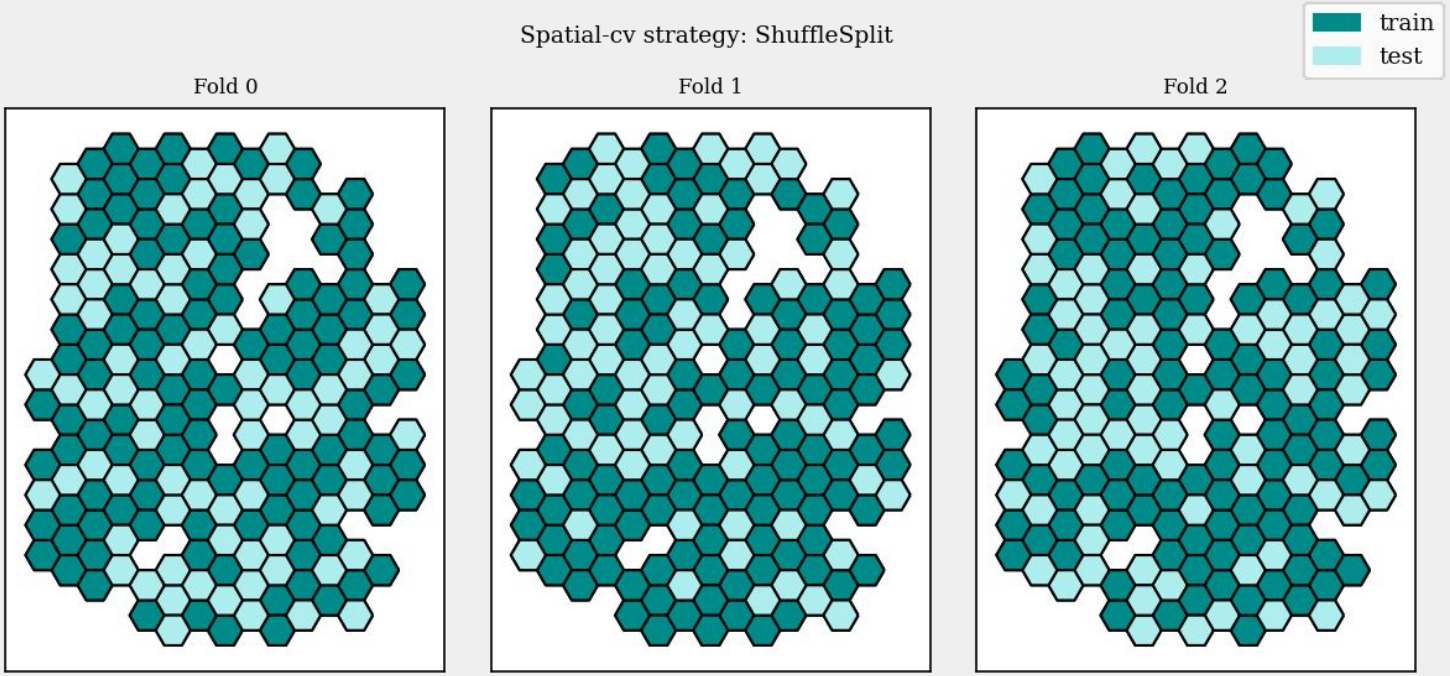
Evaluating our work

Spatial cross-validation & testing



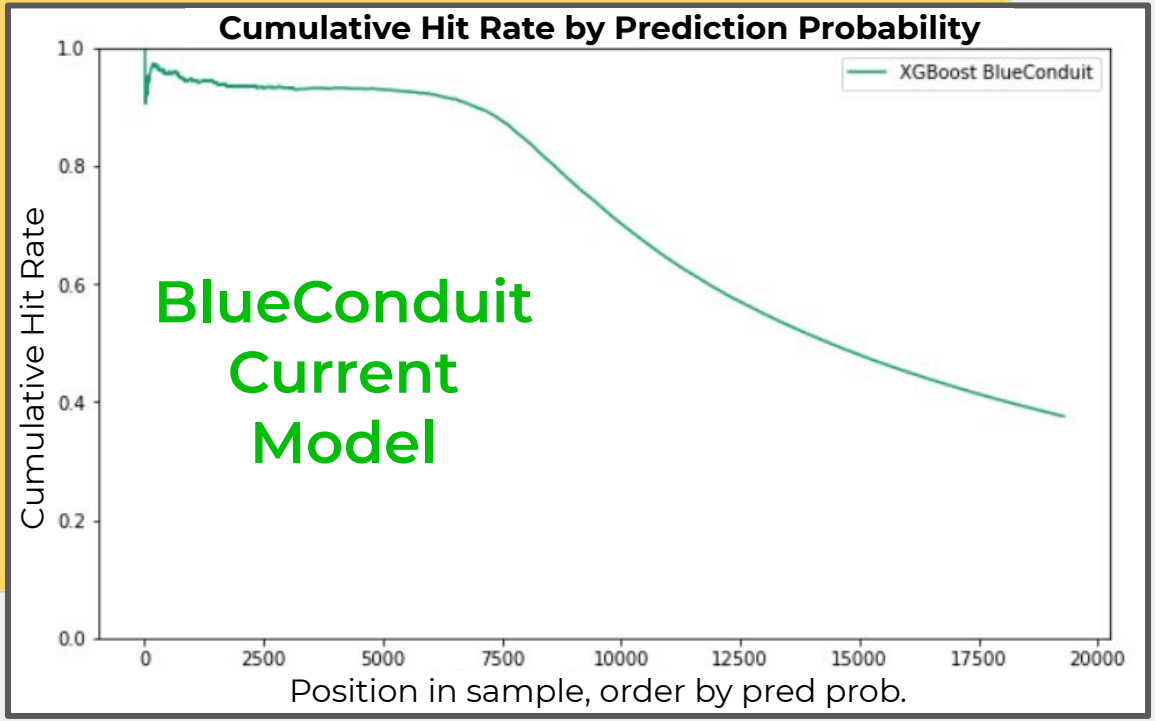
Evaluating our work

Spatial cross-validation & testing



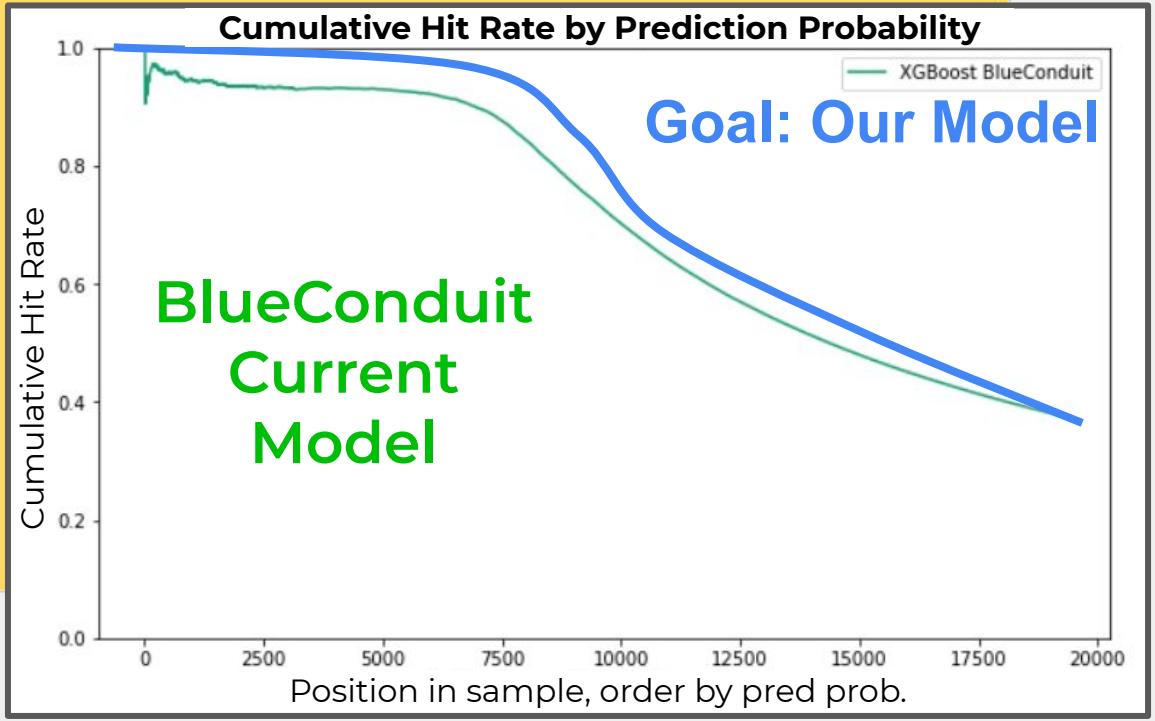
Evaluating our work

Goal

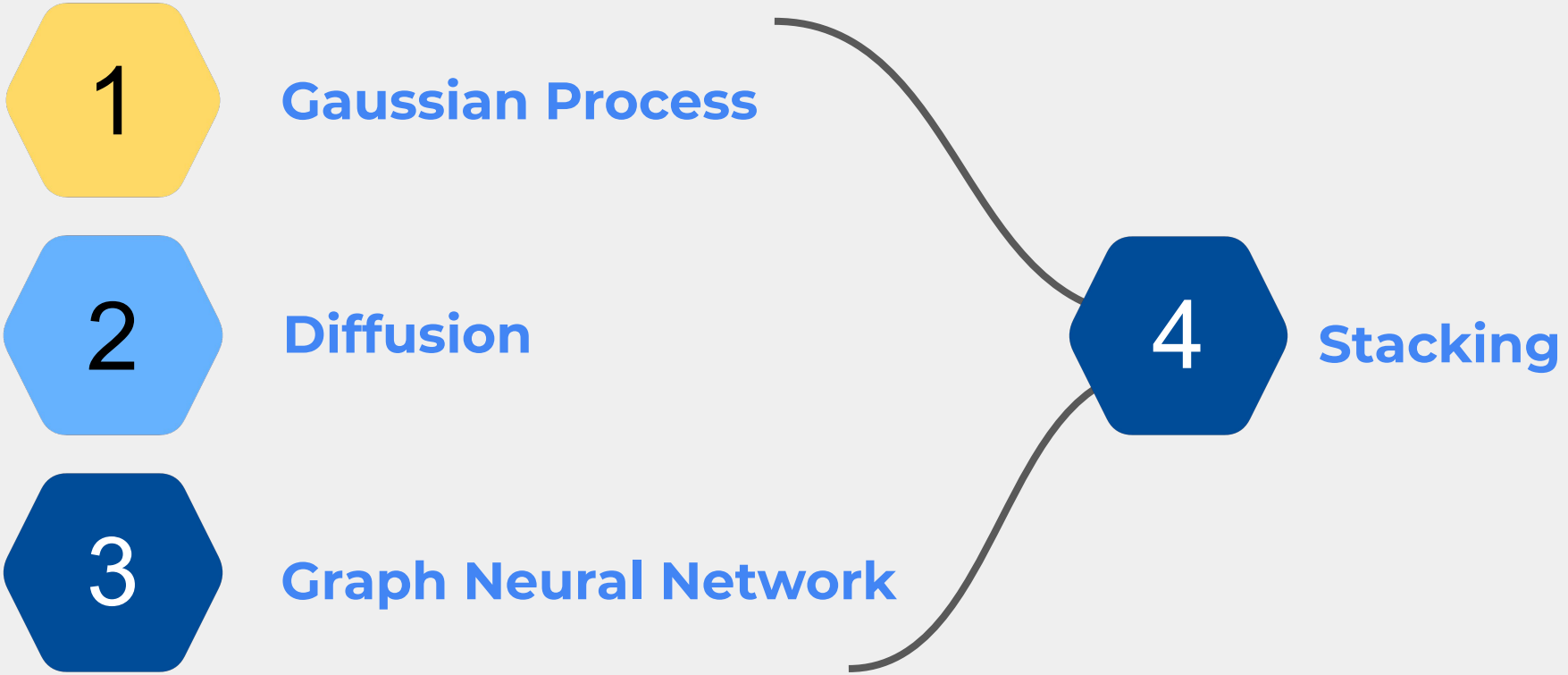


Evaluating our work

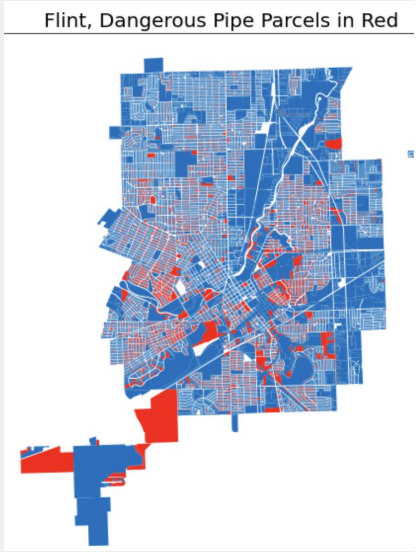
Goal

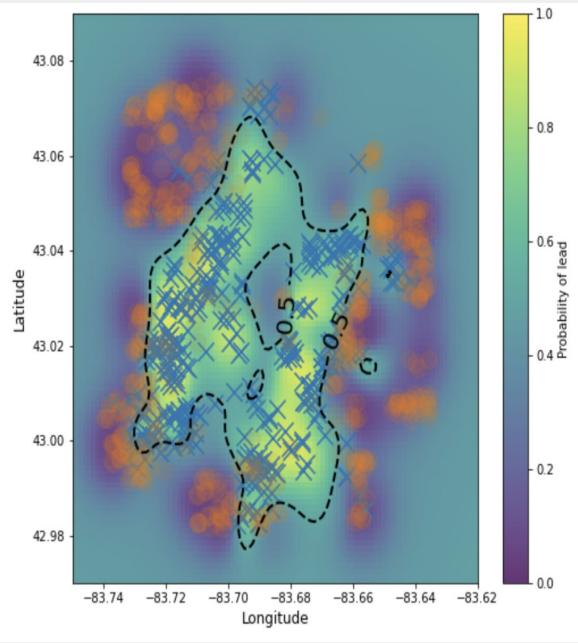


MODELS


1
Gaussian Process
2
Diffusion
3
Graph Neural Network
4
Stacking

1 *Gaussian process*





Features: Lat/Lon
Outputs: probability of lead

Upsides:

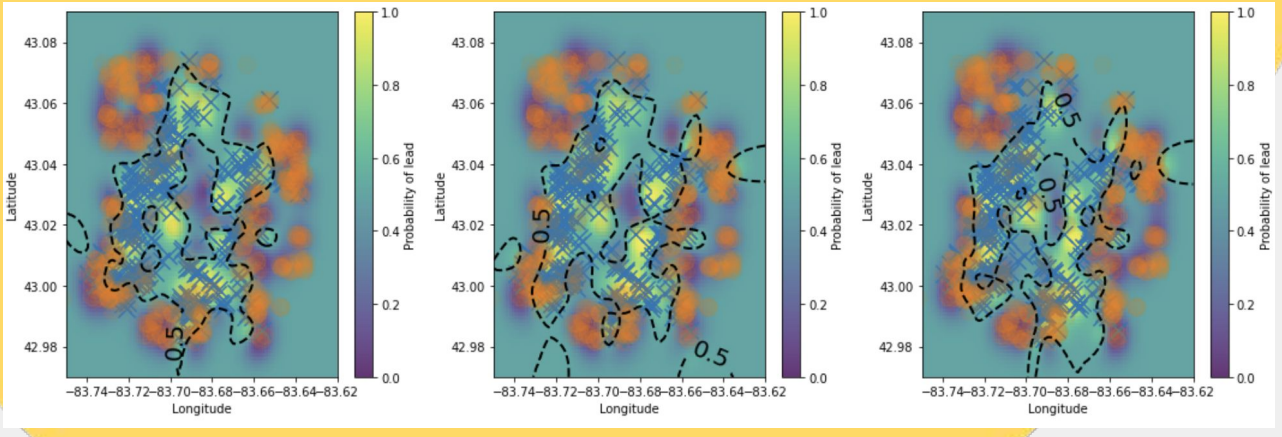
- Little data collection required
- Expresses uncertainty in unseen areas

Downsides:

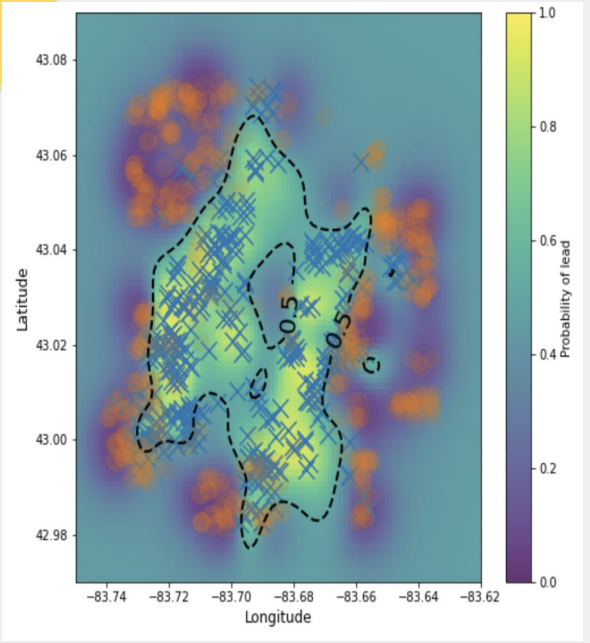
- $O(n^3)$ runtime, n = # homes
- Sensitive to hyperparameters
- Does not distinguish between two types of uncertainty:
  - epistemic (lack of data) & aleatoric (inherent noise)

1. Gaussian process

Ensemble of 50 GPs on subsets of 1000 homes

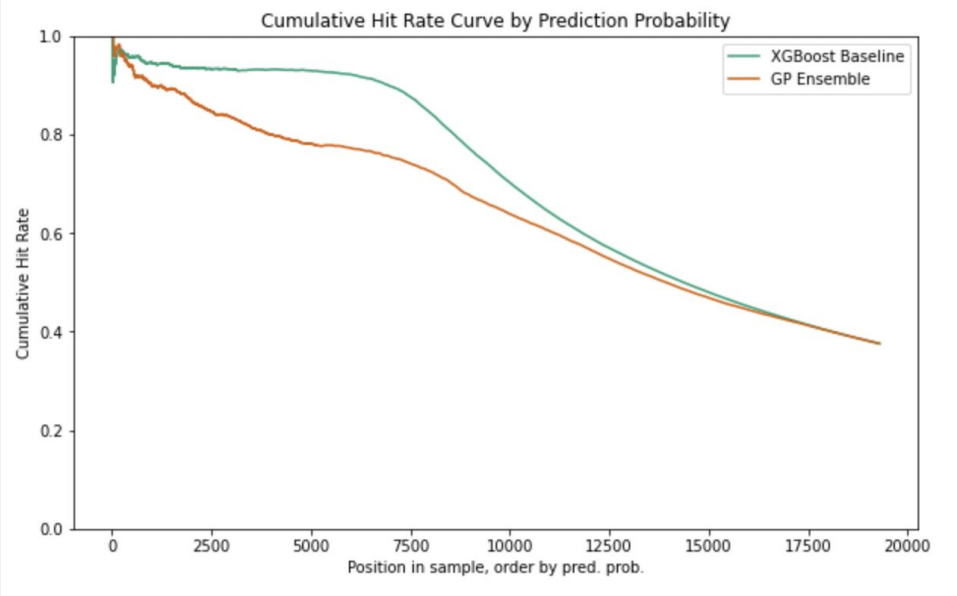


Ensemble Average



1 *Gaussian process*

Improvement over baseline only for the first few hundred homes



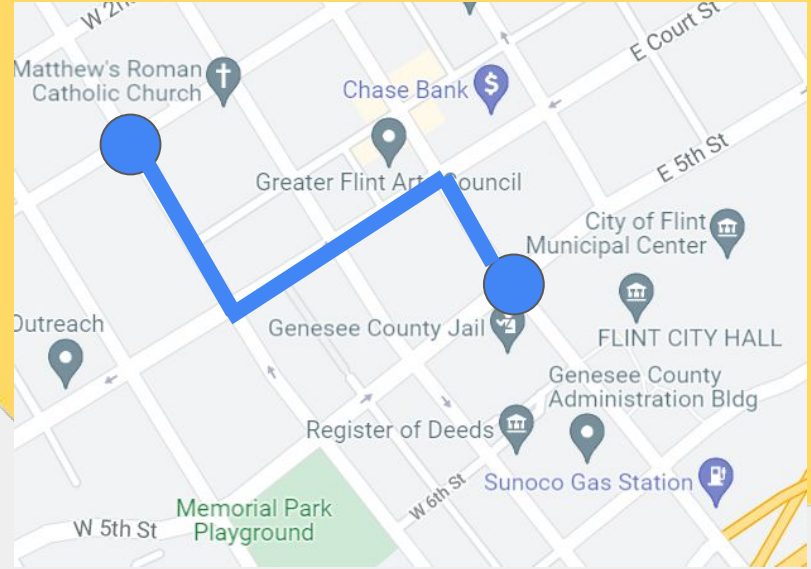
2 *Diffusion across a graph*

- **Intuition:** Homes near one another typically share characteristics (i.e. era of construction, builders, etc.)

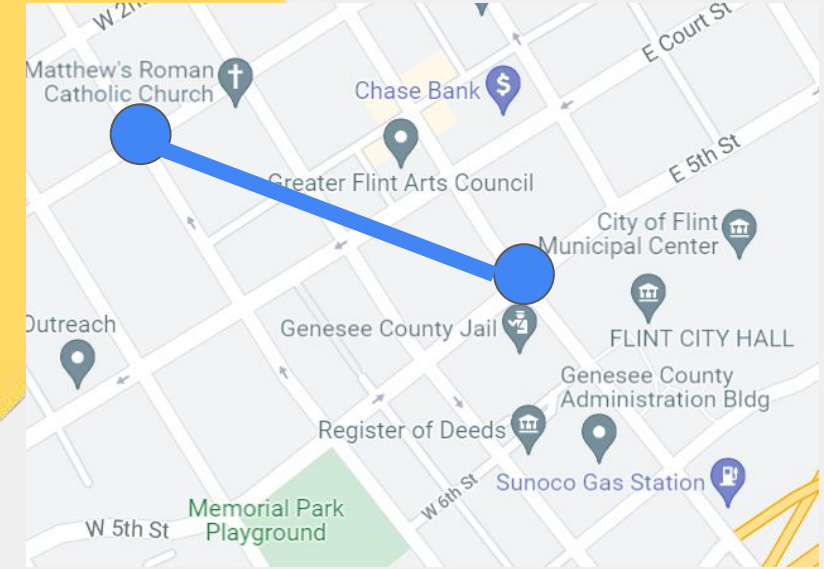
- kNN, literally:
  - “Smooth” out prediction probabilities

Getting Distances

Street vs. Euclidean Distances

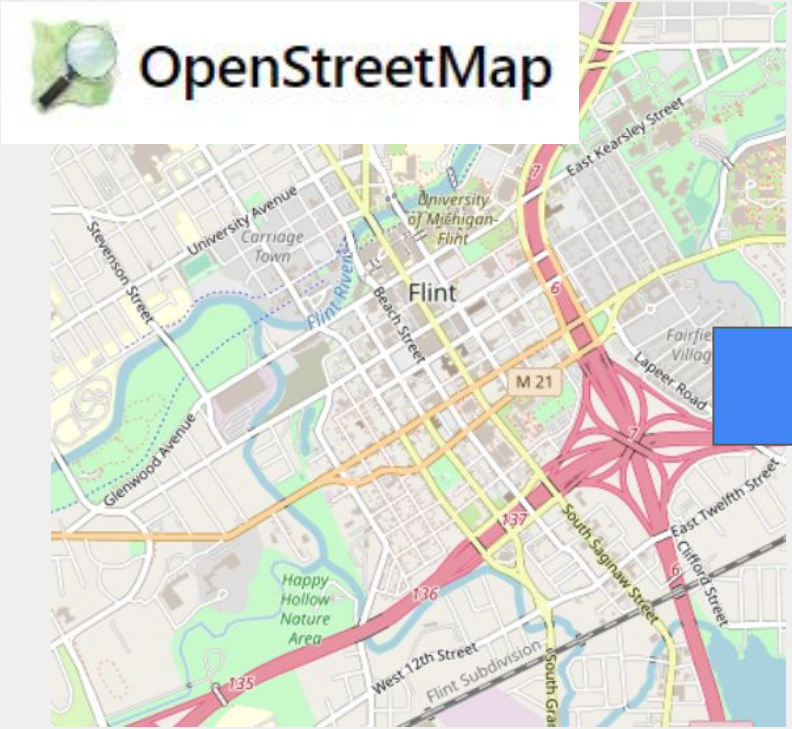


vs.



Getting Distances

Street vs. Euclidean Distances

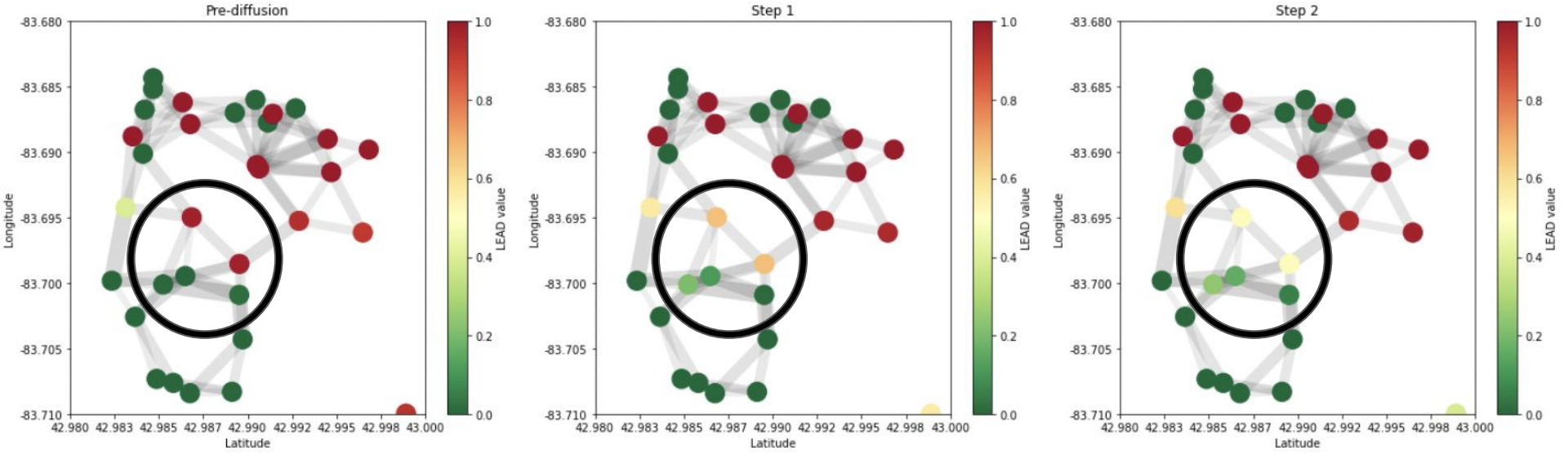


Parcels

Parcels	1	2	3	…	n
1	t_{11}	t_{12}	t_{13}	…	t_{1n}
2	t_{21}	t_{22}	t_{23}	…	t_{2n}
3	t_{31}	t_{32}	t_{33}	…	t_{3n}
…	…	…	…	…	…
n	t_{n1}	t_{n2}	t_{n3}	…	t_{nn}

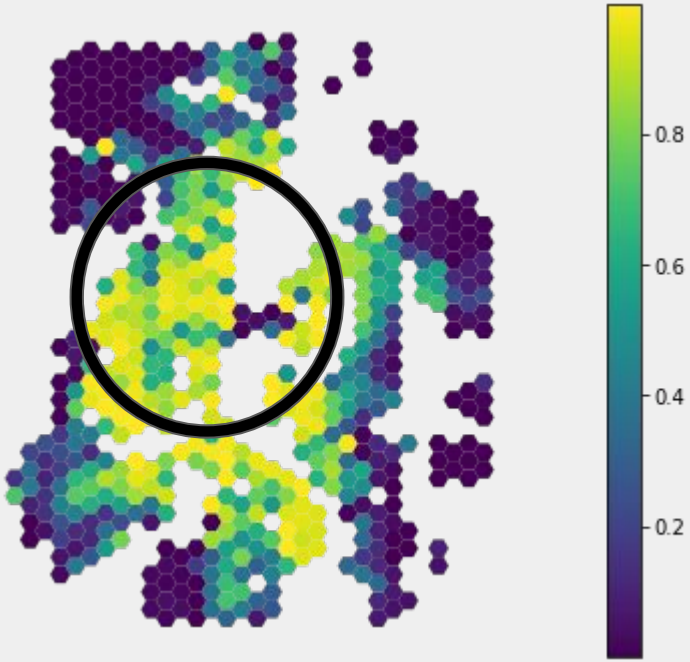
t_{ab} = walking time from a to b

Diffusion across a graph

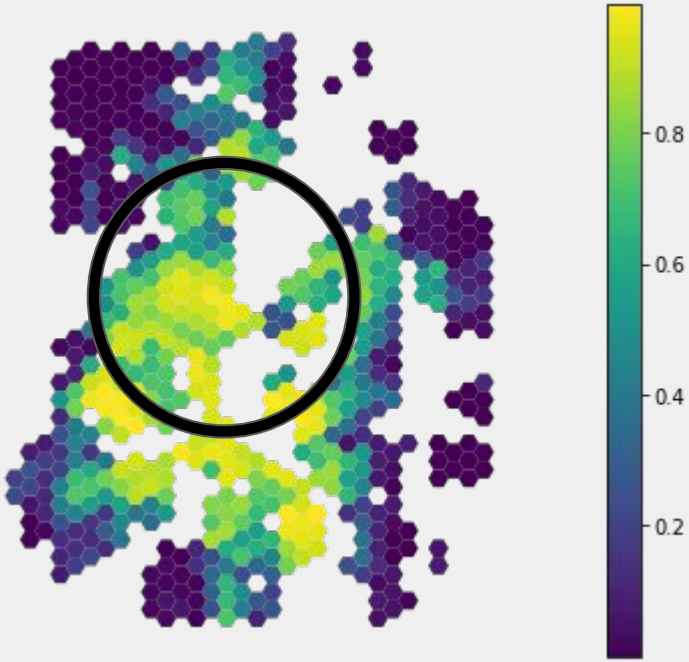


What does this look like on a map?

Baseline Prediction



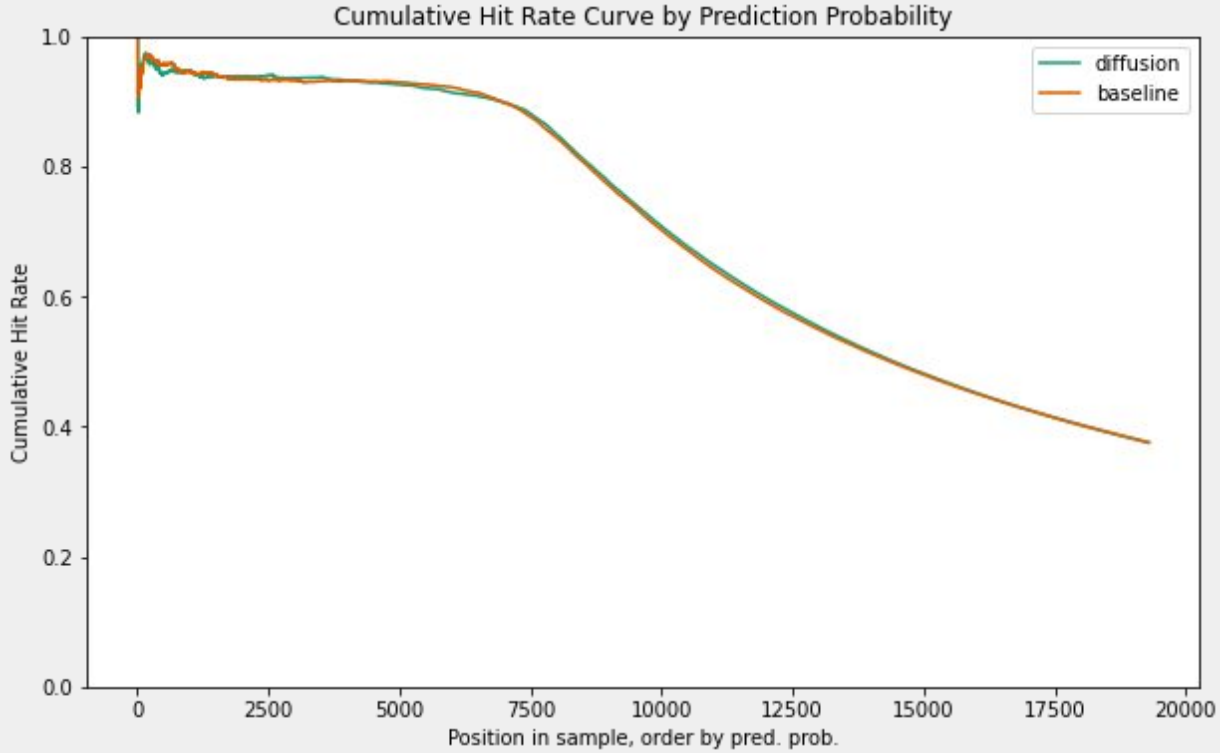
After 100 Iterations



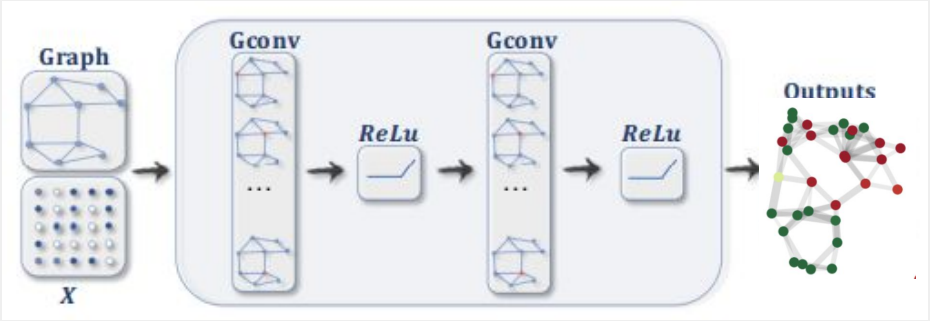
Diffusion hyperparameters

- *Who is a neighbor?*
  - kNN
  - Radius nearest neighbors
- *How many neighbors?*
- *Distance metric*:
  - Haversine (Euclidean) distance
  - Walking time
- *What is the kernel? (i.e. weighted average)*

Diffusion Results (1 Iteration)



3 Graph Neural Network



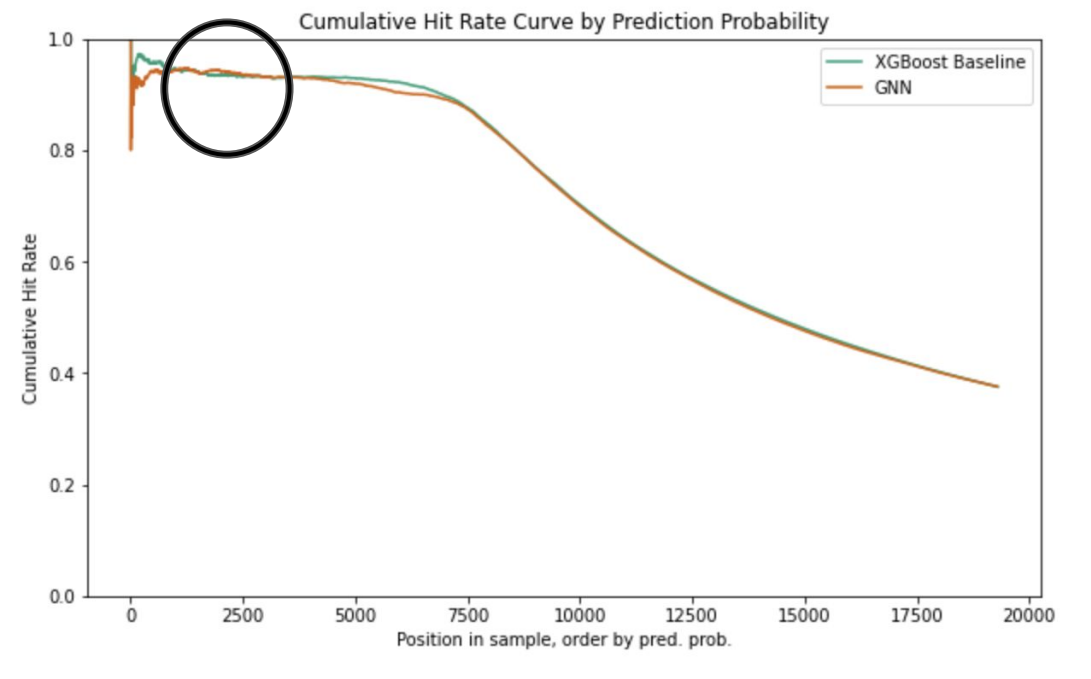
https://arxiv.org/abs/1901.00596

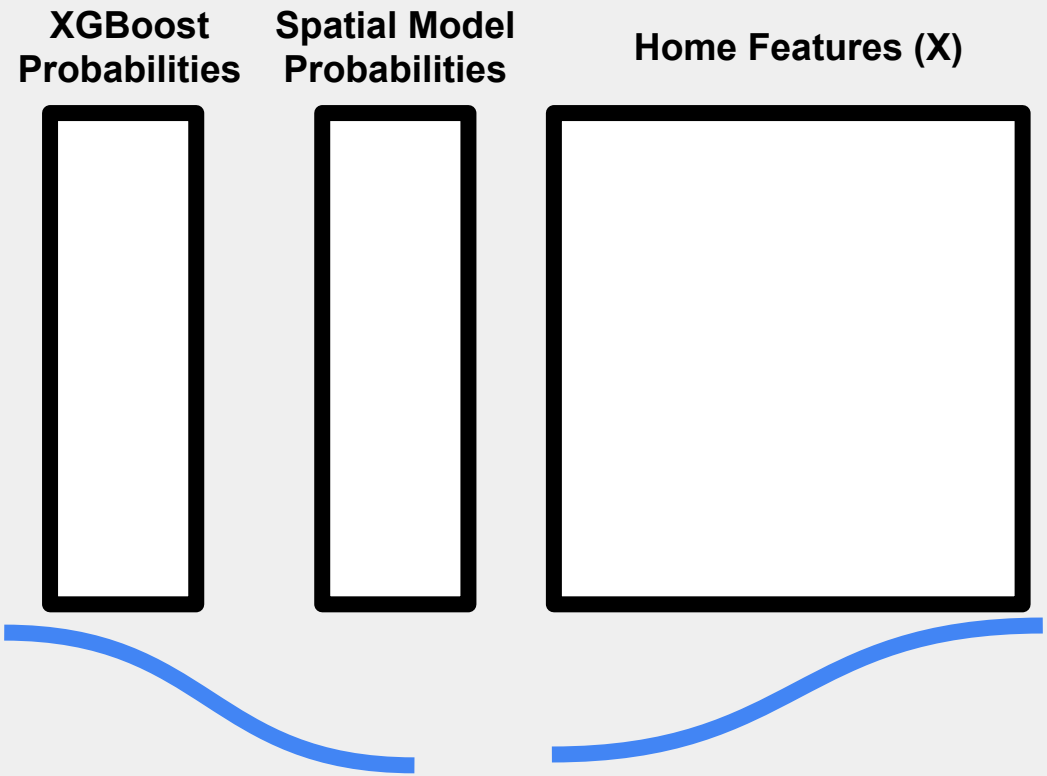
Features: important XGboost features & road distance matrix
Outputs: probability of lead

Upside:
Promising performance
Downside:
Hard to interpret

3 Graph Neural Network

Temporary improvement over baseline after ~2000 homes
However, seems to be just luck





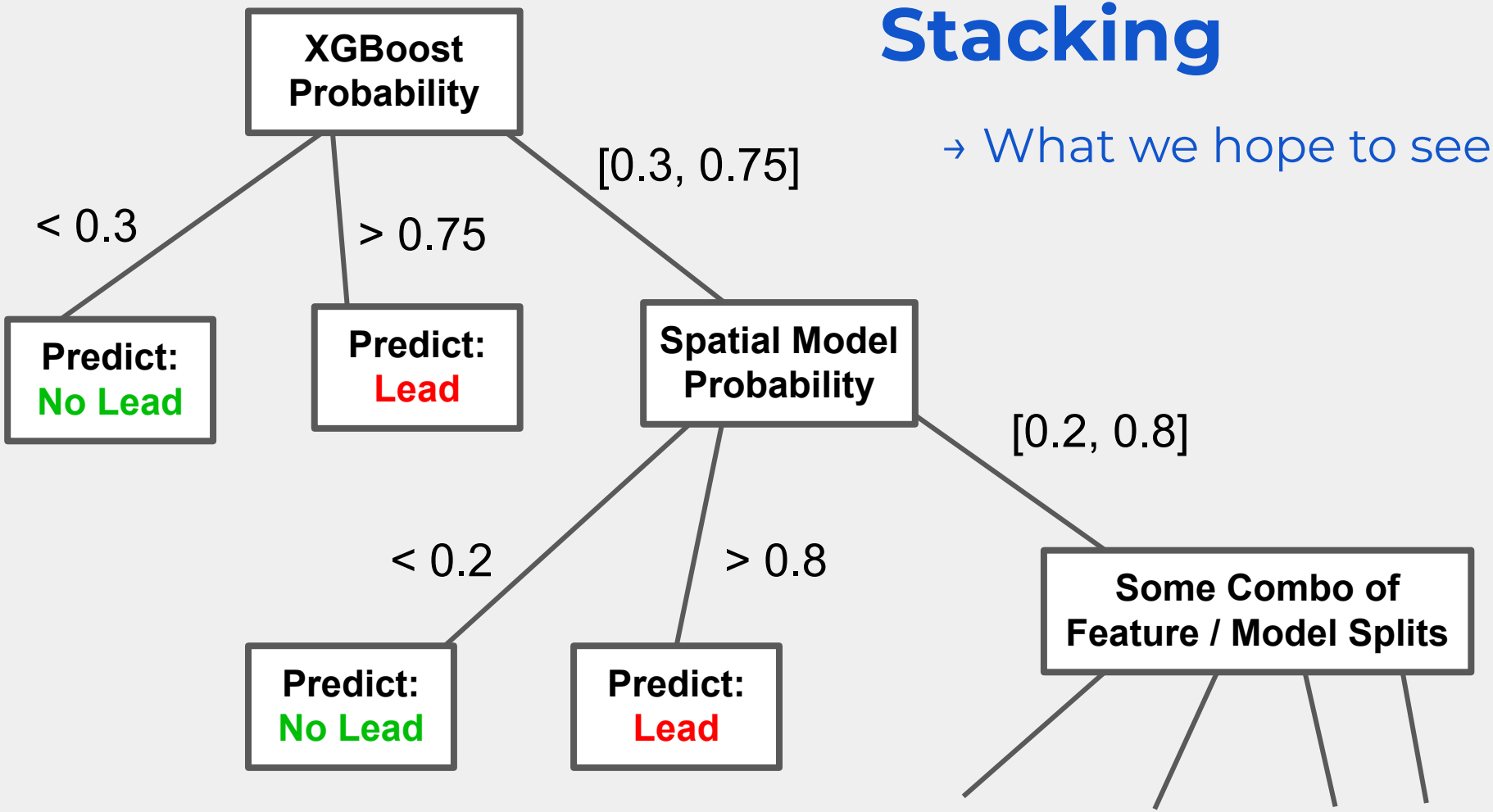
4 Stacking

Motivation: Only use spatial info *when it's helpful.*

Meta Model → Learns **when to use** XGBoost vs. Spatial predictions

Stacking

→ What we hope to see



Future directions

1. *Training / Test Split + Cross-Validation*
   a. Multiple train/test split validations.
   b. Test out different spatial resolutions.

2. *Use stacking with multiple spatial models at once*

3. *With more model tuning and steps #1-2, hopefully improve on BlueConduit's current hit rate curve.*

4. *Simulate excavation by neighborhood.*

What's At Stake?

Save lives in Flint

Save $$$ in Flint

Save lives and $$$ elsewhere

Thank you!

Kevin, Javiera, Max, Dash