Spatial Machine Learning for Lead Service Line Detection
Harvard IACS - BlueConduit
Capstone Project, Fall 2021
Javiera Astudillo, Max Cembalest, Kevin Hare, and Dashiell Young-Saver

Problem Description

Most cities around the United States have a lurking health hazard spread out across thousands of homes: lead pipes. Lead has a long history of being featured in pipes of all sorts. As a malleable and leak-resistant material, it was a popular choice for thousands of years.[1] Throughout the 20th Century, however, the health effects of lead became more apparent, with the detrimental effect on young children's development taking center stage.[2] To combat these effects, American regulators sought to remove lead from prominent environmental locations, namely household paint, service lines, and gasoline.[3] In this project, we are concerned with the second of these environments: lead service lines.

If lead were easily and inexpensively removed, though, this project would have begun and concluded around the time that the U.S. Environmental Protection Agency (EPA) banned lead from new service lines in 1986. Of course, that is not the case. The lurking danger element of lead service lines manifests due to the fact that low levels of lead in water sources may not be immediately obvious. Moreover, the largest challenges occur when lead *leaches* into the water supply from corroded service lines, highlighting that minor environmental changes such as a change in water supply or failing solder can suddenly introduce significantly higher concentrations of lead into a home's water supply.[4] Finally, there are substantial data challenges in this domain. Cities typically have poor records of lead service lines, and many of these service lines are privately owned and without a public record at all. The cost associated with false positives and false negatives are both high, compounding this problem. Because failure to locate lead pipes could lead to lead poisoning for thousands, there is an obvious desire to limit the number of dangerous homes which are never investigated. On the other hand, verifying whether a home has lead pipes requires extensive excavation and incurring a cost of roughly $2,500, regardless of the outcome.

In this project, we will be assisting BlueConduit, an Ann Arbor, MI-based startup that develops bespoke machine learning algorithms to assist municipalities and water utilities find lead service lines so that they can be replaced. In particular, this project will focus on ***investigating whether spatial information can be incorporated into an existing machine***

---

[1] https://www.safeplumbing.org/advocacy/health-safety/lead-in-water.
[2] https://www.cdc.gov/nceh/lead/prevention/health-effects.htm.
[3] https://www.epa.gov/lead/protect-your-family-sources-lead
[4] Service lines are the pipes that connect a residence or commercial property to the common water main. From discussions with BlueConduit, we understand that lead was not a popular choice for larger commercial-use properties due to its relative weaknesses when carrying high-flow water.

learning pipeline to improve the "hit rate", or the rate at which digs find lead. To date, BlueConduit's models focus on parcel-level features only, incorporating information such as the year the home was built and other neighborhood-level features to predict whether a particular parcel is likely to have lead. Because these observations are essentially viewed as independent and identically distributed, there is substantial spatial information loss. Intuitively, homes which are near one another are likely to share characteristics which cannot be inferred directly from home-level features. In a literal sense, this is the same intuition behind ML algorithms such as k-Nearest-Neighbors, where observations (or parcels) that are near to one another are assumed to provide information about each other as well.

The outcomes and goals of this project are twofold. First, following discussions with our partner organization, we plan to evaluate multiple approaches to incorporating spatial information to see whether it can improve the model. Importantly, from our partner's point-of-view, this investigation can be successful even if spatial information does not improve the model. BlueConduit has rarely tested these spatial methods. Indeed, they have stayed away from direct utilization of spatial features such as longitude and latitude, which are prone to overfitting and failure mode due to the lack of causal structure. During this evaluation phase, though, we would consider a model successful if it improves the "hit rate". As described above, this can be thought of as the precision over the "dangerous" homes. A higher hit rate means more excavations that find lead and fewer which do not, improving the rate at which lead can be removed and reducing the total program costs.

As a stretch goal, we hope to evaluate whether the spatial information can improve predictions for low data environments. As described below, we will work with data for Flint, MI. Flint has been plagued by a water crisis which began unfolding in 2014 and culminated in widespread distribution of bottled water to residents for years due to the extreme levels of lead within the city. As a result, there is a rich dataset for us to work with. As BlueConduit moves from Flint to other cities, they are particularly interested in whether this spatial information can provide lift when few predictors have been captured and many homes have not yet been verified.

Data

The primary data for this project have been provided by BlueConduit and focus on Flint, MI. The raw dataset is roughly 55,000 rows and contains 74 features. This proprietary data records a single observation for each parcel in Flint. Notably, however, only roughly half of the data have a determination of lead or no lead. These reflect homes which have been verified either before the Flint Water Crisis began in 2014 or were previously known to the city.

There is one notable bias here, which is that there is a real-world selection mechanism going on when each individual parcel is selected and verified. The homes in this dataset may be more likely than the average parcel in Flint to have lead, as those would

have likely been targeted for removal. We do not believe that this is necessarily an issue for the external validity of the project for two main reasons. First, our approach is model and feature agnostic. While we will use the Flint data to validate the approach and measure progress, our charge is to assess the viability of the spatial information for Flint and future cities. Second, the overrepresentation of homes with lead may contribute to more of a balanced dataset for our model's purposes. Class imbalance is a well-known and omnipresent issue in machine learning, and if the selection effect were reversed (i.e. the dataset were likely to have very few homes with lead service lines), that would be cause for greater statistical concern.

Over the previous years, BlueConduit and the city of Flint have developed many novel sources of data, including the digitization of over one hundred thousand index cards detailing citywide maintenance.[5] Due to these efforts and the scope of our project, we do not intend to develop additional features apart from the spatial information. One outside source of data which we will be incorporating is OpenStreetMaps (OSM). A natural challenge of this project is measuring distance. True Euclidean (or Haversine) distance over a map may not accurately represent the way that spatial information should flow through a city: for example, if two houses share a backyard but are on different streets, then they might have small Euclidean distance but be connected to two different water mains. Thus, we intend to query OpenStreetMaps to retrieve walking distances to better mimic the patterns of development in Flint and other cities, which are likely to be more important than a direct geographic measure. We hope that this is especially promising and it can be implemented for any city in the United States so long as the appropriate OSM tooling is used.

Exploratory Data Analysis (EDA)

The dataset has 26,863 rows, each representing a parcel (home or commercial property) in Flint, MI. Each parcel has 74 described features, such as the market value, size, location (latitude/longitude), year built, voting precinct, etc. Because our dataset size is relatively small, we are currently not too concerned about efficiency and computational issues.

[5] https://arxiv.org/pdf/1806.10692.pdf.

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

Table 1: *The head of our dataset, which shows the one-row-per-parcel structure of the data*

In addition to our features, we have a column for our target: a binary indicator of whether the home has dangerous pipes (1) or safe pipes (0). In the full dataset, we found that 10,266 parcels had dangerous pipes, or about 38% of all homes. Because this is a relatively large proportion of all homes, we don't have to treat the target as a "rare" outcome in our models.

Commercial Condition 2013	26760
B_aggregate_income	4666
B_imputed_value	1500
B_imputed_rent	1500
B_hispanic_household	1500
Housing Condition 2012	400
Residential Building Style	114
Housing Condition 2014	109
USPS Vacancy	59
Owner State	26
Longitude	6
Latitude	6
Use Type	2
Zoning	2

Table 2: *Missing value counts*

Table 2 shows the missing value counts for different features in our data. Most concerning was the high count of missing values for "Commercial Condition." However, we quickly discovered that the vast majority of parcels in the dataset are residential, and this feature is coded as "missing" for residences. So, it's not truly missing but, rather, not applicable to residences. The other features with large missingness rates are the "B_" features. These are all features that were imported from the American Community Survey (part of the US Census). We found out from the partner that the missingness is likely from homes that could not be linked to ID's from the American Community Survey. If these features are found to be important, we may go back to the census to see if we can successfully match these homes to American Community Survey data and fill in the missing values with the true values.

Papers published by BlueConduit show that "year built" is consistently the most important predictor of lead in their models. So, we explore this variable with particular interest. A histogram of "year built" is shown in Figure 1.

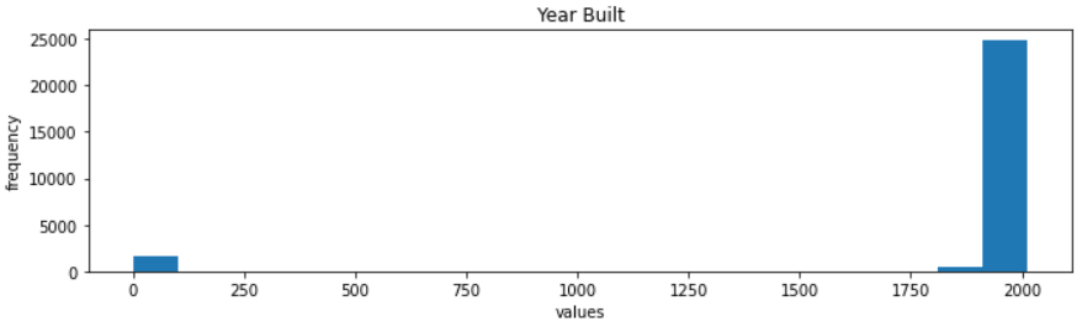


Figure 1: *Histogram of "year built," including the nonsensical values present in the data*

We saw that there were a fairly high number of nonsensically small values - homes that were built before the year 100 A.D. In total, there are 1,629 homes with nonsensically small year built values. We asked the partner why these values are present, and they reported that some city records are simply unreliable and that they treat these values as "missing." The Baseline XGBoost model is able to take this missingness in stride by splitting into different cases whether the year built feature is present or absent - we plan to use a similar approach when processing data in our spatial model as well. In future work, we may try to see if the missingness of these values corresponds with certain neighborhoods in the city or with values of other features.

When looking just as the sensical year built values, we find an interesting pattern, visualized in Figure 2.

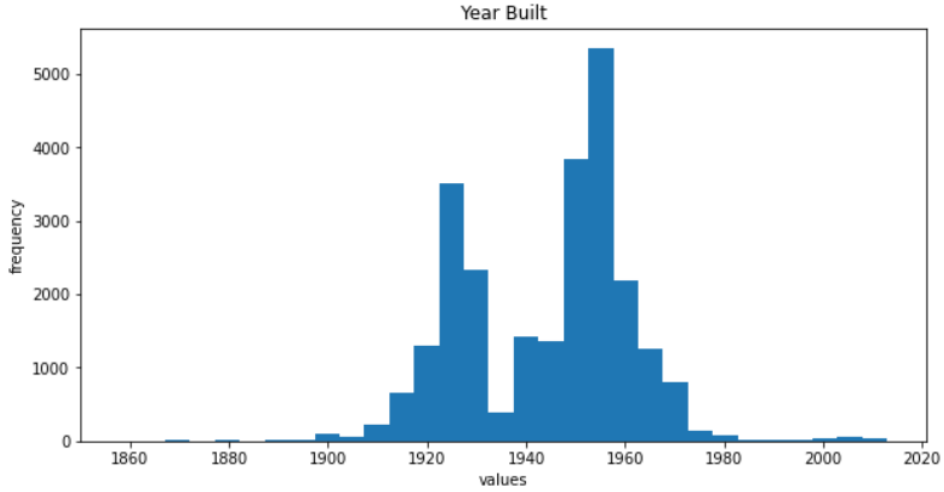


Figure 2: *Histogram of "year built," only including the sensical values*

Figure 2 shows that year built appears to be somewhat bimodal, with large buildup of homes prior to 1935 and immediately following 1945. This is likely due to historical circumstances, such as a lower amount of homes built during the Great Depression (1930s)

and with a boom after the second world war (1950s). It may end up being the case that these two eras (pre Depression, post WW2) are fairly predictive of dangerous pipes.

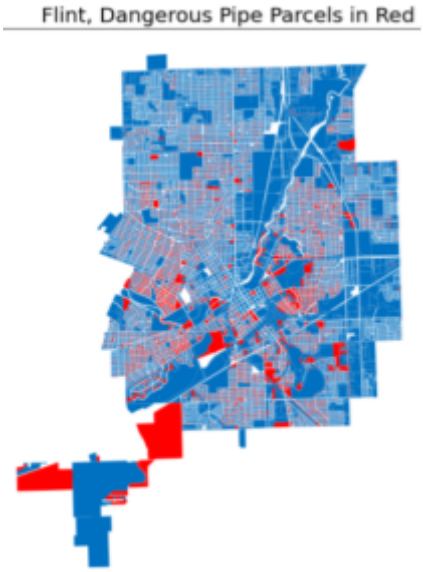


Figure 3: *Flint Michigan. Homes with known lead pipes in red.*

To test this theory, we found the rate of dangerous pipes within both eras. Among pre-Depression homes, 88.5% had dangerous pipes. Among post-World War 2 homes, only 7.2% had dangerous pipes. So, it's clear that year built is highly predictive of lead presence. This explains why it's such an important feature in BlueConduit's current models.

Finally, in Figure 3, we visualized parcels where there is known lead in Flint (lead parcels shown in red). We see that lead pipes are distributed throughout the city; however, there is some variation in lead pipe density by neighborhood. This suggests that using geospatial information could help improve the BlueConduit models. In future map visualizations, we will vary color intensity by predicted probability of lead, and we will refrain from coloring the large commercial parcels (to preserve the granularity of seeing the smaller residential parcels).

Literature Review - BlueConduit

The very first work of BlueConduit with the Flint Council dates back to 2016 and is reviewed in Chojnacki at el. (2017)[6]. There they frame Flint's water lead contamination, examine the water sampling process in detail and propose an initial model for house prediction lead presence. Their work also includes a discussion on how selection bias is a concern. Their models directly impact public health and, thereby, must address interpretability and accountability.

Their initial proposed model predicts a lead presence probability based on parcel dataset, service line dataset, and census dataset, for a total of 71 features. Their target variable uses water testing datasets based on residential and sentinel water testing programs, gathering information for around 15,000 parcels. Their findings show that the most predictive features for predicting lead levels include home value, demographic data from the census bureau and property age.

Their next work (Abernethy et al. 2018[7]) describes their model results in the context where the City of Flint took action in the water crisis. In 2016, Flint's Mayor contracted the

[6] https://arxiv.org/pdf/1707.01591.pdf
[7] https://arxiv.org/pdf/1806.10692.pdf

Flint Fast Action and Sustainability (FAST Start) team. Their task was to remove as many hazardous service lines as possible up to the funding level. This publication analyses how their model considerably surpasses the FAST Start strategy and empirically shows the resources saved through their strategy.

They update their classification framework to handle unobserved variables with a Bayesian spatial model. Further, their targets are built based on excavations in conjunction with water samples lead level (ppb) in this new setting. As of September of 2017, the FAST Start had carried out 6,506 home excavations. Consequently, they pose their strategy in an active learning framework, simulating the actual scenario process. Every time they make a new batch of discovery excavations, they update their dataset and their predictions accordingly.

They quantify the cost savings between their strategy and Flint FAST Start's actual home selection during 2016-17. Their strategy reduced the rate of costly unnecessary replacement visits from 18.8% (actual) to 2.0% (proposed). Suppose 18,000 total planned service line replacements; this would translate into savings amounting to as much as $11M from current spending. This is approximately equivalent to 2,100 additional homes in the city that would receive safe water lines.

Literature Review - External

Kernel Cookbook - University of Toronto
https://www.cs.toronto.edu/~duvenaud/cookbook/

This resource has been useful when exploring the use of a Gaussian Process (GP) as a spatial model to predict the presence of lead at a given latitude/longitude. We discuss more about our reason for investigating GPs in the *Developed Model* section below. The GP kernel is an important design choice: specifying the covariance matrix between data points represents our prior belief about the strength of spatial correlation between the parcels.

Data Dependent Distance Metric for Efficient Gaussian Processes Classification
https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.719.8106&rep=rep1&type=pdf

We conjecture that using a distance matrix calculated from walking distance in OpenStreetMap will help us track correlation between homes that lie on the same water pipe system, which are built underneath the city streets.

Bayesian Spatio-Temporal Gaussian Process
https://dl.acm.org/doi/fullHtml/10.1145/3300185

We found this paper when researching spatial models that take time into account as a feature. The temporal dimension is a very relevant consideration for our project since our

EDA and the trained XGBoost model both show the importance of YEAR BUILT as a feature.

Matern Kernels on Graphs
http://proceedings.mlr.press/v130/borovitskiy21a/borovitskiy21a.pdf

This is a resource we have investigated for future use if/when we can represent our data as the nodes/edges of a graph; cross validation over kernels of scikit-learn Gaussian Processes on our data suggested Matern kernels performed best among kernels.

Baseline Model

Our chosen baseline model consists of BlueConduit's current model, mentioned in the previous literature review. We agreed on this with our partner since this project aims to improve based on what they have built so far by incorporating spatial modelling.

More specifically, the baseline model consists of an XGBoost, with 102 input features, spanning parcel, service line and census data. Their target is a binary variable indicating whether there's a lead pipe presence or not. We are mainly concerned about this model's hit rate, representing the lead rate discovery behaviour of the samples ordered by their predicted probability of having a lead pipe. We will consistently compare our developed models with this same metric, ultimately the most significant performance evaluation method for BlueConduit. In Figure 4, we depict the hit rate of BlueConduit's baseline model alongside some basic comparison models included as a reference.

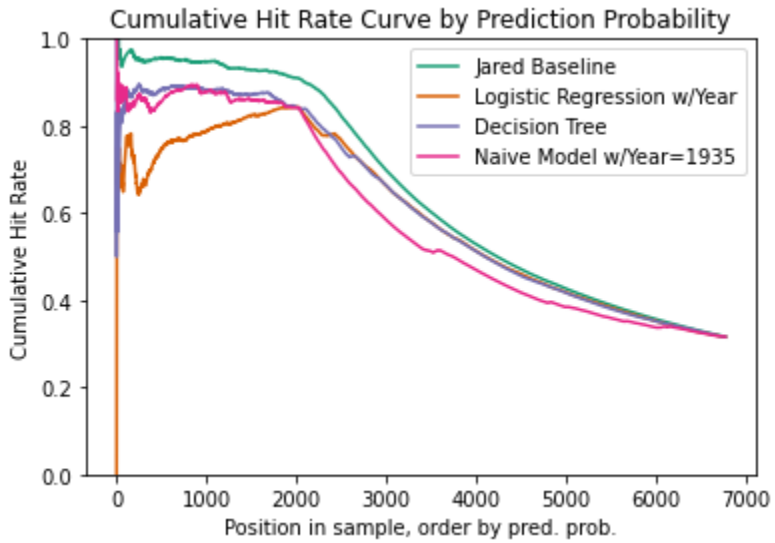


Figure 4. Baseline model and other additional model references.

In Figure 4, the BlueConduit's baseline model is represented by 'Jared Baseline'. Based on our EDA findings, we included reference models trained only on the year predictor ("Logistic Regression w/Year" and "Naive Model w/Year") to get a sense of how

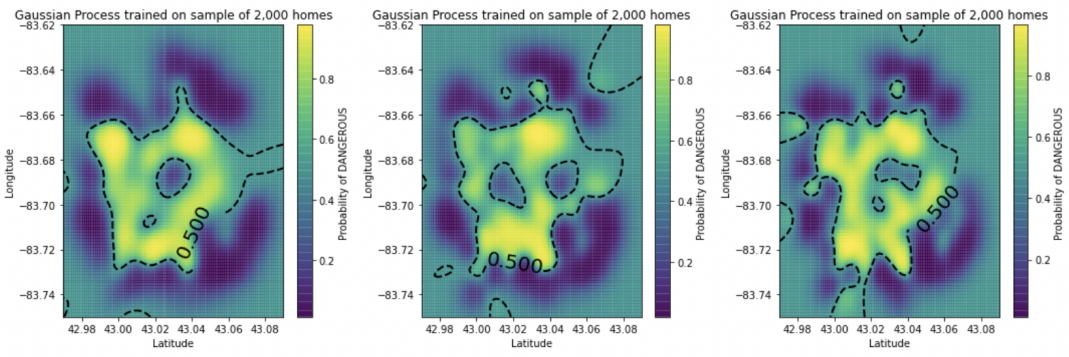
much information is conveyed by the year. Also, we included a simple decision tree as a comparative reference of the XGBoost baseline model.

Developed Model

For a first pass at a spatial model, we chose to use a Gaussian Process to predict the probability a parcel contains dangerous materials, given only latitude/longitude as features. In our future work, we discuss our ideas to use more features to measure the spatial correlation between parcels.

We chose a Gaussian Process for now because it conforms better than simpler models to complicated distributions of data in continuous space. From our EDA we could see the homes with lead are generally clustered in the middle of the city. Our spatial model would need to be able to identify the spread of these regions and construct a set of highly nonlinear decision boundaries to account for the twists and turns of the structure of the underlying streets and neighborhoods. A Gaussian Process is well equipped among spatial models to account for these requirements, but is not the only algorithm we plan to consider.

The biggest drawback of a GP we have encountered is its runtime, which is $O(n^3)$; therefore we have not trained a GP on all n=26k homes. Instead, we trained 100 different GPs on a set of bootstrapped data samples, with n=2,000 homes in each sample. Figure 5a shows to what extent the fit of the GPs varies when changing the particular sample of homes they were trained on, and Figure 5b shows what the average of the 100 GP predictions looks like.



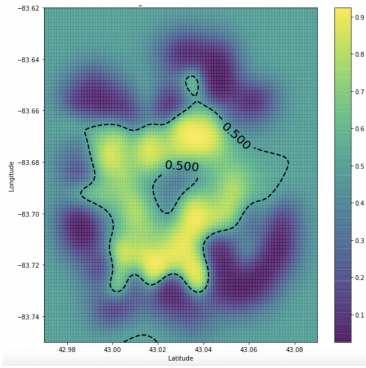


Figure 5a (left three): Three of the 100 trained GPs, plotting the predicted probabilities of lead across the latitude/longitude values spanning Flint, MI. Each of the GPs identifies that lead in Flint is mostly concentrated towards the center of the city.
Figure 5b (right): Predicted probabilities of the AVERAGE of the 100 trained GPs.

$$k_{\mathrm{SE}}(x, x') = \sigma^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$$

Equation 1: Radial Basis Function (RBF) Kernel. After cross validation (part of which is shown below in Figure 6) we found desirable parameters are in the range of sigma in [1e-8,1e-3] and L in [1e-4,1e-2]

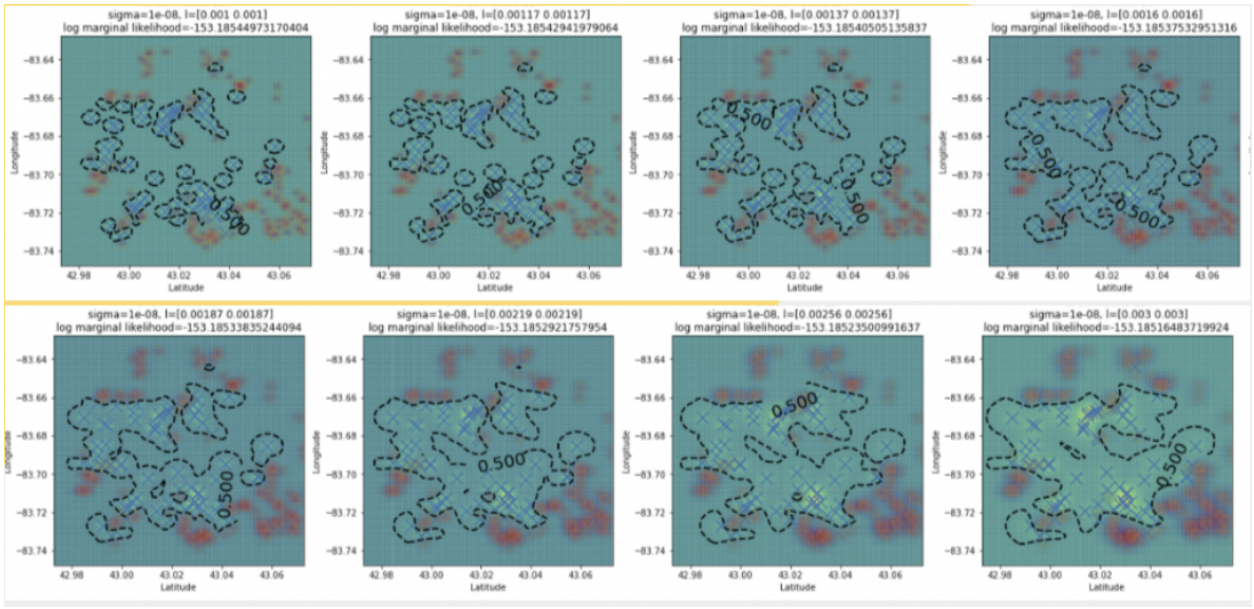


Figure 6: Increasing the L parameter in the RBF kernel broadens the area within which our model predicts lead.

Figure 6 shows how we visually evaluated the performance of the L hyperparameter, which corresponds to the distance within which the GP model would consider homes spatially correlated.

Our next spatial modeling choice will be a meta-modeling architectural decision: namely, the order of the information flow, described as follows:

> Option 1: use spatial model probabilities as extra feature in XGBoost
> Option 2: use XGBoost model probabilities as prior for spatial model
> Option 3: train ensemble model to process both XGBoost results & spatial model results

Results

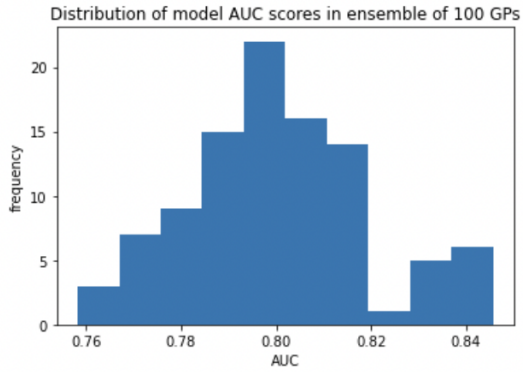


Figure 7: Distribution of the AUC scores of each of the 100 trained GPs, evaluated on the test set of 6,778 homes.

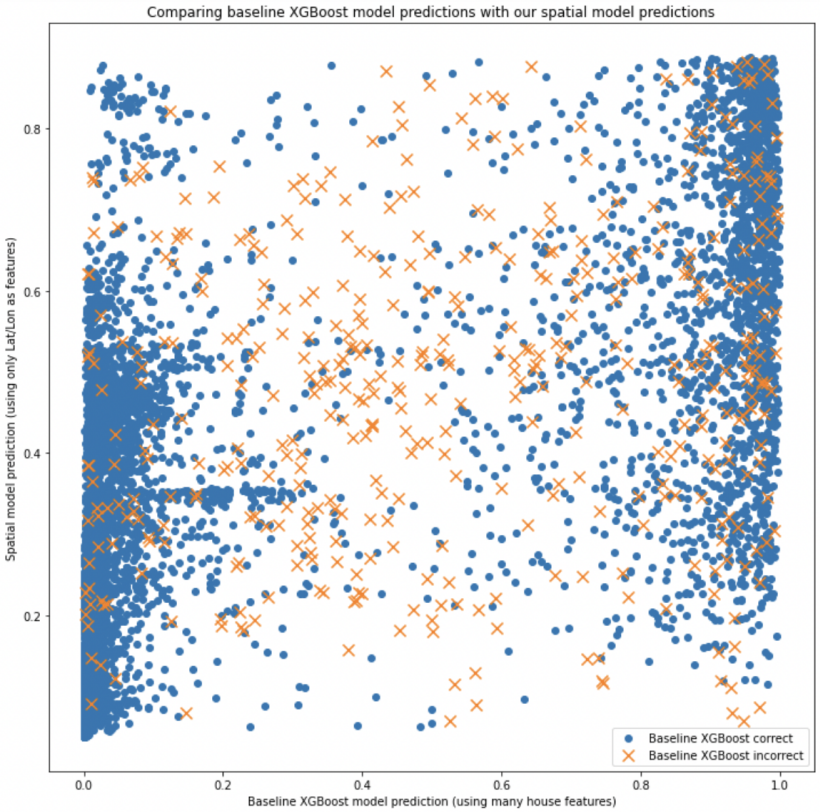


Figure 8: Comparing our GP Ensemble's predicted probabilities with those of the baseline XGBoost model. Points are marked with a blue O if they represent a home that was predicted correctly by the baseline, and marked with an orange X if they represent a home predicted incorrectly by the baseline.

The distribution in Figure 7 shows how the prediction of the 100 different GPs varies by AUC. The average model had AUC 0.8288, which is higher than almost all the individual GPs, **suggesting that combining the results of many instantiations of the spatial model will likely be better than any one instantiation.**

Figure 8 shows the average model plotted against the baseline XGBoost model. This chart shows our model is much more uncertain about many of the homes in the test set, as seen by all the points clustered close to a height of 0.5. The baseline model, on the other hand, is

very confident, with most of the points towards the extremes in the horizontal direction. Notably, however, the orange Xs - the points representing homes the baseline model incorrectly classified - are more distributed throughout the center of the plot, **suggesting we focus our modeling approach on homes for which the baseline is predicting with high uncertainty.**

Future Work

In the next few weeks, we plan to complete several tasks. First, we will compile street distance information, getting the street distances between every parcel in the dataset. Because pipes generally go under streets, we believe that street distances (road/Manhattan distances) are going to be a more powerful indicator of "pipe distance" than pure geographical distance. For example, homes on separate cul-de-sacs that happen to be geographically nearby one another could be very far from one another in terms of streets and pipes. So, we want to make sure our calculations of geospatial information are utilizing street distances.

Figure 9: *We are aiming to calculate street distances (left) rather than geospatial distances (right)*

Once we compile the distances, we will begin fitting our models. We will fit both a Gaussian Process model and an Algorithmic Bias Diffusion model. Both will aim to identify areas with concentrated high probabilities of lead, and then both will "smooth" those probabilities into surrounding homes (i.e. increase the probability of lead for nearby homes). Then, we will see if our added spatial modeling produces a higher hit rate curve than BlueConduit's current model. If we achieve a higher hit rate, we will show proof-of-concept that geospatial information contains some helpful signal for identifying lead. If we do not achieve a higher hit rate, we will show suggestive evidence that geospatial modeling may be an unproductive route for BlueConduit to explore in future work (saving them time and resources).