Predicting Lead Water Service Lines in the Pittsburgh Water and Sewer Authority Service Area

Prepared by

Dr. Michael Blackhurst, PE (TX)
Co-Director, Urban and Regional Analysis
Center for Social and Urban Research
University of Pittsburgh

Dr. Hassan Karimi
Professor, School of Computing and Information
Director, Geoinformatics Lab
University of Pittsburgh

Saeed Hajiseyedjavadi
Doctoral Student, School of Computing and Information
University of Pittsburgh

Executive Summary

In an effort to reduce uncertainty in the locations and counts of its customers' lead service lines, the Pittsburgh Water and Sewer Authority (PWSA) hired researchers at the University of Pittsburgh to estimate the probability that each of its residential customers' public and private water service line is made from lead and predict as lead or non-lead each service line.

PWSA provided observed service line materials for 8,100 of its 70,000 residential customers. We first cross referenced these observations with PWSA's historical records, which are currently used to plan service line replacements and thus serve as a point of comparison for predictive modeling. We then used these 8,100 to predict service line materials for 61,000 PWSA customers, representing 98% of PWSA's residential customers. The remaining 1.6% of customers had too little information to support predictions.

We explored three primary applications of the model: prioritizing an annual cycle of replacements (or excavations); prioritizing a larger number of replacements (or excavations); and estimating inventories by service line material. With respect to estimating inventories by material, predictions demonstrate only marginal value: the model was 73% precise whereas the historical data have been 63% precise to date. The reason for this is that too few predicted probabilities are extreme enough (high or low) to discriminate well lead from non-lead service lines for all customers. As a result, Pitt does not recommend using the current model for estimating inventories of service line by material.

However, there does appear to be value in applying the model to prioritize annual replacements. As part of the 2019 LSLR Program, PWSA primarily used a combination of historical records, curb box inspection, and meter replacements to prioritize replacements at 9,080 customers. However, only 56% of these customers had lead services lines, resulting in costly and unnecessary excavations. In contrast, the model indicates PWSA would find lead 90% of the time if it excavated at the 9,080 customers predicted as most likely to have lead (a mean probability of lead is 90%). The model identifies lead well for these customers because they represent those most likely to be lead.

Identifying lead worsens when applying predictions to more customers, as predicted probabilities decrease and therefore make material discrimination less certain. For example, we applied the model to 35,000 customers where historical data are missing or indicate lead. The mean predicted probability these customers are lead is only 44%. Historical data indicate PWSA would find lead in about 63% of these customers. Here, the model results in only a modest improvement to 66%. As a result, Pitt and PWSA do not support using the current model to plan replacements beyond an immediate planning cycle.

These results suggest reliable model applications are limited to customers where predicted probabilities are extreme (high or low). A field verification of the model conducted since July 2019 confirms these findings. For those customers with a predicted probability of lead greater than 80% and less than 20%, precisions of 80% and 85% were verified for lead and non-lead, respectively. However, the field verification indicates the model performs poorly for more moderate predicted probabilities.

Pitt and PWSA plan to retrain the model with new field data collected in hopes of identifying highly probable lead lines to plan future replacements and expanding applications of the model pending revised performance. To support training and in recognizing that diagnoses can be expensive, Pitt provided an active learning plan to sample at up to 5,000 customers strategically recommended to improve predictions.

In addition, Pitt analyzed correlations between service line materials and counts of customer water samples in which lead was detected. Customers where lead was detected at least twice are highly likely (>98%) to have lead service lines. Customers where lead was not detected at least twice are unlikely (>92%) to have lead service lines. Incentivizing customers to collect multiple samples could serve as a non-intrusive and inexpensive detection method and has the dual benefit of improving future predictive modeling.

In addition to this report, Pitt separately delivered a database and maps of select model results.

1. Introduction

Prior to the U.S. Environmental Protection Agency issuing the Lead and Copper Rule (LCR) in 1991, there was no federal regulatory basis motivating municipalities to maintain records of water service lines. In older cities like Pittsburgh, this has led to over a century of unobserved changes influencing the extents of lead service lines. As a result, many municipalities face considerable uncertainty understanding the potential health risks of lead service lines and deciding what, if anything, should be done to reduce these risks.

Uncertainty in service line materials is particularly problematic when municipalities are required to replace lead service lines in compliance with the LCR. The LCR requires that non-compliant municipalities annually replace 7% of public lead service lines until consecutive samples of lead from customers' taps meet regulatory expectations. Uncertainty in service line materials affects confidence in the overall inventory, thus making it hard to allocate the appropriate resources to meet the 7% replacement rule. Such uncertainty also makes it hard to identify customers with lead service lines, which can lead to costly excavations of lead-free service lines.

Since June 2016, the Pittsburgh Water and Sewer Authority (PWSA) has been under a replacement mandate. Given public concern over potential increases in water concentrations from partially replaced lead service lines, PWSA's board endorsed full service line replacements, offering to replace customers' private lead service line for free where the public side is scheduled for replacement. As such, PWSA needs confident information describing the materials of both public and private service line sections.

In an effort to improve understanding of service line materials in its service area, PWSA started photographing service lines through curb box inspections (CBIs) and digitizing historical records in parallel with replacing service lines. These three sources - photographs of service lines from CBI program, digitized historical records, and service line observations made during replacements - have been primary sources of information guiding PWSA's decision making with respect to lead service lines since 2016. While helpful, applications of this information to date result in costly and unnecessary excavations at homes that ultimately do not have lead service lines. As part of the 2019 LSLR Program, over 44% of excavations were made at homes where the public-side service line material record was not known or lead, but were found to not have a public-side lead service line.

Given these sampling efforts and the underlying spatial, temporal, and demographic trends driving lead service line installations and potential replacements, predictive modeling can help reduce the uncertainty in the locations of lead service lines in PWSA's service area. Previous studies of the water system in Flint, MI have applied machine learning techniques or statistical models to improve upon lead inventories, inventory discovery activities, and inventory replacement decisions (Abernethy et al., 2016; Chojnacki et al., 2017; Abernethy et al., 2018).

In May of 2019, PWSA contracted with the University of Pittsburgh (Pitt) to develop and apply models predicting the locations of its lead service line inventory. The objectives of this work are to provide estimated probabilities that the public and private service line is lead for each of

PWSA's 70,196 residential customers and predict as lead or non-lead the public and private service line materials for these customers. This report summarizes the data, methods, and results associated with this work. A database of the estimated probabilities and material classifications has been separately provided to PWSA.

2. Data Sources, Applications, and Limitations

Table 1 summarizes the project data made available to Pitt as of July 2019, which describe customer locations, records or estimates of service line material, water data, and property and lot characteristics. What is documented in this report are only those data sources containing features selected for modeling. We also explored using child blood lead level and demographics by Census tract in model development, but this information proved to be unhelpful.

Table 1: Summary of data used for predicting service line materials for PWSA's customers.

Data set name	Shorthand label	Brief description	Temporal coverage	Spatial coverage	Count	Use in modeling
Spatial locations of customers (PWSA, 2019b)	LOC	Customer address and location	Received June 2019	All customers	70,196 customers	Scope of customers subject to lead service line replacement decisions; modeling spatial correlation; 6 potential features for training and testing
Lead service line replacement (PWSA, 2019b)	LSLR	Schedule of service line replacements and materials observed	2/28/14 To 06/25/19	58 out of 72 neigh-borhoods	8,891 customers	Training and testing data describing dependent variable
Water service line database (PWSA, 2019b)	WSL	Historical data describing original service	From 1899 to 2017	All neigh-borhoods	69,903 customers	24 potential features for training and testing
Curb box inspections (PWSA, 2019b)	CBI	Service line material diagnosed using pictures taken at the curb box	12/14/2016 to 12/13/2018	60 out of 72 neigh-borhoods	21,577 customers	6 potential features for training and testing
Water sampling (PWSA, 2019b)	TWS	Levels of lead at customers' taps	5/24/17 to 04/24/19	68 out of 72 neigh-borhoods	7,886 customers; 8,505 samples (5,960 non-detect)	2 potential features for training and testing
Property tax assessments	TAX	Descriptions of property use, buildings, and lots	11/12/1827 to 4/18/19	Entire service area	69,319 parcels	86 fields, 42 of which we considered features for training and testing
Meter replacements (PWSA, 2019b)	MET	Private service line material observed at meter	File dated 04/2019	68 out of 72 neigh-borhoods	2,644 customers	Potential future training and testing data as described in narrative

The meter dataset describing private service line materials observed when replacing meters was inconsistent with the material observed during excavation as indicated in the LSLR data. For example, 626 private service lines indicated as lead in the LSLR data were indicated as non-lead in the meter data, and 10 private service lines indicated as lead in the meter data are indicated as non-lead in the LSLR data. PWSA has suggested that these differences are primarily due to differences in the times in which data were collected. As a result, we assume

that the private service line is lead if either the LSLR or meter data indicate so for the purposes of training the model given we seek predictors of lead service lines.

2.1 Water Service Lines

PWSA provided four potential sources of information describing service line material:

1. historical information describing initial service line material in the water service line database;
2. materials diagnosed visually from pictures taken through the curb box in the CBI data set;
3. materials observed during service line replacements in the LSLR data set;
4. and materials observed during meter replacements in the meter data set.

The historical records in the water service line database provide the most coverage (e.g., fewest missing values), recording values of the public service line material, private service line material, installation date, inspection data, and diameter. However, PWSA has not perfectly maintained these historical data. As such, historical material indicators were thought to be too inaccurate to serve as the dependent variable for modeling, and these data were considered only for independent model features.

As part of routine water system maintenance (outside the scope of the lead service line replacement program), PWSA has replaced publicly owned lead service lines and coded the historical materials at these locations as “non-lead.” Pitt therefore removed these locations from the testing data, as they do not need a prediction.

Of the approximately 21,577 locations for which a CBI was attempted, public and private materials were visually estimated for only 26.8% and 24.9% customers, respectively, due to an inability to locate the curb box, observe the service line, or visually diagnose the material. By cross-referencing the material diagnosed through the curb box with that observed when replacing service lines, PWSA estimates that CBI diagnoses lead at 97% true positive rate and non-lead at 72% true negative rate (PWSA, 2019a). While the CBI program was designed to be spatially representative, PWSA and Pitt felt the limitations of the CBI data were too significant to be used as the dependent variable for modeling. The CBI material diagnostics were considered as a potential independent model feature. Given the LSLR data are confident visual observations of the material, we elected to use as the dependent (predicted) variable the material observed during service line replacements (the LSLR data set).

The CBI data include binary indicators of lead or non-lead. The historical and excavation information specify the non-lead material (e.g., copper). We mapped the material indicators to binary values of 1 and 0 corresponding to “lead” and “non-lead” values.

2.2 Customer Water Lead Concentrations

PWSA provides kits for customers to collect samples of water from their household plumbing. Samples are collected for either compliance with state and federal regulations, at the request of

customers interested in lead concentrations in their drinking water, or to characterize the impact of lead service line replacements.

As of October 2019, PWSA provided three water data sets: customer requested (CR) samples from October 2013 through July 2017; CR samples from January 2017 to September 2019; and samples collected as part of lead service line replacements (LSLR) completed or planned from January 2018 to January 2019. Each of these data sets includes a unique customer identifier, customer address, select dates describing the administration and custody chain of sampling, the lab performing the analysis, and a reported lead concentration.

Given that research suggests highly variable lead concentrations immediately following a lead service line replacement, Pitt removed 8,326 LSLR samples associated with monitoring lead levels after a lead service line was replaced. We also removed 5,893 CR samples that were either cancelled, still at the lab for processing, not returned to PWSA by the customer, or otherwise flagged by PWSA or the lab as potentially problematic.

The raw data include mixed representation of equipment detection limits. Where samples fell below the detection limit, the raw data may report the detection limit, a reading of zero, a “ND” for non-detection, or a lead concentration below the detection limit. In addition to different reporting practices, the detection limits for the two labs PWSA employs for certified sampling have different detection limits or have modified their limits over the sampling period. The lab ALS reports a detection limit of 2 parts per billion (ppb), and the lab CWM modified their detection limit from 4 ppb to 2 ppb on January 22, 2019. Pitt applied the following criteria to change the reported concentration to the detection limit where the sample is expected to fall below the detection limit.

The detection limit for the labs analyzing the older CR samples (prior to July 2017) are not reported. Thus, Pitt assumes a conservatively high detection limit of 4 ppb for these samples (n = 5,493). A total of 3,594 CR and LSLR samples report the lab and chain of custody dates. For these samples, Pitt assigned the detection limit as described above. If the date at which the sample was analyzed was missing, Pitt assumed the analysis took place 6 days after the lab received the sample, a lag estimated from those samples reporting full chain of custody dates.

Where the lab analyzing the sample and analysis dates are both missing (n = 4,183), Pitt assumed detection limits of 2 ppb prior to January 22, 2019 and 4 ppb after the earliest date of record associated with the sample. Events used to determine the earliest date of record include the customer request date, all mailing dates for shipping and receiving the sample kit and sample, the date at which the sample was collected, and any dates associated with correspondence from PWSA to customers.

In the raw data, the same customer identifier can be associated with multiple samples reporting the same lead concentration on at least one duplicate date field. In many cases, these repeated samples reflect multiple valid samples taken at different locations downstream of the same service line, such as in a multi-family unit. However, Pitt removed 258 samples that appear to be genuinely duplicated as defined by repeated values for the customer identifier, the address, the

lead concentration after correcting for detection limits, and any of the date fields. For actual multiple samples from the same customer, Pitt used the maximum concentration for machine modeling and counts of samples in which lead was detected for linear modeling.

The final sample of water data consists of 18,661 samples for which 4,662 customer identifiers and 10 addresses are missing. Of these samples, 13,574 report lead levels below detection limits.

2.3 Property Assessment Data

Allegheny County property assessments include 86 fields describing information used for the purposes of taxing real estate. Exemplary information includes property use (e.g., single family residential), age, floor space (for residential properties only), assessed value, building quality, and lot size. After merging this dataset with the customers dataset, the portion of the missing values of property assessment fields ranges between less than 1% to about 7%.

3. Sample Preparation

3.1 Joining Data

Data provided by PWSA are observed by customer, whereas the property tax assessment data are prepared by parcel. There are several reasons why these units of observations may not align. Approximately 1,000 PWSA customers are multiple addresses served by the same water service line, a situation referred to as a “party line.” In Allegheny County, multiple parcels may be associated with the same billing address, even if not served by a “party line.” Finally, historical multi-parcel sales may introduce discrepancies between the official addresses on record at the County and PWSA.

PWSA previously hired a consultant to match 69,203 (98.5%) of their customers subject to LCR compliance with a County tax assessment record. We were able to match an additional 116 customers to County tax assessment records by address strings. The remaining 877 customers (1.2%) are unmatched to a County record.

3.2 Handling Missing Values

Figure 1 summarizes the missing values in model features. The property development decisions influencing initial service line installation and potential replacements are expected to demonstrate spatial correlation. Table 2 indicates a positive spatial autocorrelation in the spatial structure of the CBI data as measured by Moran’s I statistics with queen contiguity-based spatial weights (Moran, 1950). Based upon the results in Table 2, we used Kriging technique (Matheron, 1973) to spatially estimate the missing values for the year of construction and CBI public and private material.

Figure 2 demonstrates the estimation of the missing values of CBI data for the public service line, where values closer to 0 and 1 indicates higher and lower probability of lead in the material of the public side of the service line. Other features were missing for less than 2% of the observations. For these continuous and discrete values, we applied the median and mode, respectively.

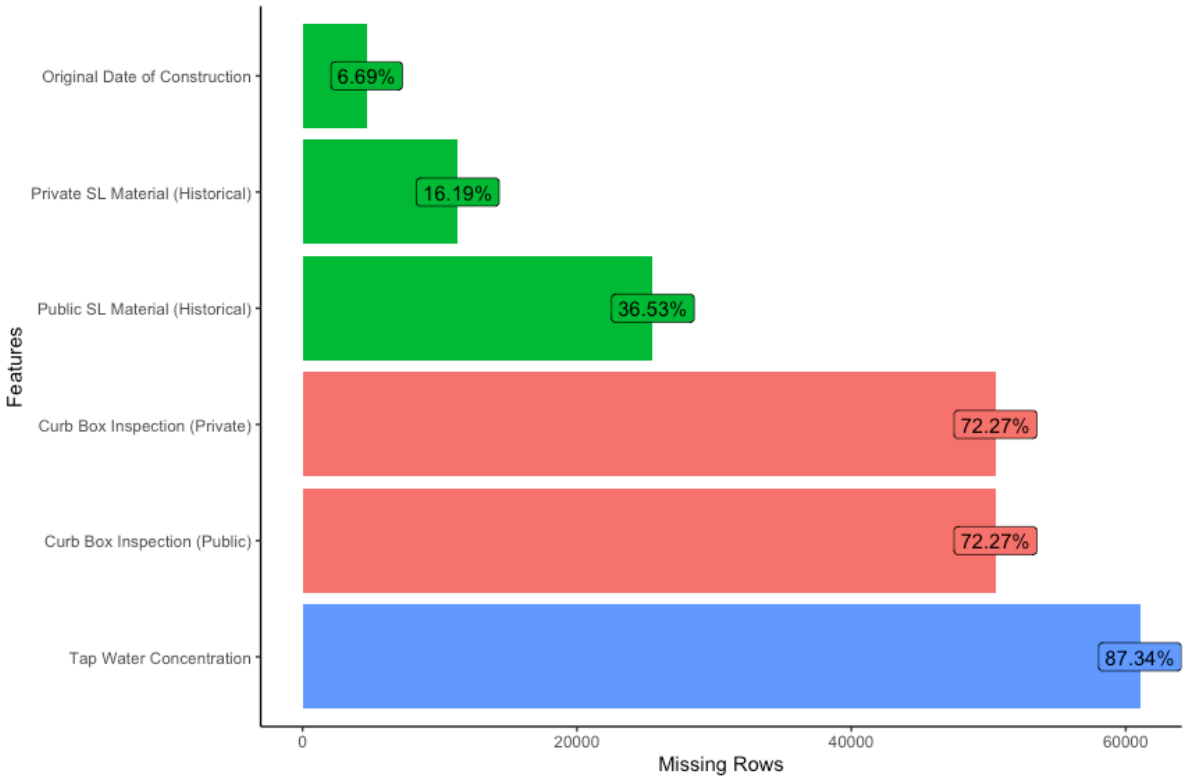

Figure 1. Missing values for model features considered for predicting the service line material for PWSA customers. Features missing less than 5% of their values not shown for clarity.

Table 2: Moran's I calculated by Monte Carlo simulations over a thousand permutations for three independent variables.

Feature	Moran's I	P-value
CBI (public side)	0.124	< 0.001
CBI (private side)	0.122	< 0.001
Origin year of the construction	0.667	< 0.001
Water samples	0.081	<0.001

3.3 Feature Selection

Starting with the potential predictive features summarized in Table 1, we use Recursive Feature Elimination (RFE) (Guyon et al., 2002) as an effective feature selection technique that at each step recursively eliminates feature(s) with the weakest predictive power until no feature is left. The importance of the predictors is calculated at each iteration so that eventually, RFE selects and returns a subset of features with the highest predictive power. We paired RFE with random forest (Tin Kam Ho, 1995) to reduce our feature space from more than 88 to 24 features to be considered for both predicting lead public and private service lines. Figures 3a and 3b show the selected features selected for public and private service lines, respectively.

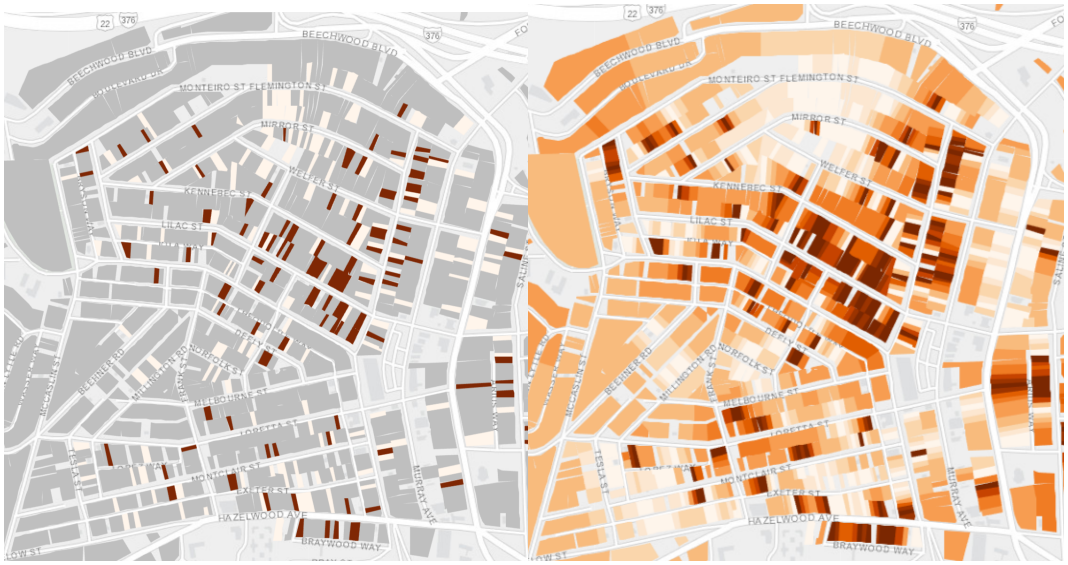

Figure 2: An example of interpolating missing curb box inspection results using inverse distance weighting methods with a power of 2. Training data are on the left and interpolations are on the right. Results are shown for the public portion of water service lines. Darker and lighter coloring indicates a higher and lower probability of lead, respectively.

3.4 Final Sample

Our training sample consists of 8,142 observations of public and private service line materials. Importantly, our sample reflects conditions prior to February 2016, which is when PWSA's records of lead service line replacement began. This approach preserves as many original lead service lines as possible so that the modeling reflects the underlying data generating processes influencing the materials in service.

3.5 Sample Weighting

Most of the training data are from PWSA's lead service line replacement (LSLR) program. This program prioritizes neighborhoods where PWSA's historical data indicate lead is present and does not excavate at locations that have a historical record of non-lead, suggesting potential sampling or spatial biases. As a result, the distribution of predictive features in the training data differ from the respective distribution in the test sample, a situation referred to as *covariate shift*. Figure 4 shows an example of covariate shift for the historically recorded materials of public service lines. Covariate shift can compromise model accuracy.

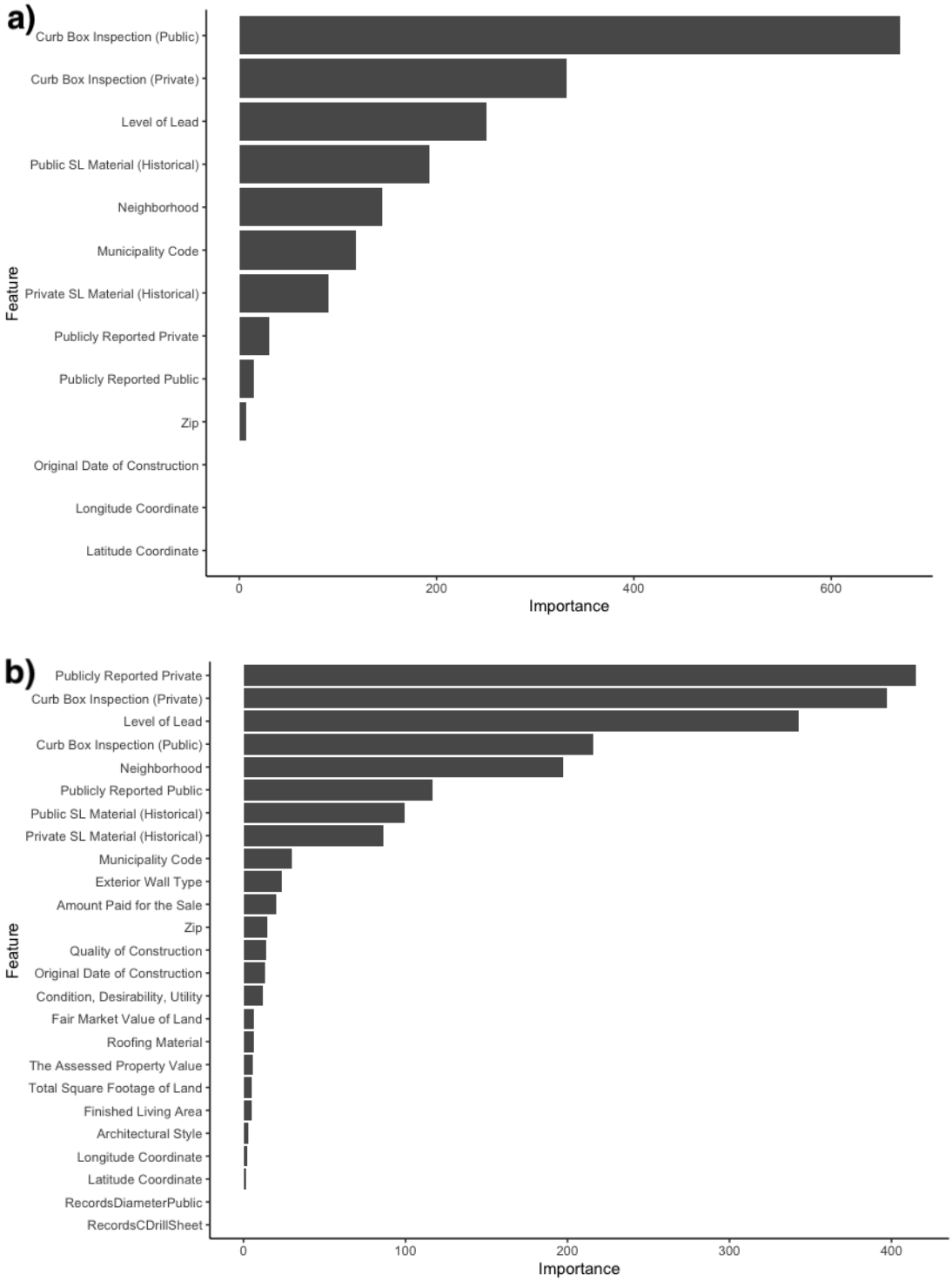

Figure 3. Feature importance scores for predicting public (a.) and private (b.) lead service lines.

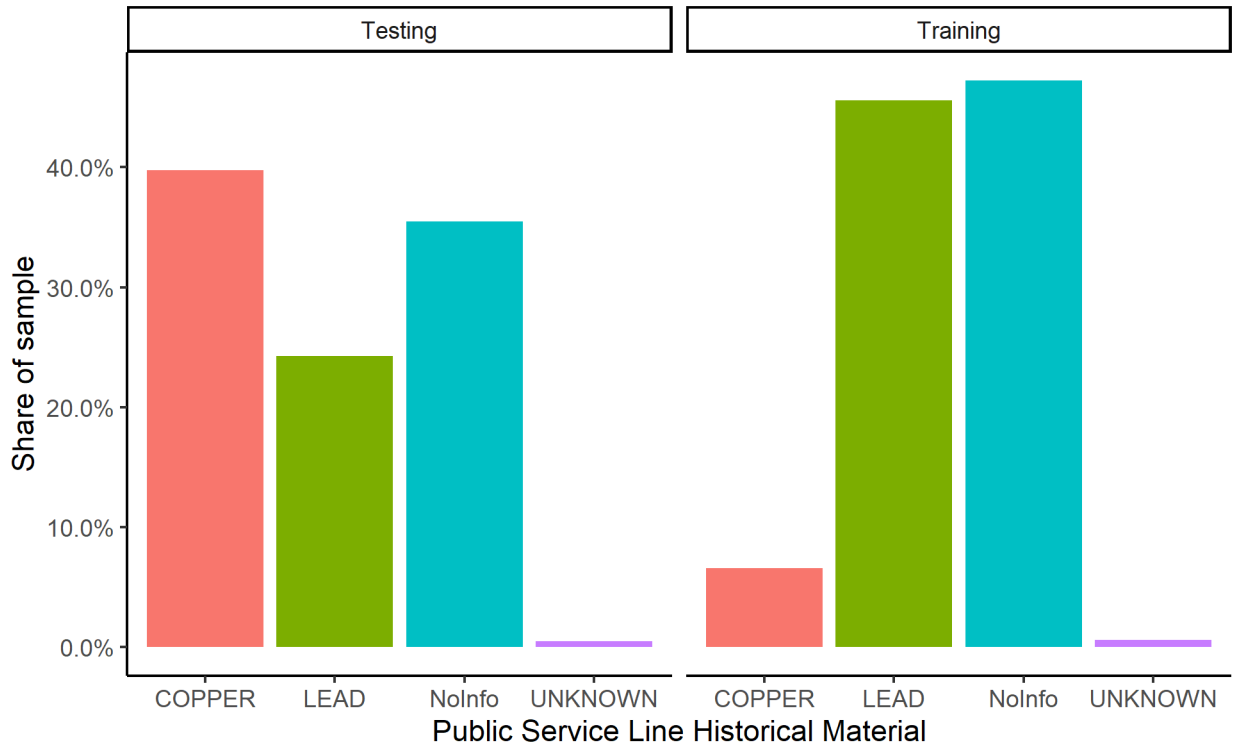

Figure 4: An example demonstrating differences in the distribution of the historically recorded materials for public service lines. Severe differences in the distribution of predictive features can cause covariate shift and compromise predictive accuracy.

Figure 5 collapses all potential predictive features into two dimensions using t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize potential distributional differences between the testing and training samples. Figure 5.a shows the instances of training and test sets embedded into two-dimensional space using the t-SNE algorithm. Figure 5.a shows a high level of distinction between the training and the prediction set, suggesting an imbalance in the distribution of the predictive features such that covariate shift is likely. Similar distributions of the predictive features between the training and testing data would appear as shown in Figure 5.b.

To alleviate the effect of covariate shift, the training data can be weighted (with values between 0 and 1) so that the distribution of training instances is more aligned with those in the testing set, where higher weights indicate more alignment. We used a *density ratio estimation* (DRE) technique called unconstrained *Least-Squares Importance Fitting* or "uLSIF" (Kanamori, Hido, and Sugiyama, 2009) to weight the training data.

4. Machine Learning Methods

Using the 8,142 customers with material diagnoses as training data, Pitt developed predictions of public and private service line materials for 60,941 PWSA customers. Our predictions are two-fold. First, we estimate the probability that any given customer's public or private service line is lead. Second, we apply a probability threshold to discriminate as either lead or non-lead each service line section. While our predictive model is relevant to all of PWSA's 70,196 residential

customers eligible for the service line replacement program, we were unable to develop predictions for approximately 1.6% of these customers due the degree of missing features describing these customers.

It should be noted that - in addition to the modeling described below - we attempted to develop deep learning models of service line materials in PWSA's system. However, the sample collected to date were insufficient for deep learning methods.

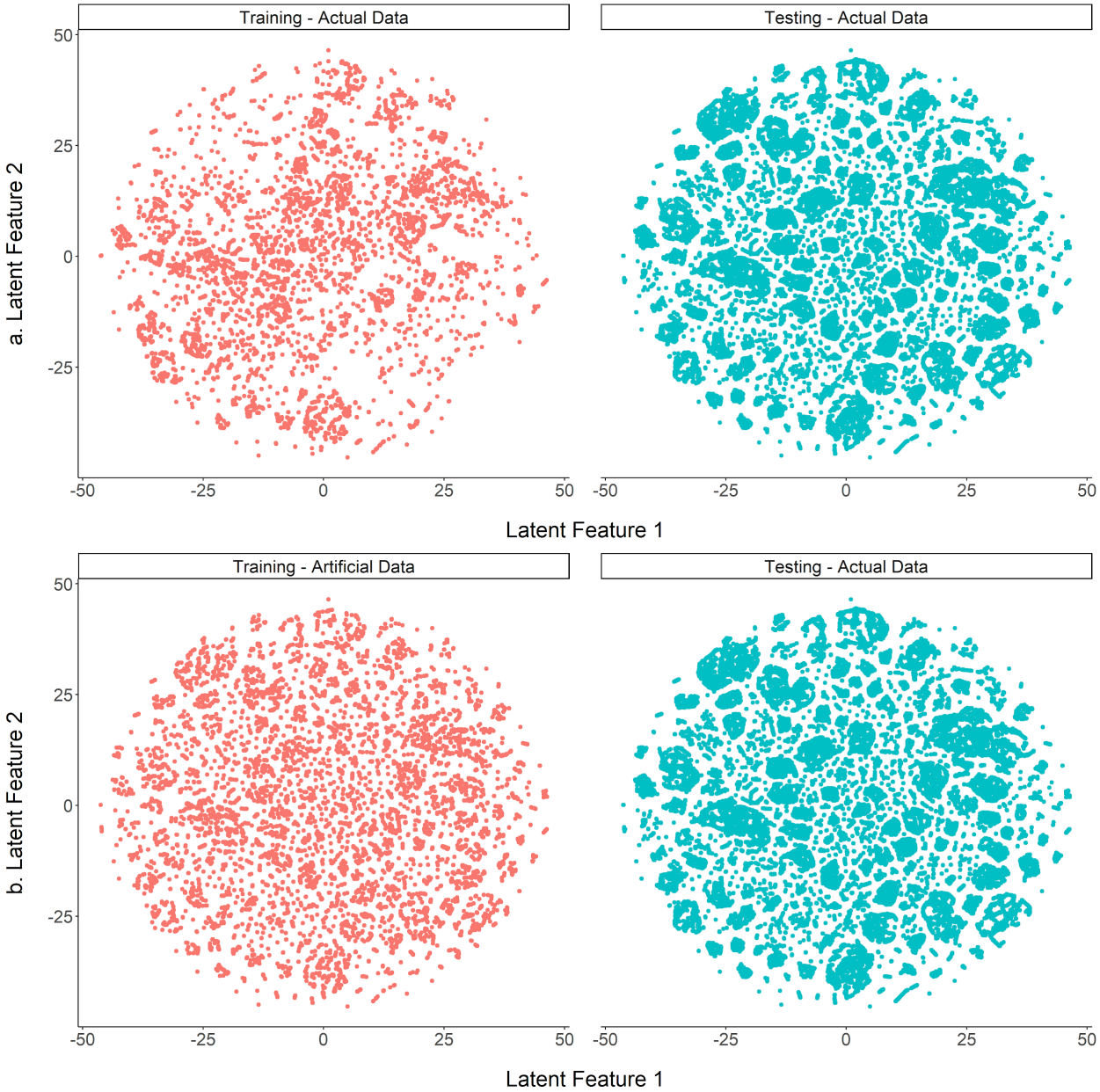

Figure 5: Model features were collapsed into two dimensions (latent feature 1 and latent feature 2) using t-Distributed Stochastic Neighbor Embedding (t-SNE) to explore the potential for covariate shift introduced by differences in the distributions of feature values across the training and testing data. The upper figure (a) demonstrate significant differences in the features, whereas the lower figure (b) does not.

4.1 Spatial Cross Validation

To evaluate model performance, we use the well-known statistical resampling procedure known as K-fold cross validation. In this procedure, the training set is split randomly into K folds (aka, subset or partitions of the training sample) in the way that each fold is put aside for validation, and the model is learned using the remaining K-1 folds. The process is repeated K times so that every fold is used once as the validation set.

Traditional cross validation would randomly select observations for training and validation as demonstrated in Figure 6.a. However, this approach does not account for spatial trends, falsely representing the training and testing samples as independent (Miller, 2004). Traditional cross validation as applied to spatial data can lead to overestimation (Roberts et al., 2017; Brenning, 2012). These issues are particularly problematic when predictions are developed outside of the geographies represented by the training data, which is PWSA's intended use.

To produce more robust and realistic predictions, we employ spatial cross validation. Spatial cross validation partitions samples based on their geographical coordination to reduce the spatial dependence of training and test samples. We first cluster the data using K-means, then treat each cluster as fold in cross validation. Figure 6.b shows the spatial clustering of the PWSA sample.

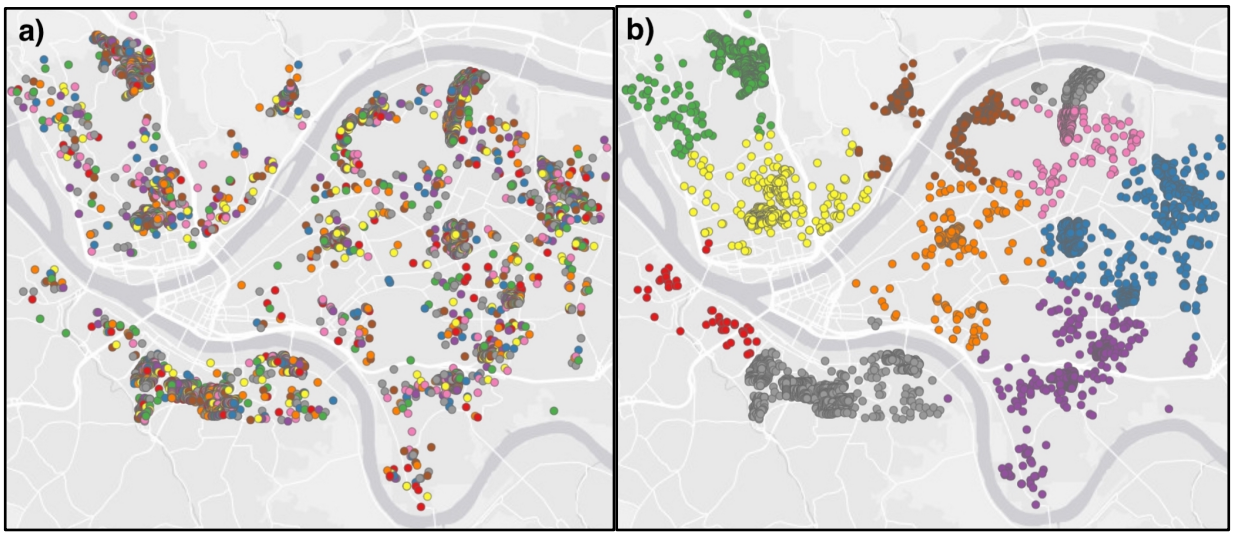

Figure 6. PWSA's training data sampled a) randomly and b) in clusters.

4.2 Model Selection

Pitt tested different classification methods for their performance in discriminating lead from non-lead service lines in the training data. Methods tested include random forest, gradient boosting machine (GBM), deep learning, support vector machine (SVM), logistic regression, and K nearest neighbors (KNN). GBM performed best using the area under the receiving operator curve (AUROC) via a grid search. The AUROC values for the GBM model are 82.9% and 81.9% for public and private service lines, respectively. The selected GBM model hyperparameters are summarized in Table 3.

Table 3: The hyperparameters associated with the selected GBM model tuned via grid search.

Hyperparameter	Public Side Value	Private Side Value
Depth	6	8
Number of trees	90	135
Shrinkage	0.1	0.1

4.3 Probabilistic Predictions

Figure 7 shows the distribution of predicted probabilities that private and public service lines are lead PWSA customers in the prediction sample (n = 60,941). As expected, the predicted probabilities are lower for those observed as non-lead and higher for those observed as lead. Distributions for the unobserved data - which largely skew low - confirm PWSA's efforts to target neighborhoods expected to have relatively high shares of lead service lines. However, the predictions for those customers for which private service line materials are unobserved are bimodal, indicating some areas partly unexplored have high counts of private lead service lines.

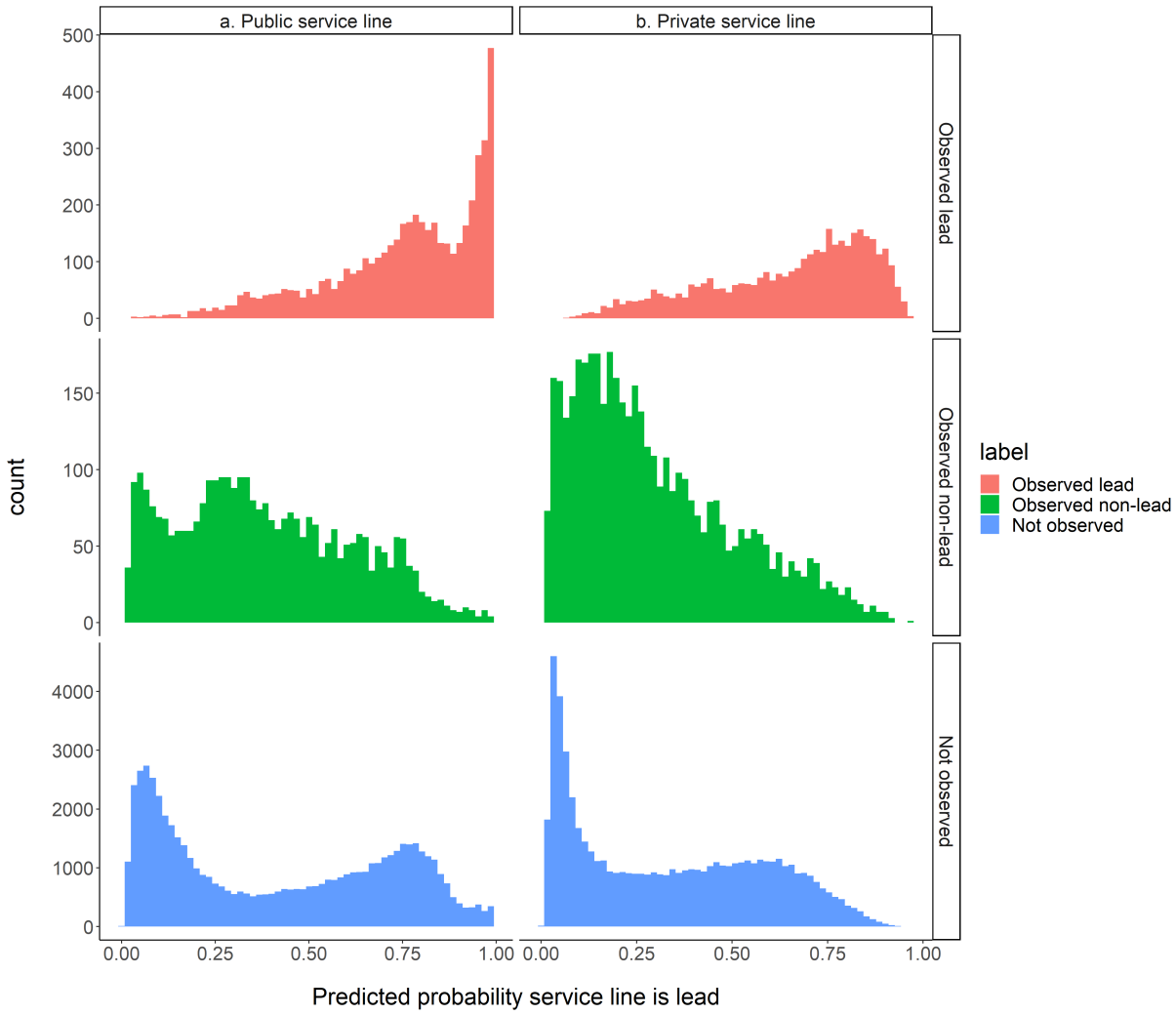

Figure 7. The distribution of predicted probabilities that PWSA customers' private and public water service lines are made from lead.

4.4 Binary Discrimination

In order to derive binary predictions from predicted probabilities, a probability threshold is applied to differentiate positive binary predictions (e.g., a lead service line) from negative predictions (e.g., a non-lead service line). Table 4 shows a hypothetical example of predicted probabilities and observed materials for 10 homes starting and ending at 101 and 110 Main St. Assuming a probability threshold of 60%, the hypothetical model would predict 5 homes have lead (103 - 107 Main St.) and the remaining 5 homes do not have lead (non-lead). As indicated in Table 4, assuming a probability threshold of 60% to assign lead and non-lead predictions to each property is not always accurate. For example, material discriminations for house numbers 101, 104, and 110 are not incorrect. Thus, we need ways to assess the accuracy of binary discrimination assuming different probability thresholds.

Table 4: Hypothetical service line observations and predicted probabilities for 10 homes.

			Threshold of 10%		Threshold of 60%		Threshold of 80%	
Address	Observed Material	Predicted prob. lead	Classification	Result	Classification	Result	Classification	Result
101 Main St	**Lead**	**20%**	**Lead**	**TP**	**Non-lead**	**FN**	**Non-lead**	**FN**
102 Main St	Non-lead	20%	Lead	FP	Non-lead	TN	Non-lead	TN
103 Main St	**Lead**	**62%**	**Lead**	**TP**	**Lead**	**TP**	**Non-lead**	**FN**
104 Main St	Non-lead	65%	Lead	FP	Lead	FP	Non-lead	TN
105 Main St	**Lead**	**71%**	**Lead**	**TP**	**Lead**	**TP**	**Non-lead**	**FN**
106 Main St	**Lead**	**90%**	**Lead**	**TP**	**Lead**	**TP**	**Lead**	**TP**
107 Main St	**Lead**	**80%**	**Lead**	**TP**	**Lead**	**TP**	**Non-lead**	**FN**
108 Main St	Non-lead	60%	Lead	FP	Non-lead	TN	Non-lead	TN
109 Main St	Non-lead	10%	Non-lead	TN	Non-lead	TN	Non-lead	TN
110 Main St	Lead	30%	Lead	TP	Non-lead	FN	Non-lead	FN
			Recall	100%		67%		17%
			Precision	67%		80%		100%
			Accuracy	70%		70%		50%
			F1	80%		73%		29%

The frequency of correct - and therefore incorrect - predictions are how binary predictions are evaluated for accuracy. Figure 8 summarizes the four possible outcomes associated with the correctness of any given binary prediction. True positives and true negatives are correct predictions, and false positives and false negatives are incorrect predictions.

	Positive observation	Negative observation
Positive prediction	True positives (TP)	False positives (FP) *Type 1 error*
Negative prediction	False negatives (FN) *Type 2 error*	True negatives (TN)

Figure 8: Confusion Matrix showing the possible combinations of predictions and observations.

The measures of correctness in Figure 8 are used to define discrimination performance. Common measures of discrimination performance are summarized in Table 5. Drawing on the example in Table 4, applying a probability threshold of 60%, the *recall* would be 4 true positives / (4 true positives + 2 false negatives) = 67%. As shown in Table 4, assuming a lower threshold of 10% captures all the positive observations (i.e., correctly predicts lead) but introduces a lot of false positives (i.e., incorrectly predicts lead). This has the effect of increasing *recall* from 67% to 100% but decreasing the *precision* from 80% to 67%. Conversely, increasing the threshold to 80% captures all of the negative observations but introduces a lot of false negatives.

Table 5: Common measures of binary discrimination performance.

Name	Equation	Interpretation
Precision	TP / (TP + FP) or TP/all positive predictions	Share of correct positive predictions relative to all positive predictions
Recall (or Sensitivity)	TP / (TP + FN) or TP/positive observations	True positive rate
Accuracy	(TN + TP) / (TP + TP + FN + TN)	Share of correct predictions relative to all observations
F-scores	$F = (1+B^2)(Recall*Precision) / (B^2 Recall + Precision)$ With B = 1 F1 = 2(Recall*Precision) / (Recall + Precision)	Harmonic mean of recall and precision (F1)

The example in Table 4 demonstrates the common challenge of selecting probability thresholds that balance correctly predicted positive and negative predictions. Low probability thresholds (i.e., probably that a service line is lead) will capture more true positive observations (i.e., observed lead) but also increase false positives. In practice, a low probability threshold will lead to conservatively high estimates of customers eligible for a service line replacement, which better ensures lead service lines are replaced but increases excavations at customers without lead service lines. In this case, the precision (or the probability of finding lead when excavating) will be low, but relatively few lead service lines will be missed. Conversely, higher probability thresholds will lead to conservatively low counts of customers eligible for a replacement, but the resulting precision - the rate of finding lead when excavating - will be higher. However, higher thresholds narrow those customers eligible for a replacement to those most likely to have lead service lines, and, as a result, increase the likelihood that lead service lines remain in service.

These trade-offs are also demonstrated in the hypothetical distributions of predicted probabilities shown in Figure 9. Figure 9.a. shows hypothetical probability distributions that perfectly discriminate between lead and non-lead. In contrast, Figure 9.b. includes regions were the probability distributions for each material overlap, representing areas of high discrimination uncertainty. In other words, in this shared region of predicted probability, the predictions do not discriminate well between lead and non-lead relative to those predictions near the extremes of 0 and 1. The assumed threshold determines the balance of how these uncertain predictions are either cast as false negatives or false positives. While the total false predictions remain the same, the count of false positives decreases as the assumed threshold increases. As a result, a higher threshold would result in a better precision, as the share of correct predictions of lead increases as the threshold increases. Conversely, as the assumed threshold decreases, the count of false negatives decreases, but the count of false positives increases.

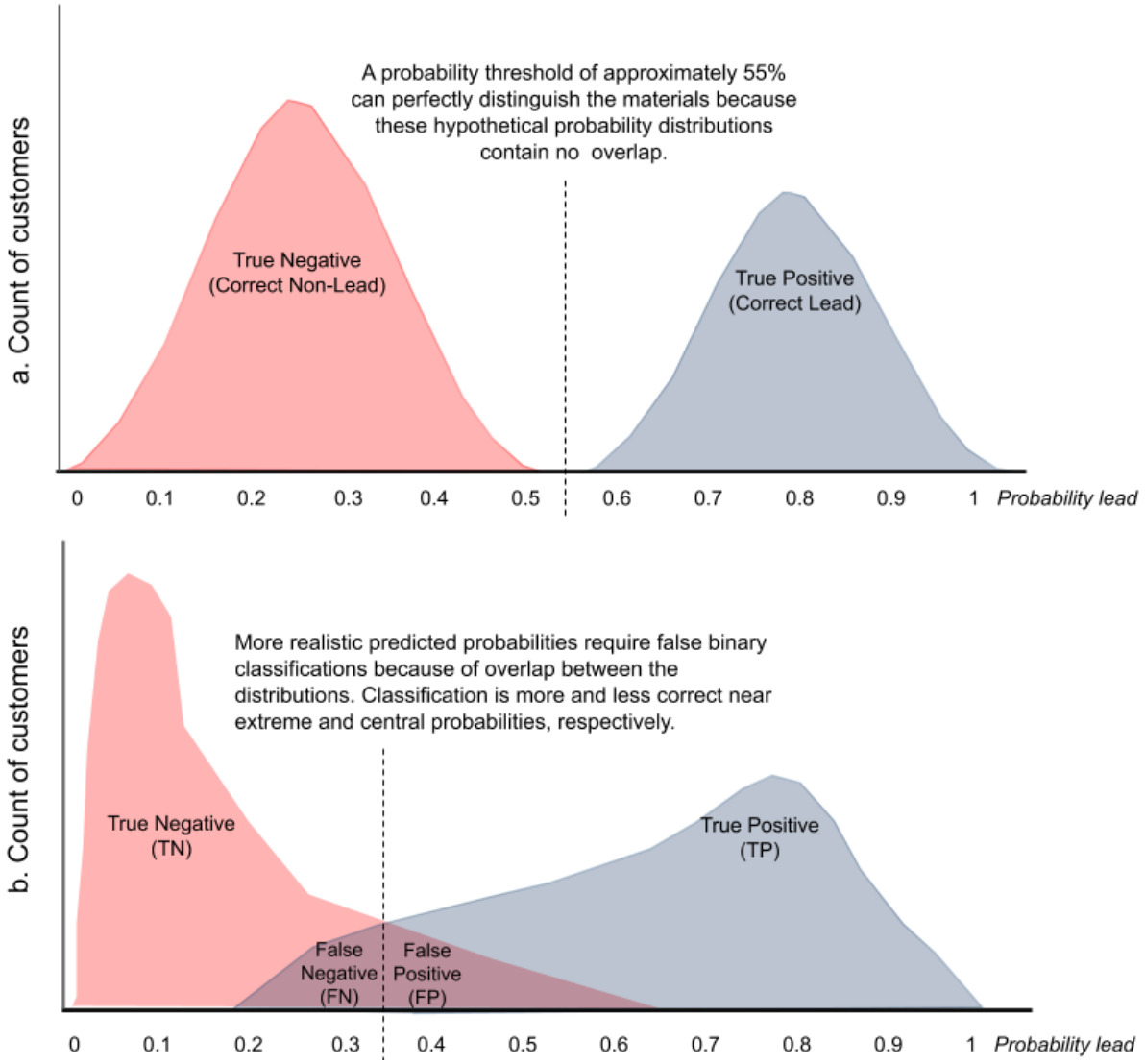

Figure 9: Distributions of predicted probabilities for hypothetical sample of service lines. Panel a. shows a perfect discrimination. Panel b. shows a more realistic situation that requires false classifications, where selected probability thresholds are selected to balance true positives and negatives.

There are no prescriptions for choosing a binary classification technique and respective means for evaluating model performance. Judgment regarding intended model applications is needed, with particular sensitivity to relative differences in the consequences of type I or type II errors. For example, estimating a lead service line inventory should balance predictions of lead and non-lead service lines (balance type I and II errors). Probability thresholds at maximum model *accuracy* or *F1* score are intended to achieve this balance. *Accuracy* is the simpler of these two, measuring the share of total correct predictions (positive and negative) relative to all observations. However, *accuracy* is a poor performance measure of unbalanced data. For example, say 95 out of 100 homes were free of lead service lines. A model that predicted all 100 homes are free of lead service lines would have an accuracy of (95 + 0)/100 or 95%. This model appears to perform well but is biased towards the most frequent observation (lead free service lines) and did not correctly predict any homes with lead service lines. Alternatively, the *F1* score (the F score in Table 5 with B = 1) is appropriate for model applications that involve large pools of customers (e.g., estimating PWSA's inventory by material), as ongoing replacements will lead to an increasingly unbalanced sample (fewer lead service lines and more non-lead service lines). Importantly, appropriate uses of thresholds do not imply the model will perform well.

Different assumptions about the term B in F-score provide a means to explicitly favor type 1 or type 2 errors. Selecting probability thresholds at maximum model F2 score (B = 2) would favor a low type 2 error (or a low false negative rate). This approach would miss few lead service lines but at the expense of poor precision, meaning finding non-lead relatively frequently when excavating. In contrast, selecting a probability threshold at maximum F0.5 (B = 0.4) score would increase precision when excavating but at the expense of leaving lead lines in service.

For this study, Pitt presents model precision, recall, F1, and F0.5 scores for different probability thresholds that represent different model applications. This provides PWSA with some flexibility in applying model results. For example, should PWSA want to use the results when estimating an inventory, we would recommend choosing a threshold at the maximum F1 score to balance false positives and negatives. Should PWSA prefer to prioritize finding lead when excavating (a low false positive rate), we would recommend choosing a higher probability threshold, such as the threshold associated with F0.5. These different model applications are demonstrated in Section 6. As such, Pitt provides customer level prediction for both F1 and F0.5.

5. Summary of Existing PWSA Programming

PWSA currently uses their historical records (WSL) to prioritize service line replacements. All customers except those with historical indicators of non-lead are eligible for a replacement. Table 6 cross references these service lines with those observed when replacing service lines. Table 6 indicates that PWSA should expect a precision – a true positive rate - of around 63% when excavating at homes with a historical label of either LEAD, UNKNOWN, or missing records (e,g., NoInfo), and that there are approximately 35,000 customers with these labels. As of the end of 2019, PWSA excavated services lines at 9,080 locations with a lead or unknown historic record. As part of this work, PWSA observed non-lead during approximately 44% of excavations.

Table 6: The performance of PWSA's historical public service line material records through July 2019.

Historical values assumed lead	n	Precision: TP/(TP + FP)	Recall: TP/(TP+FN)	F1 (see Table 5)	F0.5 (see Table 5)	Find Lead	Miss Lead
Lead	14,347	70.8%	38.8%	50.2%	60.8%	10,164	16,000
Lead, Missing Value*	***35,319***	***63.0%***	***85.2%***	***72.4%***	***66.5%***	***22,283***	***3,881***
All	60,941	42.9%	100%	60.10%	48.50%	26,164	0
* Missing values include both missing labels and a label of "UNKNOWN"							

6. Model Applications

We demonstrate three primary applications of the model:

1) prioritizing an annual cycle of replacements (or excavations) with the goal of reducing unnecessary excavations in the short-run;
2) prioritizing a larger number of replacements (or excavations) for longer term planning;
3) and estimating inventories of service line material for the entire service area.

These applications differ in the threshold assumed to distinguish lead from non-lead. For example, prioritizing replacements involves summarizing model performance for a pool of customers most likely to be lead. Here, the probability threshold discriminating lead from non-lead is implied by the count of customers targeted for replacement. In contrast, applying the model to estimate complete material inventories involves selecting a threshold that balances true (or false) positive and negative predictions. In each application, we compare the model results to either a precision of 63% (observed excavating at homes where historical data is either missing or indicate lead) or the rate at which PWSA found non-lead during excavations in 2019, which was 44%.

Each application is focused exclusively on the publicly owned service lines and draws on model results summarized in Figure 10. However, similar applications to privately owned service lines could readily be made by drawing parallel results summarized in Figure 11.

6.1 Prioritizing an Annual Cycle of Replacements

PWSA excavated 9,080 locations by the end of 2019. Non-lead service lines were observed at approximately 4,000 (44%) of these locations, resulting in costly and unnecessary excavations. In contrast, by excavating at those 9,080 customers where lead is predicted as most probable (a mean probability of lead of 90%), PWSA could improve precision to approximately 90%, which would significantly reduce unnecessary excavations. Readers can verify model performance by examining Figure 10. Figure 10.b shows counts of customers for different assumed probability thresholds. Applying the model to the 9,080 customers most likely to be lead (the 9,080 customers furthest to the right) would result in a precision of 90%, as indicated in Figure 10.a.

Excavating at these 9,080 customers, the model indicates that only about 910 (approximately 10% of 9,080) service lines would be non-lead. In this example, the model therefore could reduce unnecessary excavations by 3,090 customers. At an estimated cost of $1,300 per excavation, the savings from applying the model in this manner are nearly $4,000,000.

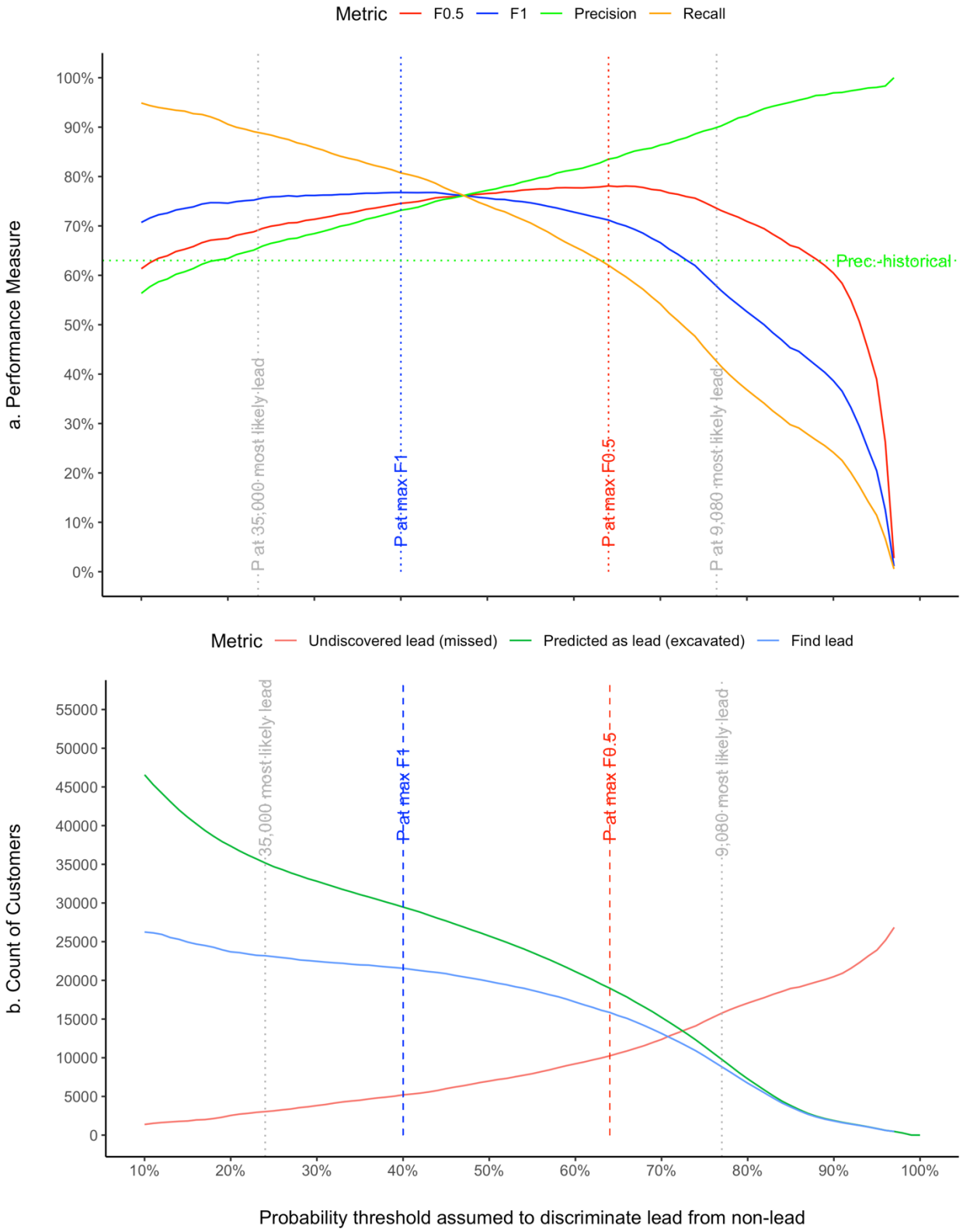

Figure 10: a. Model diagnostics and b. predicted customer counts assuming different probability thresholds for public service lines.

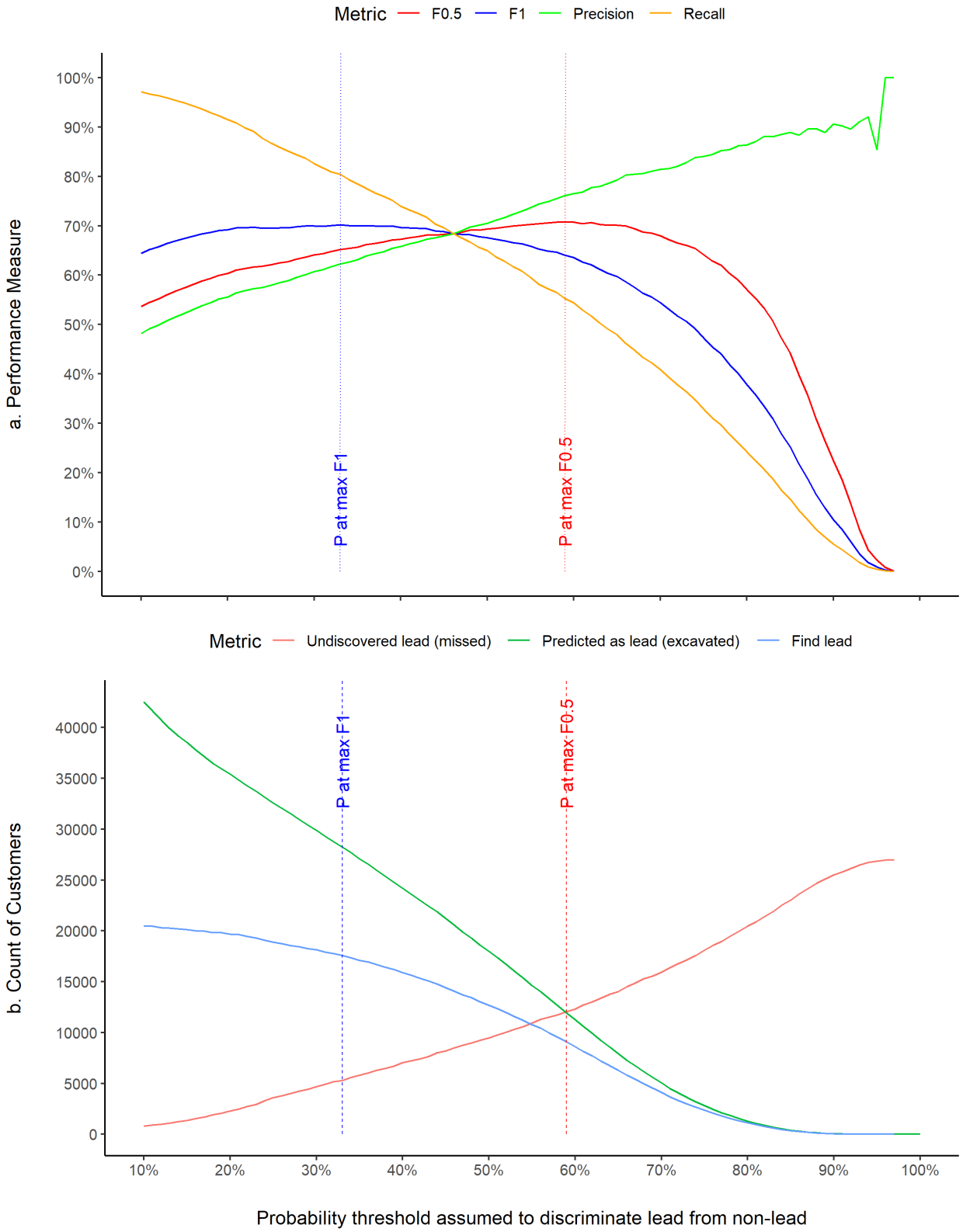

Figure 11: a. Model diagnostics and b. predicted customer counts assuming different probability thresholds for private service lines.

Figure 10.a also indicates the recall rate for these 9,080 customers is relatively low at 42%. This occurs because the threshold implied by this pool of 9,080 customers leaves many lead lines in service (a high false negative count), a necessary trade-off for a precise application of the model.

6.2 Prioritizing a Large Number of Replacements

Assume that PWSA were interested in planning replacements for a bigger pool of customers. For example, PWSA plans additional investigation or replacements at the 35,000 homes indicated as having lead or missing historical data as of July 2019. As summarized in Table 6, excavations to date indicate these historical labels are only 63% precise. As summarized in Figure 10, applying the model to these 35,000 customers leads to only a modestly more precise outcome of 66%. This application is considerably less precise than the 90% estimated when applying the model to prioritize annual replacements at fewer customers (Section 6.1). Figure 10.b indicates the implied probability threshold associated with these customers is only 24%. This low threshold falsely discriminates lead for many customers (a high count of false positives), resulting in an imprecise application of the model. Why consider replacements for a larger pool of customers? Such an approach leads to fewer lead lines in service, but at the necessary expense of imprecision and thus more unnecessary excavations.

6.3 Estimating Counts by Service Line Material

In contrast to the above applications that imply a probability threshold given a targeted pool of customers, estimating counts by service line material involves explicitly choosing a probability threshold that balances correct and incorrect predictions across all customers. Here, we summarize thresholds at maximum values of F1 and F0.5. F1 evenly balances correct and incorrect predictions. F0.5 balances correct and incorrect predictions but favors finding lead.

The model demonstrates only marginal value in estimating materials for all customers. Using maximum F1 to select a probability threshold, the model achieved a precision of 73%, whereas the historical data alone were at 63%. The reason for this is that too few predicted probabilities are extreme enough (high or low) to discriminate well between lead and non-lead service lines for all customers. Using maximum F0.5 to select a probability threshold, precision is increased from 73% (F1) to 84% (F0.5), but recall decreases from 80% (F1) to 62% (F0.5). Similar to the discussion in Section 6.2, these applications differ because a more precise model (F0.5) requires more false negatives (leaving lead in service).

6.4 Summary of Model Applications

Table 7 and Figure 10 summarize the above model applications. By applying the model to different pools of customers under different decision contexts, these applications exemplify the inherent trade-offs between precision and the count of false negatives (the lead lines left in service). These applications indicate that the value of the model is currently limited to those customers where the predicted probabilities are extreme (high or low). As such, Pitt does not currently recommend using the model to estimate inventories of service lines by material or plan replacements for customers beyond those planned for an annual planning cycle.

Table 7: Summary of model applications

Model application	Strategy	Selection of probability threshold	Customer count	Probability threshold	Mean probability lead	Precision	Strengths	Trade-off
Reduce unnecessary excavations in annual cycle	Apply model to count of customers subject to a replacement	Threshold is ***implied*** by constraining application to only those customers to be replaced	9,080 (assumed using year 2019 data)	77% (implied)	90%	90%	Precision is high because material discriminated for customers most likely to be lead	Prioritizing finding lead leads to a high share of false negatives (leaving lead in service)
Plan long-term replacements across service area			35,000 (customers where historical data are missing or indicate lead)	24% (implied)	44%	66%	The application covers all likely excavations	Lead discrimination performs poorly for customers moderately likely to be lead, resulting in more unnecessary excavations
Estimate counts of service lines by material	Select a probability threshold that balances true (or false) positive and negative predictions	Threshold at maximum F1 value balances true positive and negative outcomes	61,000 (all customers)	77%	44%	73%	The application covers all customers and balances true positives and negatives	Too few customers have extreme probabilities, making material discrimination uncertain
		Threshold at maximum F0.5 value favors finding lead	61,000 (all customers)	75%	44%	84%	The application covers all customers but favors “finding lead”	

7. Active Learning

Pitt applied active learning algorithms on PWSA's data to identify locations where additional data are most likely to improve predictions. Pitt used Hierarchical Sampling, which exploits the cluster structure of the test data to identify small subsets of customers where additional training data would be both informative and representative of the feature diversity across PWSA's customers. Twenty clusters were identified in the training data, and Pitt selected the 10 locations whose material predictions were most uncertain from each cluster to produce an active learning sample of 5,000. Predicted service line materials are most uncertain when the predicted probability the line is lead is near 0.5. For these customers, the model does no better than a chance guess. Thus, should PWSA be interested in reducing the active learning sample, Pitt recommends PWSA keep those locations closest to a predicted probability of 0.5 while sampling in the spatial clusters associated with the 5,000 locations flagged for active learning. The active learning results are submitted separately from this report as discussed in Section 11.

8. Clustering Excavation and Active Learning Locations

Pitt recognizes that PWSA can achieve cost efficiencies in their work order by spatially clustering their services. Those customers predicted as most likely to have lead service lines may or may not be spatially clustered. While not part of this scope of work, Pitt would be happy to work with PWSA to assess any spatial correlation associated with positive predictions or active learning recommendations that could assist PWSA in their programming.

9. Water Data as Non-Intrusive Material Detection

While water data are available for only 11% (n = 7,886) of customers, these data boost predictions of lead. Figure 12 shows median water samples increase as service lines are increasingly made of lead.

There are many factors other than the service line material that could influence water lead concentrations, including, but not limited to, seasonal variation in temperature and water quality, water treatment changes, premise plumbing and fixtures, customer demands, and sampling procedures. Importantly, our sample reflects stable water treatment methods with respect to corrosion control. However, repeated samples from the same customers significantly reduce the influence of this variability on predicting service line materials. Table 8 shows the number of customers where lead was detected potentially multiple times for different service line material configurations. Table 9 shows customers are much more likely to have a lead service line as the count of water samples in which lead is detected increases. Regression demonstrates a highly significant relationship between the service line material and the count of samples in which lead was detected. Table 9 shows the predicated probability that either the public or private service lines are lead given the count of samples in which lead is (a) detected or (b) not detected.

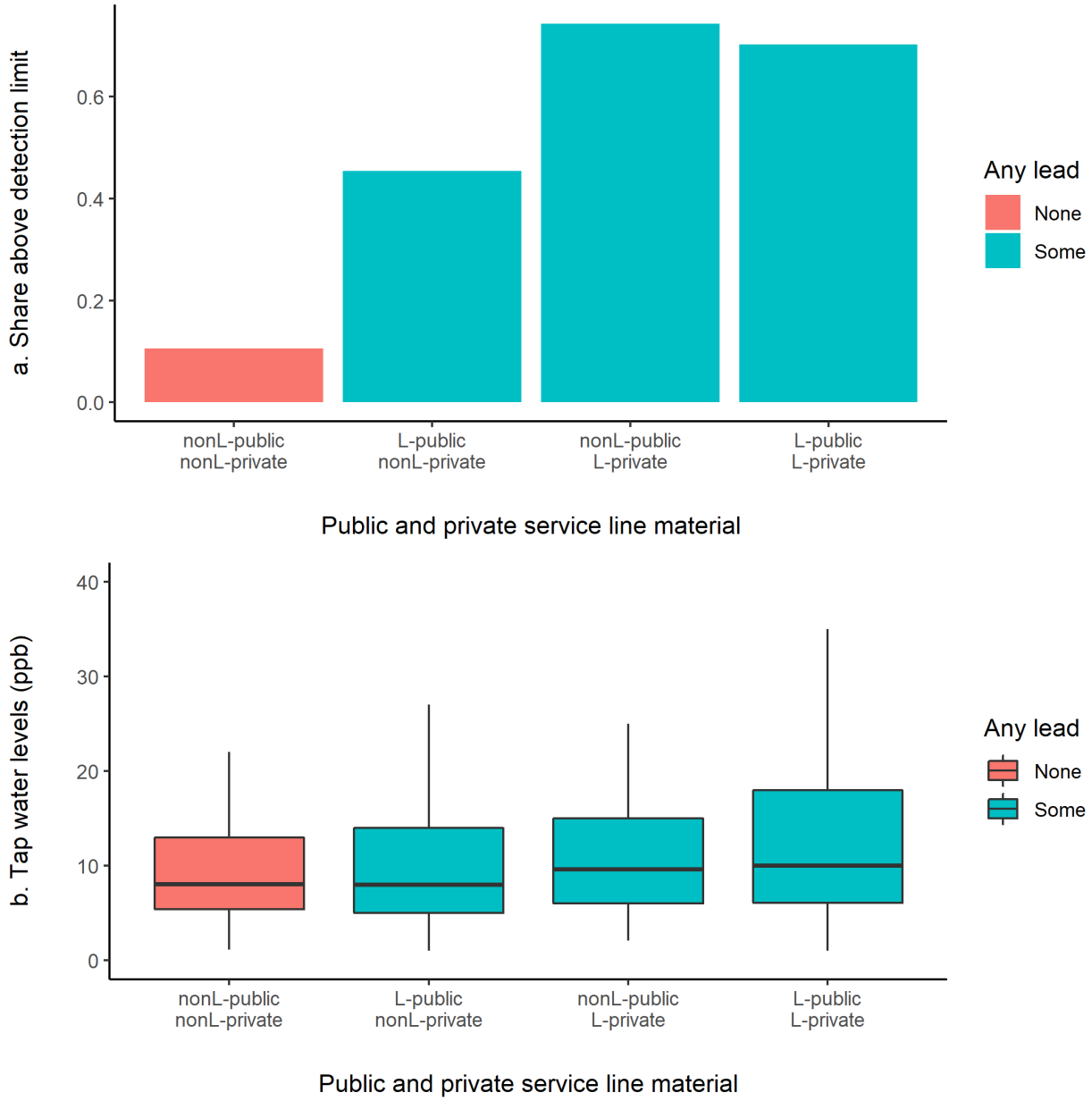

Figure 12: a. The share of customers with detectable lead concentrations and b. variation in the concentration of lead in water sample given different service line material configurations for 7,886 PWSA customers. Repeated samples are reported for 442 customers for a total of 8,505 unique samples.

Unfortunately, there are too few customers with multiple water samples to perform widespread predictions. However, targeted repeated water sampling for material diagnoses could serve as a helpful complement that reduces false negatives should PWSA prioritize replacements and excavations at customers predicted as most likely to have lead.

Table 8: Cross tabulation of customers showing counts of samples in which lead was detected in customers' taps given different service line material configurations. Each cell includes the count of customers meeting the indicated detection count and material followed by the row share in parenthesis.

Count of samples in which lead is detected	L-private L-public	L-private nonL-public	nonL-private L-public	nonL-private nonL-public	Row total
0	248 (22%)	26 (2%)	172 (15%)	682 (60%)	1,128
1	564 (65%)	75 (9%)	134 (15%)	94 (11%)	867
2	50 (74%)	3 (4%)	9 (13%)	6 (9%)	68
3	2 (50%)	1 (25%)	1 (25%)	0	4

Table 9: Predicted probabilities that either the public or private service line includes a. lead or b. is lead free given the counts of samples in which lead is detected or not detected.

a. Count of samples in which lead is detected	Count of customers	Predicted probability lead is in the service line ($logit(P_{Lead}) \sim B_0 + B_1 * count\ detects$)
0	1,128	40%
1	867	88%
2	68	98.8%
3	4	99.9%
a. Count of samples in which lead is not detected	Count of customers	Predicted probability the service line is lead free ($logit(P_{Lead\ free}) \sim B_0 + B_1 * count\ non\text{-}detects$)
0	896	11.8%
1	1,127	56.4%
2	43	92.6%
4	1	99.9%

10. Field Testing of the Model

Field verification conducted subsequent to delivering the trained model results to PWSA is summarized in Table 10. These data indicate that the model performs better at extreme predicted probabilities. For those customers with a predicted probability of lead greater than 80% and less than 20%, precisions of 80% and 85% were verified for lead and non-lead, respectively.

Table 10: Summary of field verification. Field verification data were collected after July 2019.

Predicated probability lead	Predicted lead	Predicted non-lead	Verified Lead (TP)	Verified Non-Lead (FP)	Precision
Greater than 80%	831	0	662	169	80% (lead)
40% to 80%	2106	32	1203	903	57% (lead)
20% to 40%	0	471	162	309	66% (non-lead)
Less than 20%	0	212	31	181	85% (non-lead)

11. Supplemental Resources

Supplemental to this report Pitt has separately provided the following:

a. Four maps
    i. Threshold-.41(F1 score).html,
    ii. Threshold-.41(F1 score)-Details.html,
    iii. Threshold-.61(F0.5 score).html,
    iv. Threshold-.61(F0.5 score)-Details.html) `

b. A dataset that includes the following fields by customer
    i. The customer weight to correct for potential sampling and spatial biases
    ii. The probability each customer has a lead service line as of July 2019
    iii. The predicted material as of July 2019 assuming a probability threshold that maximizes F1
    iv. The predicted material as of July 2019 assuming a probability threshold that maximizes F0.5
    v. A flag indicating the predicted material using F1 differs from PWSA's historical data
    vi. A flag indicating the predicted material using F0.5 differs from PWSA's historical data
    vii. A flag indicating the customer is recommended for active learning

References

Abernethy, Jacob, Cyrus Anderson, Chengyu Dai, Arya Farahi, Linh Nguyen, Adam Rauh, Eric Schwartz, et al. 2016. “Flint Water Crisis: Data-Driven Risk Assessment Via Residential Water Testing,” September. https://arxiv.org/abs/1610.00580v1.

Abernethy, Jacob, Alex Chojnacki, Arya Farahi, Eric Schwartz, and Jared Webb. 2018. “ActiveRemediation: The Search for Lead Pipes in Flint, Michigan.” In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 5–14. KDD ’18. New York, NY, USA: ACM. https://doi.org/10.1145/3219819.3219896.

Brenning, A. 2012. “Spatial Cross-Validation and Bootstrap for the Assessment of Prediction Rules in Remote Sensing: The R Package Sperrorest.” In *2012 IEEE International Geoscience and Remote Sensing Symposium*, 5372–75. https://doi.org/10.1109/IGARSS.2012.6352393.

Chojnacki, Alex, Chengyu Dai, Arya Farahi, Guangsha Shi, Jared Webb, Daniel T. Zhang, Jacob Abernethy, and Eric Schwartz. 2017. “A Data Science Approach to Understanding Residential Water Contamination in Flint.” In , 1407–16. ACM. https://doi.org/10.1145/3097983.3098078.

Guyon, Isabelle, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. “Gene Selection for Cancer Classification Using Support Vector Machines.” *Machine Learning* 46 (1): 389–422. https://doi.org/10.1023/A:1012487302797.

Kanamori, Takafumi, Shohei Hido, and Masashi Sugiyama. 2009. “A Least-Squares Approach to Direct Importance Estimation.” *J. Mach. Learn. Res.* 10 (December): 1391–1445.

Matheron, G. 1973. “The Intrinsic Random Functions and Their Applications.” *Advances in Applied Probability* 5 (3): 439–68. https://doi.org/10.2307/1425829.

Miller, Harvey J. 2004. “Tobler’s First Law and Spatial Analysis.” *Annals of the Association of American Geographers* 94 (2): 284–89. https://doi.org/10.1111/j.1467-8306.2004.09402005.x.

Moran, P. A. P. 1950. “Notes on Continuous Stochastic Phenomena.” *Biometrika* 37 (1/2): 17–23. https://doi.org/10.2307/2332142.

PWSA. 2019a. “Service Line Material Inventory Revised Approach.”

———. 2019b. “Varied Datasets Describing Historical Data Describing Original Service Line, Installment Date, Service Line Replacements, Etc.” PWSA.

Roberts, David R., Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, et al. 2017. “Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure.” *Ecography* 40 (8): 913–29. https://doi.org/10.1111/ecog.02881.

Tin Kam Ho. 1995. “Random Decision Forests.” In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–82 vol.1. https://doi.org/10.1109/ICDAR.1995.598994.