Water quality in Chicago, IL: Predicting lead hazards using housing assessment and socioeconomic variables

CAPP 30254: Machine Learning for Public Policy

Valeria Balza (vbalza)
Tarren Peterson (tarren)
James Midkiff (jmidkiff)

June 4, 2021

1. Executive Summary

The city of Chicago has more lead service water lines than any other city in the United States, with nearly 400,000 lead service lines connecting family homes and small apartment buildings to water mains.[1] Corrosion of these lines may expose residents to toxic lead in drinking water, which even in small concentrations can result in developmental problems in children and contribute to health complications in adults. The magnitude of this problem was recently acknowledged by the local government when Chicago Mayor Lori Lightfoot announced the first phase of the Lead Service Line Replacement Program in September 2020.[2] The program, which is estimated to cost 8.5 billion USD and take multiple decades to complete, must prioritize those communities that observe the highest risk of lead exposure and allocate resources accordingly.

We drew on a range of housing assessment, demographic, and socioeconomic variables to train classification algorithms—namely, logistic regression, random forests, and linear support vector classification (Linear SVC) models—that predict high lead exposure at the census block group level in Chicago. For all our models, we used recall as the main evaluation metric, as the long-term consequences of elevated lead exposure in children and vulnerable individuals are devastating—and are associated with public costs estimated to be over 200 billion USD nationwide.[3] In other words, we place more weight on false negatives, in which a model fails to identify block groups with elevated lead results, than on false positives, where a model incorrectly predicts that a block group has elevated lead results when in fact it does not. The best models—a weighted random forest and a logistic regression—achieved recall and accuracy scores of over 75% and 65%, respectively. While feature importance varied across models—most models deemed owner-occupancy rates and the presence of single-family homes to have the most predictive power. When interpreting and leveraging the results of this study, it is important to consider its limitations, including possible selection bias in our dataset.

Still, our analysis reveals machine learning algorithms can provide meaningful insight into which Chicago block groups may face the highest risk for lead exposure in their drinking water and may be used to inform the city's strategy as it launches the Lead Service Line Replacement Program in Summer 2021. Ultimately, it is imperative that the city encourages increased testing among low-income renters to better understand the distribution of lead exposure risks across the city.

2. Background: Water Quality in Chicago

While the U.S. Environmental Protection Agency (EPA) establishes a 15 parts per billion (ppb) action level—requiring water systems to undertake corrosion control measures and to inform the public when 10% of water samples show lead concentrations exceeding 15 ppb—the organization recognizes lead can be harmful to humans even at low exposure levels. In children, low lead exposure has been linked to damage to the central and peripheral nervous system, learning disabilities, shorter stature,

[1] Paul Caine, "Chicago Has More Lead Service Pipes Than Any Other US City, Illinois the Most of Any State," *WTTW Chicago*, March, 24, 2021, https://news.wttw.com/2021/03/24/chicago-has-more-lead-service-pipes-any-other-us-city-illinois-most-any-state.

[2] Mayor's Press Office, "Mayor Lightfoot Launches Equity-Focused Lead Service Line Replacement Program," Office of the Mayor, September, 10, 2021, https://www.chicago.gov/city/en/depts/mayor/press_room/press_releases/2020/september/EquityFocusedLeadServiceReplacement.html.

[3] Pew Charitable Trusts, "Cutting Lead Poisoning and Public Costs," https://www.pewtrusts.org/~/media/assets/2010/02/22/063_10_paes-costs-of-lead-poisoning-brief_web.pdf.

impaired hearing, and impaired formation and function of blood cells.[4] In adults, lead exposure can result in reduced growth of the fetus and/or premature birth among pregnant women as well as contribute to cardiovascular complications, decreased kidney function and reproductive problems in both men and women.[5]

In Chicago, a study conducted in 2018 found lead in the tap water of nearly 70 percent of the 2,797 homes tested between 2016 and 2017.[6] This same study reported that 30 percent of the samples had lead concentrations of higher than 5 ppb—the maximum level allowed in bottled water by the U.S. Food and Drug Administration. The gravity of this problem was recently acknowledged by the local government when Chicago Mayor Lori Lightfoot announced the first phase of the Lead Service Line (LSL) Replacement Program in September 2020. The program, which aims to replace the city's nearly 400,000 LSLs, is estimated to cost 8.5 billion USD and take multiple decades to complete.[7]

Local government officials have already recognized the importance of adopting an equity-driven implementation approach, prioritizing low-income residents who (i) own and reside in their home; (ii) have a household income below 80% of the area median income (72,800 USD for a family of four); and (iii) have consistent lead concentrations above 15 ppb in their water, as tested by the Department of Water Management.[8] Some experts, however, have already criticized this approach, noting the current prioritization framework may overlook low-income renters who are traditionally more exposed to lead because their housing is in poor condition.[9] Given the cost and scope of the LSL Replacement Program, it is imperative that Chicago's city officials adopt a data-driven approach that prioritizes vulnerable communities and allocates resources accordingly.[10]

To identify communities that may be at higher risk for lead exposure in their drinking water, we analyzed water quality data from the Department of Water Management showcasing lead concentrations (measured in ppb) of residential water samples across Chicago. We complemented our analysis by including socioeconomic and demographic indicators from the American Community Survey's Five-Year Estimates and aggregate housing variables derived from the Cook County Assessor's Residential Property Characteristics dataset. We trained machine learning classification algorithms on the final dataset combining socioeconomic, demographic and housing assessment indicators to predict high lead exposure at the census block group level. We then used our models to develop a block group risk profile for the city.

We hope our analysis can be leveraged by Chicago city officials—namely those in the Department of Water Management and the Mayor's Office—and encourages the adoption of an equity- and data-driven prioritization framework for the LSL Replacement Program as it launches in the upcoming months. We also hope our analysis helps inform Chicago city residents concerned about lead in their

[4] United States Environmental Protection Agency, "Basic Information about Lead in Drinking Water", https://www.epa.gov/ground-water-and-drinking-water/basic-information-about-lead-drinking-water#health

[5] Ibid.

[6] Michael Hawthorne, "Brain-damaging lead found in tap water in hundreds of homes tested across Chicago, results show," *Chicago Tribune*, April, 12, 2018, https://www.chicagotribune.com/investigations/ct-chicago-water-lead-contamination-20180411-htmlstory.html.

[7] Monica Eng, "What We Do — And Don't — Know About Chicago's Lead Water Problem," *NPR*, September 21, 2020, https://www.npr.org/local/309/2020/09/21/915248028/what-we-do-and-don-t-know-about-chicago-s-lead-water-problem.

[8] Mayor's Press Office, "Mayor Lightfoot Launches Equity-Focused Lead Service Line Replacement Program," Office of the Mayor, September, 10, 2021, https://www.chicago.gov/city/en/depts/mayor/press_room/press_releases/2020/september/EquityFocusedLeadServiceReplacement.html.

[9] Monica Eng, "Despite A Promise, Chicago Has Made No Progress On Removal Of Lead Pipes," *WBEZ Chicago*, April, 18, 2021, https://www.wbez.org/stories/despite-a-promise-chicago-has-made-no-progress-on-removal-of-lead-pipes/02644e18-4cd5-4e7e-b595-3ebc111c62a6.

[10] We define vulnerable communities as those neighborhoods/community areas who are at higher risk for lead exposure through lead service lines relative to other neighborhoods/community areas.

drinking water, as there are control measures that can reduce the risk of lead exposure as the city works to replace the 400,000 LSLs across the city.

3. Data

The data used in our analysis is all publicly available and can be accessed online via the relevant government organizations. Our primary analysis was conducted at the census block group level for the city of Chicago, with each row representing a block group including the aggregate features described below and an outcome variable classified as 1 if any house in that block group exceeded the lead concentration threshold.

We developed two different thresholds (tested separately): a high threshold of 15 ppb and a medium threshold of 5 ppb. The high threshold matches the action level established by the EPA described in Section 2 while the medium threshold of 5 ppb matches a lower standard that the U.S. Food and Drug Administration has established as the limit for bottled water. Other organizations also advocate for reducing the EPA lead action level from 15 ppb to 5 ppb to match the FDA's standard.[11]

Water Quality Data

Water quality data featuring lead concentrations (ppb) in water were obtained from the Chicago Department of Water Management (DWM).[12] The dataset features the results of water samples conducted across Chicago residences between 2016 and early 2021. Lead tests are initiated when a resident requests a free testing kit from the DWM. The homeowner gathers the water samples according to the directions provided; if the samples were correctly gathered, the department analyzes the results, reports them to the homeowner, and adds them to its dataset, which at the time of analysis contained over 22,000 observations.

Each observation in the dataset contains three sample readings: (i) immediately after first turning on the tap; (ii) two-three minutes after turning on the tap; and (iii) five minutes after turning on the tap. For our analysis, we considered the maximum of these three readings as the final reading for the corresponding observation. Further, if any water sample returns fewer than 1.0 ppb, the DWM replaces that value with "<1.0". We replaced any values of "<1.0" with 1.0.

Observations in the original dataset are partly obfuscated, in that the last two digits of a homeowner's address are replaced with "XX" for anonymization purposes. We imputed these values with "00" such that "13XX E Hyde Park Blvd", for instance, became "1300 E Hyde Park Blvd", yielding block-level addresses. We then geocoded these addresses to retrieve a sample's geographic location within a city block. We constructed two additional features that were used to construct the final outcome variables at the census block group level: (i) *threshold (high)*, coded as 1 if the maximum sample reading exceeded 15 ppb and 0 otherwise; and (ii) *threshold (medium)*, coded as 1 if the maximum sample reading exceeded 5 ppb and 0 otherwise. **Table 1** presents descriptive statistics for the water quality dataset at the sample level.

[11] Kristi Pullen Fedinick, "Millions Served by Water Systems Detecting Lead," *Natural Resources Defense Council,* May 13, 2021, https://www.nrdc.org/resources/millions-served-water-systems-detecting-lead.

[12] Data from Chicago Department of Water Management's Water Quality Study. Data was retrieved from https://www.chicagowaterquality.org/home#results.

Table 1. Descriptive Statistics for Water Quality Data

	1st draw	2-3 minute	5 minute	Max Reading	Threshold (H)	Threshold (M)
Mean	3.64	4.11	2.26	5.34	0.04	0.33
Std. Dev	13.72	6.83	3.0	14.70	0.20	0.47
Min	1.0	1.0	1.0	1.0	0.00	0.00
25%	1.0	1.0	1.0	1.0	0.00	0.00
50%	2.0	2.20	1.20	2.90	0.00	0.0
75%	3.80	5.40	2.50	6.40	0.00	1.00

American Community Survey Five-Year Estimates (2019)

To build our set of features, we first drew on the American Community Survey (ACS) Five-Year Estimates (2019), which contain socioeconomic, demographic, and housing variables at the census block group level. We anticipated demographic features to have predictive power because of Chicago's redlining history discriminating against Black/African American and other minority communities—making it more likely for these populations to live in lower quality housing.[13]

Indeed, a study by the Metropolitan Planning Council found Black- and Hispanic-majority communities in Illinois to be more likely to live in residences with LSLs relative to White-majority communities.[14] Demographic and socioeconomic variables in the feature set include (i) total population; (ii) median income; (iii) White population (percentage of total); (iv) Black/African American population (percentage of total); and (v) Non-White population (percentage of total, to account for Hispanic/Latino and other non-Black minority groups).

Further, since Chicago required the use of lead service lines until 1986, when the practice was banned by the federal government, individuals living in single-family or two-flat homes built before that year have a higher likelihood of being connected to a LSL (unless it was replaced during a renovation).[15] As such, we included ACS housing variables in our analysis to capture differences in living conditions across census block groups. These features include (i) average household size; (ii) number of occupied housing units; (iii) median gross rent; and (iv) percentage of owner-occupied housing units.

Cook County Assessor's Residential Property Characteristics

Finally, we complemented the ACS feature set with the 2020 Cook County Assessor's Residential Property Characteristics dataset.[16] We used aggregate indicators at the census block group level for (i) mean/median land value; (ii) mean/median building value; (iii) mean/median land size in square feet; (iv) mean/median building size in square feet; (v) mean/median building age; and a series of one-hot encoded binary variables for property type, wall material, roof material, repair condition, and renovation status of the property.

[13] Natalie Moore, "New Redlining Maps Show Chicago Housing Discrimination," *WBEZ Chicago*, October, 28, 2016, https://www.wbez.org/stories/new-redlining-maps-show-chicago-housing-discrimination/37c0dce7-0562-474a-8e1c-50948219ecbb.

[14] Justin Williams, "Data Points: the environmental injustice of lead lines in Illinois," Metropolitan Planning Council, November, 10, 2020, https://www.metroplanning.org/news/9960/Data-Points-the-environmental-injustice-of-lead-lines-in-Illinois.

[15] Lead-Safe Chicago, Lead Service Line Replacement, https://www.leadsafechicago.org/lead-service-line-replacement.

[16] Data retrieved from Cook County Assessor's Office, Cook County Assessor's Residential Property Characteristics. Data last updated on November, 27, 2020. Data retrieved via https://datacatalog.cookcountyil.gov/Property-Taxation/Cook-County-Assessor-s-Residential-Property-Charac/bcnq-qi2z.

Of the 1,995,108 total residences in the Cook County assessment dataset, we focus on the 728,543 that fall within Chicago's city boundaries to construct the features. The following table shows summary statistics for the assessment variables we utilized in our models.

Table 2. Descriptive Statistics for Housing Assessment Features

	Non-null values	Null values	Mean	Std. Dev	Median	Min	Max
Prior tax year market value estimate (land)	728,543	0	45,602	54,501	34,320	0	3,040,940
Prior tax year market value estimate (building)	728,543	0	225,590	253,792	164,960	0	13,358,380
Land square feet	728,543	0	17,032	37,369	4,495	0	7,753,680
Building square feet	438,342	290,201	1,875	1,234	1,489	0	19,992
Age	728,543	0	71	39	70	1	205
Property class	728,543	0	--	--	--	--	--
Wall material	438,329	290,214	--	--	--	--	--
Roof material	438,329	290,214	--	--	--	--	--
Repair condition	438,329	290,214	--	--	--	--	--
Renovation	1,243	727,300	--	--	--	--	--

It should be noted that the Cook County assessment data includes information on smaller residential properties that have a property class in the 200-series.[17] The dataset does not include properties categorized as multifamily housing, which will have a property class in the 300-series. As such, no apartment buildings with seven or more units appear in our dataset.

Table 3. Property Class Counts

Property class	Description	Count	Proportion
202	One-story Residence, any age, up to 999 square feet	56,905	7.81%
203	One-story Residence, any age, 1,000 to 1,800 square feet	135,822	18.64%
205	Two-or-more story residence, over 62 years of age up to 2,200 square feet	36,312	4.98%
211	Apartment building with 2 to 6 units, any age	125,638	17.25%
299	Residential condominium	277,880	38.14%
All others	--	95,986	13.17%
Total	**--**	**728,543**	**100.00%**

See *Appendix 1* for figures that provide the full distribution of residential property classes and property age.

[17] See the *Definitions for the Codes for Classification of Real Property* available at: https://prodassets.cookcountyassessor.com/s3fs-public/form_documents/classcode.pdf.

The described datasets were cleaned, wrangled and merged into one final dataset containing aggregate features and outcome labels at the census block group level. The total features and outcome variables are listed in Section 4. See *Appendix 2* for a selection of rows in the final dataset.

Figure 1, which presents the number of water samples within a census block group that exceed 15 ppb (high threshold), suggests lead concentrations are not evenly distributed across Chicago—with block groups in the north and south parts of the city showcasing a higher number of households exceeding the threshold. Figure 1, however, also reveals a slight classification imbalance in the dataset, as there are a substantial number of block groups with a count of zero. Indeed, Figure 2 shows that census block groups with zero households exceeding the threshold are overrepresented in the data, comprising 73% of the total observations. The opposite is observed for the medium threshold, with the census block groups with zero households exceeding the medium threshold comprising only 25% of the data. We discuss mitigation strategies during modeling for this imbalance problem in the following section.

Figure 1. Number of Tests exceeding 15 ppb by Census Block Group

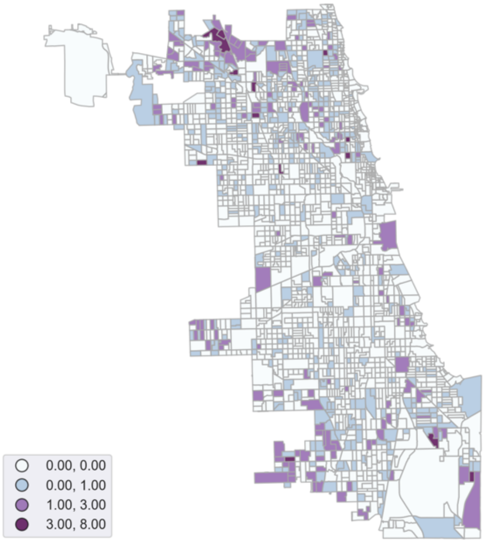
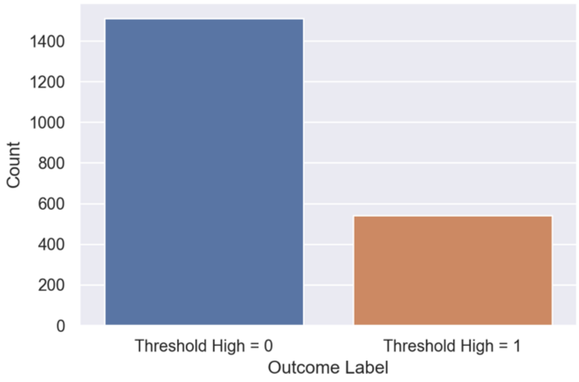


Figure 2. Threshold (high) class distribution



4. Machine Learning and Details of Solution

Drawing on the described set of socioeconomic, demographic, and housing features, we trained several classification models to predict whether or not a census block group will observe high lead exposure levels (depending on the threshold used, either (high) 15 ppb or (medium) 5 ppb). Logistic regression, random forests, and Linear SVC models were all fit to the training data. We adopted this approach as similar methods have been adopted by other scholars predicting the likelihood of lead exposure in children, including one that draws on home inspections and property value assessment data to predict lead level from blood tests in children for Chicago from 1993-2013.[18]

[18] Eric Potash et al., "Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning", KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August, 2015, https://dl.acm.org/doi/10.1145/2783258.2788629.

Our final dataset includes the following features at the census block group level:

> Total population; median income; White population (percentage of total); Black/African American population (percentage of total); Non-White population (percentage of total); average household size; number of occupied housing units; median gross rent; and number of owner-occupied housing units; mean and median land value; mean and median property value; mean and median land size in square feet; mean and median property size in square feet; mean and median property age; and a series of one-hot encoded binary variables for property type, wall material, roof material, repair condition, and renovation status of the property.

For the assessment data, we selected these variables on the basis of features that seemed most relevant for predicting the presence of LSLs (i.e. not the size of the garage), were not directly constructed from other features (such as the size of the lot, squared), were not prone to overfitting the data (such as the deed number), and that had the most potential for policy recommendations (i.e. not the basement finish). In addition, Cook County describes to a limited extent the dataset features on its website and notes that some are not suitable for analytical purposes, such as the construction quality or site desirability.

We then one-hot encoded the assessment data for two reasons. First, many of the features are categorical and have no intrinsic hierarchy; for example, 'wall material' takes on the values of 1, 2, 3, or 4, corresponding to wood, masonry, wood & masonry, or stucco materials comprising a residence's external walls. Second, many of these categorical features are missing a significant number of values for which there is no obvious replacement value. After aggregating these variables at the block group level, we converted them to proportions of the total number of residences within that block group. For any numerical features missing values, we imputed with the feature median. And while not necessary for the random forests model, we normalized all non-target variables to a mean of 0 and standard deviation of 1.

As noted earlier, our models were trained separately on two binary targets: threshold (high), classified as 1 if any house in the respective census block group exceeded 15 ppb and 0 otherwise; and threshold (medium), classified as 1 if any house in the respective census block group exceeded 5 ppb and 0 otherwise.

In order to account for the imbalance in our dataset, we chose to train two other types of random forests, a weighted random forest and a random forest with SMOTE (Synthetic Minority Over-sampling Technique). In a weighted random forest, the algorithm optimizes using a heavier penalty if it misclassifies the minority class in the unbalanced sample. In our case, the algorithm would give a heavier penalty for misclassifying the incorrect identification of a census block group above the 15ppb threshold as being below the threshold. This helps us to increase the recall metric above what was achieved in a normal random forest. Similarly, we use random forests with SMOTE as another method to account for the unbalanced dataset. SMOTE works by synthetically creating new examples of the minority class by selecting examples that are close in the feature space and drawing new sample at a nearby point. This approach is different from traditional oversampling techniques because rather than continually drawing from the same, small pool of data points in the minority class, SMOTE artificially creates new ones based on like points to feed the model new data. Balanced class weights were also included as parameters for the Linear SVC models. For all our models, we used 10-fold cross validation.

5. Evaluation and Results

We used three different metrics to evaluate the performance of all our models: recall, accuracy, and precision. The accuracy metric evaluates the fraction of predictions a model labeled correctly, recall evaluates the proportion of actual positives that were identified correctly, and precision evaluates the proportion of positive identifications that were actually correct. We evaluated all models based on the maximum average recall score achieved during the 10-fold cross validation, as a false negative (incorrectly labeling a block group as having a low risk of lead exposure) likely has a higher societal cost than a false positive. Still, incorporating accuracy and precision metrics helps to understand the tradeoffs which researchers and policymakers face when choosing among different models and parameters. **Table 4** summarizes the results of our analysis.

Logistic Regression

The best logistic regression models saw significant variation in their evaluation metrics depending on the specified target. When predicting the medium threshold (5.0 ppb), the model correctly recalled almost 98% of observations. However, the logistic model performed less well when looking to identify block groups with lead levels ≥15.0 ppb or when trying to distinguish between block groups with high lead levels and those with medium lead levels (multinomial). Part of this variation could be a reflection of the imbalance between the various target classes, as mentioned in Section 4, for which reason we considered both balanced and non-balanced class weights. Using class weights only improved the evaluation metrics of the first model in the table—likely because the target class is larger than the non-target class in the second model and because there are multiple target classes that are somewhat evenly distributed in the third.

In terms of features importance, there was some broad agreement between the various logistic models. The most predictive features *positively* associated with elevated lead levels are: (i) Property Class 205—referring to two-or-more story residences, over 62 years of age up to 2,200 square feet; (ii) percentage of homes that are owner-occupied; (iii) roof material 1—referring to shingle/asphalt; (iv) prior tax year market value estimate (land), median; and (v) property class 203—referring to one-story residences, any age, 1,000 to 1,800 square feet. The most predictive feature *negatively* associated with elevated lead levels was the median building size in square feet. In general, these features are strongly associated with single-family homes, suggesting that these residences are more likely to have elevated lead test results than apartment buildings. This could be a direct consequence of the city's plumbing code which required the use of lead water pipes until 1986, particularly for homes and two-flat buildings.[19]

Overall, these results indicate that the logistic models are good at identifying block groups where lead is present to some degree in residential water, but perform less well at predicting the severity of the lead problem (i.e., 5.0 vs 15.0 ppb). See *Appendix 4* for additional figures and tables on the most important logistic regression models, and a confusion matrix for the multinomial logistic model.

[19] See Hawthorne, 2021.

Table 4. Summary of Classification Algorithms

	Model	Target	Parameters	Recall	Accuracy	Precision
1	**Logistic Regression (Binary Classification)**	Threshold (**H**, 15 ppb)	C: 0.001 Penalty: l2 Solver: liblinear Class weight: balanced	0.75	0.65	0.41
2	**Logistic Regression (Binary Classification)**	Threshold (**M**, 5 ppb)	C: 0.01 Penalty: l1 Solver: liblinear Class weight: None	0.98	0.77	0.77
3	**Logistic Regression (Multinomial Classification)**	2 At least one result ≥ 15.0 ppb 1 At least one result ≥ 5.0 ppb and <15.0 ppb, 0 Otherwise	C: 1 Penalty: l2 Solver: lbfgs Class weight: None	0.56	0.57	0.58
4	**Random Forest**	Threshold (**H**, 15 ppb)	N_Estimators = 1000 Criterion = Gini Max Depth = 5 Min Samples Split = 2	0.20	0.76	0.64
5	**Weighted Random Forest**	Threshold (**H**, 15 ppb)	N_Estimators = 5000 Criterion = Entropy Max Depth = 1 Min Samples Split = 2 Class weight: balanced subsample	0.76	0.66	0.42
6	**Random Forest with SMOTE**	Threshold (**H**, 15 ppb)	N_Estimators = 100 Criterion = Gini Max Depth = 5 Min Samples Split = 2	0.62	0.72	0.47
7	**Weighted Linear SVC**	Threshold (**H**, 15 ppb)	C: 0.01 Class weight: balanced	0.65	0.68	0.43
8	**Weighted Linear SVC**	Threshold (**M**, 5 ppb)	C: 1.0 Class weight: balanced	0.73	0.74	0.90

Random Forests

The random forest models were all trained to predict whether a given block group contained high levels of lead (>15 ppb). Due to the significant amount of training time for each random forest model, we decided to only include the predictions for the highest levels of lead, which we felt was a more important predictive metric. In total, three different types of random forest models were optimized using a grid search. We trained random forests, weighted random forests, and random forests with SMOTE. As discussed in the data section, the threshold (high) variable is unbalanced, as only a quarter of the total census block groups have tests with more than 15 ppb of lead. As a result, the recall metric for the standard random forests were very low and no model in the grid search scored higher than 20% for recall.

When evaluating the three types of Random Forest models, it is clear that weighted random forests and random forests with SMOTE achieve significantly better results than do normal random forests. The weighted random forest with the highest recall achieved recall of 74% and the random forest with SMOTE with the highest recall achieved a recall of 64%. Since we are seeking to maximize on recall, weighted random forest was the best model of all the random forests models executed.

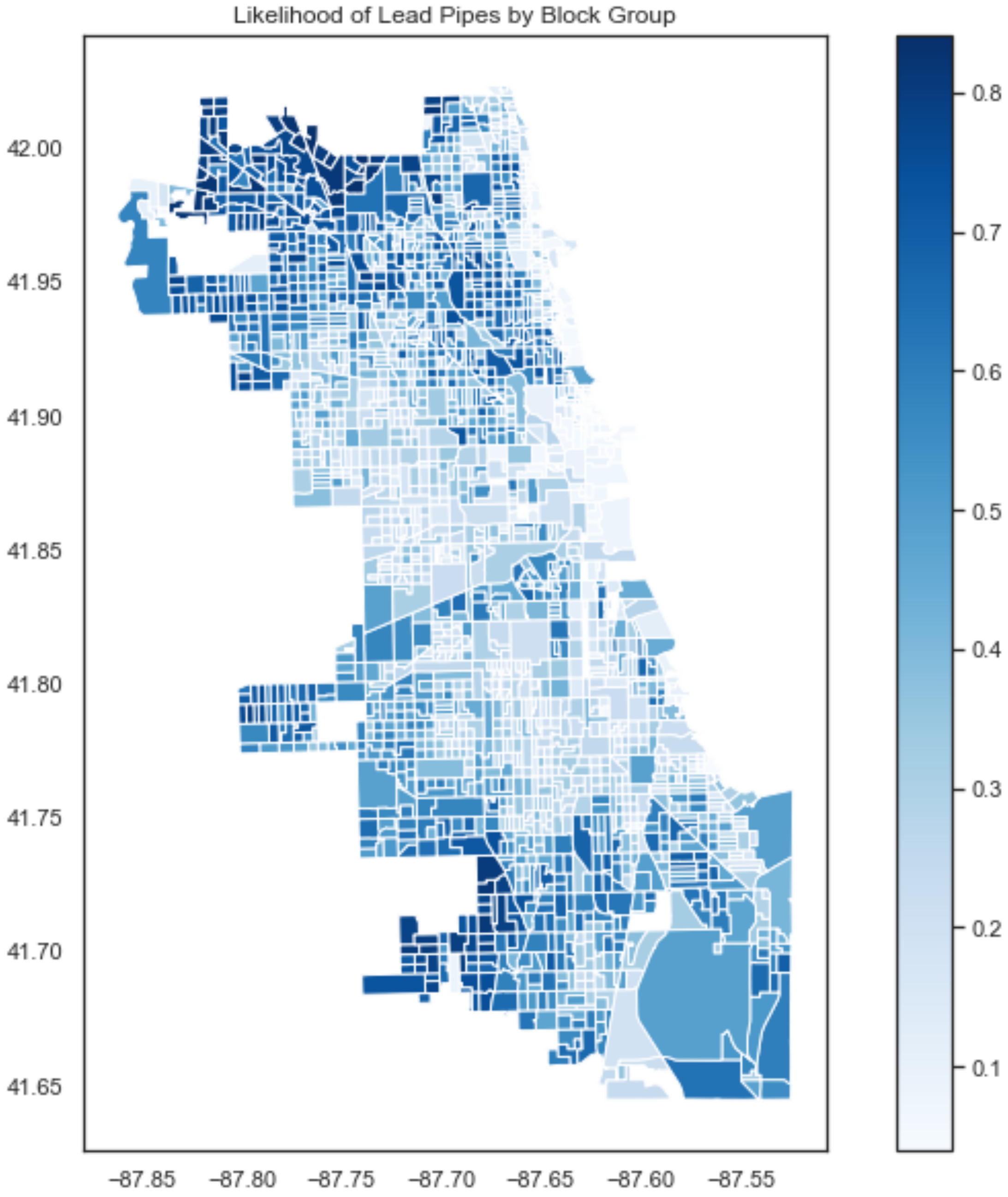
For a confusion matrix and feature importance diagram of the most important features in the best performing random forest model, see Appendix 5.

Linear SVC

Similarly, we tested our data using linear support vector classification with balanced class weights, which performed better when testing the medium threshold of 5.0 ppb. This second Linear SVC model resulted in 73% recall, 74% accuracy and 90% precision. The features with the highest importance for the best Linear SVC model comprised: (i) total population; (ii) building market value (estimate); (iii) average household size; (iv) Property Class 211 (apartment buildings of 2-6 stories); and (v) percentage of owner-occupied units.

Across all models, the weighted random forest model appears to achieve the highest evaluation metrics closely followed by the logistic regression model.

Census Block Group Risk Profile



We used a weighted random forest to predict the probability with which a given block group would have a high concentration of lead pipes. As shown above, we are most confident that communities in the southwest and northwest parts of the city would experience the high levels of lead. This makes sense given the feature importance of the random forest models, which placed heavier emphasis on rates of home ownership, the number of older, single family homes, and median income. These clusters identified on the map all have above average rates of home ownership, single family units, and median income. See the *Appendix 6* for a similar map that used a Multinomial Logistic Regression to predict the classification of block groups into low, medium, or high thresholds for lead levels.

6. Discussion

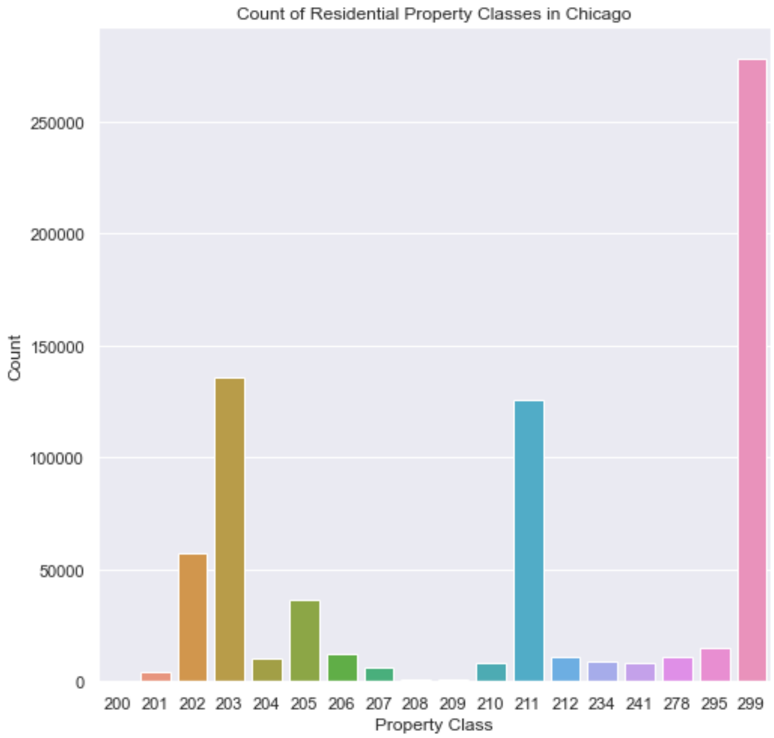
As the city prepares to launch the Lead Service Line Replacement Program in Summer 2021, our analysis reveals that lead concentrations in water are not evenly distributed across the city. As such, it is imperative that the city adopts a data- and equity-driven approach when prioritizing replacement schedules, considering the factors that may increase the likelihood of lead exposure in drinking water. As seen in the risk profile map above, communities in the northwest and southwest parts of the city appear to be at an elevated risk of high lead exposure in their drinking water. There could be a variety of explanatory factors—including the presence of older single-family homes and higher owner-occupancy rates. As mentioned in Section 2, this may be attributed to the fact that older homes may be more likely to be connected to LSLs due to the plumbing codes present until 1986.

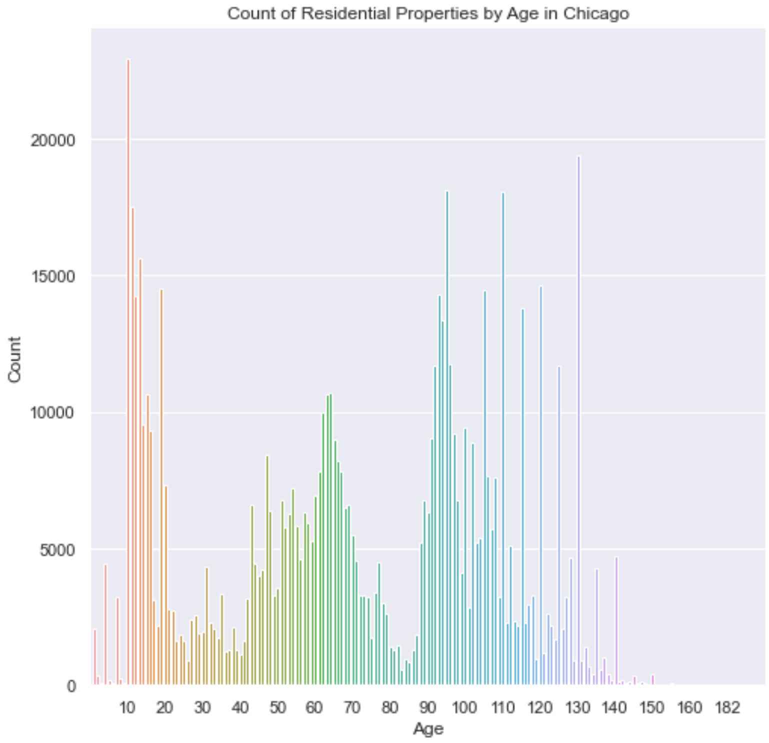
When interpreting and leveraging the results of our analysis, however, we must also consider potential selection bias in the water quality, as homes in the northwest and southwest parts of the city also observed increased sampling rates (see *Appendix 3* for a map of sampling distribution). There are features likely associated with the number of tests that occur in a particular block group, and as the number of tests increases in a particular area, it is statistically more likely that any one of those tests contains a result greater than our 'high' threshold of 15.0 ppb or our 'medium' threshold of 5.0ppb. For instance, we wonder if homeowners are more likely to test their water for lead than are renters because homeowners tend to live in the same residence for a longer period of time and thus pay more attention to their residence's condition, or renters do not test because they believe landlords would have already replaced lead service lines for liability reasons. In addition, the areas with higher testing rates also have higher median incomes, and perhaps individuals with higher incomes are better situated to test for lead in their water systems. As such, we strongly advise the reader to interpret our models' results with caution and consideration of the above. The results of our models do not suggest causal relationships between the features and outcomes analyzed but are rather indicative of the factors that may benefit from further testing.

Since the plumbing codes present until 1986 make it fairly likely that lead service lines are pervasive throughout the city, it is imperative policymakers prioritize further testing in areas of the city that have yet to see many tests, mostly in the west and south parts of Chicago. After a more even distribution of data is collected, our models could be rerun to better capture the areas throughout the city that are most likely to have lead pipes and that should therefore be prioritized in the initial phases of the Lead Service Line Replacement Program. Our models would also benefit from an open data plan that would allow water testing results to be displayed at the individual property level in order to match directly to the Cook County Residential Property data rather forcing aggregation at the block or block group level.

7. Appendix

Appendix 1: Additional Assessment Data Statistics





Appendix 2: Final Data Structure

	hh_size	med_inc	med_rent	perc_white	perc_non_white	perc_black	perc_owner_occ	tot_pop	...	Repair Condition_1.0	Repair Condition_2.0	Repair Condition_3.0	Renovation_1.0	Renovation_2.0	t_high	t_medium
0	1.95	55160.5	873.0	0.57483	0.42516	0.23427	0.49576	461.0	...	0.00404	0.19028	0.00404	0.00000	0.0	0	1
1	1.50	54297.0	1071.0	0.66336	0.33663	0.24912	0.30475	1714.0	...	0.00628	0.06918	0.00000	0.00209	0.0	0	0
2	2.30	42778.0	1097.0	0.28077	0.71922	0.43669	0.31460	1706.0	...	0.00000	0.37102	0.00353	0.00000	0.0	0	1
3	2.69	39535.0	1152.0	0.54293	0.45707	0.30063	0.24789	3925.0	...	0.00000	0.21109	0.00135	0.00000	0.0	0	1
4	2.99	52948.0	1023.0	0.46546	0.53453	0.32620	0.18657	1824.0	...	0.00000	0.30581	0.00000	0.00000	0.0	0	1
5	2.03	25962.0	956.0	0.53026	0.46973	0.29425	0.27519	1305.0	...	0.01250	0.18750	0.00000	0.00625	0.0	0	1
	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Appendix 3: Variable Distributions

Figure 1. Number of Tests (per 1,000 persons) by Census Block Group

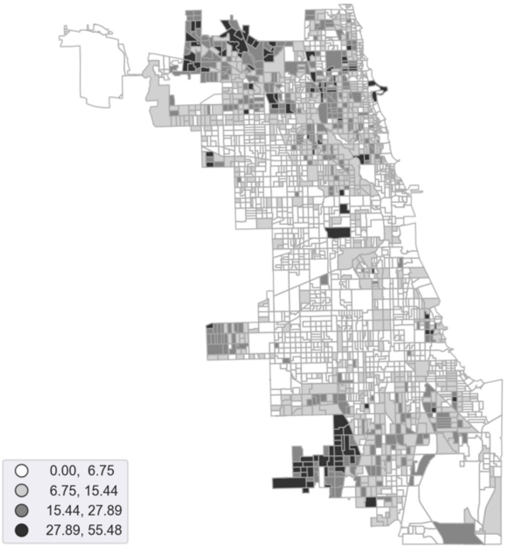


Figure 2. Median Income by Census Block Group

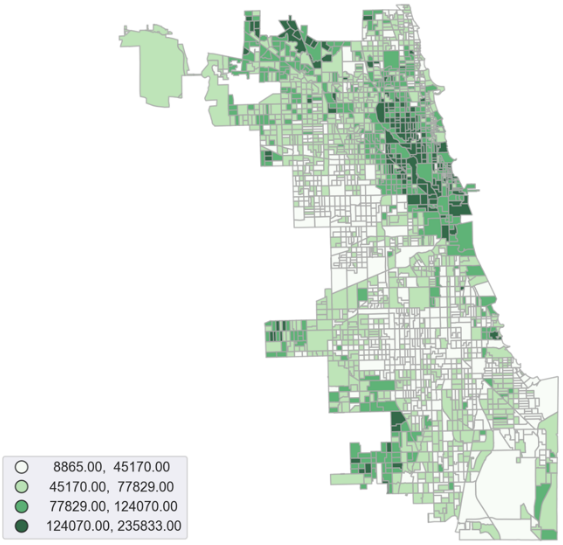


Figure 3. Population (% Black/African American) by Census Block Group

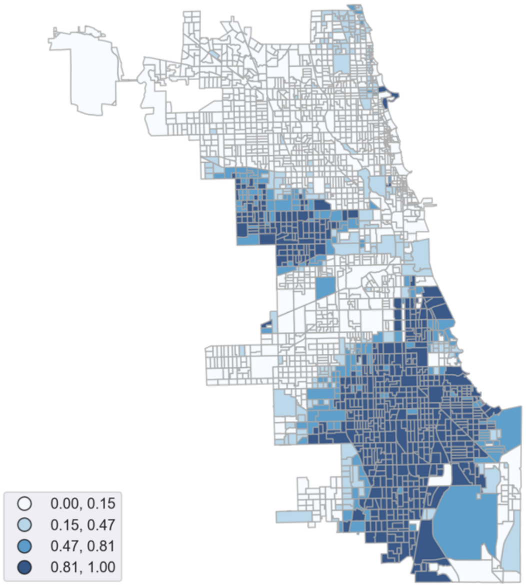
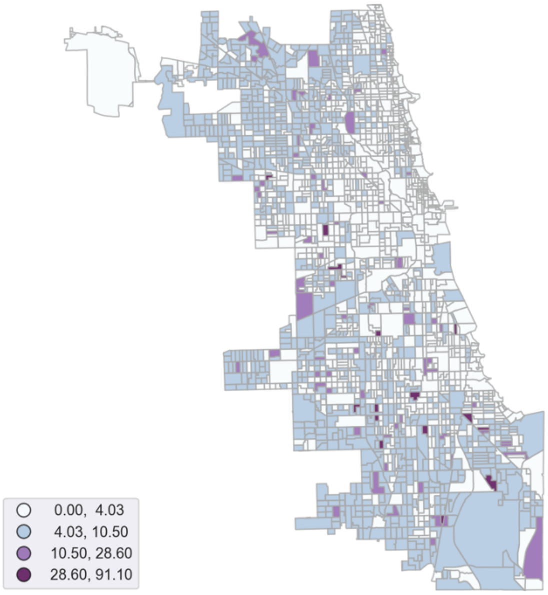
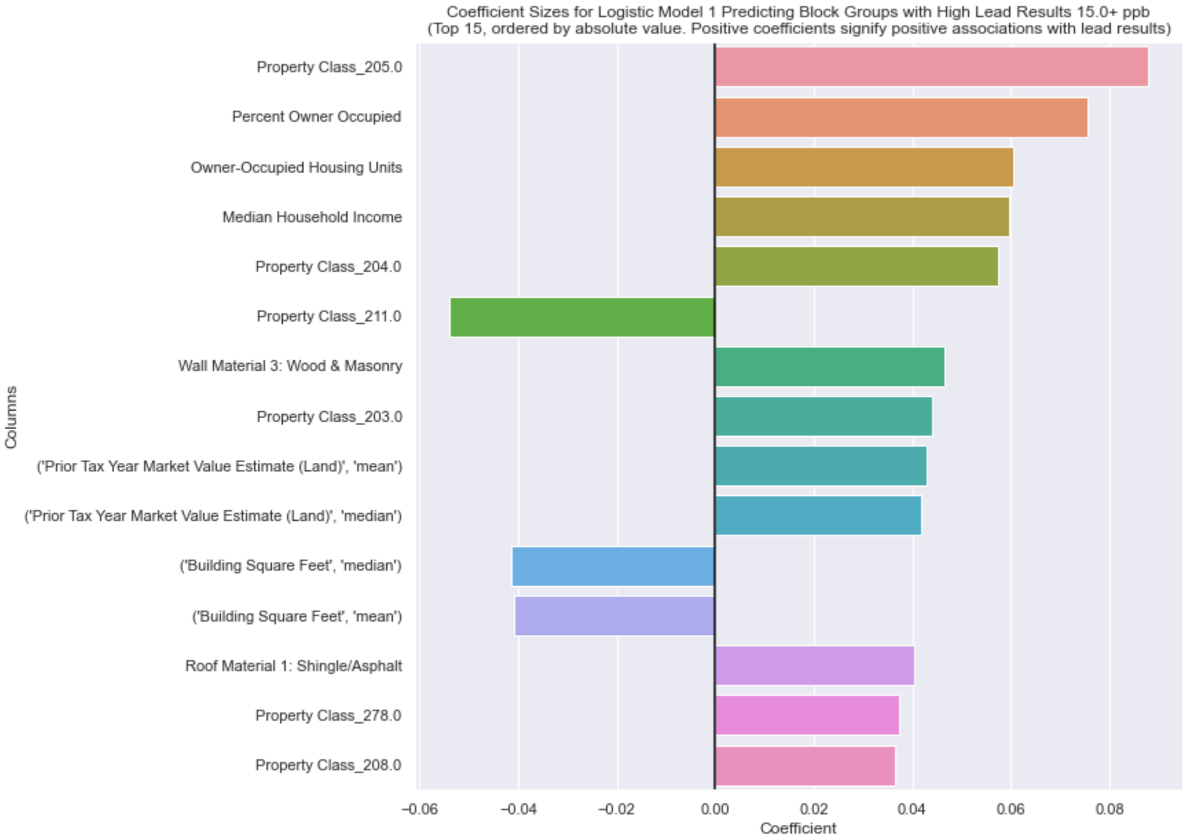


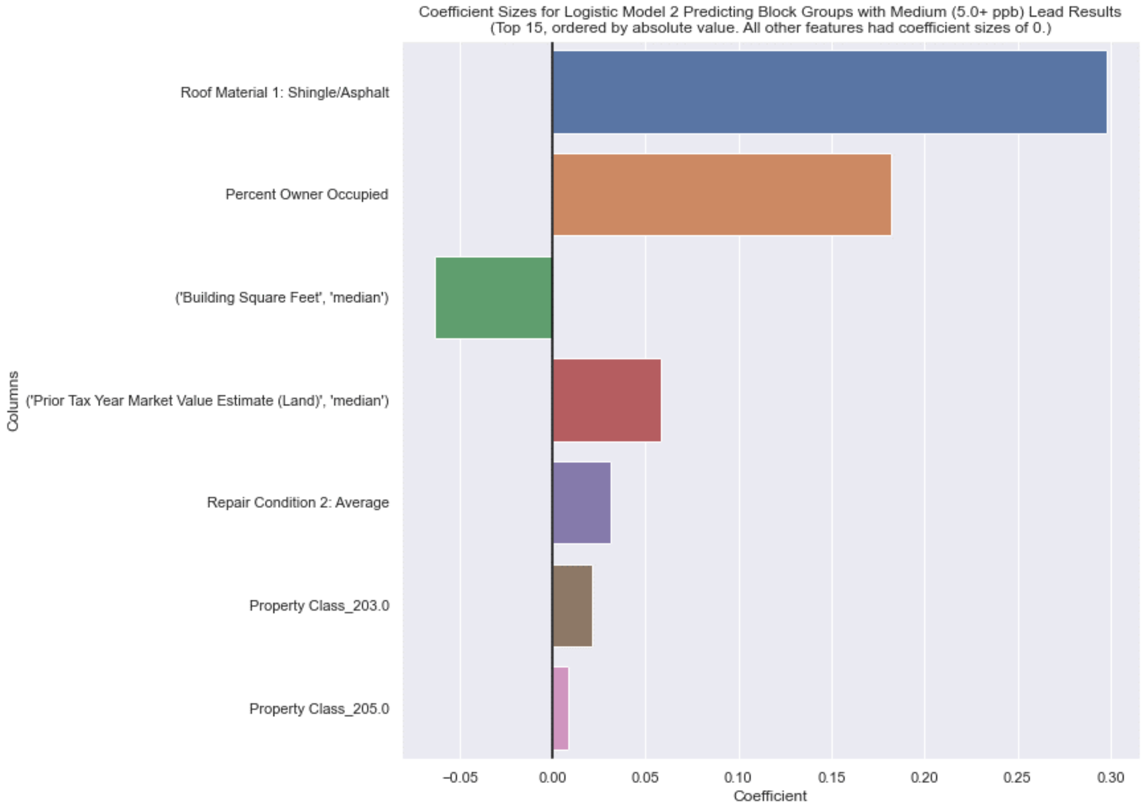
Figure 4. Mean Block Group-Level Lead Concentration (ppb)



Appendix 4: Feature Importance and Confusion Matrix – Logistic Models



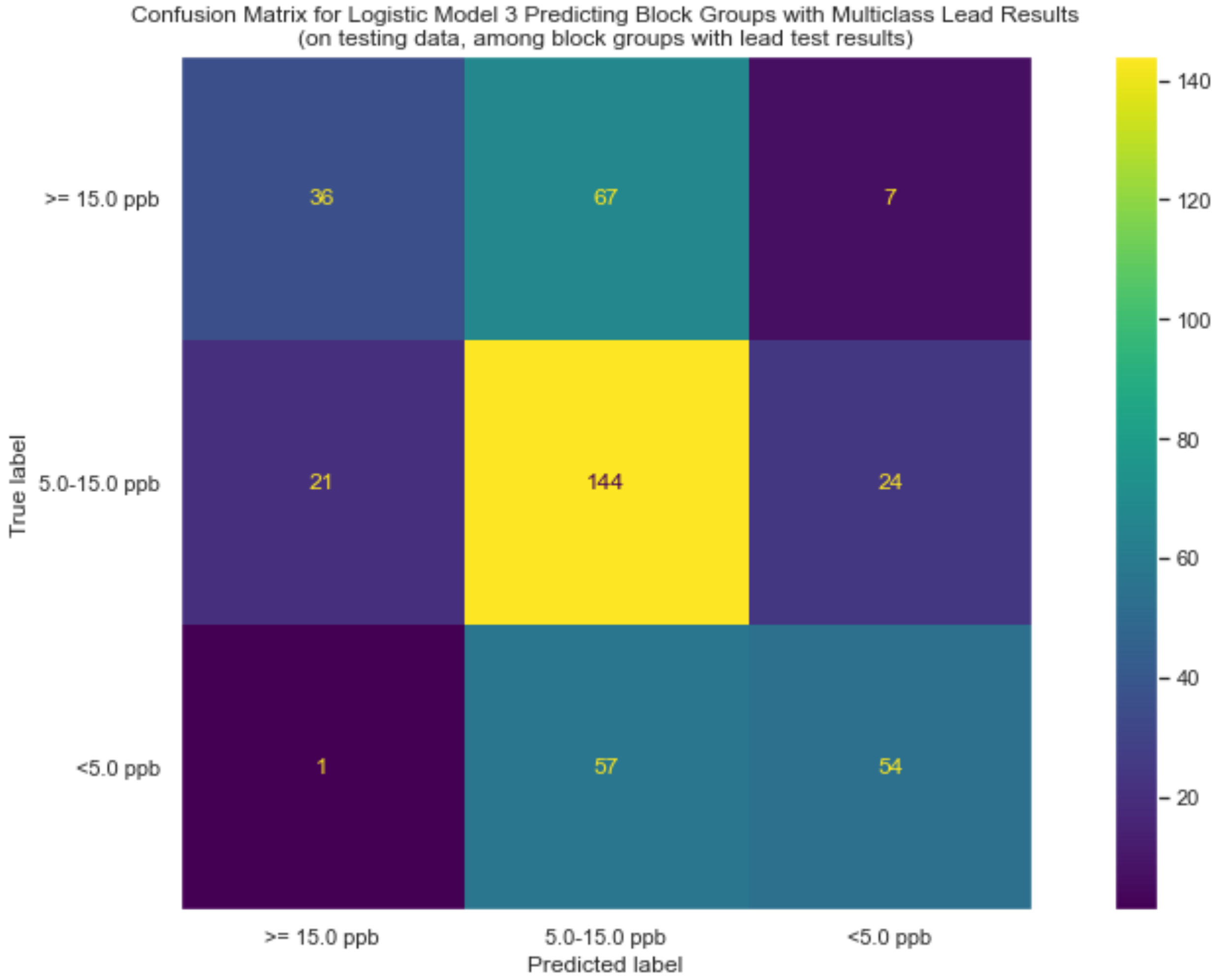
Coefficient Sizes for Logistic Model 1 Predicting Block Groups with High Lead Results 15.0+ ppb
(Top 15, ordered by absolute value. Positive coefficients signify positive associations with lead results)



Coefficient Sizes for Logistic Model 2 Predicting Block Groups with Medium (5.0+ ppb) Lead Results
(Top 15, ordered by absolute value. All other features had coefficient sizes of 0.)

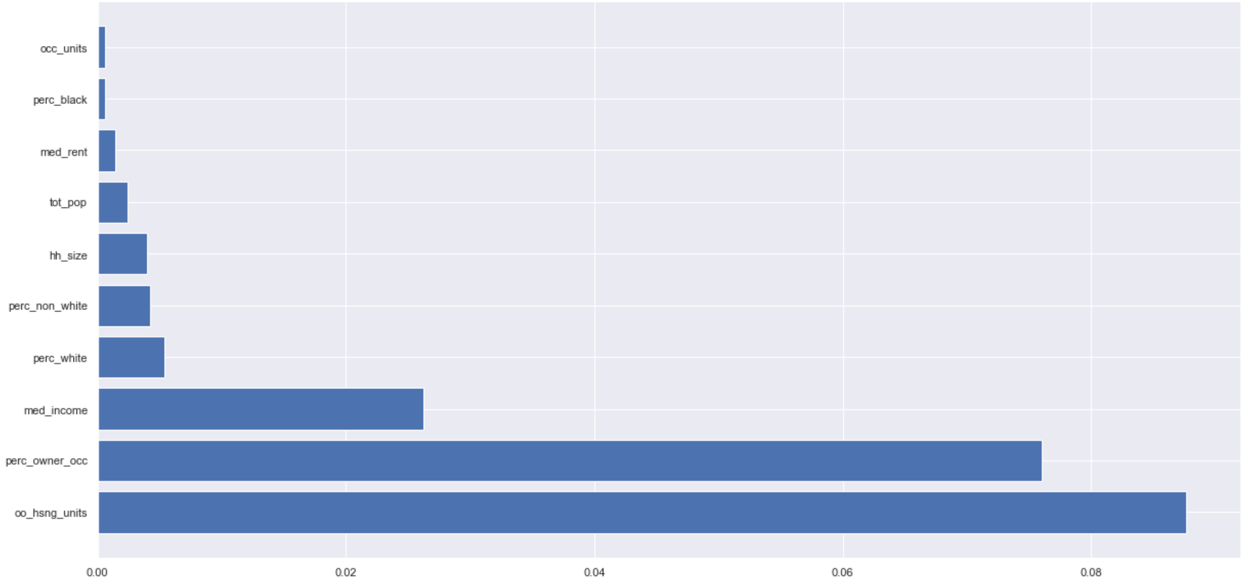
Top 15 Coefficient Sizes for Logistic Model 3 Predicting Block Groups with Highest Test Result Being Low (1.0-5.0), Medium (5.0-15.0), OR High (15.0+) Lead ppb

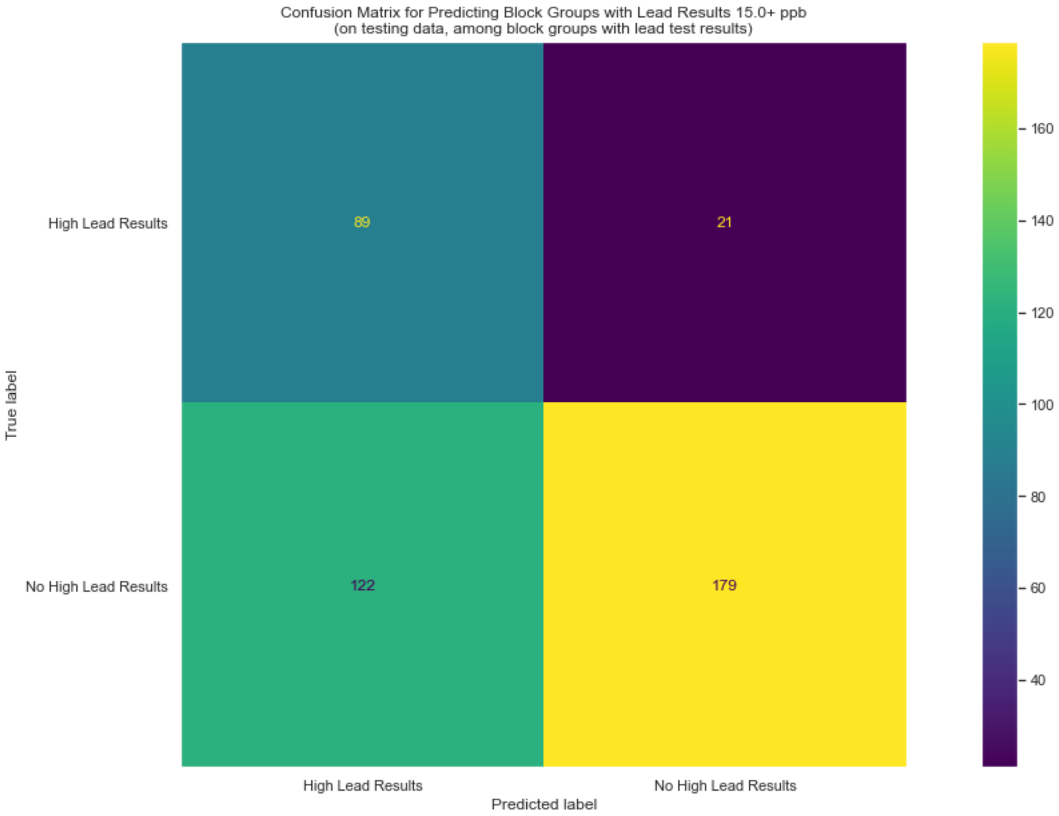
Feature	Sum of Absolute Values of Coef. Sizes	Coef. Size for Low Lead Category	Coef. Size for Medium Lead Category	Coef. Size for High Lead Category
Total Population	0.926784	-0.463392	0.122348	0.341044
(Prior Tax Year Market Value Estimate (Building), mean)	0.596959	0.29848	-0.113896	-0.184584
Percent Owner Occupied	0.592194	-0.296097	0.147656	0.148441
Household Size	0.590486	0.295243	-0.066818	-0.228425
Property Class_205.0	0.584662	-0.292331	0.055058	0.237273
Property Class_234.0	0.524771	-0.262386	0.122769	0.139616
(Prior Tax Year Market Value Estimate (Building), median)	0.493736	-0.246868	0.110057	0.136811
Owner-Occupied Housing Units	0.448366	0.112633	-0.224183	0.11155
Property Class_211.0	0.42408	0.21204	-0.058497	-0.153543
Property Class_278.0	0.410501	-0.20525	0.120129	0.085121
Percent Non-White	0.332874	0.166437	-0.030402	-0.136035
Percent White	0.332874	-0.166437	0.030402	0.136035
Property Class_295.0	0.313747	0.156874	-0.139948	-0.016925
Median Household Income	0.313664	-0.009382	-0.14745	0.156832
(Prior Tax Year Market Value Estimate (Land), (mean)	0.310838	-0.155419	0.040315	0.115104



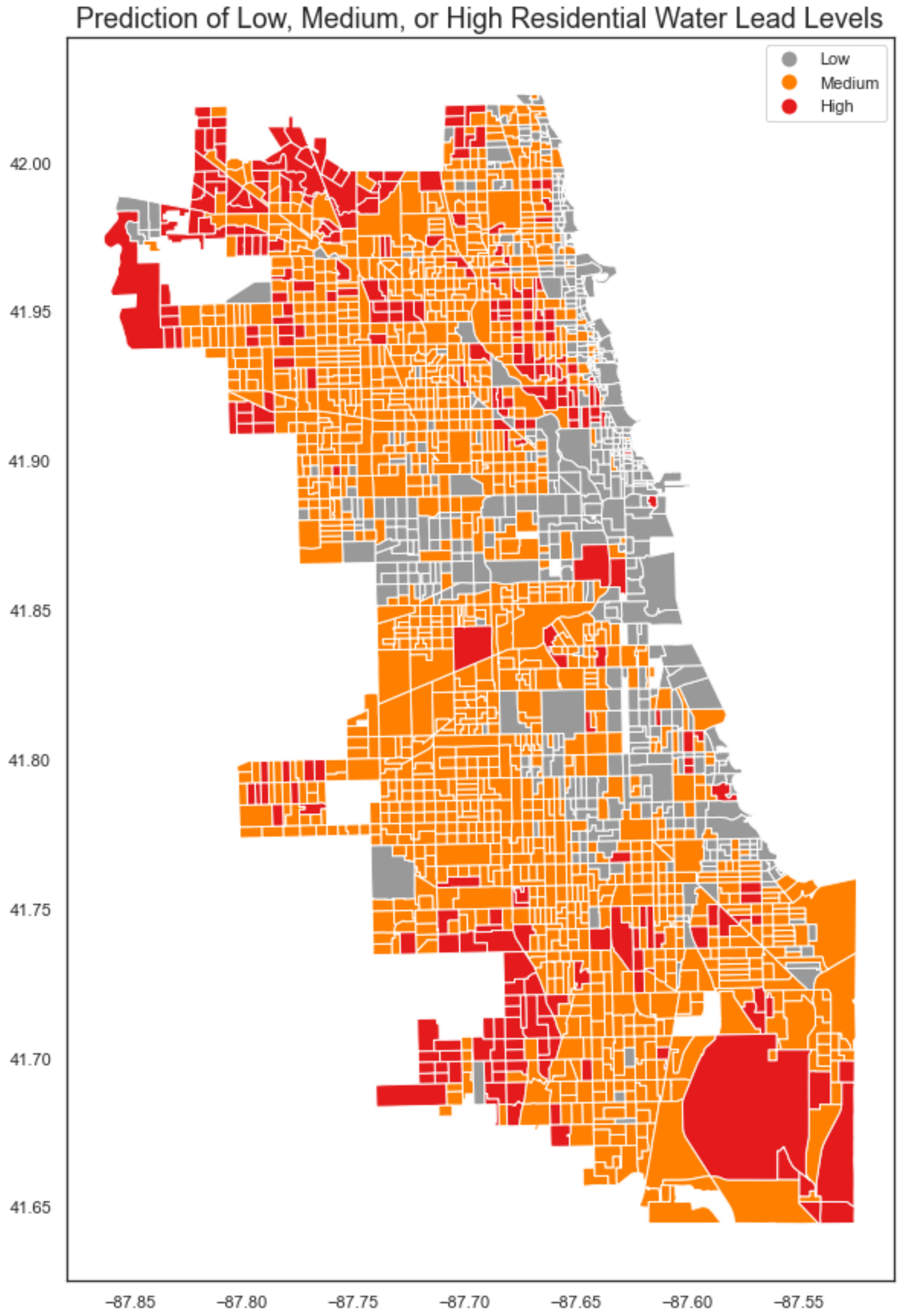
Confusion Matrix for Logistic Model 3 Predicting Block Groups with Multiclass Lead Results
(on testing data, among block groups with lead test results)

Appendix 5: Feature Importance and Confusion Matrix – Random Forests





Appendix 6: Block Group Risk Profile – Logistic Multinomial Classification


Prediction of Low, Medium, or High Residential Water Lead Levels
Low
Medium
High
42.00
41.95
41.90
41.85
41.80
41.75
41.70
41.65
-87.85
-87.80
-87.75
-87.70
-87.65
-87.60
-87.55