Lead Alert

Data Driven Guidance for Water System Improvements

Mike Amodeo, Jenn Casper, and Rob Mulla
W210 Synthetic Capstone
Final Presentation
4/18/18

Problem Definition

Lead poisoning has been a known public health issue since the 1960s, with serious health effects. However, contamination rates are still dangerously high.

Water-based lead contamination in Flint, Michigan brought renewed awareness to the problem.

How can utilities prioritize funding to identify and replace infrastructure at high risk of causing lead contamination?

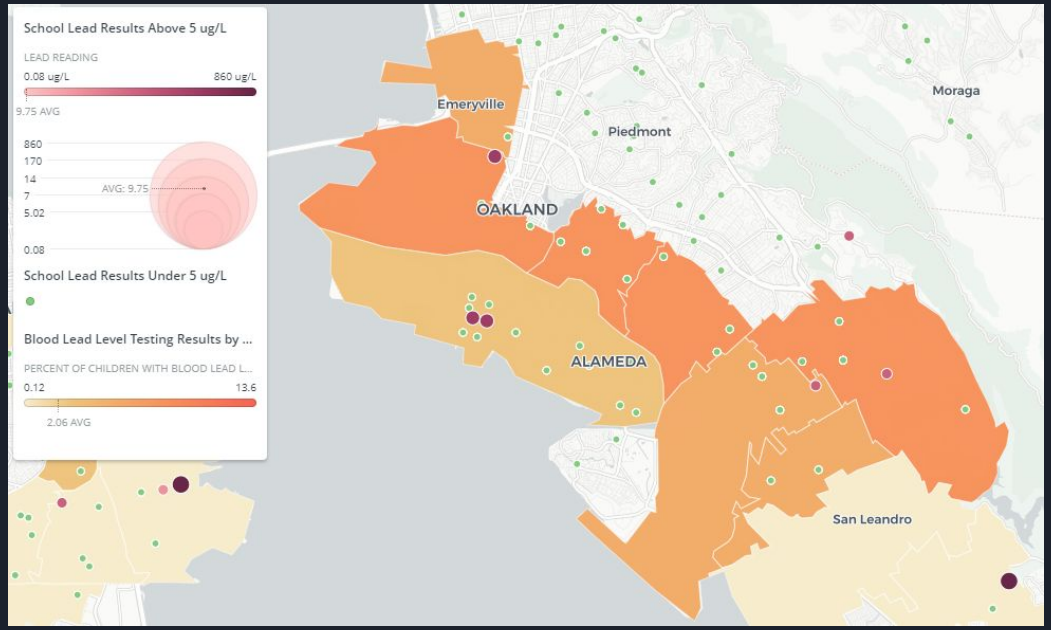
Our Solution

Predict water-based lead contamination with Machine Learning models

Difference of Our Approach

California public schools are now required to test for lead by July 2019. The beginnings of this dataset give us labeled data on the statewide scale.

Similar approaches have been used in individual cities (Flint), but not for an entire state. If extensible, this could expand to much larger areas.



leadalert.io

Contributing Factors to Contamination

	Water	Pipes
Chemical Risk Factors	Lead Corrosive Chemicals Anti-corrosion Additives Low pH (high acidity)	Lead Pipes Lead Solder on Pipe Joints
Indicators	Presence of Lead Presence of Chlorides Absence of Orthophosphates	Older Construction (pre-1984) Low Likelihood of Improvements
Data Layers	Division of Drinking Water • Chemical levels from water sampling data • Past lead violations	American Community Survey • Age of Construction • Income • Demographics • Property value

leadalert.io

Data Sources

Website Demo

Imbalanced Data

What is imbalanced data?

Scenario where the number of observations belonging to one class is significantly lower than those belonging to the other classes.

Used in:

- Fraud Detection
- Cancer Diagnosis
- Electricity Theft Detection
- High lead levels in drinking water

4.7% of the total data was labeled contaminated.

leadalert.io

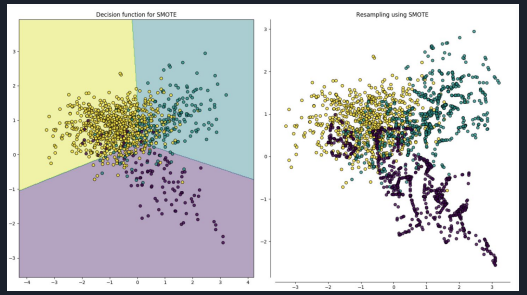
Modeling Approach

Data Leveling - Synthetic Minority Oversampling Technique (SMOTE)

Synthetic oversampling of the minority class

Ensemble Learning Method - Extreme Gradient Boosting (XGBoost)

Weak learners aggregated in a strong learner for accuracy or other metric



Model Inputs

Master data set conditioned, aggregated, and transformed to census tracts

- 8057 census tract records
- 769 total fields

Source	Topic	Number of Fields
American Community Survey (US Census)	Housing	282
	Social	164
	Economics	160
	Demographics	98
	School Enrollment	6
California Department of Water Services	School Lead Results	11
California State Water Resources Control Board	Chemical Indicators	40

Modeling Process

Ingest and Conditioning
Feature Selection
- Full (757), Subset (44)

Splitting
- Test/Train/Validation

Preprocessing
- Random Oversampling (ROS), SMOTE, None

XGBoost
- Parameter Optimization
- Early Stopping
- Cross Validation

Evaluation

	Not Lead	Lead	Total
Train	435	121	556
Test	310	100	410
Validation	224	51	275

Parameter	Values
max_depth	[5, 7, 9, 11, 13]
min_child_weight	[1, 3, 5]
colsample_bylevel	[0.1, 0.2, 0.3, 0.5, 1.0]
subsample	[0.7, 0.8, 0.9]
scale_pos_weight	[1, 1000]
learning_rate	0.1
n_estimators	1000
colsample_bytree	0.8
seed	0
objective	binary:logistic

*https://github.com/RobMulla/leadalert

Modeling Results

Feature Set	Preprocessing	Training Accuracy	Validation Accuracy	Validation Recall	Test Accuracy	Test Recall
features-subset	None	78.41%	81.82%	1.96%	75.61%	1.00%
features-subset	ROS	95.17%	80.00%	1.96%	74.15%	3.00%
features-subset	**SMOTE**	**87.59%**	**80.36%**	**13.73%**	**72.68%**	**6.00%**
features-all	None	78.05%	81.45%	0.00%	75.61%	0.00%
features-all	ROS	96.78%	81.45%	0.00%	75.61%	0.00%
features-all	SMOTE	91.95%	79.64%	0.00%	74.88%	3.00%

Parameter Optimization

Parameter	Features-subset /None	Features-subset /ROS	Features-subset /SMOTE	Features-all/ None	Features-all /ROS	Features-all /SMOTE
learning_rate	0.1	0.1	**0.1**	0.1	0.1	0.1
n_estimators	1000	1000	**1000**	1000	1000	1000
colsample_bytree	0.8	0.8	**0.8**	0.8	0.8	0.8
seed	0	0	**0**	0	0	0
objective	bin:log	bin:log	**bin:log**	bin:log	bin:log	bin:log
colsample_bylevel	0.1	0.1	**0.1**	0.1	0.2	0.1
max_depth	11	7	**11**	5	7	5
min_child_weight	1	1	**1**	1	1	1
scale_pos_weight	1	1	**1**	1	1	1
subsample	0.9	0.9	**0.9**	0.9	0.8	0.7

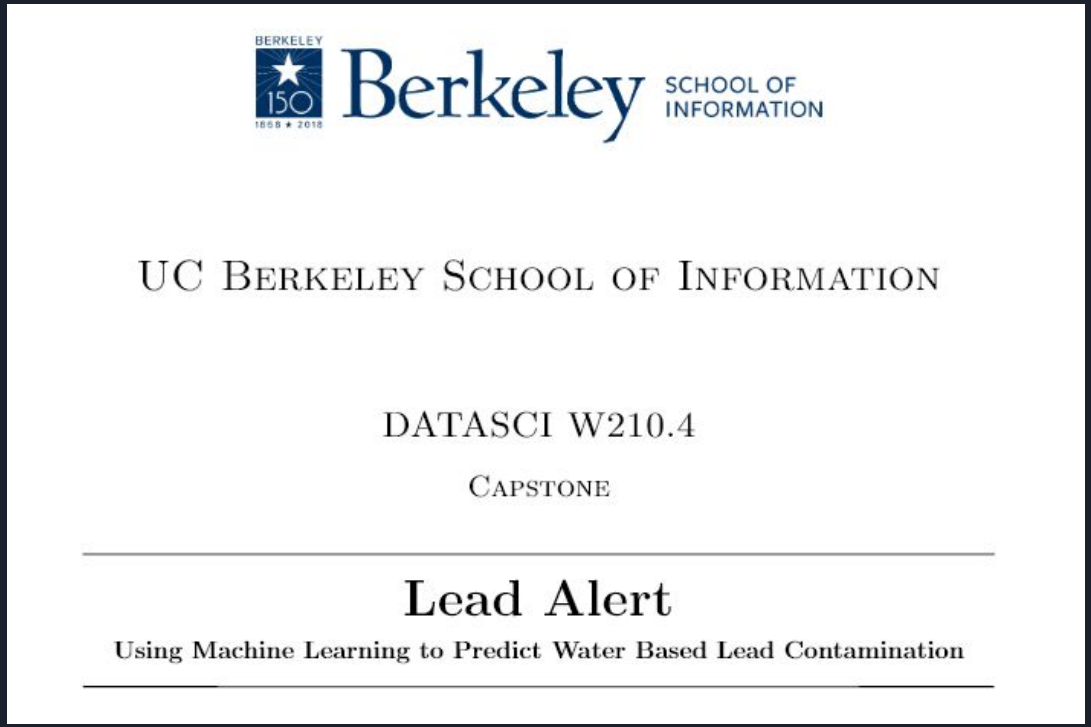
Feature Importance

ACS Feature Name	F1 Score
Percent; YEAR STRUCTURE BUILT - Total housing units - Built 2000 to 2009	184
Percent; YEAR STRUCTURE BUILT - Total housing units - Built 1939 or earlier	179
Estimate; INCOME AND BENEFITS (IN 2016 INFLATION-ADJUSTED DOLLARS) - Total households - Mean household income (dollars)	176
Estimate; YEAR HOUSEHOLDER MOVED INTO UNIT - Occupied housing units - Moved in 1990 to 1999	175
Estimate; INCOME AND BENEFITS (IN 2016 INFLATION-ADJUSTED DOLLARS) - Total households - Median household income (dollars)	175
Percent; YEAR STRUCTURE BUILT - Total housing units - Built 1960 to 1969	173
Estimate; SEX AND AGE - 60 to 64 years	168
Estimate; SEX AND AGE - 85 years and over	166
Estimate; YEAR HOUSEHOLDER MOVED INTO UNIT - Occupied housing units - Moved in 2000 to 2009	161
Estimate; SEX AND AGE - 10 to 14 years	157

Feature Selection

Threshold	Number of Features	Training Accuracy	Test Lift	Test True Positives
0.000	757	74.88%	1.64	6/100
0.002	287	74.88%	1.64	6/100
0.003	137	75.12%	1.54	3/100
0.005	80	74.63%	1.37	4/100
0.006	48	74.63%	1.54	6/100
0.008	**27**	**75.12%**	**1.82**	**8/100**
0.010	18	72.93%	1.27	9/100
0.011	11	73.90%	1.29	6/100
0.013	8	73.17%	0.77	3/100
0.014	4	74.63%	1.46	5/100
0.016	**2**	**75.12%**	**1.76**	**6/100**

Lead Alert Github & Website



Berkeley SCHOOL OF INFORMATION

UC Berkeley School of Information

DATASCI W210.4

Capstone

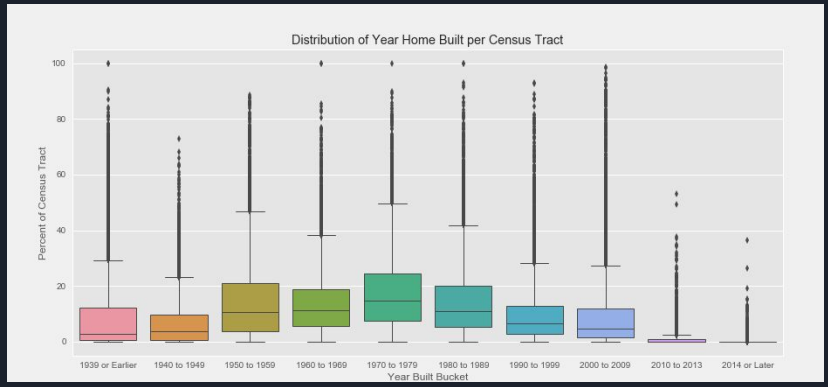
Lead Alert

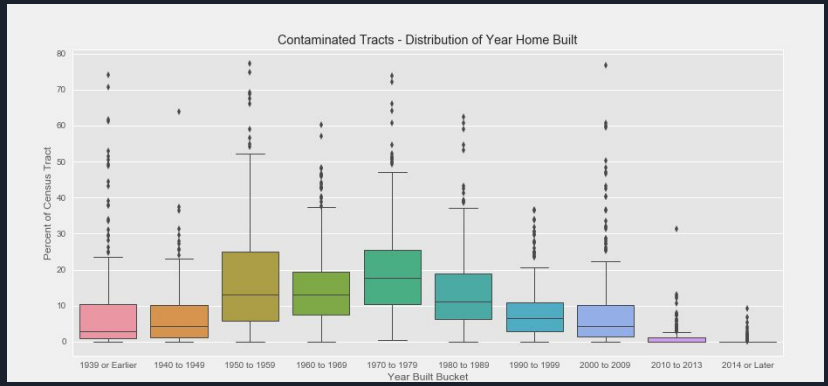
Using Machine Learning to Predict Water Based Lead Contamination

leadalert.io

Conclusions & Future Work

- Characteristics at the census tract level did not exhibit strong predictive power
- Better accuracy may be obtained by prediction of individual buildings using spatial clustering techniques
- Building a dataset of individual structures across many jurisdictions requires much effort. This would be helped by state (and federal) data standards for county assessor data storage
- Demand for product exists among water system managers





leadalert.io

Lead Alert will provide data-driven guidance for water providers and regulators to prioritize water quality system improvements

leadalert.io

Questions?

Mike Amodeo amodeo@berkeley.edu

Jenn Casper jenncasper@berkeley.edu

Rob Mulla robmulla@berkeley.edu

leadalert.io