UC Berkeley School of Information

DATASCI W210.4

Capstone

Lead Alert

Using Machine Learning to Predict Water Based Lead Contamination

Authors:
Michael Amodeo
Jennifer Casper
Robert Mulla

Emails:
amodeo@berkeley.edu
jenncasper@berkeley.edu
rob.mulla@berkeley.edu

April 16, 2018

Contents

1 Introduction

Lead has been known to be harmful to humans even in small doses for over 50 years[1]. Exposure to even low levels of lead can result in damage to the central and peripheral nervous system, learning disabilities, impaired function of blood cells, stunted growth, cardiovascular effects, and many other problems. Despite regulations restricting lead in building materials, children are still experiencing lead poisoning at alarming rates. While the Flint water crisis has brought renewed attention to this issue, investigations have shown many communities across the country with rates of lead poisoning in children exceeding that of Flint[2]. The persistence of this major public health problem and the creation of new relevant datasets create an opportunity to apply new thinking and techniques to solve it.

The creation of new datasets give us a chance to redefine the prioritization method of infrastructure improvements. Prior studies have largely focused on an individual city - Flint, Michigan. In this paper, we attempt to using machine learning to predict areas of California where communities are at highest risk of lead exposure.

This project was undertaken as part of W210: Synthetic Capstone within the Master of Information and Data Science program at the UC Berkeley School of Information. The project was conceived as a product delivered to water system managers throughout California via a website, `LeadAlert.io`. The website contains additional results and descriptions of the work, as well as visualizations of the data used. The public Github repository, `https://github.com/RobMulla/leadalert`, contains the collection of aggregated data sets, source code, modeling artifacts, and associated reference material.

1.1 Chemical Process Background

While it is possible for lead to be in the water supply source, lead typically enters the water system from pipes and soldered joints either in service lines or within the home. This is the result of a chemical corrosion process where chemicals in the water react with the pipe materials to leach lead into the water.

[1] `https://www.epa.gov/ground-water-and-drinking-water/basic-information-about-lead-drinking-water`

[2] `https://www.reuters.com/investigates/special-report/usa-lead-testing/`

Orthophosphate is commonly added to the water supply at treatment plants and distribution locations to build a layer on pipe walls to protect against corrosion, thereby sealing the lead into the pipe. Where the orthophosphate layer breaks down, dissolved oxygen attaches to the pipe wall. The oxygen then reacts with the lead in the pipe wall to oxidize the lead. Oxygen then bonds with hydrogen, leaving the lead in its oxidized state. The oxidized lead will now dissolve into the water stream, contaminating the water on its path of delivery. Other chemicals like chloride will accelerate corrosion[3].

From this chemistry, we were able to identify several risk factors - both the chemicals in the water and the properties of the pipes. Statewide California water supply testing data gave us levels of these chemicals (including lead, orthophosphates, and chloride) in the water at various distribution sampling locations. This was tracked per water district.

We can also identify locations that are likely to have pipes with more lead as potential risk factors. Lead pipes are more common in communities that were built a long time ago and have not been renovated since the Environmental Protection Agency (EPA) banned lead materials in construction. We are using census data from the American Community Survey (ACS) to represent many risk factors including age of construction, property value, and demographics. The factors have been shown in previous studies to be predictive of lead risk.

1.2 California Water Systems

Water supply systems in California are typically operated by private companies. They are not public utilities. Because of this, they often cross jurisdictions of cities and towns. The State Water Resource Control Board (SWRCB) regulates these water suppliers, providing oversight and review of water quality and allocation. EPA regulations specify water quality standards and testing procedures. This testing data is submitted to the EPA, but the Division of Drinking Water (DDW) within the SWRCB maintains much of this data specifically for California.

[3] https://cen.acs.org/articles/94/i7/Lead-Ended-Flints-Tap-Water.html

2 Data

2.1 Lead Sampling of Drinking Water in Schools

In 2017, the state of California passed Assembly Bill 746 requiring all public schools in the state to test their water fixtures for lead by July 2019. A previous law had offered financial assistance for testing, but not all schools took advantage of it. The test results from the previous program and early returns on the mandatory program are available in a public database providing some coverage of the state. This newly available dataset offers up recorded data of lead levels at the downstream end of water supply systems all throughout the state.

Within the California State Water Resource Control Board, the Division of Drinking Water is responsible for maintaining the data related to Assembly Bill 746. Lead Alert received a copy of this data in Excel spreadsheet format in February, 2018. More information on AB746 can be found on the state?s website[4].

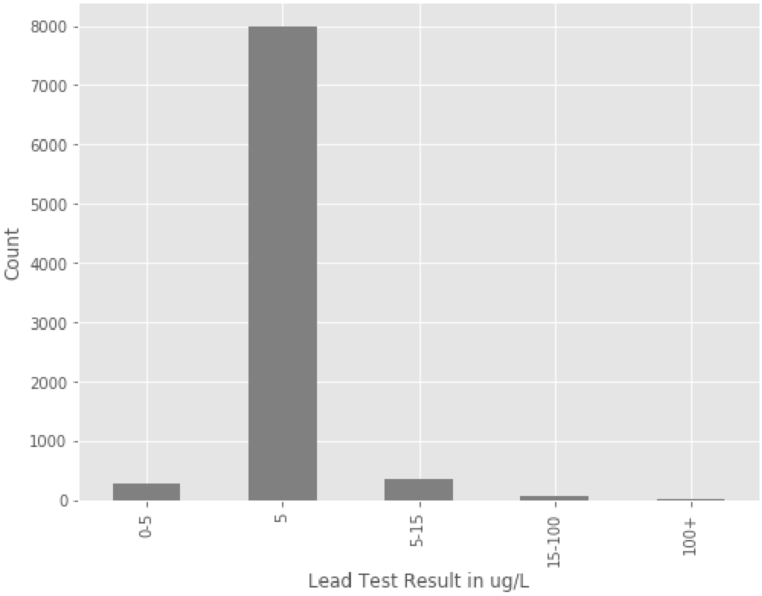

Figure 1: Distribution of lead test results.

[4] https://www.waterboards.ca.gov/drinking_water/certlic/drinkingwater/leadsamplinginschools.html

The severe consequences of lead poisoning are what makes it a public health hazard, not the frequency of poisoning. Overall, only a small percentage of test results show high levels of lead. Of the 8,688 readings in the California database, only 415 (4.7%) had levels above 5 µg/L, and 72 (< 1%) were above the EPA's action level of 15 µg/L. For machine learning prediction, this presents an imbalanced dataset.

Additionally, much of the data showed only that the reading was below a threshold, primarily the common threshold of 5 micrograms of lead per liter of water. This complicated how we could use the data, in that we could not calculate mean lead values or other statistical values. Instead, we classified each point as below 5 µg/L, between 5 and 15 µg/L, and above 15 µg/L. Figure 1 shows the distribution of records based on these thresholds.

2.2 Water Supply Testing

The Division of Drinking Water also maintains the water quality testing database[5] used for chemical indicators. Water supply sampling locations were linked to water districts and ultimately census tracts served. Indicators explored included:

- Lead
- Orthophosphates, used as anti-corrosion additives in treatment
- Chlorides, corrosive chemicals that accelerate the process of lead contamination

Many records within this data also used thresholds. For modeling purposes, we created aggregated statistics per census tract of maximum value, minimum value, and mean value. All were included in the model to see if any were predictive.

2.3 American Community Survey (ACS)

Lead Alert utilizes several common tables from the American Community Survey[6]. All data was from the 2016 5-year estimates, the most recent ACS data available. Tables used include the following:

[5] https://www.waterboards.ca.gov/drinking_water/certlic/drinkingwater/EDTlibrary.html

[6] https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

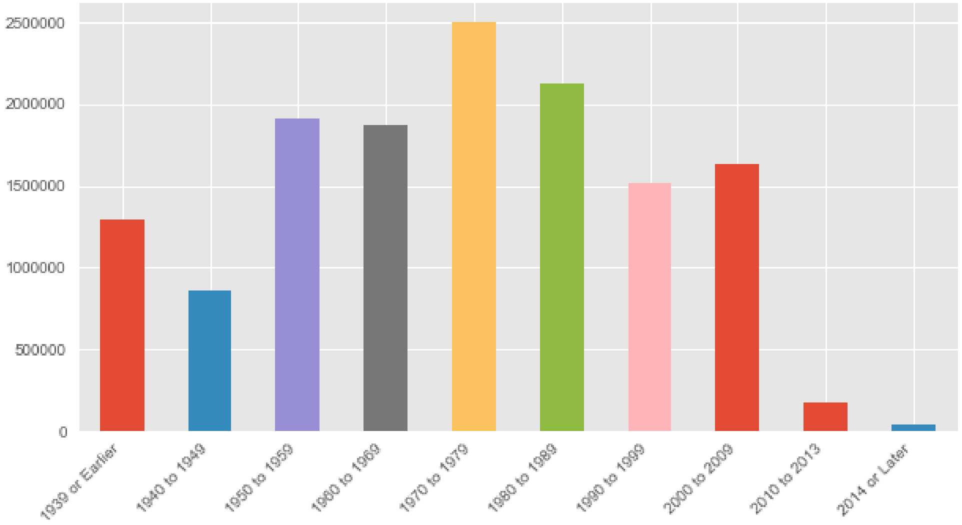

Figure 2: Estimated number of homes in California by year built.

- DP02 Selected Social Characteristics
- DP03 General Economic Characteristics
- DP04 General Housing Characteristics
- DP05 Demographic and Housing Estimates
- S1401 School Enrollment

These tables include demographic data as well as housing data. The housing data includes estimates of the age of construction for buildings in the census tract among many other indicators. Age of construction has been shown to be a major indicator of risk of lead contamination. The majority of homes in California were built before 1980 (see Figure 2).

2.4 Data Aggregation and Transformation

The data came in many formats, and the primary challenge was to aggregate all of this information across a common geographical unit. The census tract was the predominant unit across the sets and represented the underlying infrastructure and construction information. The initial challenge was associating the lead sampling data with a geographical location in order to connect with a water source. The sampling data contained the school name and water system name that were complete. A publicly available California schools list, complete with address, latitude, and longitude, and a California water system shapefile provided the additional geographical information needed.

We iteratively employed a series of text matching methods across the school name, water system name, and county information to align best matches within the California schools list. This provided the majority of matches for associating a latitude and longitude, while also identifying redundant school names. If unmatched schools included addresses, the addresses were geocoded using the Google Geolocation Service. The remaining schools were manually searched to uncover the most probable addresses for geolocating.

Now with a geolocated lead sample data collection, we used a similar approach when mapping each record with a water system identifier found in the shapefile. The majority of records had a latitude/longitude and water system name that aligned with the shapefile. For records that matched multiple water systems, due to the overlap of water system service areas, county information was used to deconflict.

The California census tract shapefile provided the geographical outline for each tract with the unique tract identifier. The latitude/longitude for each lead sample record was used to pair the record with the tract identifier. The tract identifier aligned the gelocated lead sample data with associated water system identifier. In turn, the water system identifier aligned each record with the water supply testing data.

The combined lead sampling, water supply testing, and ACS data required a series of joins and calculations to capture max, mean, min, percentage, etc. for each census tract. The final transformation resulted in the training data set that contained water supply testing and ACS features with a lead contamination result for each census tract in California.

3 Architecture

We utilized a series of applications and Amazon Web Services (AWS) throughout the iterative phases of this effort: data acquisition, exploratory data analysis (EDA), conditioning and aggregation, modeling and testing, and information serving. Figure 3

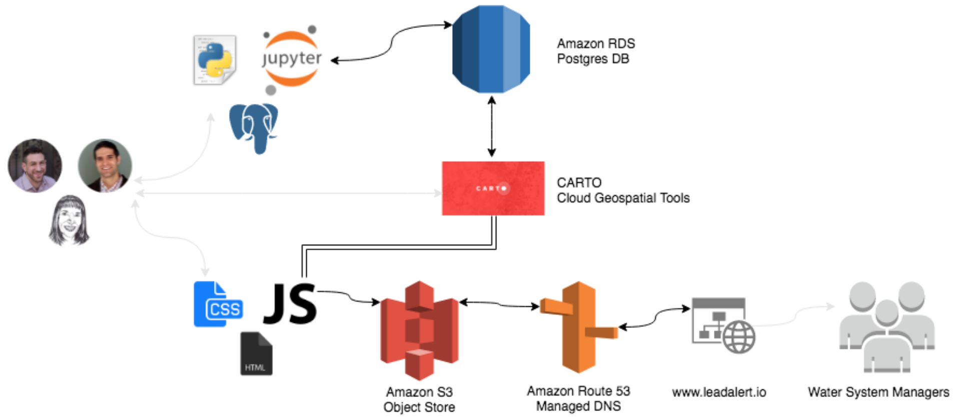

Figure 3: Architecture of resources used to support: 1) data acquisition, EDA, conditioning and aggregation, 2) modeling and testing, and 3) serving information.

shows how the resources are connected for use by the developers and consumption by the users.

Data was acquired manually from the sources described above and explored using a combination of Python, Jupyter, and Postgres applications connected to an Amazon Relational Database Service (RDS) Postgres database instance. The Postgres database provided a way to manage and track the individual sources throughout conditioning and aggregation using table versions. Modeling was conducted using Python libraries within Jupyter. The modeling and testing results are stored in Postgres in preparation for information serving.

CARTO, a service for cloud-based geospatial tools, provided the means to create geospatial views of the data versions and modeling results for the creators. Final geospatial views were created using the desired set of modeling and testing results stored in the Postgres database. The views were then linked from the Lead Alert website for the water system managers to consume.

The Lead Alert website is served up using AWS S3 and Route 53. The HTML and Javascript code, stored in S3, utilizes the w3.CSS framework. Route 53 manages the domain name service for the registered leadalert.io domain. As more data becomes available in future work, updates can be distributed through the Postgres database, CARTO, and the web interface.

4 Modeling

The following subsections outline the modeling process. Further details may be obtained in the Jupyter notebooks found in the Lead Alert Github repository. Each notebook relies on access to the aggregated and transformed data collection stored in RDS. It contains the 8,057 records, one for each California census tract, with 762 fields.

4.1 Approach

The data being collected for AB 746 provided a new dataset that had not previously been available at such a large scale. Water supply testing is done by water providers, but it is typically performed upstream. Downstream testing is only performed once a problem is identified, and the data is not usually publicly available. Lead Alert recognized these test results as potential labels for a machine learning model, hoping that they would allow us to predict where lead contamination may be entering the system from the conveyance system itself.

Previous studies[7] have shown that characteristics related to housing and construction, demographics, and property values are indicators of lead contamination for individual buildings or properties. These studies were mostly focused on a single city - Flint, Michigan. Lead Alert tried to increase the footprint of prediction to a much larger area, starting with the state of California and potentially proceeding to the United States. To do this, Lead Alert used data from the ACS as feature data to predict risk of lead contamination at the census tract level, a standard geometry defined by the United States Census Bureau to roughly represent a neighborhood. Census tracts have large amounts of aggregated data readily available. Because this data is standardized throughout the United States, if the features were shown to be predictive, the model would potentially be very easily extensible to the entire United States.

Lead Alert also attempted to use the upstream water supply testing data as features in prediction. This data includes upstream water quality tests for chemicals that are key elements in the lead contamination process, like orthophosphates and chlorides.

The use of the census tracts and schools relied on the assumption that each were representative of their communities. Schools are built along with the rest of a neighborhood

[7] “1707.01591] A Data Science Approach to Understanding Residential ....” 5 Jul. 2017, https://arxiv.org/abs/1707.01591. Accessed 10 Apr. 2018.

or town and presumably would have similar construction age, methods, and materials. The census tract is a standard unit of aggregation designed by the Census Bureau to represent a community, varying in size but representing a somewhat fixed number of people. Schools also can serve a population of a certain size. Also, we are not relying on the people in the tract to enroll in the school, as the testing data is the testing of the pipes in the school, not the students.

Lead Alert did not use any available data regarding historical lead violations per water district or blood testing results because these are the sort of results we intended to predict. We deemed these not to be valid inputs that would possibly corrupt the model.

4.2 Ingest and Conditioning

After acquiring the data from Postgres, we conditioned it in preparation for modeling:

- Drop records with null in the prediction field, `lead_result_max_bucket`
- Convert strings to numerical values
- Code categories as numerical values
- Add a percentage threshold field for each of the seven water chemicals
- Convert null to 0 as XGBoost does not handle null values

4.3 Feature Selection

Two sets of features were used for the modeling: the full feature set and a feature subset. The full feature set contained 757 features, which included the chemical threshold percentages but did not include the lead result field. The feature subset contained a total of 44 features, also including the chemical threshold percentages but no lead result data. The subset fields were selected during the initial exploratory data analysis. The selected fields were observed as having higher performance when comparing ensemble models. These fields include data related to age of construction, population age, and the year residents moved into their current residence.

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

Table 1: Record and prediction value counts for split data sets.

4.4 Splitting

The resulting data set was randomly split into training and test data with the test set being a third of the total size. The training data was then split again into training and validation with validation being a third of the total training size. The validation split further reduced the training size, although this split allowed for the comparison of validation and test results for model stability. Table 1 summarizes the data set sizes and prediction value counts.

4.5 Data Imbalance and Preprocessing

To work with the data, Lead Alert employed both data leveling and ensemble learning techniques[8]. Given the imbalanced nature of the data, preprocessing options were considered to help highlight the characteristics of the minority class that would otherwise be treated as noise[9]. Depending upon the size of the data set, over-sampling and under-sampling methods are available.

Data leveling techniques like synthetic minority over-sampling technique (SMOTE) create synthetic instances of the minority class (in this case high lead test results) based on the features of actual instances in the minority class. This avoids overfitting that would result from strictly over-sampling. The relatively small size of the aggregated lead contamination data lent itself to over-sampling of the minority class. Random over-sampling (ROS) and SMOTE were employed on the training data using the Imbalanced-Learn library[10], along with no preprocessing for comparison[11].

[8] https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/

[9] https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/

[10] https://github.com/scikit-learn-contrib/imbalanced-learn

[11] https://beckernick.github.io/oversampling-modeling/

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

Table 2: Record and prediction value counts for split data sets.

4.6 XGBoost with Parameter Optimization, Early Stopping and Cross Validation

Boosting based ensemble modeling methods tend to perform well on imbalanced datasets as they create many weak learners that in aggregate create a strong learner capable of making accurate predictions. XGBoost (Extreme Gradient Boosting) is an efficient implementation of a boosting method whereby the loss function of previous learners is used when building additional learners, increasing both speed and accuracy.

XGBoost has several parameter options for optimization[12]. Using GridSearchCV, the XGBoostClassifier was trained on the parameters and values listed in Table 2 using a 10-fold cross validation scored by accuracy.

Once the best parameters were calculated, a XGBoost classifier was trained using the training data on the best parameters. Ten-fold cross validation and early stopping using the error metric were employed to help prevent overfitting. The boosted round number from the optimized XGBoost classifier as well as the best parameters were captured for defining final model to be used in evaluation.

[12] `https://www.dataiku.com/learn/guide/code/python/advanced-xgboost-tuning.html`

4.7 Evaluation

With the final model ready for evaluation, we predicted using the validation and test sets. The prediction outcome was a set of probabilities. We evaluated several probability thresholds but have chosen a 0.75 prediction probability cut-off for indicating lead contamination. A 0.50 cut-off is more common, but we wanted to remain conservative given the nature of the problem space. We wanted locations that we predicted as potentially at risk to be very likely at risk, not just more likely than not. The true positive and false positive rates may be used to calculate the optimal threshold for the training data, although this did not prove beneficial in testing.

The predictions were then compared to the target values for the validation and test sets. We chose accuracy with error rate, recall, confusion matrix, lift, and area under the curve (AUC) metrics to inspect model performance. Accuracy and error rate provide evaluation of overall performance while the recall and confusion matrix provide insight into how well the predictions matched the target values. Lift is a scoring function to understand the overall ratio of predicted positive data to actual positive data.

5 Results and Findings

Aggregation and transformation methods may be extended to other geographical units providing greater predictive power. The census tract unit was chosen as it was the smallest unit for which the standardized housing data was available at the statewide scale. The census tract is representative of the population and provided a pseudo population normalizing view. For instance, a tract in an urban area will be geographically smaller than one in a rural area due to population density. The ACS data was already provided at the census tract level. The chemical and lead results required transformation. Transforming the data to different geographical levels may reveal alternative insights and understanding of the underlying data.

Modeling the aggregated ACS, water chemical and lead results data using the methods described above failed to produce metrics indicating confidence in a predictive model. The most performant method in terms of validation and test accuracy was the model using a subset of features without preprocessing. The recall for this model was poor, however, at 1.96%. The model using a subset of features with SMOTE had a slightly lower validation and test accuracy although the recall

Feature Set	Pre-processing	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

Table 3: Resulting metrics for the six models.

results were significantly higher. This model proved to be the most performant given the methods and options described above and the aggregated data collection. While performant under the circumstances, results do not provide confidence in prediction. Again, if 80% of tracts did not show any testing results above 5 micrograms of lead per liter of water, naively predicting all tracts as not contaminated would yield an 80% accuracy. Therefore, 80% became our baseline for performance. See Table 3 for the resulting metrics from the six models.

Although XGBoost parameter optimization was consistent across the models compared in the table above, there exist several additional options to explore that may improve performance[13]. Table 4 presents the given and optimized parameters for the XGBoost modeling. The parameters were chosen for optimization given lessons and recommendations from references on ensemble methods for imbalanced data. Across the models compared, the `max_depth`, maximum tree depth for base learners, saw the greatest variation. `Max_depth` and `min_child_weight` impact the model complexity. `Colsample_bylevel` and `subsample` add randomness to impact model robustness. Additional parameters to consider for an imbalanced data

[13] https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

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

Table 4: Optimized parameter values for the six models.

set include `gamma` and `max_delta_step`[14].

Feature importance varied over all models. However, the most performant model revealed the influence of building age, income level, occupation year, and population age. Table 5 presents the list of the top ten most influential features for the performant model. The performant model was created using the feature subset (44 out of 757 total features). The feature importance profiles for the subset feature models varied in terms of influential features and F1 score profiles. The complete feature models were widely inconsistent for influential features and results in low F1 scores spread across the hundreds of features.

Feature selection may prove to be another area of improvement for predicting lead contamination risk. We spent time studying feature selection as an impact on modeling. Using XGBoost without preprocessing or discrete parameter op-

[14] `http://xgboost.readthedocs.io/en/latest/how_to/param_tuning.html`

ACS Feature Name	
Model Feature Name	F1
Percent; YEAR STRUCTURE BUILT - Total housing units - Built 2000 to 2009	
pct_yr_blt_units_built_2000_to_2009	**184**
Percent; YEAR STRUCTURE BUILT - Total housing units - Built 1939 or earlier	
pct_yr_blt_units_built_1939_or_earlier	**179**
Estimate; INCOME AND BENEFITS (IN 2016 INFLATION-ADJUSTED DOLLARS) - Total households - Mean household income (dollars)	
est_inc_tot_hshld_mean_household_inc_dol	**176**
Estimate; YEAR HOUSEHOLDER MOVED INTO UNIT - Occupied housing units - Moved in 1990 to 1999	
est_year_householder_moved_into_unit_occ_unts_1990_to_1999	**175**
Estimate; INCOME AND BENEFITS (IN 2016 INFLATION-ADJUSTED DOLLARS) - Total households - Median household income (dollars)	
est_inc_tot_hshld_median_household_inc_dol	**175**
Percent; YEAR STRUCTURE BUILT - Total housing units - Built 1960 to 1969	
pct_yr_blt_units_built_1960_to_1969	**173**
Estimate; SEX AND AGE - 60 to 64 years	
est_sex_and_age_60_to_64_yrs	**168**
Estimate; SEX AND AGE - 85 years and over	
est_sex_and_age_85_yrs_plus	**166**
Estimate; YEAR HOUSEHOLDER MOVED INTO UNIT - Occupied housing units - Moved in 2000 to 2009	
est_year_householder_moved_into_unit_occ_unts_2000_to_2009	**161**
Estimate; SEX AND AGE - 10 to 14 years	
est_sex_and_age_10_to_14_yrs	**157**

Table 5: Top ten most influential features by F1 score for the performant model.

timization, the feature selection was based on feature weight. Table 6 shows how the XGBoost model performed when the threshold increased from 0 (all features) to 0.016 (top 2 influential features). The two most performant models were similar in terms of accuracy and lift, noting that the lift metric was still poor overall. The number of features needed to produce similar performance were wide - 2 versus 27. Optimizing on feature selection during the modeling process may produce a more performant model over the existing feature subset which was selected based on observation.

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**

Table 6: XGBoost model performance given feature selection threshold.

SMOTE preprocessing resulted in the greatest performance, in terms of recall, as compared to random oversampling and no preprocessing. Oversampling and under-sampling methods are highly recommended for imbalanced data such as our aggregated input data set of ACS, chemical, and lead results. Given the small data set, under-sampling would have resulted in even poorer performance as the majority class would be reduced to meet the minority. Over-sampling methods remained the focus with SMOTE performing the best. Additional methods to consider for the future include Modified Synthetic Minority Over-sampling Technique (MSMOTE). It is similar to SMOTE except MSMOTE categorizes minority class samples as safe (good for the classifier), border (unknown), or noise (poor for the classifier) and handles the nearest neighbor selection different.

6 Conclusions

Ultimately, the chosen feature data was not predictive of the selected label data. The factors that show some predictive power at the building level did not have the same relationship to lead contamination once aggregated to the census tract level. Figure 4 shows the distribution of percent of homes built in every tract over different time periods. Figure 5 shows the distribution of percent of homes built in contaminated

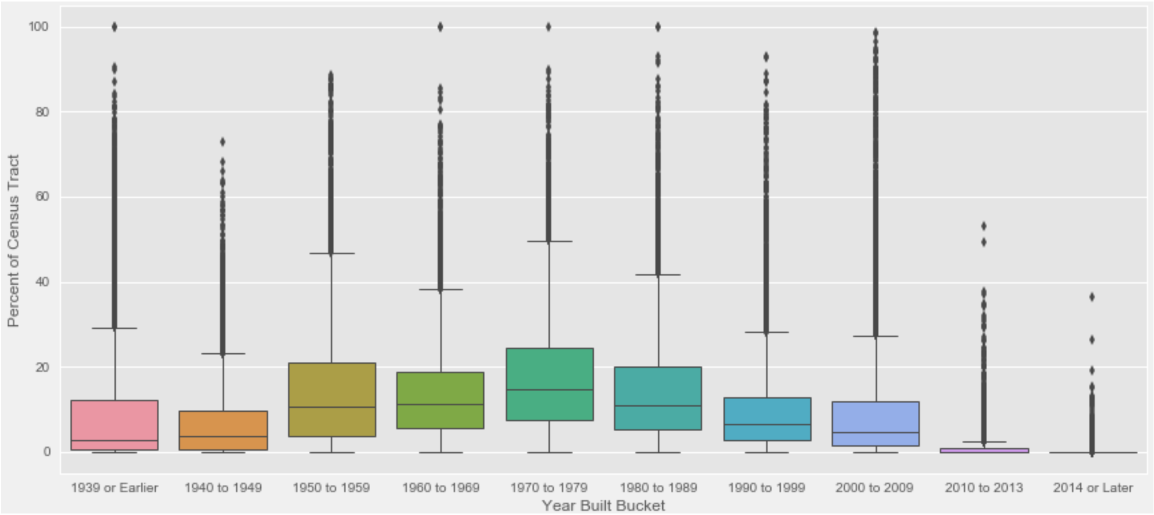

Figure 4: Distribution of year home built per census tract.

tracts (lead readings above 5 µg/L) over different time periods. While the tracts with readings above 5 µg/L show some variation, the distributions are largely similar. This could one reason be why the model had difficulty making predictions.

Also, the upstream water supply testing data did not prove to be predictive of lead contamination either. Lead contamination is most heavily influenced by the materials within the building itself.

To pursue similar predictive modeling in the future, it will be critical to use building property data as feature data. This is typically contained within county assessor data. We did not pursue this data because our project timeline did not allow for the time necessary to acquire data from every county in California and merge them into the same schema. Additionally, not all county assessor offices in California have their data online, and some counties charge fees for the data. If the state were able to formalize a structure and portal for assessor data throughout the state, it would allow more analysis to occur.

If a future study were to occur using building level feature data and the school testing database as label data, we would recommend following many of our steps around data leveling and modeling. The data will be imbalanced, requiring data leveling. Also, if feature data is at the building level, spatial clustering methods should be evaluated. Not every building is a school, and schools will have different properties than other

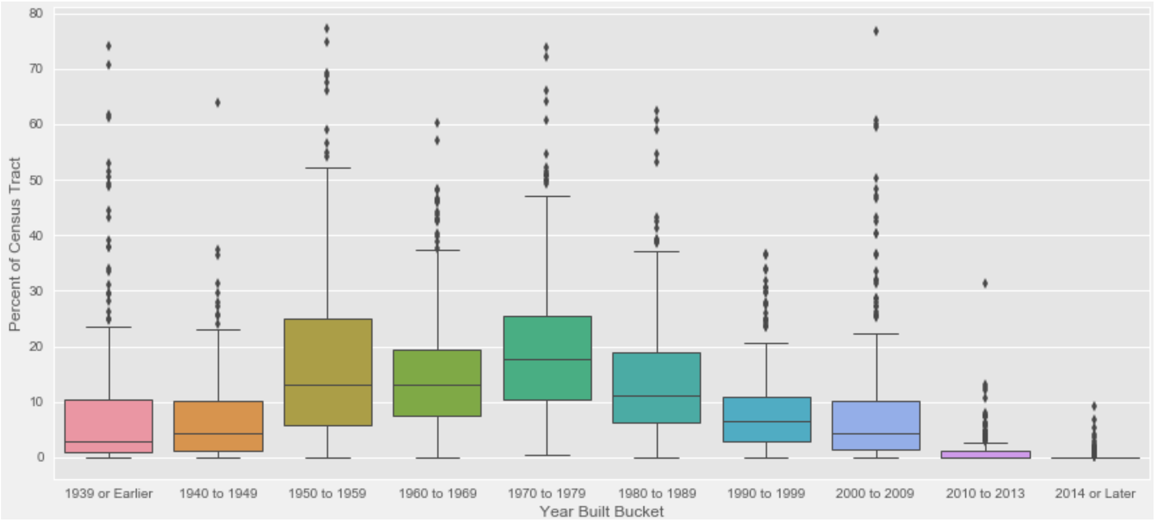

Figure 5: Distribution of year home built per contaminated tracts with lead readings above 5 µg/L.

building types, but this is still the most complete data that can be used to train a model.

Acknowledgements

We would like to thank our advisers Joyce Shen and Alberto Todeschino for their input and feedback throughout our process, including in idea development. Joyce and Alberto pushed us to define our project and attempt to solve a valuable problem. We would also like to thank Capstone Section 4 for their valuable insights as we continuously tweaked our approach and communication materials.

Further, we would like to thank the many individuals at the California SWRCB for making the school testing data available to us as well as providing guidance on several other datasets and our project overall.

List of Figures

List of Tables