Feature Selection for Water Lead Contamination in Flint, MI

Ajay Kumar [1]

[1]Neuroscience and Behavior Program, University of Massachusetts, Amherst

GRiD Hack2O, 2017

Outline

1. Introduction
   - Flint Water Crisis
   - Problem
   - Project Question
   - Hypothesis
2. Feature Selections
   - Water Quality
   - Weather
   - Feature Processing
3. Models
   - XGBoost
   - Model Selection
   - Hyperparameter Optimization
4. Results
   - Accuracy Metrics
   - Feature Importance

Flint Water Crisis

- Flint Water Crisis began in April 2014 when water source was switched from Lake Huron to Flint River.
- Proper anti-corrosive actions were not taken, resulting in high lead levels.
- High lead levels are a major health hazard for children, highlighting importance of accurate measurements and corrective actions.

Problem

- Aging infrastructure combined with low socioeconomic status of region has lenghtened the crisis.
- While logitudinal EPA tracking has shown decreasing lead levels, tracking still continues to ensure water quality.
- Lead testing methods are costly, complex and only performed at certified labs, leaving room for data methods to use alternative measures to classify/quantify lead contamination.

Project Question

Which cost-effective water quality indicators can be used to accurately predict drinking water lead contamination status?

Hypothesis

Temperature and pH are hypothesized to have the highest feature importance for predicting classification of a property as contaminated with lead or not.

Water Quality Features

[Copper](ppb)	Maximum **copper** levels in water across all collected sequential samples on a specific day by property.
[Iron](ppm)	Maximum **iron** levels in water across all collected sequential samples on a specific day by property.
pH	Average **pH** of water measured on a specific day by property
[Chloride](ppb)	Average **chloride** levels in water on a specific day by property

Features Correlation

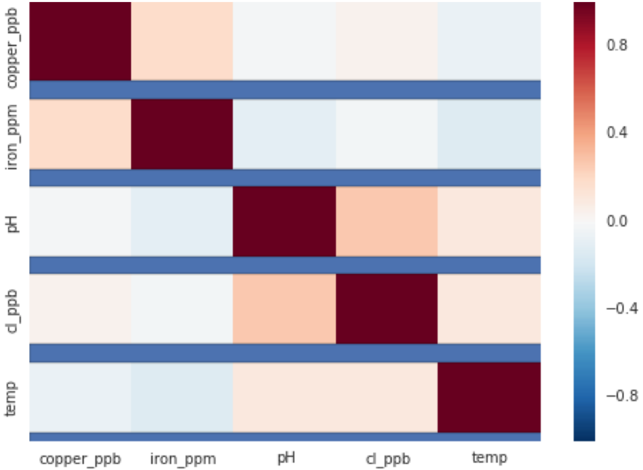


Figure: Correlation Matrix across all features

Temperature

- Previous research has indicated that air temperature has an effect on lead release from pipes [Masters et al., *Environ. Sci. Technol.*, 2016.]
- Historical temperature data for sequential samples on specific day obtained via Dark Sky API [https://darksky.net/poweredby/]

Effect of temperature on lead release

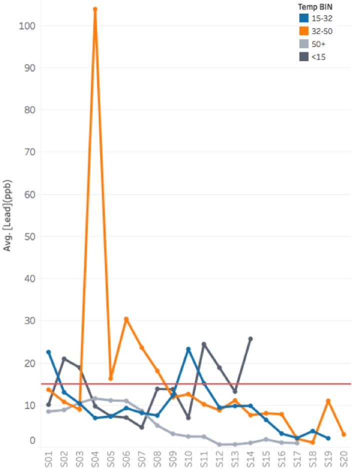


Figure: 'First Flush' phenomenon of lead by temperature [Dark Sky API]

Feature Processing of Lead Concentration

Sequential samples of lead concentration in drinking water was converted to a classification problem by assigning **'No Contamination'** to a property where $> 95\%$ of samples obtained had $< 15ppb$ of measured lead. Properties with $< 95\%$ of samples with $< 15ppb$ of lead were classified as **'Contaminated'**.

Feature Processing of Lead Concentration

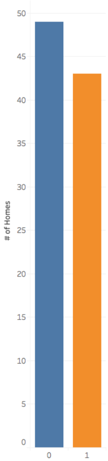


Figure: No. of homes classified as 'Not Contaminated' vs. 'Contaminated'

Model Selection

- Boosting is a sequential technique which works on the principle of ensemble. It combines a set of weak learners and delivers improved prediction accuracy.
- The XGBoost algorithm builds Gradient boosted trees in parallel fashion providing much faster grid search for optimizing hyperparameters in model tuning.
- Hyperparameters selected by 10-fold cross-validation on a 70/30 train/test split of data.

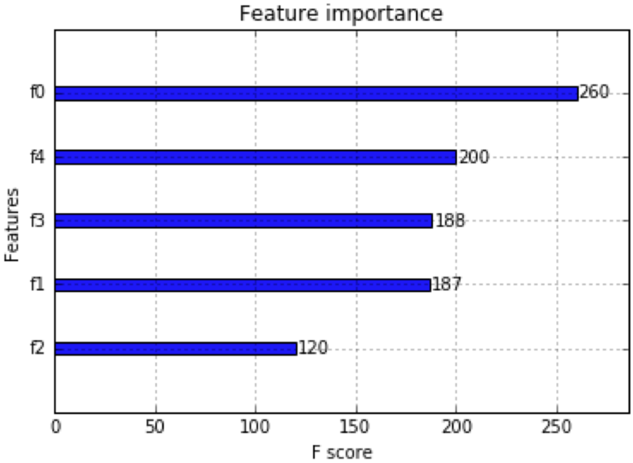
Hyperparameter Optimization

Hyperparameter	**Values**
Max Depth	3
Learning Rate	0.1
subsample	1
Min-child-weight	1
Num-boost-round	166

Accuracy Metrics

Metric	**Values**
Accuracy	72.41%
Error	27.59%
AUC	0.73

Feature Importance



f0-[Copper]; f4-Temp; f3-[Chlorine]; f1-[Iron]; f2-pH

Summary

- Air temperature affects pattern of lead flushing observed in sequential water samples, identifying additional factors to be considered during testing.
- Copper levels and Temperature were identified to be the most important features in classifying a property as either being lead contaminated or not.
- Outlook
  - Increase size of test data to fit classifiers and verify feature importance identified. Highlights importance of running water before use.
  - Test different classifiers to test for classification accuracy/AUC and subsequently test feature importance.