

Drinking Water: Predicting quality violations in New England

François Delavy



Contents

1. Data and Research Question
2. Strategy for Model Selection
3. Results
4. Discussion, Practical Use and Future Development

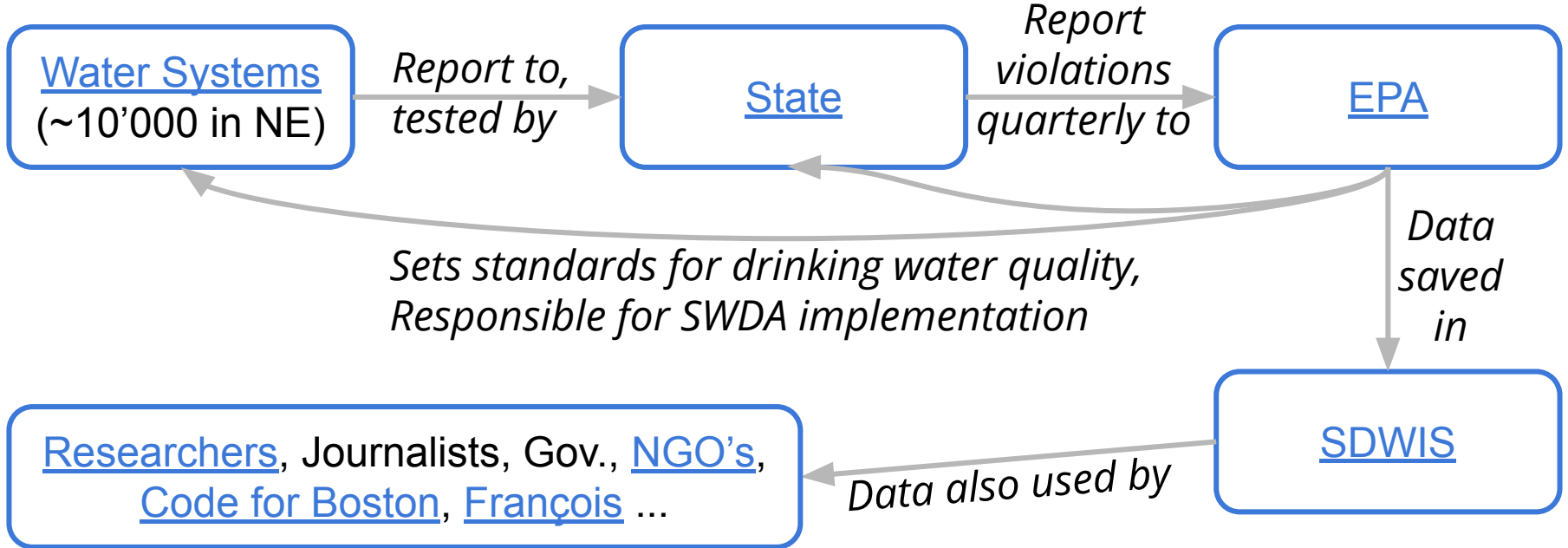


1. Data and Research Question



Data and Information Flow

The **Safe Water Drinking Act** ([SWDA](#)) aims to ensure safe drinking water for the public.



Research Question

How predictable are drinking water violations in water systems?

- Which factors are good predictors?
- Is using only SDWIS data sufficient?
- How predictable are specific types of violations? e.g. pesticides

Potential Biases and Issues with the SDWIS Data

- We only know what the EPA knows (+ EPA is aware of inaccuracies in SDWIS)
- The methodology, sampling rate, completeness of the data might differ from state to state or water system to water system
- Sampling rates might differ from contaminant to contaminant, giving more weight to certain violations
- Only violations above a Maximum Contaminant Level are reported to the EPA
- Localization of the water systems: the WS might be located in another county than the source of the water or the treatment facilities
- The water is sampled (usually) at the entry point to the distribution system and not at the exit point: the taps.
- The types of the violations reported in SDWIS are numerous (e.g. legal, lead, coliforms, pesticide, disinfectants, ...) and thus their causes are numerous too!
- Time lag between observation and appearance in SDWIS (~3 months and more): some violations have maybe already been addressed at the time they appear in SDWIS.
- ...

2. Strategy for Model Selection

Approach: Classification

Aim: Predict which water systems are likely to be subject to violations of water quality in a given year.

Binary classification of water systems:

1. likely to have a violation of the drinking water quality this year
2. not likely to have a violation of the drinking water quality this year

Classifier trained on data from previous years:

Training set: 2013 - 2015, Validation set: 2016, Testing set: 2017

Anticipated Issue

- Multi-causality of the violations (e.g. legal, lead, coliforms, pesticides, disinfectants, ...)
- The features available in SDWIS are limited

Solution (?)

- Link external data to SDWIS in order to add relevant contextual information about each type of violation
 - *Complication: finding datasets, understanding causes*
- Focus on a particular type of violation
 - *Complication: increase of class imbalance*

⇒ 2 models, both with the [estimated pesticide use](#) by county added to SDWIS:

1. Outcome variable is *had any violation*
2. Outcome variable is *had any pesticide violation*

Steps

1. SDWIS for NE, 2013-17, add pesticide use by county
2. Features Engineering
3. Split data in train, validation and test sets
4. Handling Class imbalance (~90% | 10%):
 - SMOTE, class weights, under-sampling, score = area under roc curve: AUROC
5. 3 models: Logistic Regression, Gradient Boosting, Random Forest
6. Grid search with cross-validation to tune parameters on train set
7. Evaluation of models on validation set: generalizability?
8. Testing one selected model on test set \Rightarrow report results

iter.



3. Results

3.1 All Types of Violations

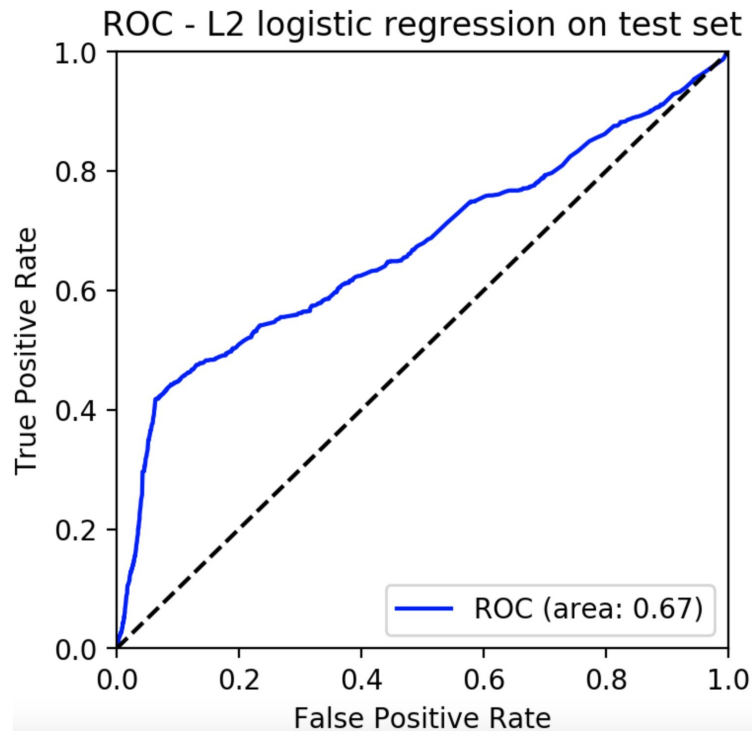
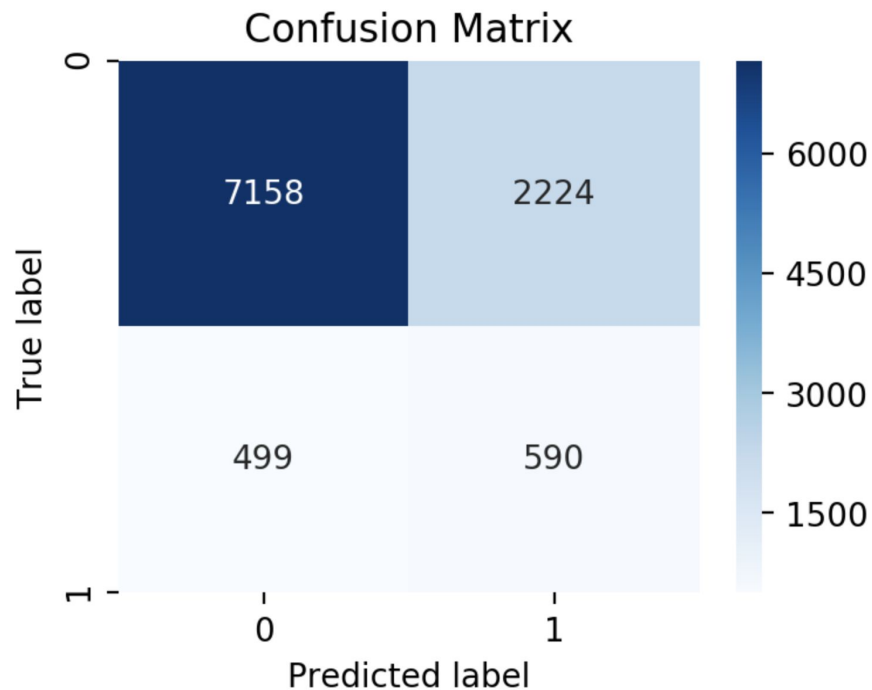
Model Selected: L2 Logistic Regression

Because:

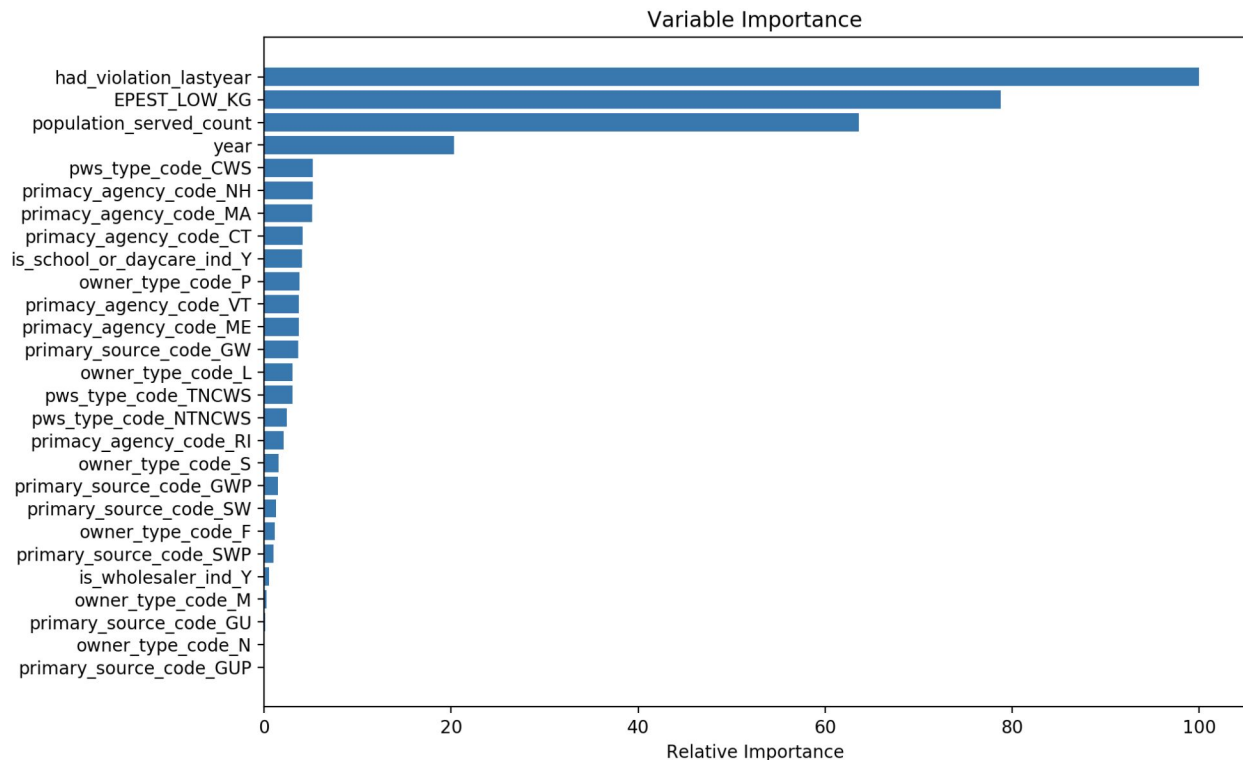
- It achieves a similar AUROC than the gradient boosting and random forest, while...
- ... being simpler and faster to run, and...
- ... it is an explanatory approach.

Inability to improve the lack of generalizability of the gradient boosting and random forest (due to overfitting) in a way so that they outperform logistic regression.

Classification on the Test Set: Poor...



Features Importance (Gradient Boosting)



Results are similar for the L2 logistic regression.

Key Points

- Inability to train a performant classifier of water systems likely to be subject to water quality violations
- Joining external data to SDWIS is necessary, as few features in SDWIS are good predictors
- Last year's violations are the most important predictor

3.2 Pesticide Violations

Contaminations by Pesticides

On average, per year, only ~10 out of the ~10'000 water systems in New England had a pesticide violation*

⇒ not enough to train a classifier

Conclusion:

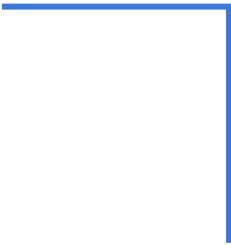
“it was an entertaining waste of time!”

**for the contaminants that I could match to the NAWQA pesticides list*

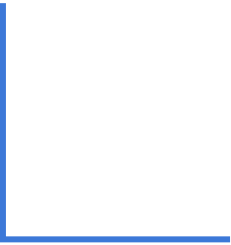
Key Points

Was it really a waste of time?

- Estimated pesticide use by county is useful external data to complement SDWIS, even to predict all types of violations
- Maybe some other regions in the US are more subject to drinking water contaminations by pesticides?



4. Discussion, Practical Use and Future Development



Weak Points and Shortcomings

- Obviously, the poor performance
- Weak ability to identify new violations (last year violations is most important feature)
- Going from classification to prediction (recency of the data)
- Granularity in time and location in SDWIS (quarter, county)
- Dichotomization of the outcome variable induces loss of information

Probable Future Direction

Focus on a particular (but not too rare!) type of contamination and add related external contextual data

- e.g. coliforms

Potential Practical Use and Audience

- Prioritization of resource allocation:
 - of the agents testing water quality
- Risk exposure map:
 - for water systems owners: the contaminations the most likely to occur in their (type of) water system
 - for private wells owners: the contaminations in neighbouring or similar public water systems
- ...

Lessons Learned

- Class imbalance:
 - Over-sampling (SMOTE), under-sampling and cost-sensitive training (class weights) led to similar results *in this case*
 - Use area under roc curve (AUROC) instead of accuracy for training the models (evtl. recall)
 - R2 Logistic Regression is not performing less well than “more complex” gradient boosting and random forest *in this case*
- Joining external contextual data to SDWIS is necessary
 - There are not many other features in SDWIS that are correlated to the outcome (service connections, previous violations, evtl. type of water system)
 - Not always easy to find relevant and joinable external data...
 - Estimated pesticide use per county and year is useful for the model, even for predicting all types of violations. Proxy for rurality?

Many Thanks To

Abhimanyu Mittal, my mentor at Thinkful, for his time and always pertinent advices.

All the members of the *Code for Boston* [safe-water project](#).

This presentation is my attempt to predict drinking water violations. I benefited enormously from exchanges with the dedicated volunteers coming to the hack-nights every Tuesday. I hope that the lessons learned from this small project will help the team!



FIN.

[link to the code and data](#)

