Predicting Drinking Water Quality Violations in New England

Using Data Collected by the EPA

François Delavy

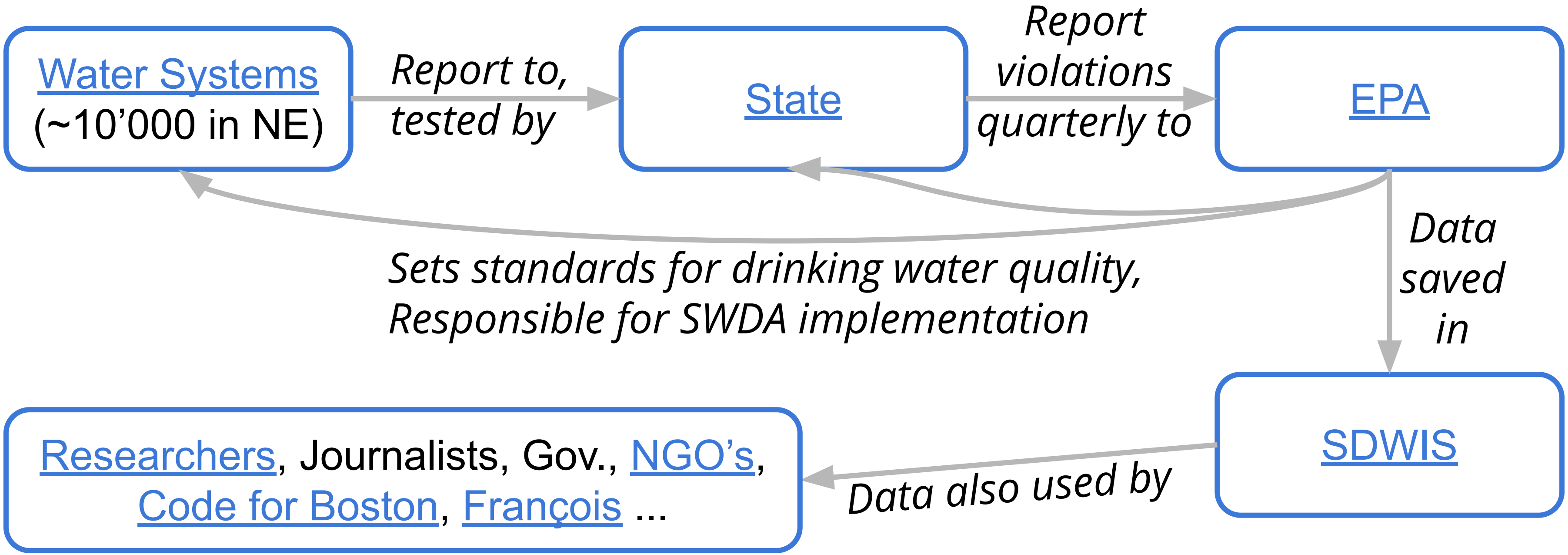
Contents

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 at ensuring safe drinking water for the public.



Water Systems
(~10’000 in NE)

Report to, tested by

State

Report violations quarterly to

EPA

Sets standards for drinking water quality, Responsible for SWDA implementation

Data saved in

SDWIS

Data also used by

Researchers, Journalists, Gov., NGO’s,
Code for Boston, François ...

Research Question

Can we Predict Violations of Drinking Water Quality?

i.e. predict which water systems are likely to see a violation in a given year

- … and some more specific types of violations? (e.g. pesticides)

- … only with SDWIS data?
- … or by linking to other external data (e.g. estimated pesticides use)

- … and explain why? (explanatory VS predictive models)

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): some violations have maybe already been addressed at the time they appear in SDWIS.
- ...

2. Strategy for Model Selection

Approach: Classification

Classify the water systems into two classes:

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

Predict which water systems are likely to be subject of a violation of drinking water quality this year, based on data from previous years.

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

Anticipated Issue

The types of the violations reported in SDWIS are numerous (e.g. legal, lead, coliforms, pesticide, disinfectants, ...) and thus their causes are numerous too!

Will a classification model based on the features available in SDWIS only be able to capture this diversity?

Solution: (?)

- Link external data to SDWIS, to add information relevant for the classification of each types of violation (but not always doable)
- Focus on a particular type of violation (but increases class imbalance)

⇒ try a focus on pesticide contaminations, using estimated pesticide use

Steps

1. Download SDWIS for NE, 2013-17, add pesticide use by county
2. Features Engineering (notably last year’s violations)
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 selected model on test set ⇒ report results

iter.

3. Results

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 and untertaining waste of time!”

Follow-up:
The idea is promising though, maybe some other regions in the US are more subject to drinking water contaminations by pesticides?

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

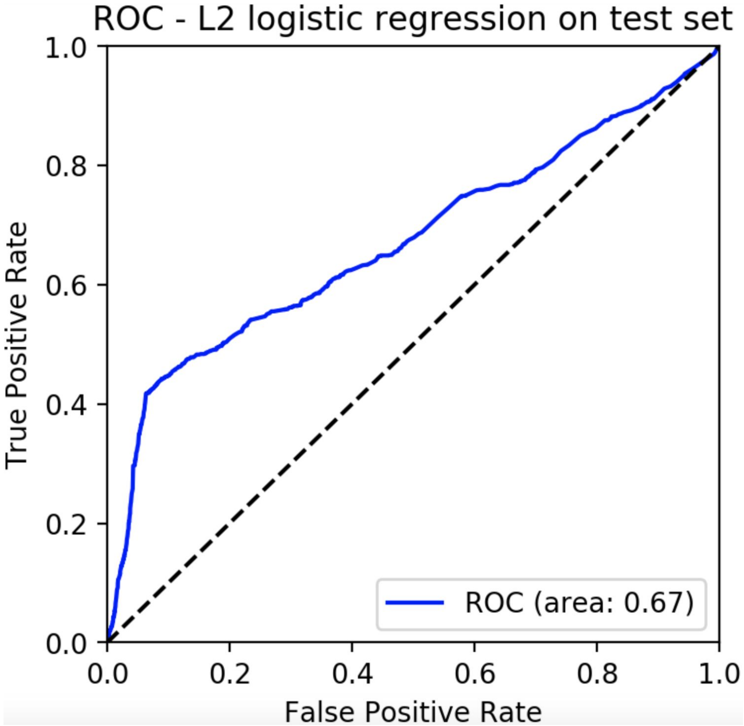
Model Selected: L2 Logistic Regression

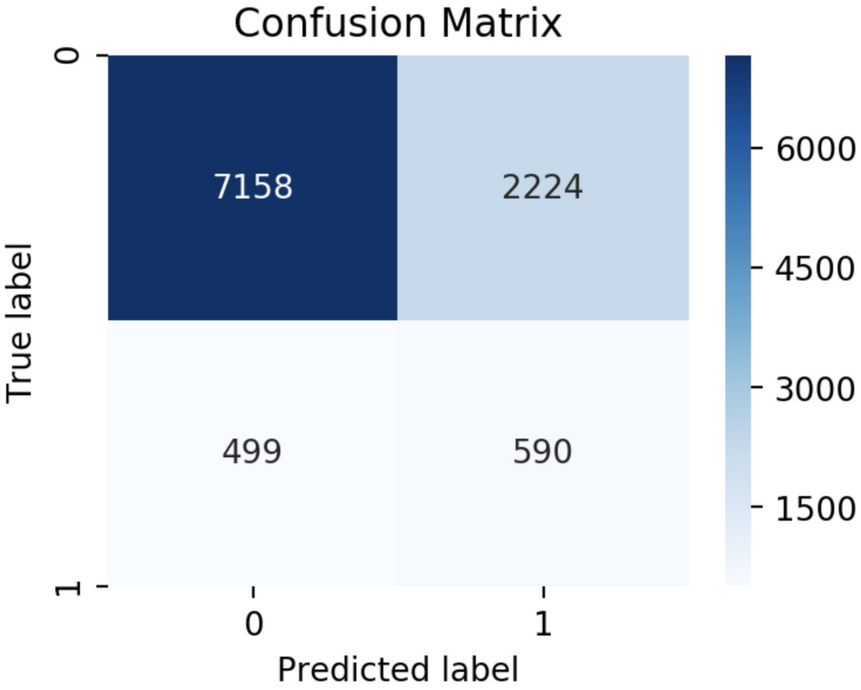
To predict “general/all” water violations, because:

- 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.

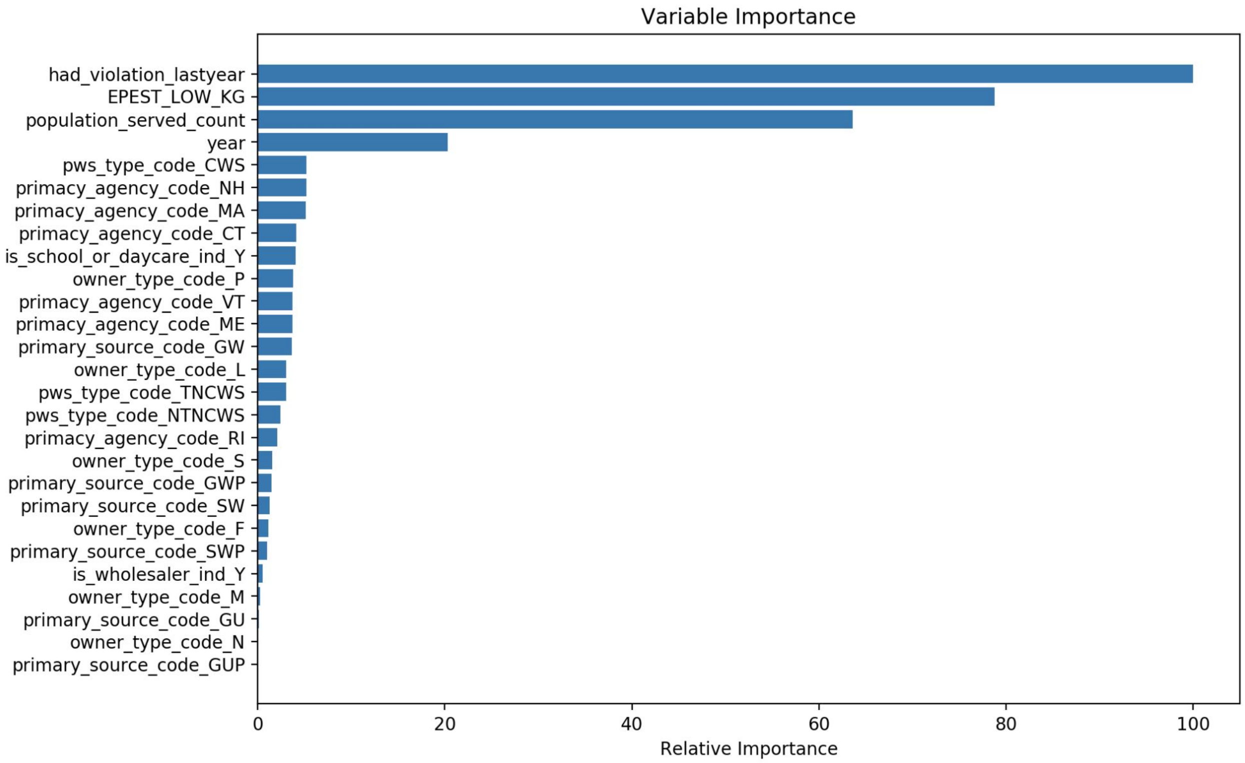
I could not control the overfitting and classification performance of the gradient boosting and random forest in a way so that they outperform logistic regression.

Classification on the Test Set: Poor...





Features Importance



On the left, the relative importance of the features used by the gradient boosting algorithm trained.

Results are similar for the L2 logistic regression

4. Discussion, Practical Use and Future Development

Weak points, shortcomings

- Obviously, the poor performance
- Most important feature = past violation
- By dichotomizing the outcome variable, we lose information
- Predicting the past? (currently most recent violation was on 2018-09-21)
- Probable presence of some bias that could be relevant, depending on the practical use (see next slide)

Probable Future Direction:

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

(potential) Practical Use and Audience

As there is a focus on explanatory machine learning:

- Directing the agents testing water quality to the water systems the most likely
- Informing water systems owners on the likely contaminations of their wells
- Informing private wells owners of the water systems contaminations in neighbouring 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,

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 at predicting some drinking water violations. This would have not been possible without all the fruitful exchanges with and advices from the dedicated volunteers coming to the hack-nights every Tuesday. I hope that the lessons learned from this small project will help back the team!

FIN.

link the code and data